Estudi i aportacions a l'aprenentatge de distàncies parametritzades per matrius mètriques

VNIVERSITAT DE VALÈNCIA

Departament d'Informàtica

Programa de Doctorat:
Informàtica i Matemàtica Computacional

11 de desembre de 2014

Autor: Adrián Pérez Suay
Directors: Francesc Josep Ferri i Rabasa
Miguel Arevalillo Herráez

Tinc el plaer d'agrair l'inestimable dedicació per part dels meus tutors de tesi i grans companys d'aventura Francesc J. Ferri i Miguel Arevalillo. Moltíssimes gràcies als dos, no sabeu quant aprecie els moments de discussió que hem tingut i sobretot la vostra ajuda. Gràcies per haver estat sempre al meu costat.

Gràcies Ana, per compartir amb mi tots els moments, t'estime amb tot el cor. Gràcies al meu pare Daniel per haver-me estés sempre la mà inclús en els moments més difícils, t'estime pare. Gràcies a ma mare Reme per voler-me tant, jo també et vull. Gràcies als meus germans Dani i Paula perquè amb vosaltres he compartit gran part de la meua vida i us estime de veres. També estic especialment agraït a la meua iaia Reme que m'ha cuidat des de ben petit i encara em fa les millors paelles. Als meus iaios Rafael, Daniel i Amalia que ja no hi són però que també els recorde i estime com si hi foren.
Gràcies a Amparo i Sento perquè formeu part de la meua vida.
Gràcies Paco per tot el que has fet per mi. Per haver-me deixat fer i desfer, però sobretot per la teua confiança. Gràcies de tot cor.
També vull agrair a Ricardo Ferris (Ferris l'immortal), que ha sigut un gran mestre d'informàtica i amb qui he passat molt bons moments.

Agraïments

Aquest treball ha sigut parcialment finançat per FEDER, el Ministeri d'Educació i Cultura del govern Espanyol i també pel govern Valencià a través dels projectes DPI2006-15542-C04-04, TIN2009-14205-C04-03, TIN2011-29221-C03-02, TIN2012-38604-C05-02, ACOMP/2010/287, GV/2010/086 i Consolider Ingenio 2010 CSD2007-00018.

Resum

Alhora de construir sistemes intel·ligents capaços de sensoritzar, reconèixer i comprendre el seu entorn, una de les opcions consisteix a fer servir una mesura de distància o de dissimilitud per tal de comparar els objectes detectats. En alguns casos, la utilització de distàncies estàndard pot ser acceptable i fins i tot convenient. No obstant això, sempre existeix la possibilitat d'aprendre una distància adaptada al problema inductivament a partir d'exemples particulars. Precisament, l'aprenentatge de distàncies és un cas particular d'aprenentatge automàtic el objectiu del qual consisteix a trobar aquella funció distància que satisfà una sèrie de condicions que tenen a veure amb l'adaptació de la distància al problema. En altres paraules, es tracta de trobar aquella funció distància que fa òptims els corresponents procediments de reconeixement.

En aquesta tesi s'aborda el problema de l'aprenentatge de distàncies des de dues perspectives diferents. Una d'elles considera el cas en què tots els exemples a l'abast es fan servir conjuntament i a l'hora en el procés d'aprenentatge (aquest paradigma es coneix com aprenentatge per lots). L'altra contempla el cas en què els exemples es fan servir de manera seqüencial, un a un, de tal forma que en cada pas es disposa d'un determinat model de distància que poc a poc va millorant (aquest procés rep el nom d'aprenentatge continu o en línia).

Respecte a l'aprenentatge per lots, en aquesta tesi es planteja com alternativa l'elecció d'exemples amb diferents criteris per tal de construir un subconjunt reduït i representatiu en contraposició a la utilització del conjunt d'exemples originalment disponibles. Aquesta aportació redueix el temps d'execució i manté l'eficiència en la classificació d'un mètode d'aprenentatge de distàncies àmpliament conegut en la literatura relacionada, que està fonamentat en la idea que els elements de la mateixa classe estiguen a distància zero.

Respecte de l'aprenentatge en línia s'han desenvolupat diferents alternatives d'un mètode que aprén a través d'un model, i que realitza una adaptació només si prediu incorrectament. Entre les diferents aproximacions presentades, s'ha introduït una formulació mitjançant mínims quadrats que defineix un nou mètode d'aprenentatge en línia de distàncies. En les versions desenvolupades, els resultats obtinguts per a classificació mostren un rendiment equiparable a alguns mètodes per lots i un requeriment computacional reduït en relació també a aquest tipus de configuració.

Índex

Índex de figures

Índex de taules

Taules de símbols i acrònims

Símbols

Notació	Descripció
$\mathbb{R}$	Cos dels reals
$\mathcal{M}$	Conjunt de mostres d'entrenament
d	Dimensió de l'espai d'entrada
n	Quantitat d'elements d'entrenament
c	Nombre de classes en el conjunt d'entrenament
$\mid \mathcal{M} \mid$	Cardinal del conjunt $\mathcal{M}$, $\mid \mathcal{M} \mid = n$
Y	Conjunt de Punts Base
p	Nombre de Punts Base
$\mathbb{R}^d = \mathbb{R} \times \cdots \times \mathbb{R}$	Producte cartesià de d espais Reals
x_i	Mostra $i-$èssima/Vector columna de $\mathbb{R}^d$
$\mathcal{I}$	Conjunt de classes
c_i	Classe associada a l'element x_i
n_k	Nombre d'elements corresponent a la classe k
y_j	Punt Base $j-$èssim
$\tilde{c}_j$	Classe del Punt Base y_j
$\|.\|_p$	Norma $p = 0, 1, 2, \ldots, \infty$, de l'espai $\mathbb{R}^n$
$\|.\|_{Frob}$	Norma de Frobenius
M	Matriu mètrica
PSD	propietat de M, matriu Semi Definida Positiva
W	Matriu transformació lineal
r	Rang de W
d_{ij}	Distància genèrica entre x_i, x_j
d_{ij}^M	Distància quadràtica parametritzada per M entre x_i, x_j
μ	Mitjana
σ	Desviació estàndard
Σ	Matriu de covariància
$p^M(j\|i)$	Distribució condicional parametritzada per M
$p_0(j\|i)$	Distribució ideal

Símbols (continuació)

Notació	Descripció
$KL[\cdot\|\|\cdot]$	Divergència de Kullback-Leibler
x^T	Vector x transposat
M^T	Matriu M transposada
$Tr(M)$	Traça de la matriu M
$\mathcal{S}$	Conjunt d'índexs sobre $\mathcal{M}$ de parells similars
$\mathcal{D}$	Conjunt d'índexs sobre $\mathcal{M}$ de parells dissimilars
$\mathcal{T}$	Conjunt d'índexs sobre $\mathcal{M}$ de triplets

Acrònims

Abreviació	Descripció
ITML	*Information Theoretic Metric Learning*
LMNN	*Large Margin Nearest Neighbor*
LS	*Least Squares*
LEGO	*Logdet Exact Gradient Online*
MCML	*Maximally Collapsing Metric Learning*
MLSVM	*Metric Learning Support Vector Machines*
PA	*Passive Aggressive*
POLA	*Pseudo-metric Online Learning Algorithm*
SVM	*Support Vector Machine*

Part I

Contextualització del treball

Capítol 1
Introducció

Resum – En aquest capítol s'introdueix de manera general el context i els conceptes fonamentals que són necessaris per enunciar la motivació i els objectius d'aquesta tesi. La part final del capítol descriu l'estructura global del document.

Contingut

1.1 Introducció general

L'aprenentatge és una habilitat de tots els organismes vius i en particular dels éssers humans. Des del moment en què naixem comencem a aprendre sota la supervisió de la nostra mare i pare. Ells ens transmeten el seu coneixement per tal d'assegurar-nos una bona adaptació al medi i aquest aprenentatge es finalitza juntament amb la vida.

L'*aprenentatge automàtic* (de l'anglès *machine learning*), s'enfronta al repte d'aconseguir que les màquines siguen capaces d'aprendre models que generalitzen i s'adapten a certs tipus de comportaments. Per això, cal tindre representacions adients de diferents conceptes que són objecte d'estudi. Aquests conceptes poden arribar a ser molt complexos. Per exemple, pot tractar-se d'imatges digitalitzades, seqüències d'ADN o formes d'ona que admeten una representació en un espai com el vectorial Euclidià. Així doncs, l'aprenentatge es realitza a través d'un conjunt de mostres concretes que són diferents exemples d'un concepte particular el qual rep el nom de *conjunt d'entrenament*.

En la vessant supervisada, l'aprenentatge automàtic parteix d'un conjunt d'entrenament format per exemples correctament etiquetats, mitjançant el qual s'a-

prén o entrena un model. Posteriorment, aquest model s'utilitza per a realitzar prediccions sobre la pertinença a una classe d'elements d'etiqueta desconeguda.

Si les dades d'entrenament no estan disponibles totes al mateix temps, sinó que arriben en *línia* o de manera seqüencial, parlem *d'aprenentatge en línia* (en anglès *online learning*). En aquest cas, les actualitzacions del model es realitzen cada vegada que es processa una nova mostra, depenent de la coincidència entre la predicció i l'etiqueta. Aquest aprenentatge té com a particularitat que pot enfrontar-se a problemes en els quals el conjunt d'entrenament és relativament gran.

L'*aprenentatge de distàncies* (de l'anglès *Distance Metric Learning*), és una disciplina relativament nova i un cas particular d'aprenentatge automàtic. En aquest cas, es tracta d'utilitzar les dades disponibles per a construir una funció de proximitat que permeta inferir un valor de distància per a qualssevol parell d'elements.

1.2 Motivació

La necessitat de tindre mesures apropiades de distància o similitud entre les dades és omnipresent en el camp de l'aprenentatge automàtic, però l'ajustament manual de bones mesures per a problemes específics és generalment difícil.

Quan les dades són representades en un espai, ocorre sovint que aquesta representació no és la idònia per a poder establir relacions entre els diferents objectes d'estudi. El problema pot ser major si s'empra explícitament algun tipus de distància per a establir relacions entre elements. En particular pot ocórrer que, sobre aquest espai de representació, elements que haurien de ser conceptualment dissimilars siguen representats pròxims. Però igualment, elements similars poden tindre representacions molt llunyanes. Ací és on rau la importància de l'aparició del camp de l'aprenentatge de distàncies, que té per objecte la construcció automàtica de mesures de distància a partir d'exemples.

En molts casos, el concepte de similitud entre elements pot ser subjectiu o difícil d'establir. Però les relacions entre els elements poden vindre donades per la informació a priori aportada per un usuari o un expert. Generalment, aquesta informació té la forma d'etiquetes sobre un conjunt d'entrenament i es pot utilitzar per a formular un problema d'aprenentatge que permeta obtindre una mesura de distància més adequada per al cas particular.

L'interès per aquest tipus de tècnica és demostrat per les nombroses publicacions realitzades durant els últims anys. La seua aplicació directa en algorismes estàndard de classificació ha incentivat el desenvolupament de diferents mètodes d'aprenentatge de distàncies.

1.3 Objectius

L'objectiu principal d'aquesta tesi és l'estudi de diferents enfocaments del problema de l'aprenentatge de distàncies, des de la seua formulació com a problema

d'optimització fins a la resolució, l'avaluació de l'aprenentatge i l'anàlisi de la complexitat de diferents propostes. En particular, es defineixen els següents objectius específics:

1. Estudi dels mètodes d'aprenentatge de distàncies existents, basats en diferents formulacions.

2. Identificació de possibles millores o extensions.

3. Formulació i desenvolupament de noves propostes per a l'aprenentatge de distàncies.

4. Avaluació empírica de les diferents aportacions en relació a l'estat de l'art.

1.4 Resultats de la investigació

Entre els resultats, es proposa una millora sobre un mètode concret i en certa manera representatiu: "*Maximally Collapsing Metric Learning*"(MCML) [40]. La intenció del MCML és col·lapsar les classes: aconseguir que els elements de la mateixa classe siguen propers i, alhora, lluny dels de diferent classe. Aquesta situació geomètrica referent al col·lapsament de les classes és la idea al voltant de la qual es desenvolupa el MCML. La nova proposta fa servir un conjunt reduït d'elements representatius situats estratègicament (els anomenats punts base), en contraposició al MCML que utilitza tots els disponibles. La conseqüència principal és la reducció de la complexitat computacional del mètode original, a l'hora que es manté la seua capacitat de generalització [73, 72].

Altre resultat rellevant de la tesi és la proposta de noves tècniques d'aprenentatge de distàncies inspirades en l'aprenentatge en línia amb component passiu-agressiu, de l'anglès *"Passive-Aggressive"* (PA) [82]. El nom PA es deu al fet que es diferencien dues possibilitats durant l'aprenentatge: el cas passiu, que no actualitza el model quan coincideixen la predicció i l'etiqueta; i l'agressiu, que si que l'actualitza en altre cas. En particular, es presenta una formulació unificada de diferents alternatives passives-agressives en el context de l'aprenentatge de distàncies en línia. Aquesta formulació no només generalitza treballs anteriors [83], sinó que també permet la introducció d'una nova proposta inspirada en mínims quadrats [69, 74].

1.5 Estructura de la Tesi

En funció dels objectius enumerats i resultats obtinguts, la tesi s'ha organitzat en dues parts diferents, formades per un total de 6 capítols i 5 apèndixs.

La Part I, recull els fonaments teòrics que permeten la contextualització i comprensió de les aportacions de la tesi. En particular, aquesta part consta de tres capítols:

- Al present Capítol 1 (Introducció), es dóna una visió general de la problemàtica tractada en aquesta tesi, se'n detalla la motivació, se n'enumeren els objectius concrets i se'n resumeixen els principals resultats obtinguts.

- Al Capítol 2 (Fonaments teòrics), s'introdueixen diversos aspectes conceptuals referents a l'aprenentatge automàtic que contextualitzen l'aprenentatge de distàncies. En concret, es presenten diferents famílies de distàncies i els fonaments matemàtics indispensables per al desenvolupament de la investigació. També es presenten classificadors concrets i es formula el problema a tractar d'una manera general. Per finalitzar, es presenten algunes representacions gràfiques emprades durant l'estudi.

- El Capítol 3 (Estat de l'art), conté una revisió de la literatura existent al voltant del problema de l'aprenentatge supervisat de distàncies. Els diferents mètodes es categoritzen en base a l'enfocament del problema i s'enumeren de manera aproximadament cronològica. La selecció particular de mètodes s'ha fet a partir de la seua relació amb les aportacions principals presentades.

La Part II consta de dos capítols. Cada un d'aquests recull una de les aportacions originals juntament amb la corresponent validació empírica.

- El Capítol 4 (Aprenentatge basat en el col·lapsament de classes mitjançant un conjunt d'elements representatius), presenta un ampli estudi del mètode MCML [40], i s'inclou una aportació original que representa una millora computacional significativa.

- Al Capítol 5 (Aprenentatge en línia passiu-agressiu i extensió per mínims quadrats), es formula el problema d'aprendre una distància mitjançant maximització del marge. A partir d'aquesta formulació, es proposa un esquema general d'aprenentatge en línia sobre el qual es presenta una col·lecció de variants.

Finalment, el Capítol 6 (Conclusions i treballs futurs) recull les principals conclusions que es deriven de tot l'estudi desenvolupat en la tesi. A més a més, es resumeixen les aportacions realitzades i es plantegen futures línies d'investigació a seguir com a possibles extensions.

Els 5 apèndixs tenen la intenció de complementar la informació que s'aporta en els diferents capítols al llarg del document. L'Apèndix A recull els detalls del conjunt de bases de dades emprades en l'experimentació, incloent les seues característiques fonamentals. L'Apèndix B conté les gràfiques complementàries dels experiments duts a terme al Capítol 4. Els apèndixs C, D i E, es refereixen a detalls sobre el Capítol 5. En concret, les gràfiques que complementen l'experimentació realitzada, es recullen a l'Apèndix C. A l'Apèndix D, es resumeix la derivació del problema dual sobre la formulació clàssica. Finalment, l'Apèndix E presenta la derivació de la regla d'actualització de l'aportació original basada en mínims quadrats.

Capítol 2

Fonaments teòrics

Resum – En aquest capítol es presenten els continguts teòrics necessaris per al plantejament del treball. Es comença introduint diferents conceptes sobre l'aprenentatge automàtic i es dóna pas a la vessant supervisada per tal de contextualitzar el problema a tractar. També es presenta la teoria fonamental, així com els conceptes i esquemes que ajuden a l'hora de comprendre l'aprenentatge de distàncies.

Contingut

2.1 Introducció

A mesura que la nostra societat evoluciona a una era post industrial, l'automatització de la producció i la necessitat del maneig i recuperació d'informació són cada vegada més importants. Aquesta tendència ha impulsat l'ús de diverses disciplines científiques relacionades amb les ciències de la computació com ara l'aprenentatge automàtic [62, 80, 8], el reconeixement de patrons [88, 28, 36], la mineria de dades [31, 100] o la intel·ligència artificial [99].

L'aprenentatge automàtic és una disciplina relativament nova, inspirada principalment en la capacitat d'aprendre dels humans. El seu objectiu és la construcció de models que més tard es faran servir en tasques com ara la classificació o la regressió. En general, l'aprenentatge consisteix en la utilització d'un conjunt d'exemples per ajustar inductivament els paràmetres del model de manera que s'obtinga un valor òptim d'una mesura de bondat adequada.

L'aprenentatge automàtic té diferents vessants. Es pot parlar d'aprenentatge supervisat quan tota la informació emprada en el procés d'aprenentatge està etiquetada (i s'utilitza). En un graó intermedi estaria l'aprenentatge semi-supervisat

[13], que fa servir tant la informació etiquetada com sense etiquetar per a realitzar el procés d'aprenentatge. Quan pel contrari s'assumeix la no existència d'informació d'etiquetes de classe s'està fent un aprenentatge no supervisat [39]. En aquest cas l'objectiu és modelitzar les dades d'entrada directament, en lloc de tractar de predir les etiquetes corresponents.

Existeixen mètodes d'aprenentatge que tenen una gran dependència quant a la distància que relaciona les mostres. Com a exemples podem citar l'algorisme dels $k-$veïns [27, 20], en el cas de l'aprenentatge supervisat; o l'agrupament per les $k-$mitjanes [58], en el cas no supervisat. L'efectivitat d'aquests mètodes depèn directament de com d'adequada siga la distància en el context del problema al qual s'enfronta. Per tant, apareix la necessitat d'aprendre, de manera automàtica i a través de les mostres, una distància que satisfaça les necessitats particulars de cada problema. En aquesta tesi ens centrarem en l'estudi de l'aprenentatge supervisat de distàncies, encara que també es poden trobar treballs en la literatura en la línia de l'aprenentatge no supervisat [1], semi-supervisat [56], o aplicats a problemes de regressió [57].

2.2 Aprenentatge automàtic

En aquesta secció s'introdueix de manera general el problema d'aprendre un model, revisant nocions relacionades amb la teoria de l'aprenentatge estadístic [91, 93, 9].

L'objectiu de l'aprenentatge supervisat pot resumir-se com l'obtenció d'un model de manera automàtica mitjançant un conjunt de mostres etiquetades que s'anomena comunament *conjunt d'entrenament*. Aquest model serveix per a realitzar prediccions sobre noves mostres sense etiquetar. Aquesta noció de conjunt d'entrenament es formalitza a continuació juntament amb altres definicions i formalitzacions al voltant de l'aprenentatge supervisat.

2.2.1 Definicions i mesures de risc

Per a centrar la problemàtica a tractar, primer presentarem les bases de l'aprenentatge des del punt de vista de la classificació que consisteix en assignar valors d'etiqueta (pertanyents a un conjunt d'eixida) a elements d'un conjunt d'entrada (o original). Per a simplificar l'exposició se suposa en aquesta secció el cas de la classificació binària sobre el conjunt d'entrada de dimensió d, $\mathbb{R}^{\mathrm{d}}$, de forma que el conjunt d'eixida és $\mathbb{I} = \{-1, 1\}$.

L'objectiu de l'aprenentatge automàtic és trobar un model que donat un element $x_i \in \mathbb{R}^{\mathrm{d}}$, siga capaç de produir una predicció $h(x_i) \in \mathbb{R}$ que finalment es correspon amb un valor en $\mathbb{I}$ mitjançant la funció signe $\mathrm{sgn}(h(x_i))$. Fent un abús de notació ens referirem tant al valor real com al valor en $\mathbb{I}$ amb la mateixa expressió, $h(x_i)$. De manera general, el predictor h (també anomenat hipòtesi),

$$h : \mathbb{R}^{\mathrm{d}} \longrightarrow \mathbb{R}, \tag{2.1}$$

es defineix com una aplicació amb domini l'espai original $\mathbb{R}^{\mathrm{d}}$ i codomini l'espai d'eixida $\mathbb{R}$.

El conjunt d'hipòtesis que cal considerar per a un problema particular rep el nom d'*espai d'hipòtesis*, $\mathcal{H}$.

Conjunt d'entrenament

S'ha fet servir amb anterioritat del concepte de conjunt d'entrenament, i el definim formalment com el següent

$$\mathcal{M} = \left\{ z_i = (x_i, c_i) \mid x_i \in \mathbb{R}^{\mathrm{d}}, c_i \in \mathbb{I}, 1 \leq i \leq n \right\}, \tag{2.2}$$

de grandària $n \in \mathbb{N}$. En particular, $\mathcal{M}$ està format per n observacions independents i idènticament distribuïdes (i.i.d.) d'acord amb una distribució (potser desconeguda) $p(z_i)$, sobre l'espai $\mathbb{R}^{\mathrm{d}} \times \mathbb{I}$. Donada una observació $z_i \in \mathcal{M}$, $x_i \in \mathbb{R}^{\mathrm{d}}$ es refereix a la instància (o exemple) i $c_i \in \mathbb{I}$ a la seua etiqueta associada. De vegades, el parell z_i s'anomena instància etiquetada.

Aprenentatge supervisat

L'aprenentatge supervisat consisteix en construir una hipòtesi $h_{\mathcal{M}}$

$$h_{\mathcal{M}} \ : \ \mathbb{R}^{\mathrm{d}} \longrightarrow \mathbb{R}, \tag{2.3}$$

a partir del conjunt d'entrenament $\mathcal{M}$. Aquesta funció, $h_{\mathcal{M}}$, es pretèn que siga la que prediu de manera òptima l'etiqueta no només dels elements de $\mathcal{M}$, sinó de qualsevol altre sempre que estiga distribuït segons p.

Funció de pèrdua (*loss*)

Per a obtindre una hipòtesi $h_{\mathcal{M}}$ adequada, és necessari un criteri que en mesure la qualitat. Açò es pot fer mitjançant una funció no negativa, ℓ, anomenada pèrdua (en anglès *loss*), que es defineix com a funció del resultat de la predicció $h(x_i)$ i de la classe c_i, de x_i, de la manera següent

$$\ell \ : \ \mathbb{R} \times \mathbb{I} \longrightarrow \mathbb{R}^{+} \cup \{0\}, \tag{2.4}$$

de manera que

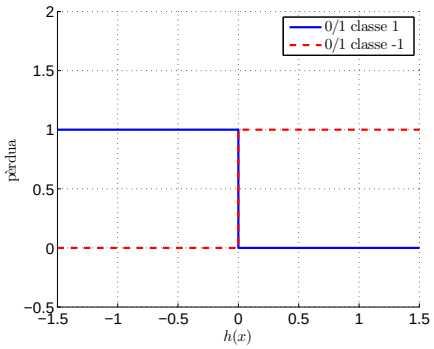
$$\ell(h, z_i) = \ell(h, x_i, c_i) = \ell(h(x_i), c_i). \tag{2.5}$$

Prenent x_i com una mostra qualsevol i c_i la seua corresponent etiqueta vertadera de classe, la funció pèrdua mesura el grau de "satisfacció" entre la predicció $h(x_i)$ i c_i.

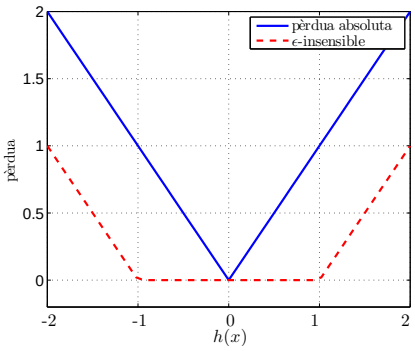
En la Figura 2.1 s'il·lustren algunes de les funcions de pèrdua més emprades en tasques de classificació i regressió. En 2.1(a) apareix la pèrdua 0/1 per al cas de la classificació binària. Aquesta és 0 si s'encerta i 1 si es falla en la predicció de l'etiqueta i pot escriure's de la manera següent

$$\ell_{0/1}(h, z_i) = \ell_{0/1}(h, x_i, c_i) = \begin{cases} 1, & \text{si } c_i h(x_i) \leq 0, \\ 0, & \text{en altre cas.} \end{cases} \tag{2.6}$$

En 2.1(b) s'il·lustren altres mesures de pèrdua per a la regressió com són la pèrdua absoluta (el valor absolut entre la predicció i el valor desitjat) o la pèrdua $\epsilon-$insensible [92] (per al cas $\epsilon = 1$).



(a) Classificació amb 2 classes



(b) Regressió

Figura 2.1: Funcions de pèrdua.

Risc vertader (*true risk*)

El risc vertader, també anomenat com error de generalització, $R^{\ell}(h)$, d'una hipòtesi h, respecte d'una funció de pèrdua ℓ, es defineix com el valor esperat de la pèrdua comesa per h donada la distribució de mostres, p. Pot expressar-se formalment com a

$$R^{\ell}(h) = \mathbb{E}_{z\sim p}\left[\ell(h,z)\right]. \tag{2.7}$$

El risc esperat amb la pèrdua 0/1, $R^{\ell_{0/1}}(h)$ es correspon amb la probabilitat d'error del predictor o error de generalització.

En general, el risc vertader és la mesura més adequada per estimar la bondat d'una hipòtesi. El problema per fer-la servir és que normalment no es té tot el coneixement sobre l'estructura estadística del problema, que ve donada per la probabilitat conjunta, p.

Risc empíric (*empirical risk*)

Donat el conjunt d'entrenament $\mathcal{M}$, una hipòtesi h i una funció de pèrdua ℓ, el risc empíric, $R^{\ell}_{\mathcal{M}}(h)$, es defineix com la mitjana de la pèrdua comesa per h sobre tots els elements de $\mathcal{M}$:

$$R^{\ell}_{\mathcal{M}}(h) = \frac{1}{n}\sum_{i=1}^{n}\ell(h,z_i). \tag{2.8}$$

Aquesta mesura sempre pot ser calculada, i pot considerar-se una estimació del risc vertader. No obstant això, la seua principal deficiència és que depèn de què els mesuraments realitzats sobre el conjunt $\mathcal{M}$ siguen extensibles al cas general.

Risc estructural (*structural risk*)

Donat un espai d'hipòtesis $\mathcal{H}$, és possible definir una mesura concreta de complexitat basada en la dimensió de Vapnik-Chervonenkis associada a cada un dels seus

elements [92]. El risc estructural [93, 92] agrupa la capacitat d'ajust del model i la seua complexitat donat un conjunt d'entrenament finit. L'ús del risc estructural és interessant per que és possible relacionar-lo directament amb la capacitat de generalització de les hipòtesis.

2.2.2 Criteris d'optimització

Aquesta secció se centra en diferents enfocaments clàssics per a trobar hipòtesis òptimes sota diferents consideracions.

La minimització del risc empíric sobre totes les hipòtesis seria una bona estratègia en el cas en què disposem d'infinites mostres i.i.d. segons p. Per contra, en les situacions dels problemes reals, el conjunt d'entrenament $\mathcal{M}$, està limitat (sol ser finit). Llavors, existeix una hipòtesi trivial h, i moltes altres en funció del conjunt $\mathcal{H}$, que realitza prediccions sobre $\mathcal{M}$ sense cometre cap error, és a dir, $R^{\ell}_{\mathcal{M}}(h) = 0$, però que no té la capacitat de generalitzar. És a dir, h té un risc vertader no zero (potencialment gran). Aquesta situació, en la qual $R^{\ell}(h) >> R^{\ell}_{\mathcal{M}}(h)$, es coneix com a sobreentrenament o sobreajustament (en anglès *overfitting*) [76]. La idea intuïtiva al voltant d'aquesta situació és que aprendre sobre el conjunt d'entrenament d'una manera excessiva no implica necessàriament una bona generalització per a la resta de dades.

El sobreajustament porta normalment a una major complexitat de les hipòtesis associades. Així doncs, existeix un equilibri entre minimitzar el risc empíric i la complexitat de les hipòtesis considerades. En particular, existeixen dues alternatives principals per a evitar el sobreajustament

1. restringir adequadament l'espai d'hipòtesis (minimització del risc empíric),

2. afavorir les hipòtesis "simples", sobre les de major complexitat (minimització del risc estructural i del risc regularitzat).

A continuació, es presenten tres estratègies clàssiques diferents per a trobar hipòtesis amb un valor de risc vertader que siga petit.

Minimització del risc empíric

La principal idea del principi de la minimització del risc empíric (ERM, de l'anglès *Empirical Risk Minimization*), és fixar un espai d'hipòtesis $\mathcal{H}$ i seleccionar finalment la hipòtesi que minimitza el risc empíric, és a dir, ser capaç de resoldre el següent problema

$$h_{\mathcal{M}} = \arg\min_{h \in \mathcal{H}} R^{\ell}_{\mathcal{M}}(h). \tag{2.9}$$

Aquest plantejament és efectiu en la pràctica, però depèn de l'elecció del conjunt d'hipòtesis (per exemple classificadors lineals, etc). Essencialment, resulta desitjable que $\mathcal{H}$ siga suficientment gran com per a incloure les hipòtesis amb risc petit, però evitant el sobreentrenament. Sense coneixement previ sobre la tasca, l'elecció d'un $\mathcal{H}$ apropiat resulta dificultós.

Minimització del risc estructural

La minimització del risc estructural (SRM, de l'anglès *Structural Risk Minimization*) és un principi general que permet assolir l'equilibri ideal entre complexitat de les hipòtesis i la seua capacitat d'ajustament a les dades. Per a dur-lo a terme, cal introduir una família de subconjunts niats d'hipòtesis de complexitat creixent. Aquesta estructuració de l'espai d'hipòtesis juntament amb algunes propietats que s'han de satisfer [93] permeten la caracterització de la hipòtesi òptima en el sentit que generalitza el millor possible, donat un conjunt d'entrenament finit. Quan l'espai d'hipòtesis és suficientment senzill, es pot arribar a expressions fàcilment computables que donen lloc a procediments eficients. En el cas particular dels predictors lineals, el principi de minimització del risc estructural implica el conegut principi de maximització del marge [19, 11], que s'utilitza en les màquines de vectors suport (SVM, de l'anglès *Support Vector Machines*).

Minimització del risc regularitzat

En la minimització del risc regularitzat (RRM, de l'anglès *Regularized Risk Minimization*) es tria un espai $\mathcal{H}$ suficientment gran, un regularitzador (habitualment algun tipus de norma sobre la hipòtesi, $\|h\|$) [94] i se selecciona aquella hipòtesi que obté el millor equilibri entre la minimització del risc empíric i la regularització mitjançant un paràmetre d'equilibri C

$$h_{\mathcal{M}} = \arg\min_{h\in\mathcal{H}} R^{\ell}_{\mathcal{M}}(h) + C\|h\|. \tag{2.10}$$

El paper del terme regularitzador és evitar el sobreentrenament mitjançant la promoció de les hipòtesis matemàticament més suaus. La suavitat de les hipòtesis està en certa manera relacionada amb la seua complexitat [19, 80].

Regularitzadors d'ús comú

La regularització s'utilitza en molts mètodes d'aprenentatge exitosos, tant en tasques de classificació com de regressió. Com la minimització de la pèrdua sol ser sotadeterminada (la informació disponible no implica la unicitat de la solució), es pot trobar una solució única si s'afegeix algun terme regularitzador. L'elecció del regularitzador és de vital importància i depèn de la tasca particular considerada i de l'efecte desitjat. A més a més, diferents regularitzadors poden donar lloc a processos d'optimització més o menys senzills. Algunes de les propietats que permeten caracteritzar de manera adequada els diferents regularitzadors són la suavitat i la convexitat.

- La suavitat o derivabilitat d'una funció és una propietat que assegura la no existència de canvis bruscs. A més, s'obtenen desenvolupaments simplificats quan intervé el càlcul de la derivada de manera explícita en la resolució del problema.
- La convexitat és una propietat que, en cas de preservar-se en tots els factors que intervenen en el procés d'optimització, pot assegurar la unicitat de la

solució. Una funció qualsevol

$$f : \mathbb{R}^d \longrightarrow \mathbb{R},$$

és convexa si: $\forall x, y \in \mathbb{R}^d$, i $\forall \theta$ tal que $0 \leq \theta \leq 1$, satisfà que

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y). \tag{2.11}$$

Geomètricament, aquesta inequació pot interpretar-se com que el segment lineal entre $(x, f(x))$ i $(y, f(y))$, es queda "per dalt" de la gràfica de f.

A continuació s'introdueixen alguns dels regularitzadors més comuns per a models basats en matrius i vectors.

- La norma 0, que es correspon amb la quantitat d'elements (del vector o matriu) no nuls. Aquesta norma promou la dispersió de la solució (quantitat de zeros elevada). És un regularitzador no convex i és també no suau [89].
- Norma 1, és la suma dels valors absoluts dels elements (del vector o matriu). És una condició que manté la convexitat i que afavoreix la dispersió. Per contra, no és una funció suau.
- Norma 2 (elevada al quadrat), és la suma de les components al quadrat (del vector o matriu). Aquest regularitzador és convex i suau.
- Traça (d'una una matriu M), és la suma dels elements de la seua diagonal. És convexa, suau i promou les solucions de rang baix.

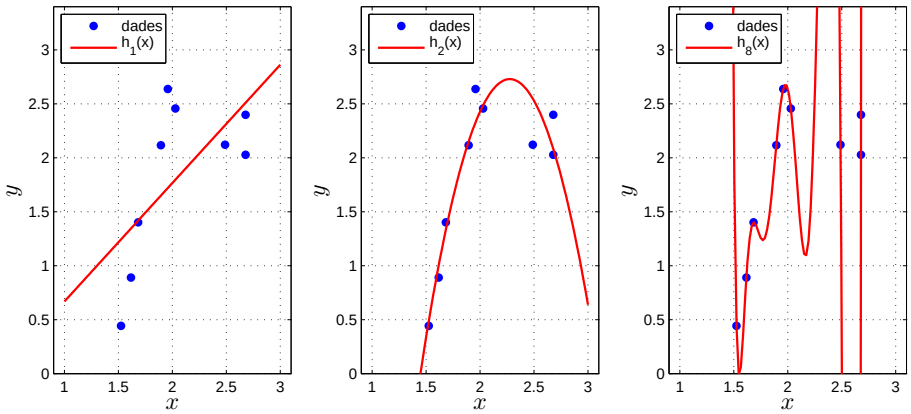
Existeixen també altres regularitzadors com per exemple l'emprat en el treball [103], que utilitza la norma matricial $\|M\|_{2,1}$ (suma de les normes "L_2", de les files/columnes). Aquesta norma és convexa però no és suau.

Exemple **2.1.** *En la Figura 2.2, apareix com exemple una tasca simulada de regressió. En aquest cas particular el conjunt d'entrenament $\mathcal{M}$ està format per parells ordenats de valors $(x, y) \in \mathbb{R}^2$, i s'estudien 3 hipòtesis diferents: h_1, h_2 i h_8 que es corresponen amb polinomis de grau 1, 2 i 8, respectivament. Per això, es considera el valor del risc regularitzat per a cada hipòtesi, i s'utilitza com a risc empíric el valor de la pèrdua absoluta i com regularitzador la norma zero del polinomi (la quantitat d'elements no nuls que conté). Els valors resultants es presenten en la Taula 2.1.*

Hipòtesi	risc $\mathcal{R}^\ell(h)$	Regularitzador $\|\cdot\|_0$
h_1	4.093	2
h_2	1.397	3
h_8	10^{-6}	9

Taula 2.1: Valors de risc i regularitzador de les diferents hipòtesis.

La hipòtesi h_1 (Figura 2.2(a)) no té la suficient llibertat per a poder ajustar-se a les dades i es pot considerar com un cas de sotaajustament. La hipòtesi h_2, (Figura 2.2(b)) s'ajusta amb menor risc que h_1 i possiblement tinga una major capacitat de generalització sobre dades noves que h_8, (Figura 2.2(c)) que representa un cas de sobreajustament (la hipòtesi s'ajusta amb precisió a les dades però pot no generalitzar correctament per a altres valors).



(a) (b) (c)

Figura 2.2: Tasca regressiva sobre les dades representades mitjançant punts. a) polinomi de grau 1, b) polinomi de grau 2, c) polinomi de grau 8.

Funcions de pèrdua

En general, els mètodes d'aprenentatge acaben plantejant la minimització directa o indirecta del risc empíric. No obstant açò, degut a la no convexitat del risc basat en la pèrdua 0/1, la seua minimització és un problema difícil i costós [5]. Açò fa que s'utilitzen altre tipus de funcions de pèrdua a l'hora d'enfrontar-se a aquests problemes. Les eleccions més prominents en el context de la classificació binària són:

- la funció de "pèrdua de frontissa" (o el *hinge loss* en anglès),

$$\ell_{\text{hinge}}(h, z) = \max\left(0, 1 - c \cdot h(x)\right),$$

  emprada per exemple en les màquines de vectors suport [92, 6, 19]. En alguns casos [60, 22], l'*hinge loss* al quadrat, $(\ell_{\text{hinge}})^2$, és una expressió més senzilla de minimitzar.

- La funció de pèrdua *logística* [45],

$$\ell_{\text{logístic}}(h, z) = \ln\left(1 + e^{-c \cdot h(x)}\right).$$

- La funció de pèrdua exponencial [34]

$$\ell_e(h, z) = e^{-c(h(x))}.$$

En la Figura 2.3 s'il·lustren les funcions de pèrdua per a classificació i només per a $c = +1$ descrites amb anterioritat. L'elecció d'una funció de pèrdua apropiada pot ser una tasca difícil i té una forta dependència del problema particular a tractar.

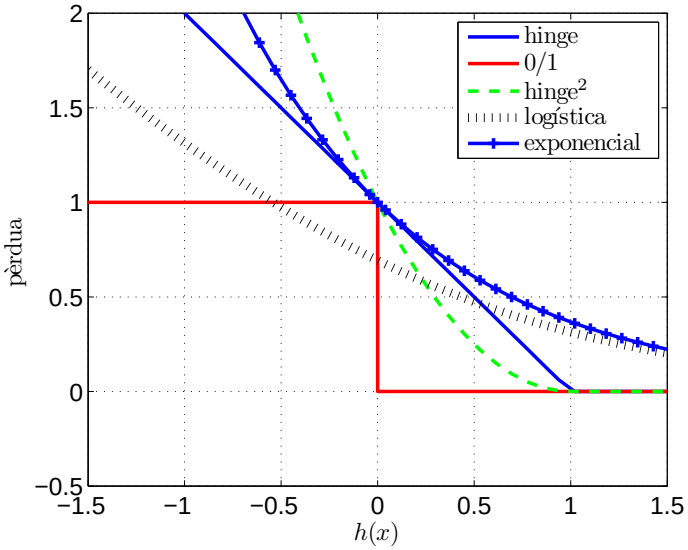


Figura 2.3: Diferents funcions de pèrdua.

2.3 Distàncies i mesures de similitud

En aquesta secció, es presenten les definicions i ferramentes necessàries per a tractar el problema de l'aprenentatge de distàncies supervisat.

En particular, el problema concret amb el qual s'enfronta aquesta tesi està definit sobre un conjunt d'entrenament $\mathcal{M}$, d'acord amb l'Equació 2.2, però a on l'etiqueta de classe està definida sobre el conjunt dels nombres naturals ($c_i \in \mathbb{N}$). La definició de conjunt d'entrenament s'estén al cas de més de dues classes de la manera següent

$$\mathcal{M} = \{(x_i, c_i) \,|1 \leq i \leq n\}.$$

En cas de ser necessari ens referirem a n_k com el nombre d'elements corresponent a la classe k.

2.3.1 Funció de distància

Una funció distància en $\mathbb{R}^d$, o simplement distància, és una funció

$$d : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^+ \cup \{0\}\,. \tag{2.12}$$

Al valor particular de la distància sobre els elements $i-$èssim i $j-$èssim l'abreviarem com $d_{ij} = d(x_i, x_j)$, de manera que $\forall x_i, x_j, x_k \in \mathbb{R}^d$ satisfà les propietats

1. no negativa, $d_{ij} \geq 0$,
2. simètrica, $d_{ij} = d_{ji}$,
3. desigualtat triangular, $d_{ik} \leq d_{ij} + d_{jk}$,
4. unicitat, $d_{ij} = 0 \Longleftrightarrow x_i = x_j$.

Si només satisfà les propietats 1, 2 i 3 anteriors, aleshores d és una pseudodistància en l'espai $\mathbb{R}^d$. En el context de l'aprenentatge de distàncies, és més habitual l'ús de pseudodistàncies.

2.3.2 Distàncies particulars

Distàncies de Minkowski

Les distàncies de Minkowski són una família de funcions de distància parametritzades per un paràmetre $p \geq 1$. Si s'escriuen les coordenades de x_i com $x_i = (x_i(1), \ldots, x_i(d))$ i donats $x_i, x_j \in \mathbb{R}^d$ pot definir-se la distància de Minkowski de la següent manera

$$d_p(x_i, x_j) = \|x_i - x_j\|_p = \sqrt[p]{\sum_{k=1}^{d} |x_i(k) - x_j(k)|^p}. \tag{2.13}$$

La distància de Minkowski és una generalització de la distància Euclidiana ($p = 2$), de la coneguda distància de Manhattan ($p = 1$) i també de la distància de Chebyshev ($p \to \infty$). En la Figura 2.5, apareixen diferents discs unitaris per

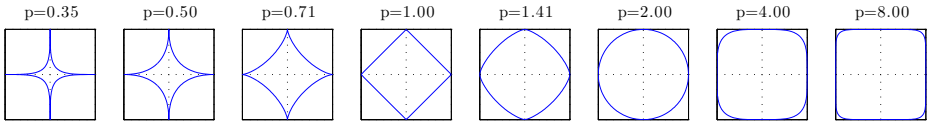


Figura 2.4: Discs unitaris de la distància de Minkowski amb diferents valors de p.

a diferents valors de p definits per la igualtat $\|x_i\|_p = 1$. Cal notar que, quan $0 \leq p < 1$, d_p no és una distància (no satisfà la desigualtat triangular) i la corresponent (pseudo) distància és no convexa. També cal dir que quan $p \to 0$, el disc s'acosta als eixos i quan $p \to +\infty$, el disc unitari és un quadrat.

Distàncies de Mahalanobis

En el nostre treball farem servir pseudo distàncies de Mahalanobis [59], encara que, tal com es fa habitualment en part de la bibliografia, i fent un abús de notació ens referirem a aquestes com a distàncies.

La distància més famosa i utilitzada és l'Euclidiana, que es defineix com

$$d_{ij} = \sqrt{(x_i - x_j)^\top (x_i - x_j)}. \tag{2.14}$$

Una generalització d'aquesta distància és la família de les pseudo distàncies de Mahalanobis, d^M parametritzades per una matriu $M \in \mathbb{R}^{d\times d}$, que anomenarem matriu mètrica o simplement mètrica. Aquesta família de distàncies té la forma

$$d_{ij}^M = d^M(x_i, x_j) = (x_i - x_j)^\top M (x_i - x_j). \tag{2.15}$$

L'Equació (2.15) s'ha introduït sense l'arrel quadrada perquè és com més sovint es troba en la literatura i serà també l'expressió dominant en aquesta tesi. La distància Euclidiana (al quadrat) és un cas particular d'aquesta família de distàncies (2.15) quan $M = I$, on I és la matriu identitat. La distància de Mahalanobis pròpiament dita, es correspon amb $M = \Sigma^{-1}$, a on Σ és la matriu de covariància de les dades.

Relació entre distància i transformació

Una condició necessària i suficient per a que (2.15) siga una pseudo distància és que la matriu M ha de ser quadrada de grandària $\mathbb{R}^{d\times d}$, simètrica amb coeficients reals i amb la propietat de ser semidefinida positiva (PSD). Aquesta condició sol representar-se com $M \succeq 0$. El conjunt de totes les matrius PSD forma l'anomenat *con PSD* [10]. La definició de matriu PSD és

$$M \succeq 0 \Leftrightarrow \forall x \in \mathbb{R}^d \backslash \{\vec{0}\},\ x^\top M x \geq 0. \tag{2.16}$$

Les descomposicions matricials són una ferramenta àmpliament emprada per a simplificar problemes [42], alguns tan coneguts com la resolució de sistemes lineals. Les matrius semidefinides positives es poden descompondre com a $M = W^\top W$, on $W \in \mathbb{R}^{r\times d}$ és una transformació lineal a un subespai lineal de rang r. Cal notar que quan M és de rang baix ($\text{rang}(M) = r < d$), llavors indueix una transformació lineal de les dades a un subespai de dimensió reduïda r, relacionant l'aprenentatge de distàncies amb tècniques de reducció de la dimensionalitat [79, 68, 44]. Fent ús d'aquesta descomposició particular sobre l'Equació (2.15) s'obté

$$d_{ij}^M = (x_i - x_j)^\top W^\top W (x_i - x_j) = (Wx_i - Wx_j)^\top (Wx_i - Wx_j) = d(Wx_i, Wx_j). \tag{2.17}$$

L'Equació (2.17), exhibeix la relació que existeix entre la distància en l'espai original i la distància Euclidiana en l'espai transformat mitjançant la transformació lineal W.

Exemple **2.2.** *Donada la parametrització per a la distància de l'Equació 2.15 següent:*

$$M = \begin{pmatrix} a & t \\ t & b \end{pmatrix},$$

i fixant diferents valors per als paràmetres $a, b, t \in \mathbb{R}$ *anteriors, en la Figura 2.5, es mostra un exemple en el qual es representen diferents discs unitaris centrats en* $(0, 0)$*. En primer lloc (esquerra de la Figura 2.5), tenim el cas* $a = 1$*,* $b = 1$*,* $t = 0$ *(o* $M = I$*). Fent variar significativament el paràmetre* a *i mantenint* $b = 1$*,* $t = 0$ *s'observa la transformació del disc unitari a una el·lipse orientada verticalment. De la mateixa manera, quan fem variar* b*, tenim una el·lipse horitzontal. I a la dreta del tot tenim una el·lipse en diagonal.*

Com s'ha il·lustrat, diferents valors de M *donen lloc a entorns que canvien la geometria de l'espai. De tal manera que l'estudi de tècniques d'automatització sobre la parametrització de la matriu* M *té especial interès a l'hora d'acarar problemes en els quals la geometria existent requerisca ser adaptada d'alguna manera particular.*

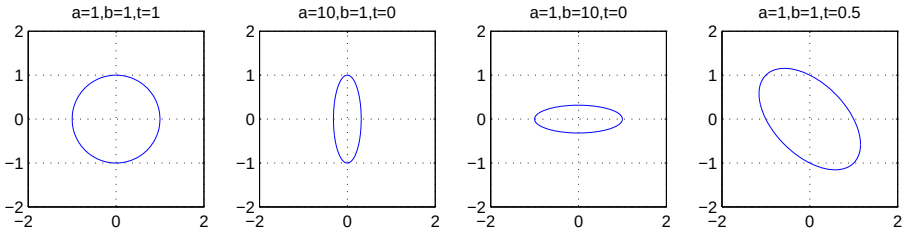


Figura 2.5: Discs unitaris d'acord a distàncies de Mahalanobis, amb diferents valors de M.

2.3.3 Funció de similitud

Mentre que una distància és un concepte matemàtic ben definit, no existeix una clara concordança al voltant de la definició de funció de (dis)similitud que pot ser essencialment qualsevol funció com la següent

$$f : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R} \tag{2.18}$$

Una funció de (dis)similitud sol mostrar un valor gran per a entrades similars i petit per a les dissimilars. Un exemple és una funció kernel [80], que és un tipus especial de funció de similitud que assigna valors reals a parells d'elements.

De forma alternativa a les mesures de distància es poden considerar funcions de puntuació (en anglès *score functions*) [16, 17, 18, 25]. En aquestes funcions, l'important no és el seu valor concret sinó l'ordenació a què aquesta funció pot

donar lloc. Un exemple d'aquest tipus de funció és

$$s_{ij}^{M} = x_i^{\top} M x_j, \tag{2.19}$$

que a diferència de la funció definida en (2.15), no ha de ser PSD, ni tan sols simètrica.

2.3.4 Avaluació de les mesures de distància

El mètode dels $k-$veïns més pròxims ($k-$NN, *k Nearest Neighbours*) [32], forma part d'una família de tècniques d'aprenentatge basat en exemples (*instance-based learning*). L'aprenentatge associat a aquests algorismes es limita a emmagatzemar en memòria els exemples d'entrenament presentats. La idea bàsica sobre la qual es fonamenta aquest paradigma estableix que els membres d'una població solen compartir propietats similars i certes característiques amb els individus que l'envolten. Per tant, la classificació d'una nova mostra es realitza tenint en compte els k elements més pròxims del conjunt d'entrenament d'acord amb una certa mesura de distància.

Aquests algorismes són coneguts com a classificadors de distància mínima. Formalment, una classificació pel veí més pròxim (1-NN) per a un nou element y de classe desconeguda, pot definir-se com:

$$y \in c_i \Leftrightarrow \arg\min_j d(y, x_j) = i \tag{2.20}$$

a on d és la (pseudo)-distància definida en l'espai de representació.

La regla de els $k-$veïns més pròxims està considerat com un mètode fonamental dins dels mètodes d'aprenentatge basats en distàncies. En la Figura 2.6 s'il·lustra el concepte per a diferents valors de k.

Si prenem $k = 1$, l'element al centre de la Figura 2.6, (pintat amb un cercle negre i el símbol (?)) es classificarà en la classe associada a l'element més proper (+). Per a valors de $k \geq 2$, l'element s'assignarà a la classe més representada entre els k elements més propers a la mostra. I en aquest exemple, per al valor $k = 3$ li assigna la classe amb el símbol (*).

2.3.5 Histogrames de distàncies

Resulta difícil definir una manera objectiva de mostrar la bondat de les distàncies. Una possible representació d'aquestes sobre un conjunt concret és útil per mostrar com es comporten. La Figura 2.7, és un exemple d'histogrames de distàncies [50]. Aquests consisteixen en mostrar els valors de distància i la quantitat de vegades que apareixen . En el nostre cas particular, diferenciem entre distàncies d'elements similars i dissimilars. Per tal de facilitar les comparacions, aquests histogrames es normalitzen i representen la distribució dels valors de distàncies tant similars com dissimilars d'una mostra fixada sobre una distància particular.

Una distància ens dóna una manera de comparar objectes. Per tant, la funció hauria de produir valors petits per objectes similars, i més grans per parells dissimilars. Donat un conjunt de parells d'objectes que són etiquetats com similars

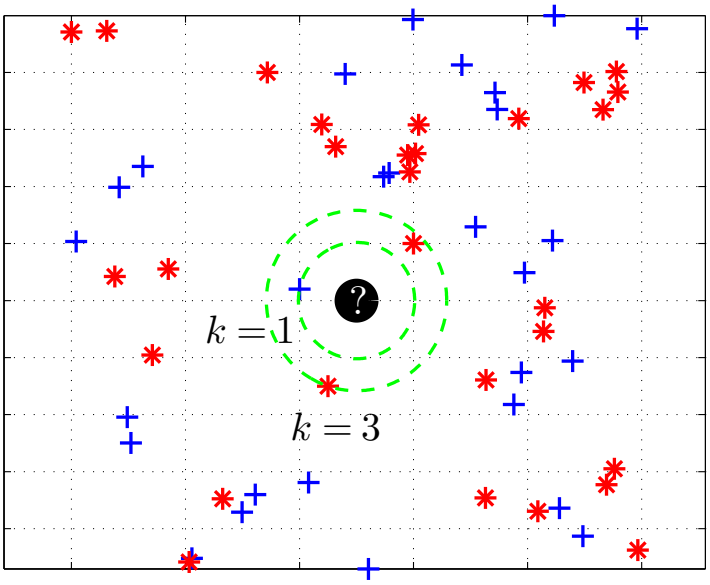
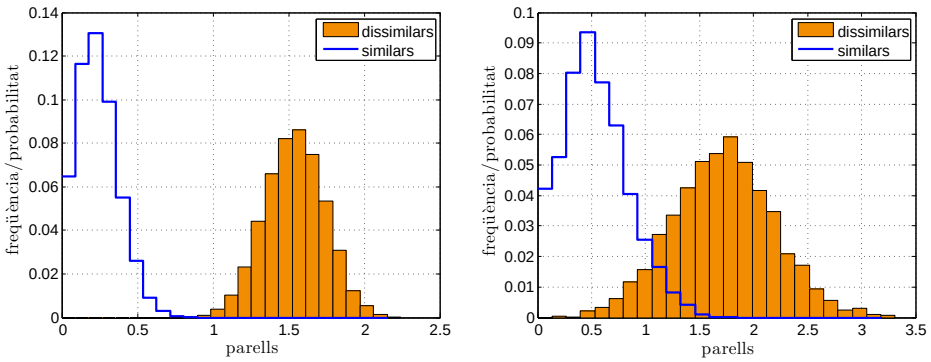


Figura 2.6: Classificador basat en la regla dels k veïns més propers.



(a) (b)

Figura 2.7: Histograma de distàncies normalitzat. Distàncies entre parells d'objectes similars (blau) i dissimilars (taronja) corresponents a Gaussianes bivariades amb separació entre mitjanes de $2,5$. a) cas separable, $\sigma = 0.02$, b) cas no separable, $\sigma = 0.1$.

o dissimilars, podem plantejar el problema d'obtindre la millor funció de distància que és compatible amb aquest etiquetat. Idealment, la funció distància obtinguda ha de mantenir tots els parells similars estrictament més a prop que tots els parells dissimilars. Aquesta situació es mostra a la Figura 2.7a). Quan aquest tipus de separació és possible, en aquesta tesi ens referirem a aquesta configuració com el cas separable. En un cas més realista en què les mostres no poden ser separades completament d'aquesta manera, encara és possible cercar una funció de distància que fa que la majoria dels parells similars siguen més a prop que els parells dissimilars. Ens referim a això com el cas no separable (veure Figura 2.7b).

Aquest tipus de representació ha sigut emprada en les diferents investigacions com a mesura il·lustrativa de la bondat de les diferents distàncies. Encara que no es pot establir com una ferramenta per a comparar objectivament distàncies, si que serà emprada durant la tesi per mostrar idees i conceptes que poden representar-se mitjançant aquests histogrames de distàncies.

2.4 Formulació del problema

De la mateixa manera que per a aprendre un classificador s'utilitza informació a priori en forma d'etiquetes de classe, per a aprendre distàncies es fa servir habitualment informació a priori en forma de restriccions entre parells de punts. Una forma convenient de representar aquestes restriccions és mitjançant una sèrie de conjunts definits sobre els elements de $\mathcal{M}$. Definim el conjunt $\mathcal{S}$ com els índexs associats a parells de mostres del conjunt d'entrenament $x_i, x_j \in \mathcal{M}$ que estan etiquetats com similars, de la manera següent

$$\mathcal{S} = \{(i,j)|x_i, x_j \text{ són similars}\}. \tag{2.21}$$

Anàlogament es defineix el conjunt de parells dissimilars

$$\mathcal{D} = \{(i,j)|x_i, x_j \text{ són dissimilars}\}. \tag{2.22}$$

L'efecte desitjat és que les representacions dels parells d'elements de $\mathcal{S}$ es troben relativament propers, si es comparen amb les representacions dels parells d'elements en $\mathcal{D}$. Açò és equivalent a que les distàncies siguen majors per als elements de $\mathcal{D}$ que per als elements de $\mathcal{S}$.

Formulacions alternatives del problema utilitzen un conjunt de ternes d'elements que contenen informació de similitud relativa referida a un mateix element x_i:

$$\mathcal{T} = \{(i,j,k)|x_i \text{ és més similar a } x_j \text{ que a } x_k\}. \tag{2.23}$$

En aquest treball i com és habitual en la literatura, els conjunts $\mathcal{S}$ i $\mathcal{D}$ solen ser referits com a conjunts similar i dissimilar, respectivament i $\mathcal{T}$ com el conjunt de triplets. Els conjunts de parells/triplets es fixen habitualment tenint en compte relacions conegudes entre els elements del conjunt d'entrenament. És comú formar aquests conjunts a partir d'un conjunt d'entrenament (el conjunt $\mathcal{M}$ en l'Equació 2.2 amb el corresponent $\mathbb{I} = \{1, 2, \cdots, c\}$ multi classe), format per elements etiquetats. En aquest cas, es pot definir el conjunt $\mathcal{S}$ prenent parells

d'elements de la mateixa etiqueta; el conjunt $\mathcal{D}$ prenent parells d'elements de diferent etiqueta; i el conjunt $\mathcal{T}$ prenent triplets d'elements de manera que 2 tenen la mateixa etiqueta i un tercer de diferent etiqueta a la d'aquests. No obstant això, existeixen en la literatura diverses formes de definir aquests conjunts, bé d'una manera aleatòria [24] o amb una noció de veïnatge [96]. En algunes aplicacions la pertinença d'un parell o un triplet concret a aquests conjunts ve donada a la bestreta.

L'aprenentatge de distàncies se sol definir partint d'una interpretació geomètrica, en la qual se cerca un espai de característiques on les restriccions definides se satisfan de manera òptima. A més a més, l'avaluació d'aquestes distàncies sol fer-se en combinació amb un classificador (habitualment un basat en distàncies com el dels $k-$veïns), mitjançant el càlcul de l'error de classificació sobre un conjunt independent al d'entrenament, conegut com conjunt de *test* (o de validació).

Els problemes d'aprenentatge de distàncies es formulen amb el suport d'algun tipus de problema d'optimització matemàtica [10]. Aquesta formulació ve donada per la definició d'alguna funció criteri (o objectiu) i un conjunt de restriccions sobre $\mathcal{S}, \mathcal{D}$ o $\mathcal{T}$ que estableixen les relacions del problema.

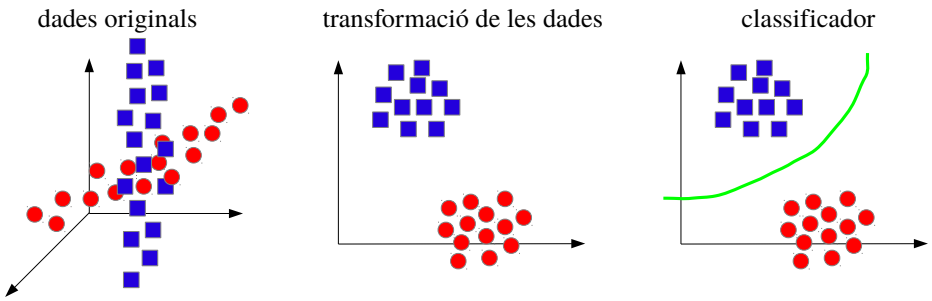


Figura 2.8: Esquema de classificació mitjançant aprenentatge supervisat de distàncies.

En la Figura 2.8, apareix un esquema que resumeix el procés d'aprenentatge automàtic de distàncies. En aquest es poden diferenciar dos fases, primerament les dades d'entrenament etiquetades estan representades en l'espai original, en la primera fase s'aplica una tècnica d'aprenentatge de distàncies amb l'objectiu de trobar una distància que complisca les relacions imposades entre els elements, de manera que la representació obtinguda satisfà de manera òptima les restriccions. En la segona fase, s'aprofita la nova representació de les dades per a entrenar un classificador que permeta poder inferir les classes de nous elements.

L'aprenentatge de distàncies també es pot entendre des d'un punt de vista més local, arribant fins al nivell d'entorns de punts. Dins d'aquesta aproximació resulta interessant que els punts més pròxims siguen de la mateixa classe i també que els punts de diferent classe estiguen sempre més lluny que els de la mateixa classe. En la Figura 2.9 apareix un exemple on inicialment no es dóna aquesta situació, però on resulta útil l'aprenentatge de distàncies per aconseguir que els

elements de la mateixa classe estiguen més propers. L'aprenentatge de distàncies pot adaptar l'entorn de manera que els elements més propers en compartesquen l'etiqueta.

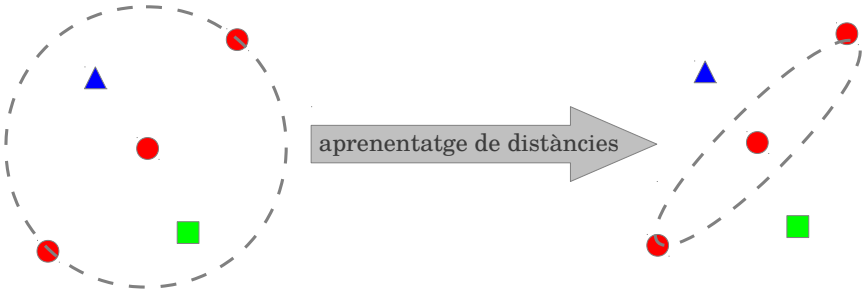


Figura 2.9: Entorn original (esquerra) i entorn idealitzat després d'aplicar aprenentatge de distàncies. La línia discontínua representa punts equidistants al punt central representat amb un cercle en l'espai original.

Capítol 3

Estat de l'art

Resum – En aquest capítol es revisa la literatura relacionada amb l'aprenentatge supervisat de distàncies. Primer s'introdueixen els principals conceptes en aquest tòpic d'investigació. Després, es cobrirà l'aprenentatge de distàncies a partir de vectors de característiques (també denominat aprenentatge de distàncies de Mahalanobis). Finalment, es conclou amb la discussió de les limitacions generals de la literatura actual que motiven el nostre treball.

Contingut

3.1 Introducció

La utilització d'una distància apropiada és un factor clau per al rendiment de molts algorismes d'aprenentatge. L'ajustament manual de distàncies (quan permeten algun tipus de parametrització) per a problemes donats del món real és sovint un problema difícil i tediós. Una gran quantitat de treball està dirigit a l'aprenentatge automàtic a partir de dades etiquetades, la qual cosa porta directament a l'aparició de l'aprenentatge supervisat de distàncies.

En termes generals, els enfocaments d'aprenentatge supervisat de distàncies es basen en la intuïció raonable que una bona funció de similitud (o distància) ha d'assignar un valor gran (respectivament petit) a parells d'elements que pertanyen a la mateixa classe (respectivament diferent classe). Seguint aquesta idea, l'objectiu és la recerca dels paràmetres (habitualment una matriu) d'aquesta distància que millor satisfaça les relacions locals construïdes a partir de la informació sobre el conjunt d'entrenament. Aquestes relacions estan típicament basades en restriccions en la forma de conjunts de parells $\mathcal{S}$, $\mathcal{D}$ o triplets $\mathcal{T}$, que solen construir-se

a partir de les etiquetes de classe dels elements del conjunt d'entrenament. L'aprenentatge de distàncies també pot contemplar-se des de la perspectiva de la reducció de la dimensionalitat. En aquest cas, es pot veure com trobar un nou espai de característiques per a les dades on les restriccions se satisfan en la seua majoria.

El nombre de mètodes diferents d'aprenentatge de distàncies és molt gran i variat. Hi ha algunes revisions recents que en fan una anàlisi exhaustiva [51, 102, 95]. En el present treball, no s'han considerat algunes aproximacions que no estan directament relacionades amb les aportacions presentades com per exemple els treballs de R. Jin [47] o algunes adaptacions d'algorismes més generals com per exemple [53, 66]. Entre els treballs considerats es poden distingir algunes aportacions com a treballs de referència entre tota la literatura relacionada amb el camp de l'aprenentatge de distàncies. Cada un dels treballs que es revisa presenta una formulació del problema que es recolza en un formalisme teòric, i exhibeix el seu funcionament al cas pràctic sobre un conjunt de bases de dades públiques. En la seua majoria estan formulats com a problemes d'optimització amb restriccions. Moltes de les idees contingudes en aquests treballs han motivat les aportacions originals d'aquesta tesi.

L'organització del capítol segueix l'estructura que es descriu a continuació. En la Secció 3.2, es descriuen mètodes d'aprenentatge de distàncies. En la Secció 3.3, es recullen els mètodes en línia, necessaris quan l'aprenentatge s'enfronta a problemes en els quals les dades estan disponibles de forma seqüencial. Finalment, en la Secció 3.4, es discuteix sobre l'estat actual del camp i algunes de les característiques que presenta l'aprenentatge de distàncies.

3.2 Mètodes d'aprenentatge de distàncies

En aquesta secció es revisen els mètodes amb la característica comú de realitzar un aprenentatge a través del conjunt d'entrenament, assumint que estiga disponible en la seua totalitat.

L'esquema de la Figura 3.1 il·lustra el procés d'aprenentatge de distàncies. El conjunt d'entrenament etiquetat ($\mathcal{M}$) s'utilitza per a generar les relacions de proximitat que determinen les restriccions aplicables al problema particular. Habitualment es defineix un problema d'optimització format per un criteri o objectiu juntament amb les restriccions. Aquest problema dóna una solució que sol ser una matriu M.

A continuació presentarem mètodes d'aprenentatge de distàncies importants en la literatura.

3.2.1 Aproximacions mitjançant agrupament amb informació

Els orígens de l'aprenentatge de distàncies es remunten a alguns treballs anteriors [37, 85, 36, 35, 43, 4], però l'aprenentatge de distàncies s'introdueix per

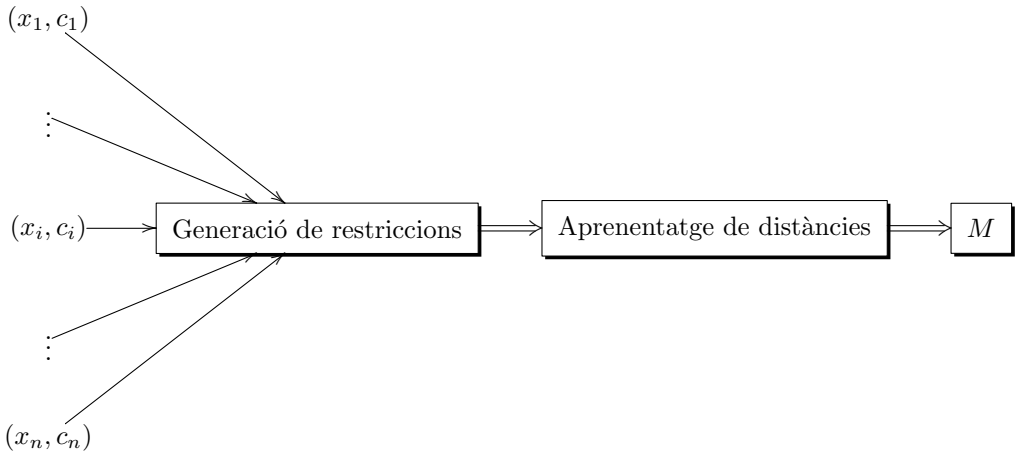


Figura 3.1: Esquema genèric d'aprenentatge de mètriques.

primera vegada tal i com es coneix en l'actualitat en el treball de Xing [101], que es presenta a continuació.

El mètode de Xing [101], es diferencia de les aportacions anteriors per ser el primer a formular el problema de l'aprenentatge de distàncies com un problema d'optimització convex. Aquest tipus de formulació assegura l'existència i la unicitat d'un mínim (o el que és el mateix, l'existència d'un mínim global).

L'objectiu immediat és agrupar els parells del conjunt de parells similars $\mathcal{S}$. Com a exemple particular, en la Figura 3.2 es mostren a l'esquerra dues Gaussianes linealment separables en 3 dimensions. A la dreta s'il·lustra com l'aprenentatge de la distància M pot donar lloc a un espai de característiques en el qual les dades estan agrupades òptimament, en un espai de dimensionalitat inferior.

Més formalment, la idea principal d'aquest mètode és construir un agrupament dels parells d'elements en el conjunt $\mathcal{S}$, de manera que la suma de les seues distàncies siga mínima. A l'hora es defineix una restricció per tal d'evitar la solució trivial $M = 0 \in \mathbb{R}^{d\times d}$. S'exigeix doncs, que la suma de les distàncies dels elements en el conjunt de parells dissimilars siga major o igual que 1. És a dir

$$\begin{aligned} &\min_{M\succeq 0} \sum_{(i,j)\in\mathcal{S}} d_{ij}^M, \\ &\text{tal que} \sum_{(i,j)\in\mathcal{D}} \sqrt{d_{ij}^M} \geq 1. \end{aligned} \tag{3.1}$$

Finalment, els autors [101] formulen un problema equivalent a l'anterior (3.1), bescanviant la funció de la restricció i la funció objectiu, i maximitzant la nova funció criteri. Aquest nou problema s'optimitza mitjançant ascens per gradient sobre la funció objectiu, juntament amb projeccions alternades [30] sobre els conjunts de restriccions següents

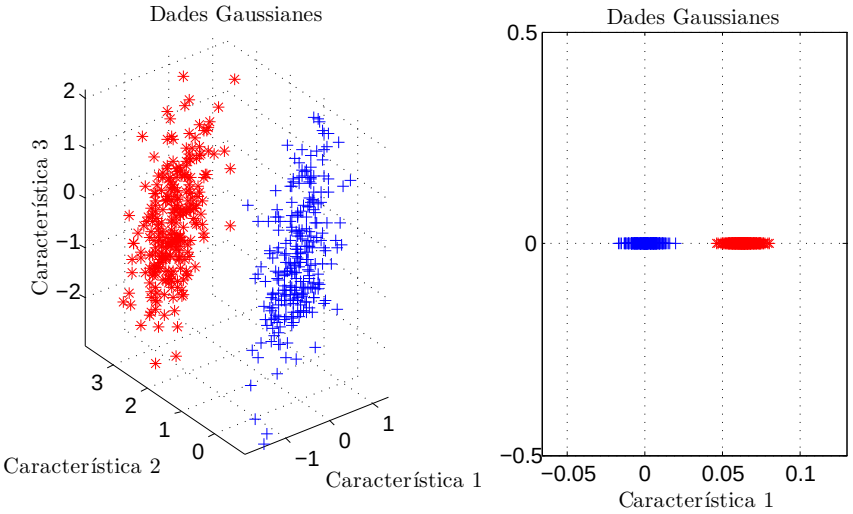


Figura 3.2: Espai original (esquerra) i espai de característiques en què la suma de les distàncies similars és mínima (dreta).

- $R_1 = \left\{M | \sum_{(i,j)\in\mathcal{S}} d_{ij}^M \leq 1\right\}$, que assegura que la suma de totes les distàncies en $\mathcal{S}$ (respecte M), siga menor o igual que 1,
- $R_2 = \{M | M \succeq 0\}$, per a mantenir la condició PSD de M.

Cal destacar que el mètode de Xing ha sigut recentment revisat per Ying i Li [104], proposant la següent modificació del problema a resoldre

$$\begin{aligned} &\max_{M\succeq 0} \min_{(i,j)\in\mathcal{D}} d_{ij}^M, \\ &\text{tal que} \sum_{(i,j)\in\mathcal{S}} d_{ij}^M \leq 1. \end{aligned} \tag{3.2}$$

La principal diferència és que (3.2) maximitza la distància mínima entre parells dissimilars, mentre que el mètode de Xing (3.1), minimitza la suma de les distàncies [101]. Una de les millores és que el problema (3.2) pot resoldre's eficientment calculant només el major valor propi en cada iteració, en contraposició al mètode de Xing que requereix d'una descomposició espectral completa. Un altre mètode relacionat dins del marc de l'aprenentatge de distàncies mitjançant l'agrupament és el de Bilenko et al. [7]. Aquest realitza un aprenentatge semi-supervisat, emprant una quantitat petita de dades etiquetades per a facilitar l'aprenentatge no supervisat. Al seu treball es proporcionen mètodes alternatius per a l'enfocament de l'agrupament i de l'aprenentatge de distàncies. Finalment, els resultats experimentals demostren que un enfocament unificat produeix millors agrupaments.

3.2.2 Col·lapsament de les classes

Independentment del contingut dels conjunts $\mathcal{S}$ i $\mathcal{D}$, l'agrupament es pot considerar des d'un punt de vista individualitzat. És a dir, com a relacions d'un element fix del conjunt d'entrenament, x_i, respecte d'un subconjunt de la resta d'aquests. En els treballs [40, 41] es relacionen els elements x_i amb el conjunt complementari (respecte a $\mathcal{M}$), és a dir $\mathcal{M} \setminus \{x_i\} = \{x_j | j \neq i\}$, mitjançant distribucions de probabilitat associades a cada x_i, que reflecteixen les relacions de cada punt respecte d'aquest conjunt. La intenció és maximitzar el nombre esperat de punts correctament classificats sota aquestes probabilitats, que es refereixen al fet de ser o no similars a x_i.

En el mètode MCML de Globerson et al. [40], s'introdueix la idea intuïtiva de "*col·lapsar les classes*". Aquesta idea consisteix en aconseguir una situació geomètrica on els punts d'una mateixa classe siguen pròxims entre ells, mentre que els de diferent classe se situen a valors de distància arbitràriament grans. Es pot deduir immediatament que es tracta d'una situació idealitzada, i que en la pràctica es podrà tindre només d'una manera aproximada.

La situació de les diferents classes de la Figura 3.3 està en la línia del concepte de col·lapsament de les classes. Però per tractar de visualitzar-la realment, caldria que pensarem en tres punts (cadascun representaria una classe diferent), a distància arbitràriament gran.

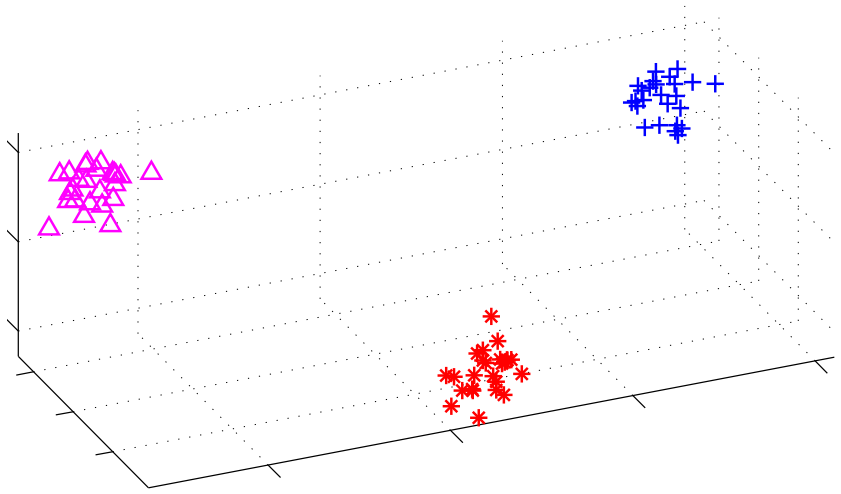


Figura 3.3: Tres classes diferents en què els elements de la mateixa classe estan pròxims i els de diferent classe llunyans.

MCML convex

En aquesta variant del mètode, per a cada x_i, es defineix la distribució condicional sobre tots els elements $j \neq i$, del conjunt d'entrenament. Aquesta distribució de probabilitat es pot entendre com la probabilitat de què x_j siga similar a x_i si el

primer es pren aleatòriament d'acord amb una distribució Gaussiana centrada en x_i i amb covariància M^{-1}, i s'escriu de la manera següent

$$p^M(j|i) = \frac{1}{Z_i} e^{-d_{ij}^M}, \text{ si } j \neq i. \tag{3.3}$$

Cal dir que el terme $Z_i = \sum_{k \neq i} e^{-d_{ik}^M}$, és el valor normalitzador per tal de satisfer la condició $\sum_j p^M(j|i) = 1$.

Es pot considerar el cas ideal en el qual tots els elements en la mateixa classe siguen col·lapsats en un únic punt i, al mateix temps, infinitament allunyats dels punts de classes diferents. Aquesta idea intuïtiva d'agrupament de les classes s'il·lustra gràficament en la Figura 3.3.

En el límit, la situació idealitzada de classes col·lapsades es correspon amb la distribució de probabilitat

$$p_0(j|i) = \begin{cases} \frac{1}{n_k - 1}, & \text{si } c_i = c_j, \\ 0, & \text{si } c_i \neq c_j, \end{cases} \tag{3.4}$$

on n_k és el nombre d'elements de la classe k a què pertany l'element $i-$èssim.

Així doncs, resulta natural cercar la matriu M de manera que $p^M(j|i)$ siga el més pròxim possible a $p_0(j|i)$ per a tot i. Al tractar-se de distribucions de probabilitat, l'objectiu es planteja com la minimització de la suma de les divergències de Kullback-Leibler [52], entre la distribució ideal p_0 i la distribució que s'ha d'aprendre p^M per a cada element $x_i \in \mathcal{M}$:

$$\min_{M \succeq 0} f_{\text{MCML}}(M) = \min_{M \succeq 0} \sum_{i=1}^{n} \text{KL}\left[p_0(j|i) || p^M(j|i)\right]. \tag{3.5}$$

Es pot comprovar [40], que el problema (3.5) és convex. Conseqüentment, pot assegurar-se l'existència d'un mínim global. La funció objectiu finalment es pot escriure equivalentment com

$$f(M) = -\sum_{i,j:c_i=c_j} \log p^M(j|i) = \sum_{i,j:c_i=c_j} d_{ij}^M + \sum_{i=1}^{n} \log Z_i, \tag{3.6}$$

i la seua minimització es du a terme mitjançant descens per gradient i projeccions alternades sobre el con de les matrius PSD.

MCML no convex

La formulació del MCML en funció de la matriu M pot ser considerada alternativament respecte d'una descomposició d'aquesta. En particular es fa servir que la matriu $M \in \mathbb{R}^{d \times d}$ pot escriure's en funció d'altra matriu $W \in \mathbb{R}^{r \times d}$ que satisfà la igualtat $M = W^\top W$, on la matriu W és una transformació lineal a un subespai vectorial de dimensió $r \leq d$. A continuació, es tenen en compte els pros i els contres d'aquesta parametrització.

Com a avantatge, tot el plantejament anterior és vàlid i es pot formular el MCML respecte de W. A més, la utilització d'aquesta nova parametrització no

requereix projectar sobre el con PSD (en combinació amb el descens per gradient), sinó que el fet d'emprar aquesta matriu W assegura la condició de PSD per a M (per construcció).

No obstant açò, l'ús d'aquesta parametrització a través de W fa que el problema siga dependent del rang i la formulació que apareix en l'Equació (3.5) expressada en termes de W resulta ser no convexa. Conseqüentment, l'existència d'un mínim global no pot ser assegurada i el problema passa a tindre mínims locals. Per tant, el procés de minimització és sensible a les condicions inicials i a l'elecció del mètode d'optimització.

3.2.3 Aprenentatge mitjançant maximització del marge

L'aprenentatge de distàncies pot ser també definit des del punt de vista de la maximització del marge [92]. Aquesta aproximació és la que minimitza el risc estructural per al cas dels predictors lineals i resulta interessant la seua utilització per al cas de l'aprenentatge de distàncies.

En la Figura 3.4, podem comparar conceptualment la maximització del marge en les màquines de vectors suport amb la maximització del marge del veïnatge d'un punt en un context d'aprenentatge de distàncies. En les màquines de vectors suport, la idea fonamental és definir un hiperplà de manera que el marge de separació entre les classes siga màxim. Per al cas de l'aprenentatge de distàncies es pretèn satisfer que en l'entorn d'un element concret només apareguen elements de la mateixa classe, i fora d'un entorn de radi major (del mateix element) estiguen elements de diferent classe.

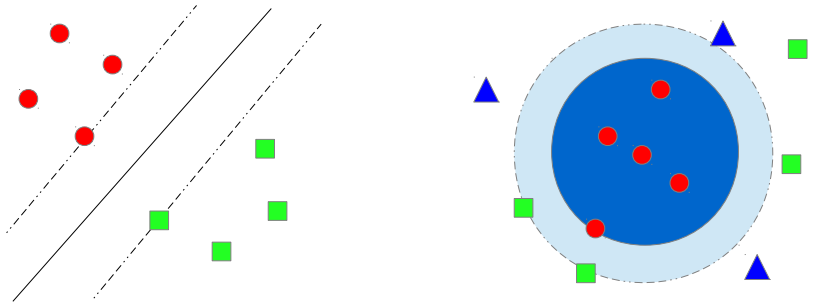


Figura 3.4: Maximització del marge. a) basat en un hiperplà, b) en un context d'aprenentatge de distàncies.

Un altre enfocament consisteix en assignar valors de distàncies petits als elements similars i valors grans als dissimilars. El concepte de gran i petit pot establir-se mitjançant l'ús d'un marge efectiu entre aquests rangs de distàncies i tractar de maximitzar la separació o marge entre aquests. La Figura 3.5, mostra aquest concepte fent servir un histograma de distàncies típic per a aquest cas.

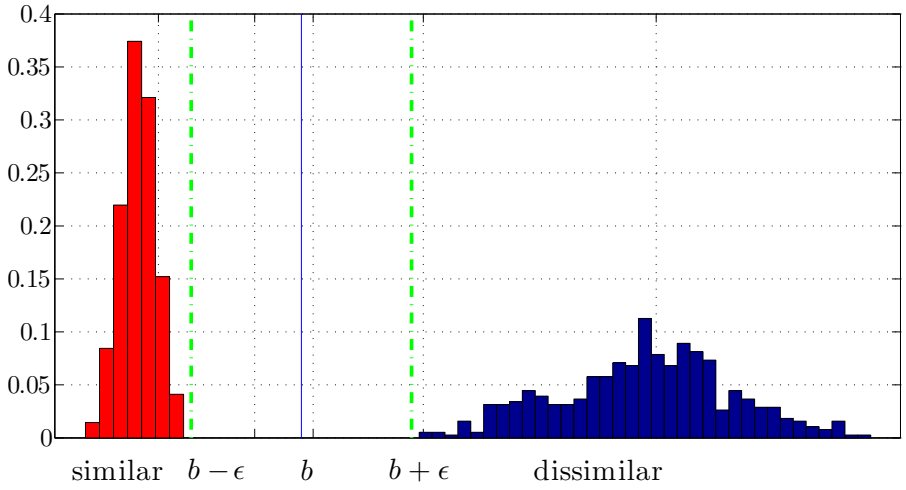


Figura 3.5: Histograma conjunt dels valors de distància etiquetats com a similars i dissimilars separats idealment per un marge d'amplada 2ϵ i centrat en b.

Els valors de les distàncies similars i dissimilars estan separats per un marge de grandària 2ϵ. Les distàncies similars estan fitades superiorment pel valor $b-\epsilon$, mentre que les distàncies entre parells dissimilars queden fitades inferiorment pel valor $b+\epsilon$.

Aquest esquema dóna lloc a la idea intuïtiva de tractar de maximitzar el marge que separa les distàncies similars de les dissimilars. D'aquesta manera, l'aprenentatge de distàncies es converteix en la cerca dels paràmetres que aconseguisquen aquesta maximització.

L'aprenentatge de distàncies sota el paradigma de la maximització del marge pot realitzar-se mitjançant diversos mètodes. A continuació es descriuen dos mètodes representatius. El primer d'ells està basat en l'aprenentatge de distàncies amb una formulació inspirada en les màquines de vectors suport. El segon tracta d'aconseguir la situació d'entorns menuts de la mateixa classe amb marge, com s'ha explicat anteriorment.

Aprenentatge de distàncies amb SVM

En el context dels mètodes de maximització del marge, Nguyen & Guo [64] van presentar el mètode MLSVM (*Metric Learning Support Vector Machine*), en què el problema es formula de manera anàloga al de les conegudes màquines de vectors suport.

Per a obtindre un veïnatge ideal com el de la Figura 3.4 b), on els punts similars estan dins d'un entorn de menor radi que els punts dissimilars, Nguyen & Guo definiren el conjunt dels $k-$veïns més propers a cada element x_i com

$\mathcal{N}^k(x_i)$, plantejant k com un paràmetre d'entrada.

Sota aquesta formulació es tracta d'aconseguir que, per a cada element x_i, la distància a tots els elements de la mateixa classe en el seu veïnatge local $\mathcal{N}^k(x_i)$, siga menor que les distàncies a qualsevol element de diferent classe en $\mathcal{N}^k(x_i)$, amb almenys un marge de valor 1.

Açò es transforma en restriccions del tipus

$$\min_{\substack{x_j \in \mathcal{N}^k(x_i) \\ c_j \neq c_i}} d_{ij}^M \geq \max_{\substack{x_j \in \mathcal{N}^k(x_i) \\ c_j = c_i}} d_{ij}^M + 1. \tag{3.7}$$

Finalment, fent ús del terme regularitzador donat per la norma de Frobenius i introduint les variables de folgança ξ_i, es formula el següent problema amb marge tou [19]:

$$\begin{aligned} \min_{M \succeq 0} \quad & \frac{C}{2} \|M\|_{Fro}^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i, \\ \text{tal que} \quad & \min_{\substack{x_j \in \mathcal{N}^k(x_i) \\ c_j \neq c_i}} d_{ij}^M \geq \max_{\substack{x_j \in \mathcal{N}^k(x_i) \\ c_j = c_i}} d_{ij}^M + 1 - \xi_i, \\ & \xi_i \geq 0, \ \forall i, \end{aligned} \tag{3.8}$$

on C és un paràmetre ($C \geq 0$) que determina el pes de la minimització de la norma de la matriu sobre les restriccions.

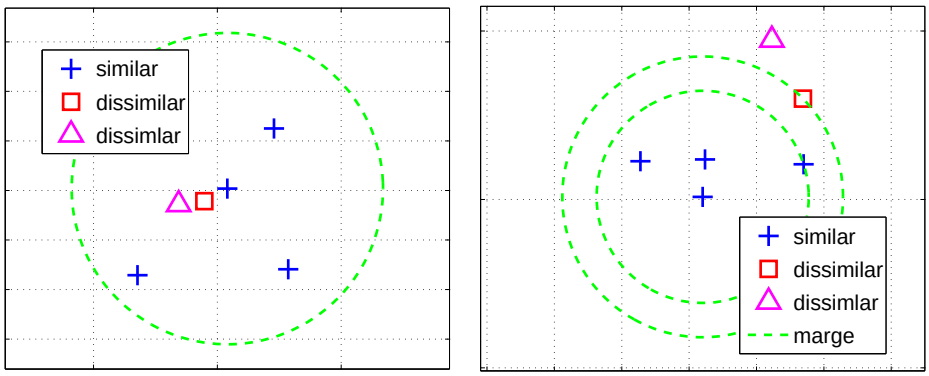
Per a resoldre aquest problema, anomenat MLSVM, s'estén el mètode Pegasos [84] que soluciona la versió primal de les màquines de vectors suport. En el treball es presenten experiments comparatius amb altres mètodes representatius de l'estat de l'art, mostrant-se competitiu en l'avaluació de l'error de classificació sobre bases de dades públiques.

Veïnatge local amb marge (LMNN)

El mètode LMNN (de l'anglès *Large Margin Nearest Neighbor*) és un dels més referenciats en la literatura i ha sigut presentat en els diferents treballs de Weinberger [96] i [97, 98]. El seu objectiu principal és que, per a cada element de la mostra, el seu veïnatge siga de la seua classe alhora que els elements de diferent classe, estiguen a distància major que 1.

La Figura 3.6, recull una situació esquemàtica a nivell d'entorns respecte a un element central (representat al centre de l'entorn amb el signe (+)). En 3.6(a), l'element central té dins del seu entorn elements que són similars, però també conté elements dissimilars. La intenció d'aquest mètode és aconseguir una situació com la de 3.6(b), en què els elements similars estan dins de l'entorn i al mateix temps, els elements dissimilars estan en un entorn de radi major que l'anterior.

Per aconseguir tindre que els k-veïns més propers a cada element siguen de la seua mateixa classe es defineixen els k veïns objectiu. És a dir, per a cada element x_i, de classe c_i, es consideren k elements x_j, $j \in j_1, \ldots, j_k$, tots de classe c_i. Aquests han d'estar a distància mínima de x_i. I per a indicar aquesta relació



(a) (b)

Figura 3.6: Il·lustració esquemàtica del veïnatge de l'element (etiquetat amb el símbol (+) al centre de l'entorn) abans d'entrenar 3.6(a), i després de l'entrenament 3.6(b). La distància està optimitzada de tal manera que, els 3 veïns similars se situen dins d'un entorn de radi petit després de l'entrenament; els elements dissimilars, queden fora d'aquest radi, amb un marge.

de veïnatge desitjat amb x_i, s'introdueix la variable $\eta_{ij} \in \{0, 1\}$. Altra variable binària, β_{ij}, indica quan els elements x_i i x_j pertanyen (o no), a la mateixa classe.

Finalment, es defineix el problema formulant-lo amb amb les pertinents variables de folgança, ξ_{ijl}, com

$$\begin{aligned} \min_{M \succeq 0} \quad & \sum_{ij} \eta_{ij} d_{ij}^M + C \sum_{ijl \in \mathcal{T}} \eta_{ij}(1 - \beta_{il})\xi_{ijl}, \\ \text{tal que } & d_{il}^M - d_{ij}^M \geq 1 - \xi_{ijl}, \\ & \xi_{ijl} \geq 0. \end{aligned} \tag{3.9}$$

El paràmetre C, determina els pes de la minimització del criteri sobre les restriccions. Amb aquesta formulació, les restriccions estan formades per triplets de punts de manera que s'aprenen distàncies relatives entre els punts. Aquest problema (3.9), constitueix un cas particular de SDP (Programació Semi Definida) [96], i pot resoldre's amb metodologies estàndard però els autors presenten el seu propi mètode de resolució.

Cal comentar que al treball [67] es desenvolupa un algorisme alternatiu per a resoldre el problema (3.9). D'altra banda, al treball de Do [26], s'emfatitza la relació entre LMNN i les SVM. En el treball [3], es proposa una versió modificada del LMNN per a enfrontar-se a problemes de regressió. Per últim, els mateixos autors han estès el LMNN dins del context multi tasca, presentat en el treball [65].

3.2.4 Aproximacions basades en teoria de la informació

Divergència de Bregman (ITML)

L'ITML és un treball d'importància proposat en [24]. En ell s'estableix un paral·lelisme entre les distribucions Gaussianes i les distàncies (parametritzades per una matriu) a aprendre dins d'un marc teòric. Per això, els autors introdueixen una formulació del problema amb distribucions Gaussianes i desenvolupen un problema equivalent per a l'aprenentatge de distàncies amb un regularitzador que és una divergència de Bregman. La seua intenció final és separar les distàncies dels parells de punts etiquetats com a similars dels parells dissimilars. Per a això, s'estableix una fita superior, u, i una altra inferior, l, que són valors de referència per a establir les distàncies dels parells similars i dissimilars, respectivament. Aquests valors llindar es fixen prèviament a partir de la mostra d'entrenament. En la Figura 3.7, s'il·lustra aquesta idea mitjançant una representació en forma d'histograma de distàncies.

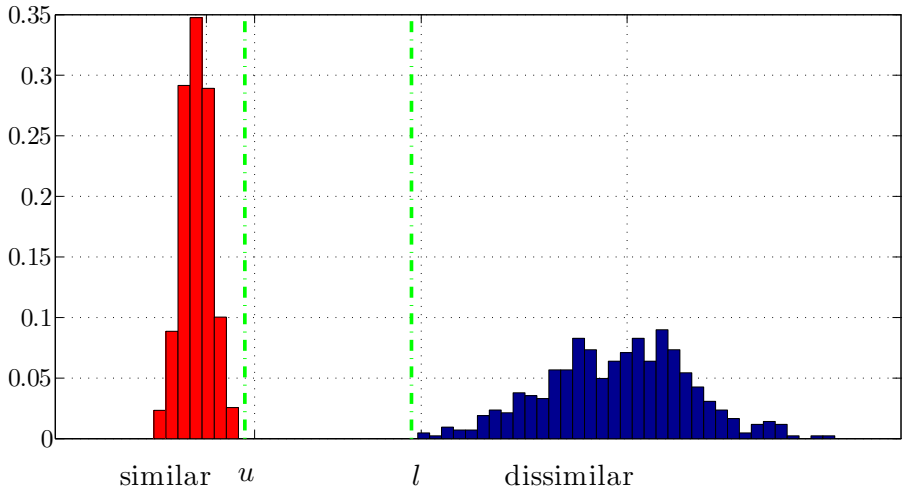


Figura 3.7: Representació mitjançant histograma de distàncies. Les distàncies entre elements dels conjunts $\mathcal{S}$ i $\mathcal{D}$ queden per sota de u i per dalt de l, respectivament. Aquests valors es fan servir com a fites (superior i inferior) dels valors de les distàncies.

ITML fa servir l'existència d'una bijecció entre el conjunt de distribucions Gaussianes multivariades amb mitjana μ i el conjunt de distàncies de Mahalanobis. Així, donada una matriu M, es pot escriure

$$p(x|M) = \frac{1}{Z} e^{-\frac{1}{2} d^M(x,\mu)}, \tag{3.10}$$

on Z és una constant de normalització, i M^{-1} és la matriu de covariància de la

distribució. Emprant la bijecció definida per la funció de probabilitat (3.10), es pot mesurar la distància entre dues distàncies parametritzades per les matrius M i M_0 respectivament, mitjançant la divergència de Kullback-Leibler de la manera següent

$$\mathrm{KL}\left[p(x|M_0)||p(x|M)\right] = \int_{R^d} p(x|M_0) \log\left(\frac{p(x|M_0)}{p(x|M)}\right) dx. \tag{3.11}$$

Aquesta mesura (3.11), proporciona una relació de proximitat entre dues distribucions de probabilitat i és l'objectiu a minimitzar, de manera que deriva en el següent problema amb restriccions (sobre els valors de distància):

$$\begin{aligned} \min_{M} \quad & \mathrm{KL}\left[p(x|M_0)||p(x|M)\right], \\ \text{tal que } & d_{ij}^M \leq u, (i,j) \in \mathcal{S}, \\ & d_{ij}^M \geq l, (i,j) \in \mathcal{D}. \end{aligned} \tag{3.12}$$

Les restriccions del problema anterior són els valors de les distàncies i M_0 és una matriu regularitzadora. La funció objectiu del problema (3.12) pot expressar-se com un tipus particular de divergència de Bregman. En particular, la divergència que es deriva de la funció convexa $\phi(M) = \log\det(M)$, definida sobre el con de matrius positives definides, que genera la divergència de Bregman entre dues matrius $M, M_0 \succ 0$ següent

$$D_{ld}(M, M_0) = \mathrm{Tr}(MM_0^{-1}) - \log\det(MM_0^{-1}) - d, \tag{3.13}$$

on d és la dimensió de l'espai original. Aquesta divergència 3.13, ha sigut emprada també en altres treballs [46, 78]. En la pràctica, M_0 sol ser la matriu identitat I, de manera que la matriu apresa M estiga pròxima (sota la noció de proximitat definida per (3.13)) a I. Aquesta divergència permet preservar de manera eficient la condició de semi definida positiva de la matriu. La funció divergència log det conserva el rang, de manera que si la matriu inicial té rang r, la matriu final apresa tindrà també rang r.

Fent ús de la següent igualtat (que es demostra en [23])

$$\mathrm{KL}\left[p(x|M_0||p(x|M)\right] = \frac{1}{2} D_{ld}(M_0^{-1}, M) = \frac{1}{2} D_{ld}(M, M_0), \tag{3.14}$$

l'Equació (3.14), permet passar del problema 3.12, definit en termes de distribucions Gaussianes al següent problema

$$\begin{aligned} \min_{M \succeq 0} \quad & D_{ld}(M, M_0), \\ \text{tal que} \quad & d_{ij}^M \leq u, \forall (i,j) \in \mathcal{S}, \\ & d_{ij}^M \geq l, \forall (i,j) \in \mathcal{D}. \end{aligned} \tag{3.15}$$

Els valors $u, l \in \mathbb{R}$ són paràmetres del problema. I com és habitual, es resol afegint variables de folgança. L'ITML tracta de satisfer les restriccions dels parells similars i dissimilars mentre alhora tracta d'estar el més prop possible de la matriu regularitzadora, M_0.

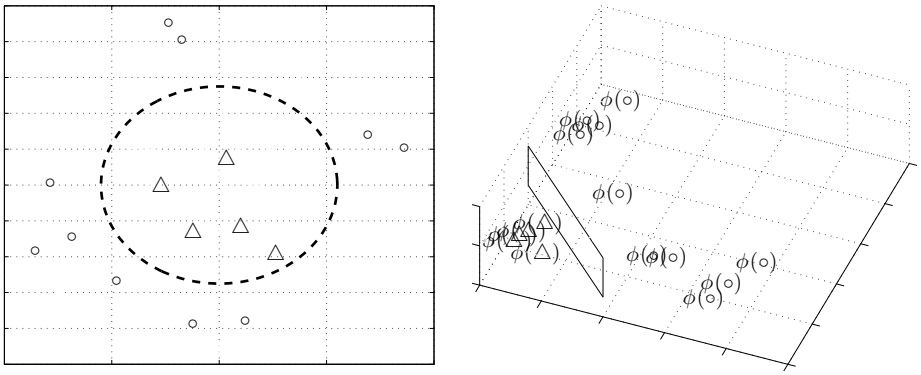
3.2.5 Extensions no lineals dels mètodes

La majoria de les aproximacions d'aprenentatge de distàncies s'han estès al cas no lineal [81, 83, 24, 48, 77, 15, 14] mitjançant la utilització de kernels. La idea fonamental rau a suposar que existeix una funció

$$\phi : \mathbb{R}^d \longrightarrow H \tag{3.16}$$

que transforma les dades a un espai H de Hilbert a on les dades sí són separables linealment [80].

En la Figura 3.8 s'il·lustra aquesta idea: en 3.8(a) les dades estan formades per dues classes representades amb els símbols cercle ($\circ$), i triangle ($\triangle$). Apareixen en un hipotètic espai original (representat amb 2 dimensions per a fer referència a una dimensió més baixa). Com pot observar-se, les dues classes no poden separar-se mitjançant un hiperplà (una recta en aquest cas), sinó que és necessària una funció alternativa (com una circumferència). En 3.8(b), s'ha aplicat ϕ sobre les dades d'entrada, i es representen en un espai de major dimensió que l'original, on les dades si poden ser separades per un hiperplà.



(a) (b)

Figura 3.8: a) Dades en l'espai original, b) dades transformades a H mitjançant ϕ.

Habitualment, no està disponible l'expressió explícita de ϕ per a transformar les dades. Per això, es fa servir una funció anomenada kernel que realitza de manera implícita productes escalars entre els elements transformats en un hipotètic espai de Hilbert segons la següent relació

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \,. \tag{3.17}$$

Mitjançant aquest procediment, és possible l'obtenció de solucions més generals que les distàncies quadràtiques.

3.3 Aprenentatge de distàncies en línia

En els mètodes exposats fins ara, es feia us de totes les mostres del conjunt d'entrenament simultàniament. Quan per les característiques del problema les mostres es generen d'una en una, és necessari utilitzar enfocaments alternatius.

En l'aprenentatge en línia [55], l'algorisme rep mostres d'entrenament d'una en una, prediu la seua classificació i actualitza el seu model, (si és necessari). Encara que l'efectivitat dels algorismes en línia és típicament inferior a la dels algorismes que treballen amb totes les mostres d'un cop (o per lots), aquests són molt útils per a enfrontar-se als punts febles dels algorismes per lots. Per exemple, pot ocórrer que el conjunt $\mathcal{M}$ puga arribar de manera seqüencial, que el nombre d'elements en $\mathcal{M}$ supere les capacitats físiques de la màquina o que l'etiquetat evolucione al llarg del temps, de manera que el concepte de similitud ho faça també. Es tornarà a parlar sobre l'aprenentatge en línia en la Part II d'aquesta tesi, junt amb les aportacions fetes a l'aprenentatge de distàncies en línia.

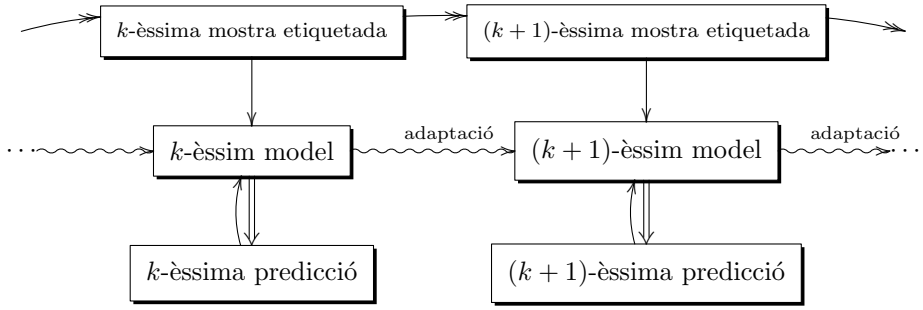


Figura 3.9: Esquema d'aprenentatge en línia.

L'esquema de la Figura 3.9, recull el procés d'aprenentatge en línia. Les mostres arriben de manera seqüencial (o en línia), juntament amb l'etiquetat corresponent. Aleshores, s'utilitza el model disponible en l'instant k, per a inferir l'etiqueta d'aquesta k-èssima mostra. Després, el model s'adapta (si és necessari), en base a l'encert de la k-èssima predicció. Com a resultat, s'obté el nou model que serà emprat per a inferir l'etiqueta de la següent mostra.

Cal dir que en l'esquema anterior es parla d'una mostra en al·lusió tant al cas en què arriba un element únic x_k, com el cas de l'aprenentatge de distàncies en què pot arribar un parell (o triplet) juntament amb la seua etiqueta de relació (similar o dissimilar, per exemple).

Aquest paradigma permet enfrontar-se a problemes de grandària relativament gran (o infinita). L'aprenentatge en línia pot realitzar prediccions i les conseqüents adaptacions d'una en una, minimitzant el cost computacional requerit per a l'actualització (o adaptació) del model.

L'aprenentatge de distàncies en línia pot enfrontar-se a problemes que requereixen d'una major adaptació. Per a evitar que una única mostra puga alterar significativament el model, habitualment estan formulats a partir d'un regularit-

zador que manté la proximitat al model anterior i una condició per tal de satisfer la nova adaptació del model.

A continuació presentarem mètodes d'aprenentatge de distàncies en línia importants en la literatura. Es comença presentant els treballs pioners en l'aprenentatge en línia de distàncies. Després, presentarem la versió en línia de l'ITML.

3.3.1 Aproximacions en línia basades en el còmput de projeccions

El treball de Shalev-Schwartz [83], al 2004 és el primer enfocament en línia d'aprenentatge de distàncies. En aquest s'aprèn la matriu M, així com un valor $b \geq 1$ que actúa com a llindar entre les distàncies similars i dissimilars sota un esquema com el de la Figura 3.5. En cada pas, l'algorisme rep un parell d'elements x_i, x_j del conjunt d'entrenament i una etiqueta de similitud associada $y_{ij} = \pm 1$. Així doncs, per a cada instant k pot descriure's una seqüència similar a la següent

$$\ldots, (x_{i_k}, x_{j_k}, y_{i_k j_k}), \ldots \tag{3.18}$$

La utilització d'aquest índex k només es farà servir quan siga necessari, de manera que cada element dóna lloc a una pèrdua, donada per

$$\ell_{ij}^{(M,b)} = \max\left\{0, 1 - y_{ij}\left(b - d_{ij}^{M}\right)\right\}. \tag{3.19}$$

Finalment, el mètode POLA (*Pseudo-metric Online Learning Algorithm*) basa la seua regla d'actualització en projeccions sobre els conjunts de restriccions següents

- $C_{ij} = \left\{(M, b) \in \mathbb{R}^{n^2+1} | \ell_{ij}^{(M,b)} = 0\right\}$, que assegura que el valor de la pèrdua siga zero.
- $C_a = \left\{(M, b) \in \mathbb{R}^{n^2+1} | M \succeq 0, b \geq 1\right\}$, el conjunt de restriccions sobre la matriu M (que ha de ser PSD) i el llindar, necessàriament major o igual que 1.

Una extensió d'aquest mètode és el treball [63], on s'estudia el mètode POLA amb un terme de regularització afegit a la formulació original.

3.3.2 Aproximacions en línia basades en teoria de la informació

El mètode ITML presentat amb anterioritat, també disposa d'una formulació en línia, sota un esquema de model regressiu [49]. L'algorisme rep en cada instant k, un parell d'elements, x_{i_k}, x_{j_k} juntament amb la distància desitjada d_k. A diferència del cas per lots, que fa servir els valors u i l. La seqüència amb la qual aprén l'algorisme és similar a la següent

$$\ldots, (x_{i_k}, x_{j_k}, d_k), \left(x_{i_{(k+1)}}, x_{j_{(k+1)}}, d_{(k+1)}\right), \ldots \tag{3.20}$$

Després de rebre en l'instant k el parell de punts juntament amb la distància relativa com a condició, l'algorisme fa servir la matriu actual M^k per a predir la distància $\hat{d}_k = d^{M^k}_{i_k j_k}$ entre el parell de punts donats.

En cada pas, l'algorisme minimitza la suma de la regularització log det respecte a la matriu anterior i una funció de pèrdua quadràtica $\ell_k(M^k) = (d_k - \hat{d}_k)^2$, resolent en cada instant el problema d'optimització següent

$$\arg\min_{M \succeq 0} D_{ld}(M, M^k) + \eta_k \ell_k^M \tag{3.21}$$

El paràmetre η_k ajusta l'equilibri entre la proximitat del model i el compliment de la restricció. La distància final sol ser lleugerament pitjor que la versió per lots però l'algorisme pot ser més ràpid.

Jain va presentar el mètode LEGO [46], que consisteix en una variant de l'ITML. És un algorisme en línia que també està basat en un terme de regularització log det que mesura la proximitat entre la matriu i una pèrdua que penalitza valors de distància diferents a l'objectiu. En cada pas de l'algorisme se subministra un parell, que dóna lloc a un nou problema d'optimització que es resol mitjançant descens per gradient. Al mateix treball s'estudia l'ús d'altra funció de pèrdua i es mostren experiments comparant el mètode amb l'ITML per lots/en línia, LEGO i POLA. Els experiments mostren que l'ITML per lots ofereix millors resultats en classificació, seguit per LEGO, ITML en línia i finalment POLA.

3.4 Conclusions

En aquest capítol s'ha tractat d'oferir una visió general dels treballs recents en aprenentatge de distàncies, amb un enfocament particularment basat en els models, extensions i aplicacions. Un dels objectius principals d'aquest estudi ha estat el de presentar les diferents tècniques d'aprenentatge de distàncies en un marc comú. Malgrat que aquesta revisió de mètodes no fa justícia amb tota la literatura sobre l'aprenentatge de distàncies, si que s'ha presentat un subconjunt de mètodes representatius de l'estat de l'art.

L'aprenentatge de distàncies promet seguir sent un camp ric d'investigació. Mentre els mètodes lineals semblen ser bastant ben estudiats i compresos, els mètodes no lineals segueixen atraient noves investigacions i idees. En particular, l'ampliació de mètodes no lineals per a l'aplicació sobre grans conjunts de dades segueix sent un problema difícil, i el desenvolupament de bases teòriques sòlides per als mètodes que impliquen distàncies locals roman oberta.

Les aplicacions d'aprenentatge de distàncies continuen apareixent, i s'espera que continue avançant. Si bé l'aprenentatge de distàncies ha estat molt utilitzat per a tasques de visió per computador, les aplicacions segueixen apareixent en biologia, música, multimèdia, etc. S'espera que aquest tipus d'aplicacions ajudaren a impulsar un major desenvolupament teòric i algorítmic per a l'aprenentatge de distàncies.

Part II

Aportacions

Capítol 4

Aprenentatge basat en el col·lapsament de classes mitjançant un conjunt d'elements representatius

Resum – El col·lapsament de les classes resulta ser una aproximació efectiva per a l'aprenentatge de distàncies. El MCML és un mètode representatiu que es planteja al voltant d'aquesta idea geomètrica. En aquest capítol es tracta de generalitzar el mètode utilitzant només alguns elements en l'espai original, en lloc de tots dels elements disponibles. Aquests elements, que anomenem punts base, actuen com a representants i es fan servir com a elements de referència per a establir relacions de proximitat. La seua utilització, atorgarà millores computacionals preservant l'efectivitat de les solucions del mètode original.

Contingut

4.1 Introducció

Al Capítol 3.2.2, es va presentar el treball de Globerson & Roweis, [40], on s'introdueix un algorisme per a l'aprenentatge de distàncies basat en el col·lapsament de

les classes. En general, els mètodes de col·lapsament de les classes intenten aconseguir una situació geomètrica ideal fonamentada en el fet que una bona distància és aquella que fa que els elements de la mateixa classe estiguen propers i simultàniament a major distància dels elements de diferent classe. Per aconseguir-ho, es formula el problema convex d'optimització (3.5), que està definit sobre les distribucions de probabilitat $p^M(j|i)$ (parametritzades per una matriu M) i $p_0(j|i)$ (la distribució idealitzada que representa la situació de col·lapsament de les classes). El problema original es resol minimitzant la suma de les divergències de Kullback-Leibler entre les distribucions p_0 i p^M definides per a cada element del conjunt d'entrenament.

En la resta del capítol es profunditza en el mètode i es descriu una proposta pròpia que condueix a una millora computacional del mateix. En la Secció 4.1.1, s'il·lustra el mètode utilitzant dades Gaussianes generades sintèticament amb la intenció fonamental de mostrar diferents representacions per a les distribucions p_0 i p^M associades a un problema concret. En la Secció 4.2, s'introdueix el concepte de punts base i la seua aplicació directa sobre el MCML. Per a deixar clar el concepte, es repeteixen alguns dels exemples emprant els punts base. En la Secció 4.3, es presenten diverses particularitzacions dels mètodes, en concret definint diferents inicialitzacions i tècniques d'elecció dels punts base. En la Secció 4.4, es desenvolupa l'estudi teòric al voltant del cost computacional del MCML i de les propostes que fan servir punts base a través d'un algorisme comú. En la Secció 4.5, es presenta una àmplia bateria d'experiments que comparen les diferents propostes prenent al MCML com a mètode de referència. Finalment, en la Secció 4.6 es realitza una petita discussió al voltant dels resultats obtinguts i es realitzen algunes recomanacions sobre els mètodes.

4.1.1 Interpretació gràfica

Per tal d'il·lustrar les distribucions de probabilitat $p_0(j|i)$ i $p^M(j|i)$ utilitzades en la formulació del problema, s'ofereixen tres exemples amb diferents situacions específiques generades de manera sintètica, utilitzant mostres particulars de distribucions Gaussianes. Els conjunts sobre els quals es desenvolupen els exemples estan formats per tres classes. Cada una d'aquestes està constituïda per una Gaussiana diferent. Les Gaussianes generades tenen una matriu de covariància proporcional a la matriu identitat i per a construir les diferents situacions es varia la distància entre les mitjanes de cada classe. Les tres situacions que s'han seleccionat són:

- Classes col·lapsades: es correspon amb la situació idealitzada en la que les tres classes estarien col·lapsades en tres punts diferents infinitament allunyats entre ells (es representarà de manera aproximada en l'exemple).
- Classes separables: aquesta situació es refereix a que les tres classes són linealment separables entre elles, és a dir, existeixen hiperplans que separen les classes dues a dues.
- Classes solapades: en aquesta situació no és possible separar les classes de manera lineal dues a dues.

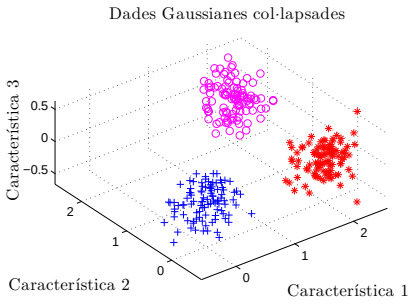
4. Aprenentatge basat en el col·lapsament de classes mitjançant un conjunt d'elements representatius

Suposant que la matriu M és coneguda i prenent les dades organitzades segons l'etiqueta de classe, es consideren per a les distribucions $p_0(j|i)$ i $p^M(j|i)$ les següents representacions:

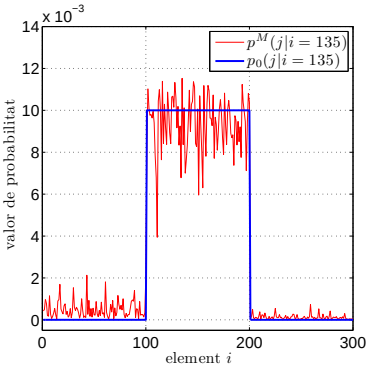
- Matricial, mitjançant una matriu que recull en la posició fila j, columna i, el valor en l'escala de grisos associat a la probabilitat corresponent a x_i, x_j.
- Gràfica (d'una funció d'una variable). Es pren un element concret de $\mathcal{M}$ i es considera com a variable la resta d'elements. Sota aquesta representació, p_0 apareix com una funció esglaonada i p^M es pot interpretar com una aproximació a p_0 amb una certa quantitat de soroll.

La representació matricial il·lustra de manera global una distribució de probabilitat discreta (p_0 o p^M) sobre les dades, i la representació gràfica il·lustra els valors de probabilitat d'un element concret contra tots els elements d'entrenament. Aquestes representacions es fan servir per a il·lustrar els exemples que es presenten a continuació.

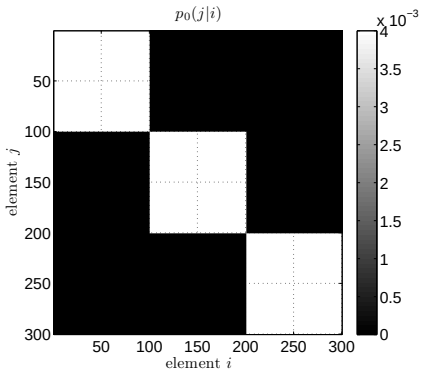
Exemple **4.1.** *La Figura 4.1 il·lustra la situació (aproximada) de col·lapsament de les classes juntament amb les representacions adients. Les dades originals estan representades en 4.1(a), i la seua situació permet discriminar amb claredat les diferents classes. Les representacions com a funció d'una variable de* $p_0(j|i = 135)$ *i* $p^M(j|i = 135)$ *de l'element* $x_{135} \in \mathcal{M}$ *(*$i = 135$*, que pertany a la classe 2 representada amb el símbol (+) en roig) apareixen en 4.1(b). En particular, els valors de* $p^M(j|i = 135)$ *estan relativament pròxims dels de* $p_0(j|i = 135)$*. Els valors de probabilitat són majors per als elements de la classe 2, la qual cosa indica una major similitud amb aquesta etiqueta. Quant a les representacions matricials,* p_0 *apareix en 4.1(c) i* p^M *en 4.1(d). El color blanc en els blocs matricials de la diagonal indiquen valors normalitzats de probabilitat pròxims a 1, que és correspon amb el color blanc, mentre que els que no estan en la diagonal són més baixos i pròxims a 0 (que es correspon amb el color negre).*



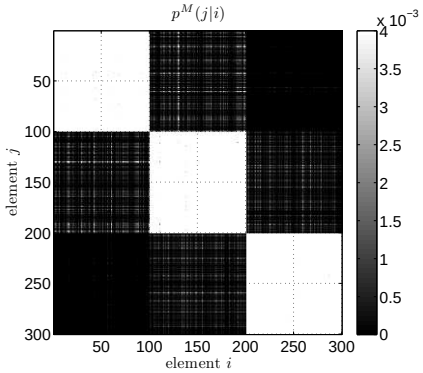
(a) Dades sintètiques generades



(b) Distribució ideal i real associada a un element de la classe cercle



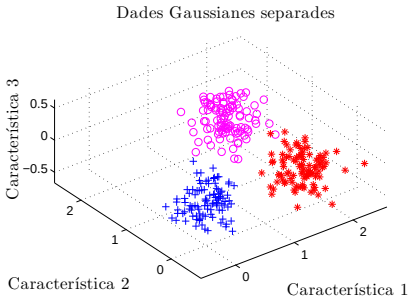
(c) Matriu p_0 que s'ha d'aproximar



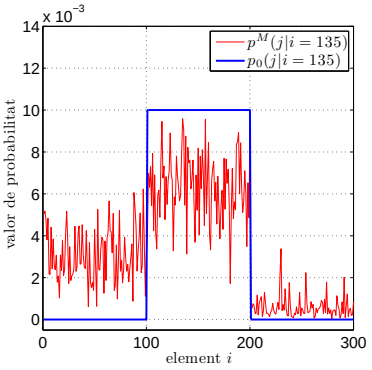
(d) Matriu amb la distribució de les dades sintètiques

Figura 4.1: (a) situació (aproximada) de col·lapsament de les classes; (b) distribucions p_0 i p^M associades l'element x_{135} respecte a $\mathcal{M}$; (c) matriu representada en escala de grisos p_0; (d) matriu p^M associada a la mostra.

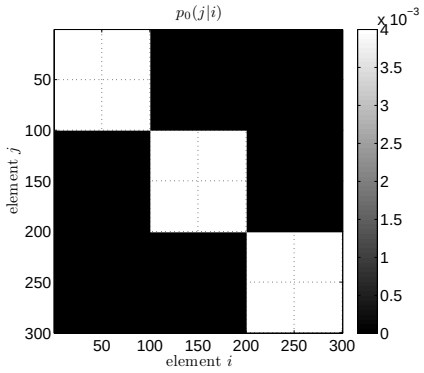
Exemple 4.2. *En la Figura 4.2 s'il·lustra la situació de separació de les classes i les representacions gràfiques de les distribucions p_0 i p^M corresponents. Les dades originals estan representades en 4.2(a), i la seua situació encara permet discriminar a mitjançant hiperplans les diferents classes 2 a 2. Les representacions com a funció d'una variable de $p_0(j|i = 135)$ i $p^M(j|i = 135)$ de l'element $x_{135} \in \mathcal{M}$ $(i = 135)$ apareixen en 4.2(b). Els valors de probabilitat de p^M han empitjorat respecte del cas de col·lapsament a pesar d'estar encara en una situació de separació de les classes. Quant a les representacions matricials, p_0 apareix en 4.2(c) i p^M en 4.2(d). El paregut entre elles ha empitjorat respecte de la situació de col·lapsament. Encara que és relativament bo, existeix confusió entre els elements de la mateixa classe.*



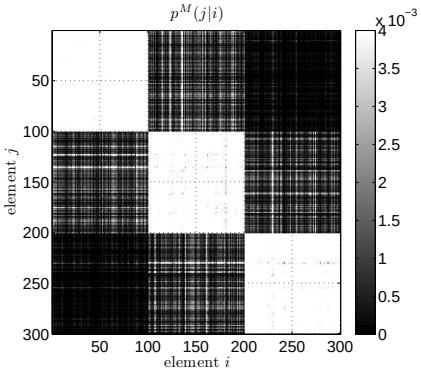
(a) Dades sintètiques generades



(b) Distribució ideal i real associada a un element de la classe cercle



(c) Matriu p_0 que s'ha d'aproximar

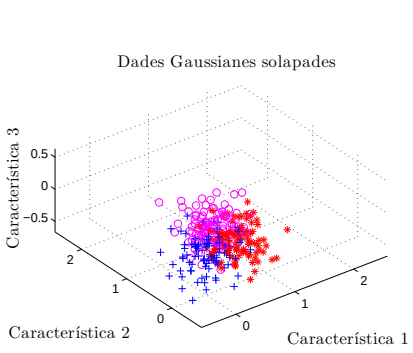


(d) Matriu amb la distribució de les dades sintètiques

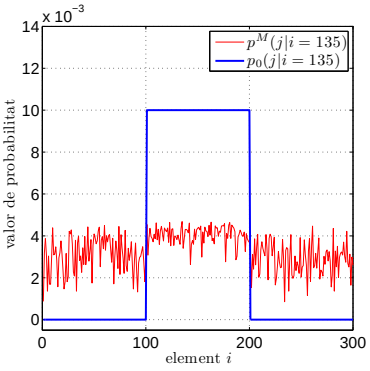
Figura 4.2: (a) separació de les classes; (b) distribució associada a l'element x_{135} juntament amb l'ideal; (c) matriu p_0; (d) matriu associada a la mostra 4.2(d).

Exemple 4.3. *En la Figura 4.3 s'il·lustra una situació en què les classes*

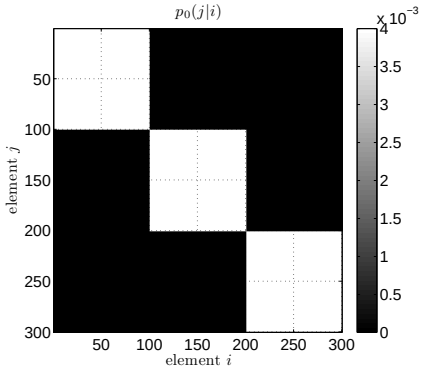
estan solapades, juntament amb les diferents representacions gràfiques de les distribucions p_0 i p^M. Les dades originals estan representades en 4.3(a), i la seua situació no permet discriminar les classes mitjançant hiperplans. La representació com a funció d'una variable de $p_0(j|135)$ i $p^M(j|135)$ de l'element $x_{135} \in \mathcal{M}$ $(i = 135)$ apareixen en 4.3(b), i la diferència entre les distribucions s'accentua més que en els exemples anteriors.Les representacions matricials p_0 i p^M apareixen en 4.3(c) i 4.3(d), respectivament. El fet que les representacions siguen tan diferents permet establir que la distribució p^M està relativament lluny de p_0.



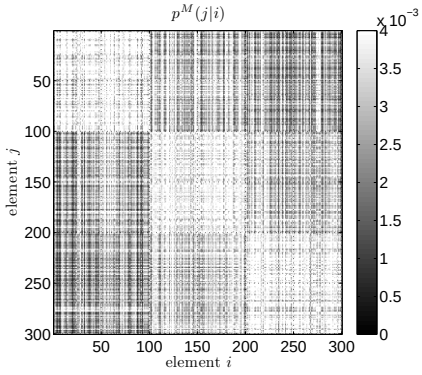
(a) Dades sintètiques generades



(b) Distribució ideal associada a un element de la classe cercle



(c) Matriu p_0 que s'ha d'aproximar



(d) Matriu amb la distribució de les dades sintètiques

Figura 4.3: (a) solapament de les classes; (b) distribució ideal i associada a l'element x_{135} de la mostra d'entrenament; (c) matriu p_0; (d) matriu p^M de la mostra.

4.2 MCML amb punts base

Un dels problemes de l'algorisme MCML, especialment en la versió convexa, està relacionat amb el cost computacional per iteració. Una estratègia per millorar aquest problema preservant l'enfocament original consisteix a no considerar tots els elements disponibles sinó un conjunt d'elements que actuen com a representants. Aquests elements els denominarem punts base (també anomenats com *landmarks* o *anchor points* en anglès) i es faran servir com elements de referència a partir dels quals prendre distàncies a la resta d'elements del conjunt d'entrenament.

La utilització d'un conjunt de punts referents, models o prototips (com els punts base) es comú en altres contexts [61, 87]. El MCML amb punts base és una generalització del mètode i depenent de l'elecció d'aquests pot obtindre's exactament el MCML original.

Fins ara, s'ha fet servir el conjunt d'entrenament $\mathcal{M}$ per establir relacions i calcular distàncies entre els seus elements. Ara, es defineix un altre conjunt de punts base

$$Y = \{y_1, y_2, \ldots, y_p\},\ y_j \in \mathbb{R}^d,\ 1 \leq j \leq p, \tag{4.1}$$

on l'element $y_j \in Y$ pertany a la classe $\tilde{c}_j \in \{1, 2, \ldots, c\}$. Els elements d'aquest conjunt Y poden pertànyer al conjunt $\mathcal{M}$, o no, segons siga convenient. S'obri així una direcció de recerca al voltant de l'elecció particular d'aquest conjunt Y.

La intenció fonamental a l'hora d'emprar punts base és situar estratègicament elements en l'espai original, $\mathbb{R}^d$, fent de referents, per tal d'apropar o allunyar els punts a aquests. Açò, transforma la idea de col·lapsar les classes a l'establiment d'agrupaments forçats al voltant dels punts base [71, 70].

D'altra banda, l'elecció del conjunt de punts base Y pot ser arbitrària i només queda restringida a l'espai en què treballem, $\mathbb{R}^d$. Però segons si la quantitat d'aquests és menor, igual o major que la quantitat de mostres en el conjunt $\mathcal{M}$ tindrem que les matrius corresponents a $p_0(j|i)$ i $p^M(j|i)$ amb $1 \leq i \leq p, 1 \leq j \leq n$ seran quadrades si $n = p$, o no quadrades quan $n < p$ (el nombre de punts base supera al nombre de mostres) o $n > p$. Aquest últim és el cas més interessant, i sobre el qual anem a fer l'estudi. Cal destacar que en el cas particular en què $Y = \mathcal{M}$ el mètode és el MCML original.

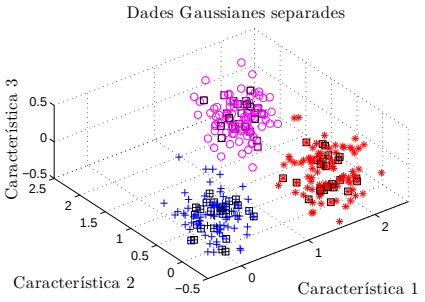
L'objectiu principal d'aquesta idea és fer $|Y| << |\mathcal{M}|$, sense perdre efectivitat. És a dir, prendre un conjunt Y de grandària inferior a la grandària de la mostra però amb la suficient representativitat per tal de preservar l'efectivitat del mètode original.

A continuació es reprodueixen les condicions de l'Exemple 4.2, fent ús de punts base triats aleatòriament dins del conjunt de Gaussianes generat. En aquest cas, les matrius corresponents a $p_0(j|i)$ i $p^M(j|i)$ ja no són quadrades de grandària $n \times n$ i passen a ser de grandària $p \times n$. També deixen de ser simètriques però conserven l'estructura original.

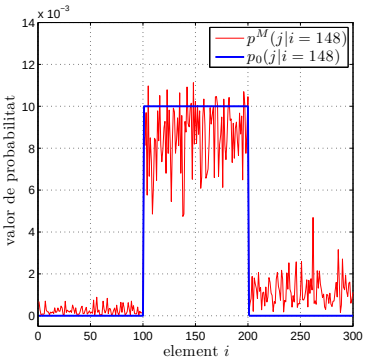
Exemple **4.4.** *Per a construir l'exemple s'ha tornat a generar un conjunt de dades sintètiques consistent en tres distribucions Gaussianes, de manera que cadascuna té assignada una classe diferent representades amb creus, asteriscs i*

cercles. En la Figura 4.4 s'ha fet servir com a conjunt Y una mostra aleatòria del 10% dels elements de cada classe.

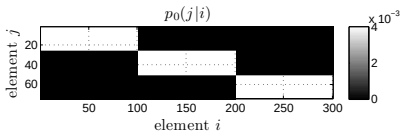
En la Figura 4.4(a), apareixen les classes separades dos a dos, i els elements marcats amb quadrats negres representen els punts base seleccionats. En la Figura 4.4(b) s'il·lustra que els valors de la distribució p^M són similars als de p_0 encara que existeix variabilitat poden diferenciar-se rangs diferents per a cada classe. En les Figures 4.4(c) i 4.4(d) apareixen les representacions matricials de les distribucions de manera global. Les matrius són a simple vista, relativament paregudes, mantenint l'estructura a blocs.



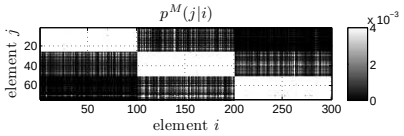
(a) Dades sintètiques generades



(b) Distribució ideal associada a un element de la classe asterisc



(c) Matriu p_0 que s'ha d'aproximar



(d) Matriu amb la distribució de les dades sintètiques

Figura 4.4: El conjunt Y està format per una mostra aleatòria del 10% de $\mathcal{M}$. En aquest cas les classes estan separades 4.4(a) i observem en 4.4(b) que p^M restringida a un punt en particular queda distant de complir l'objectiu. La matriu ideal 4.4(c) i la matriu associada a la mostra 4.4(d) tampoc no s'aproximen.

Com s'observa en aquest exemple, s'ha reduït la grandària de les matrius de les distribucions p_0 i p^M. Aquest canvi en la grandària es tradueix en una reducció del cost computacional del MCML.

4.3 Particularitzacions del mètode

Fins ara s'ha parlat del MCML i s'han introduït els punts base com una possible extensió genèrica per a aquest mètode. En aquesta secció, es fa menció a

diferents tipus d'inicialització i de tècniques per a seleccionar-los. S'ha intentat triar un conjunt representatiu de tècniques que permeten estudiar amb detall el comportament de l'algorisme.

4.3.1 Inicialitzacions

Tant el MCML com les extensions amb punts base d'aquest mètode requereixen d'una matriu PSD per a la seua inicialització. En particular, el conjunt de possibles matrius ha sigut reduït a 4 diferents:

- la matriu identitat (I),
- la de Mahalanobis (Mah.),
- la matriu resultant d'aplicar el mètode discriminant lineal de Fisher (LDA). S'utilitza la transformació lineal $W \in \mathbb{R}^{(c-1)\times d}$, a un subespai de dimensió el nombre de classes menys 1 $(c-1)$, obtinguda mitjançant aquest mètode de manera que $M = W^\top W$ és la matriu que defineix la distància d'inici.
- Per últim, una matriu amb valors propis en el rang [0,1] (Aleat.). Per a construir-la, primer es genera una matriu aleatòria de grandària $d \times d$; segon es realitza una descomposició en valors i vectors propis; i finalment es reemplaça la matriu diagonal amb els valors propis per una matriu amb d valors igualment espaiats entre 0 i 1.

4.3.2 Elecció de punts base

En aquest estudi es consideren tres alternatives, que es diferencien en la forma de seleccionar el conjunt de punts base:

- la primera, fixant el conjunt Y a l'inici de l'algorisme (mitjançant un mostreig aleatori sobre $\mathcal{M}$) i preservant l'elecció inicial durant tot el procés d'optimització. Aquesta versió rep el nom de FIX, [71].
- La segona, permet definir novament Y al llarg de l'aprenentatge de la distància. En aquest cas, es recalcula Y al llarg del procés d'optimització, realitzant un mostreig aleatori sobre $\mathcal{M}$. Aquesta versió s'anomena CAN [72], perquè el conjunt de punts base és canviant.
- Per completar l'estudi, es considera una tercera versió que també es manté fixa i la denotarem com KMN. En aquest cas, s'utilitza l'algorisme de les $k-$mitjanes per a cada classe (de $\mathcal{M}$) de la següent manera: es realitza un agrupament per classes amb tants clústers com elements té Y, després s'agafa com a únic representant de cada clúster la mitjana d'aquest.

4.4 Estudi de la complexitat

L'algorisme original del treball on es va presentar el MCML està basat en un descens per gradient combinat amb successives projeccions ortogonals sobre el con de les matrius Semi Definides Positives. Per a l'experimentació s'ha fet servir un mètode de minimització paregut a l'utilitzat pels autors i que pot trobar-se en [90]. La forma més senzilla d'estudiar el cost computacional pot fixar-se amb un descens per gradient de longitud (de pas) fixa per a tot l'algorisme. L'objectiu és tindre una versió senzilla d'interpretar i que a l'hora siga la part central de les diferents versions per tal de comparar-les posteriorment.

Abans d'introduir l'algorisme cal donar l'expressió del gradient del criteri original del MCML (Equació (3.5)). Aquest pot escriure's en funció del conjunt de punts base Y de la manera següent

$$\nabla f_{\text{MCML}}(M, Y) = \sum_{i=1}^{n}\sum_{j=1}^{p}(p_0(j|i) - p^M(j|i))(x_i - y_j)(x_i - y_j)^\top. \tag{4.2}$$

Aquesta expressió és la direcció de major descens en la superfície del criteri que s'ha de minimitzar. Una vegada definit el gradient pot introduir-se l'Algorisme 1, què és la peça fonamental de l'estudi comparatiu de les diferents versions presentades.

Algorisme 1 MCML generalitzat amb punts base

Entrada: $X = \{x_i\}_{i=1}^n, Y = \{y_j\}_{j=1}^p$. //conjunt d'objectes i punts base
Entrada: M^0. //model inicial
Entrada: versió, tol, ϵ, $iterMax$. //versió (FIX/CAN/KMN), tolerància, longitud del pas i iteració màxima
Eixida: M. //matriu apresa

1: $t = 0$
2: **repetir**
3: calcula p^{M^t}
4: calcula $\nabla f_{\text{MCML}}(M^t, Y)$
5: $\tilde{M}^{t+1} = M^t - \epsilon\nabla f_{MCML}(M^t, Y)$
6: $\{\lambda_r, v_r\}_{r=1}^d = \text{descomposar}(\tilde{M}^{t+1})$
7: $M^{t+1} = \sum_{r=1}^d \max\{0, \lambda_r\} v_r v_r^\top$ //λ_r, v_r valors i vectors propis de $\tilde{M}^{t+1}$
8: **si** versió = CAN **aleshores**
9: recalcula(Y)
10: **fi si**
11: $t = t + 1$
12: $M^{t+1} = M^t$
13: **fins que** $\|\nabla f_{\text{MCML}}(M^t, Y)\| \leq tol$ **o** $t \geq iterMax$
14: $M = M^{t+1}$

El càlcul de p^M requereix la construcció d'una matriu de grandària $p \times n$. Per

tant, el seu cost computacional depèn de la quantitat d'elements en $\mathcal{M}$, n, i de la quantitat de punts en Y, p. En el cas del MCML, pot aprofitar-se que $Y = X$, i fer servir l'estructura simètrica de la matriu $p^M(j|i)$, requerint un total de $\frac{n(n-1)}{2}$ repeticions de les operacions necessàries per calcular p^M (equació (3.3)). Aquesta xifra contrasta amb les $\frac{p(n-1)}{2}$ operacions necessàries en el cas $Y \subset \mathcal{M}$, i $\frac{p(n+1)}{2}$ en el cas $Y \not\subset \mathcal{M}$.

El càlcul del gradient $\nabla f_{\text{MCML}}(M, Y)$, consisteix en una combinació lineal de matrius simètriques de rang 1 i grandària $d \times d$. Cada una pot calcular-se amb un cost de $\frac{d(d+1)}{2}$ productes escalars de dimensió d. En el cas $Y = \mathcal{M}$ caldria repetir-les $\frac{n(n-1)}{2}$ vegades, mentre que serien $\frac{p(n-1)}{2}$ vegades en el cas $Y \subset \mathcal{M}$, i $\frac{p(n+1)}{2}$ vegades en el cas $Y \not\subset \mathcal{M}$. Finalment, la descomposició en valors i vectors propis té un cost computacional estimat en d^3.

En la taula 4.1, apareix la complexitat de les diferents etapes de l'Algorisme 1.

Taula 4.1: Complexitat de les diferents etapes de l'Algorisme 1, per al MCML i les versions amb punts base.

línia/mètode	MCML	FIX/KMN	CAN
pas 3	$\mathcal{O}(n^2d^2)$	$\mathcal{O}(pnd^2)$	$\mathcal{O}(pnd^2)$
pas 4	$\mathcal{O}(n^2d^2)$	$\mathcal{O}(pnd^2)$	$\mathcal{O}(pnd^2)$
pas 5	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2)$
pas 6	$\mathcal{O}(d^3)$	$\mathcal{O}(d^3)$	$\mathcal{O}(d^3)$
pas 9	-	-	$\mathcal{O}(p)$

Es poden escriure els costs per iteració del MCML i del MCML amb punts base, respectivament, com

$$T_{\text{MCML}}(n, d) = \mathcal{O}(n^2d^2) + \mathcal{O}(d^3), \tag{4.3}$$

$$T_{\text{FIX}}(n, p, d) = \mathcal{O}(pnd^2) + \mathcal{O}(d^3), \tag{4.4}$$

$$T_{\text{CAN}}(n, p, d, \rho) = \mathcal{O}(pnd^2) + \mathcal{O}(d^3) + \mathcal{O}(p), \tag{4.5}$$

$$T_{\text{KMN}}(n, p, d, \vec{k}, \vec{c}) = \mathcal{O}(pnd^2) + \mathcal{O}(d^3). \tag{4.6}$$

De l'Equació 4.3, es pot concloure que el MCML té una complexitat quadràtica tant per a la dimensió com per al nombre de punts del conjunt d'entrenament. D'altra banda, el mètode FIX (equació 4.4) té complexitat pn, passant d'un ordre quadràtic a un ordre lineal en n ($p << n$). El mètode CAN (equació 4.6) té un terme que depèn de p que es correspon amb la selecció de Y en cada iteració i que no modifica el cost asimptòtic. El mètode KMN requereix el càlcul de les $k-$mitjanes per a cada classe però el cost per iteració és el mateix. Per tant, l'ús dels punts base en el MCML redueix en un grau la complexitat de quadràtic a lineal sobre la quantitat d'elements amb els quals s'entrena.

4.5 Avaluació

Per tal d'avaluar les propostes presentades en aquest capítol s'ha desenvolupat

tota una bateria d'experiments. En primer lloc, es proposen experiments amb diferents inicialitzacions per tal d'avaluar els valors del criteri i tractar d'establir quina inicialització és millor a efectes pràctics. A aquests efectes s'han emprat els 4 mètodes descrits en 4.3.1 (I, Mah., LDA i Aleat.). Posteriorment, s'utilitza la millor inicialització resultant per a avaluar els mètodes de selecció de punts base proposats en la secció 4.3.2 (FIX, CAN i KMN). En tots els casos, s'han dissenyat experiments de classificació, per tal d'avaluar la bondat de les matrius obtingudes amb els diferents mètodes. A més a més, s'han analitzat els temps computacionals per a poder comparar en la pràctica les possibles diferències entre els algorismes. Finalment, les propostes s'han comparat estadísticament, en termes d'error de classificació per a tractar d'establir una recomanació.

Totes les aportacions d'aquesta tesi estan recolzades mitjançant una exhaustiva experimentació sobre un conjunt de bases de dades que es pot definir com a estàndard en la literatura relacionada amb l'aprenentatge de distàncies. Aquest conjunt de 15 bases de dades públiques s'introdueix amb més nivell de detall en l'Apèndix A.

4.5.1 Avaluació de l'efecte de la inicialització

Com a primera aproximació i per tal d'avaluar la bondat de les diferents matrius d'inici, s'han dut a terme experiments amb els mètodes FIX i CAN, sobre totes les bases de dades. S'ha establit que una solució és millor que altra en base a que el seu valor de criteri, f_{MCML}, siga menor. El nombre màxim d'iteracions s'ha fixat a 100 en base a experiments orientatius previs. En aquesta secció es mostren resultats de la mitjana de l'evolució del valor de criteri (dividit per p). Aquestes mitjanes corresponen a un total de 10 repeticions independents de la partició d'entrenament i test en un percentatge del 50%. En concret, es mostra un valor de p representatiu sobre les bases de dades malaysia i soyL. Les gràfiques corresponents a la resta de bases de dades es mostren en l'Apèndix B.1.

En la Figura 4.5, apareix l'evolució del valor mitjà del criteri per a les diferents matrius d'inici sobre la base de dades malaysia. La grandària de Y és el 25% del cardinal de $\mathcal{M}$. Les matrius amb major valor de criteri són les corresponents a LDA, seguit de Mah. Les matrius Aleat mostren un valor inicial pitjor i quasi idèntic al de la matriu I. Finalment, totes les inicialitzacions pareixen haver convergit en una solució amb un valor de criteri similar.

La Figura 4.6 il·lustra els mateixos resultats per a la base de dades soyL amb $|Y| = 0.125|\,\mathcal{M}\,|$. La inicialització amb Mah. empitjora significativament respecte a la resta. A més, els resultats mostren una convergència més lenta quan s'utilitza el mètode CAN. Finalment, destacar que en aquest cas, les inicialitzacions I i Aleat. pareixen conduir als millors resultats, independentment de la utilització de FIX o CAN.

L'evolució dels criteris en malaysia i soyL poden considerar-se com a il·lustratius de la resta de bases de dades (Apèndix B.1). Els diferents inicis es comporten de manera molt similar entre les versions FIX i CAN. A la vista dels resultats, les millors opcions d'inici són les matrius I i Aleat encara que les diferències entre els dos inicis són molt menudes: les dos arriben a un valor de criteri en el qual els

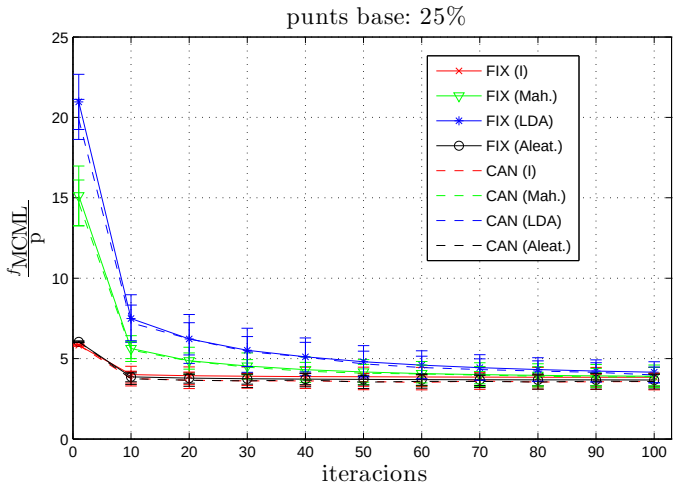


Figura 4.5: Valors de criteri amb les diferents inicialitzacions en la base de dades malaysia.

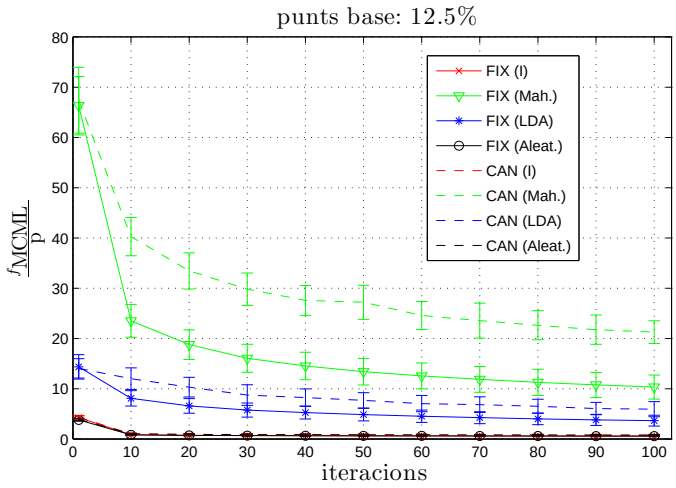


Figura 4.6: Valors de criteri amb les diferents inicialitzacions en la base de dades soyL.

mètodes pareixen haver convergit abans de les 100 iteracions fixades con a màxim. Així doncs, finalment s'ha comprovat empíricament que el nombre màxim d'iteracions fixat a 100 assegura la convergència dels mètodes i s'ha seleccionat com a matriu d'inici la identitat per a totes les versions estudiades.

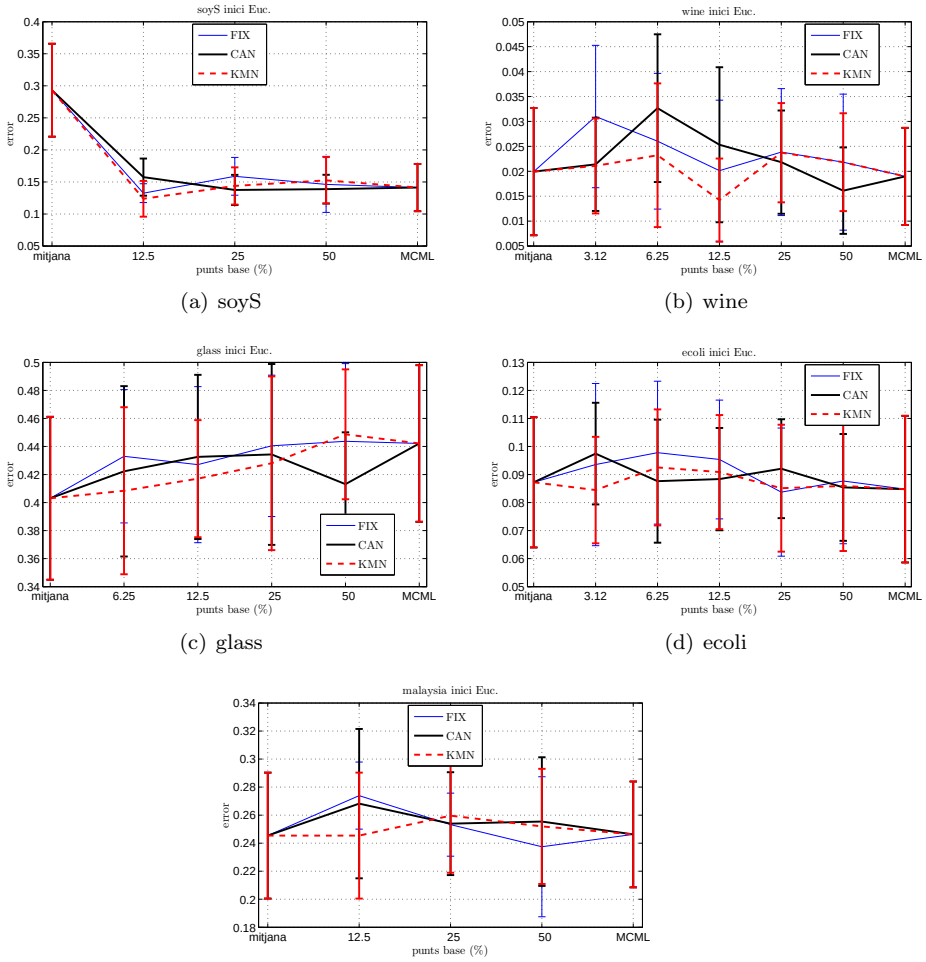
4.5.2 Avaluació del mètode de selecció de punts base

Per tal d'avaluar la bondat de les matrius apresses s'ha dissenyat un experiment basat en la classificació sobre les de bases de dades públiques seleccionades en la tesi i que estan referenciades en l'Apèndix A. La intenció és estudiar les possibles diferències entre les versions proposades: FIX, CAN i KMN, prenent al MCML com a mètode de referència. El rendiment dels algorismes s'avalua en termes de l'error de classificació sobre el veí més proper.

Totes les bases de dades s'han normalitzat linealment en el rang [0,1] independentment per a cada característica. Els resultats que es presenten són una mitjana de 10 repeticions independents. Els conjunts d'entrenament i de test s'han establit de manera aleatòria, prenent 50% per a entrenament i 50% per a test. D'acord amb els resultats de l'estudi de la Secció 4.3.1, tots els mètodes s'inicialitzen amb la matriu identitat i amb un màxim de 100 iteracions (valors que asseguren a efectes pràctics la convergència de tots els mètodes).

La diversitat en la grandària (nombre d'elements, n), de les bases de dades ha donat lloc a l'elecció d'un criteri per establir la quantitat de punts base a estudiar en cada base de dades. Aquests valors són els mateixos per a les diferents versions amb punts base: FIX, CAN i KMN. A banda del propi MCML que fa servir el 100% dels punts, la quantitat de punts base triats com a primer mètode és del 50%, el següent el 25%, i es continua dividint successivament per 2, fins arribar al cas en què alguna classe continga un únic element. També es contempla com a cas límit, un conjunt de punts base Y format per la mitjana de cada classe com a únics representants.

La Figura 4.7 il·lustra els valors d'error de classificació amb el veí més proper sobre les bases de dades soyS, wine, glass, ecoli i malaysia. En cada una de les 5 gràfiques apareixen representades com a variable independent el nombre de punts base (en percentatge per classe) fins arribar al MCML en la seua versió estàndard, i com a variable depenent el valor d'error associat. Els resultats de classificació per a la base de dades soyS, que apareixen en 4.7(a), no mostren cap diferència entre els tres mètodes presentats. El pitjor valor d'error és per al cas de la mitjana i l'augment de punts base fa millorar aquest valor d'error, que es manté quasi constant a partir del 12.5%. Açò pareix indicar que un valor relativament baix per a la grandària de Y és adequat per a aquest problema particular. Els resultats per a la base de dades wine en 4.7(b) tenen una variació relativament gran però realment es tracta de l'escala que s'està emprant. En concret, les desviacions estàndar varien entre un 0.5% i quasi un 5% d'error i a simple vista no poden establir-se diferències entre els 3 mètodes. Per a la base de dades glass, els resultats mostren una tendència a empitjorar (per als 3 mètodes) segons augmenta el nombre de punts base, veure Figura 4.7(c). La penúltima base de dades de la Figura és ecoli 4.7(d). En aquest cas, els resultats no mostren



(a) soyS (b) wine

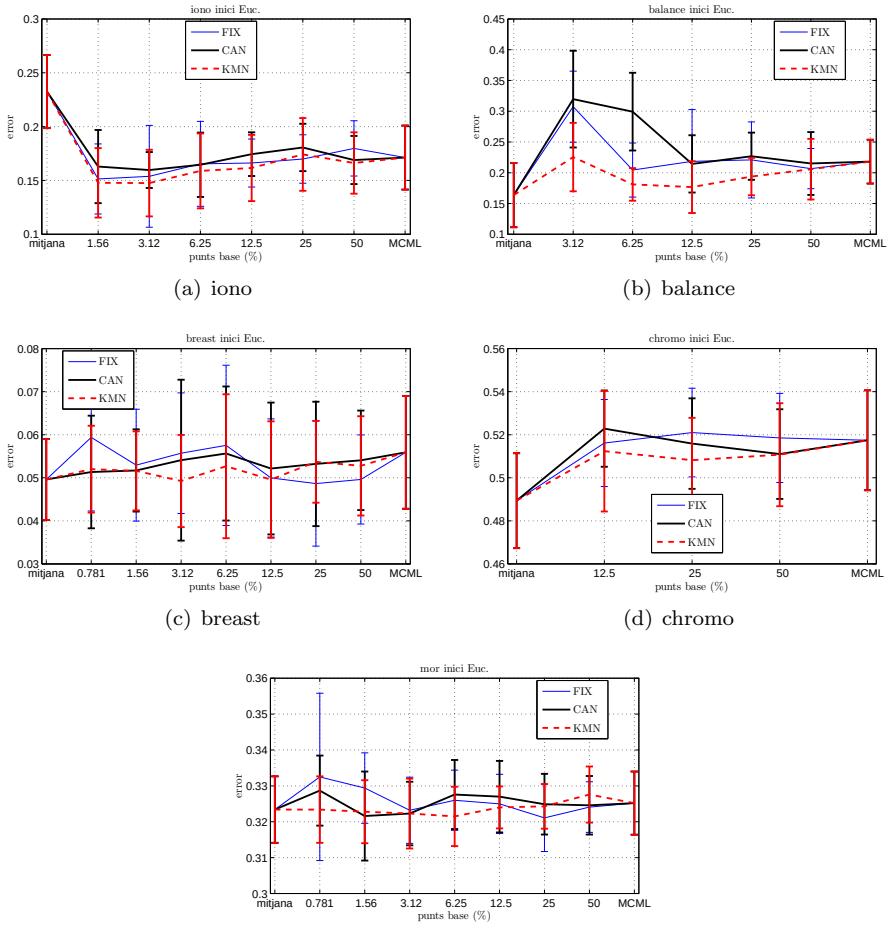
(c) glass (d) ecoli

(e) malaysia

Figura 4.7: Errors de classificació amb el 1 veí més proper de les bases de dades: soyS, wine, glass, ecoli i malaysia.

diferències entre els 3 mètodes i l'increment en el nombre de punts base tampoc mostra cap diferència important. La uniformitat dels resultats pareix beneficiar les versions amb pocs punts base, ja que, amb menor quantitat de punts s'obté el mateix valor d'error. Finalment la base de dades malaysia 4.7(e), pareix mostrar un comportament similar al d'ecoli.

La Figura 4.8, il·lustra els valors d'error de classificació amb el veí més proper sobre les bases de dades: iono, balance, breast, chromo i mor. Els resultats per a la base de dades iono 4.8(a), mostren un comportament paregut a l'anterior base de dades soyS. Es pot observar que el mètode KMN té menor valor d'error al llarg de la corba d'error però no pot establir-se com una millora significativa



(a) iono (b) balance

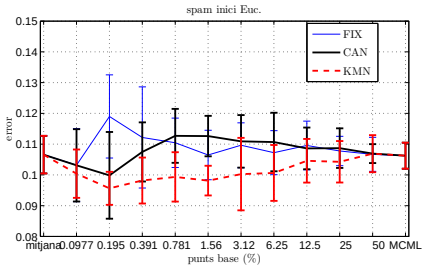
(c) breast (d) chromo

(e) mor

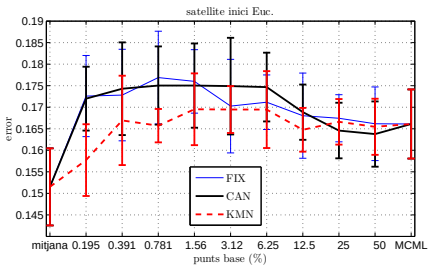
Figura 4.8: Errors de classificació amb el 1 veí més proper de les bases de dades: iono, balance, breast, chromo i mor.

respecte a la resta. Els resultats de la base de dades balance 4.8(b), il·lustren el millor resultat d'error per al cas de la mitjana, millorant l'error del MCML. L'augment del nombre de punts base empitjora l'error (per als primers valors) i mostra convergència cap al resultat del MCML segons continua augmentant. Per a aquesta base de dades, el millor valor és el de la mitjana, seguida de percentatges al voltant del 12.5%. A més, KMN és el mètode que pareix ser més robust per a aquest problema particular. L'evolució de l'error respecte al nombre de punts base per a breast 4.8(c), no mostra cap variació significativa, igual que les bases de dades wine, malaysia i ecoli. Valors menuts de p són suficients per a obtindre relativament bons valors de classificació. D'altra banda, els resultats per a la base de dades chromo 4.8(d), mostren un millor comportament en el cas

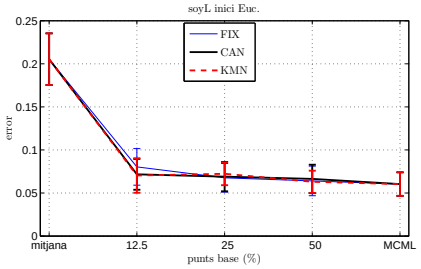
de la mitjana i segons augmenta el percentatge de punts base l'error es manté pràcticament constant. L'única apreciació és que el mètode KMN pareix mostrar millors resultats encara que no de manera significativa. En 4.8(e) s'il·lustren els resultats per a la base de dades mor, que mostren una tendència quasi constant per al valor de l'error. Pareix que en aquest problema, la grandària del conjunt de punts base no afecta a l'error i es pot destacar que en el cas de la mitjana el valor d'error és lleugerament millor que el del MCML.



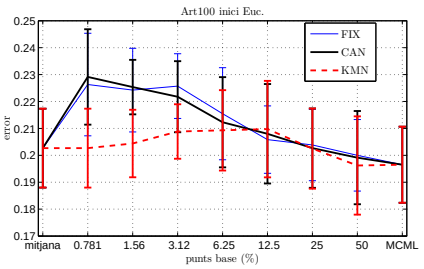
(a) spam



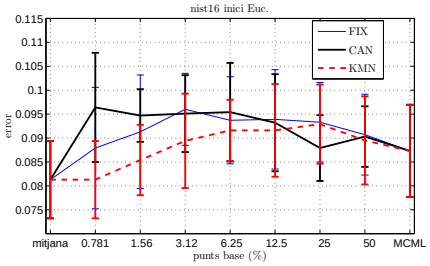
(b) satellite



(c) soyL



(d) Art100



(e) nist16

Figura 4.9: Errors de classificació amb el 1 veí més proper de les bases de dades: spam, satellite,soyL, Art100 i nist16.

La Figura 4.9, il·lustra els valors d'error de classificació amb el veí més proper sobre les bases de dades: spam, satellite, soyL, Art100 i nist16. Els resultats de la base de dades spam s'il·lustren en 4.9(a). En aquesta ocasió, la mitjana mostra un valor similar al del MCML. D'altra banda, l'algorisme KMN és el que menor

error ha comés i el seu valor mínim s'ha obtingut per a 0.195%. Els resultats de satellite que s'il·lustren en 4.9(b), tenen el millor valor d'error en el cas de la mitjana i les barres d'error quasi no se solapen amb les del MCML, de manera que esdevenen la millor opció. D'altra banda, segons augmenta el nombre de punts base, l'error creix i s'estabilitza tornant a decréixer lleugerament a mesura que s'apropa cap al MCML. De les tres versions, es pot dir que KMN torna a ser la que millors resultats dóna. Els resultats per a soyL 4.9(c), mostren el mateix comportament que les bases de dades soyS i iono: el pitjor resultat en el cas de la mitjana i millores segons s'augmenta la grandària de Y. Per a la base de dades Art100 4.9(d), els resultats són molt pareguts als de les bases de dades balance i satellite: la mitjana dóna un valor d'error comparable al del MCML i empitjora per a valors intermedis, mentre que per a valors propers al 50% s'estabilitza i s'acosta més als resultats del MCML. Finalment, en la base de dades nist16 4.9(e), els resultats mostren un bon comportament per al cas de la mitjana però aquest empitjora segons augmenta el nombre de punts base i no torna a un bon valor d'error fins al cas del MCML.

Com a conclusió i resumint, es pot dir que la utilització de punts base manté el comportament del MCML (millorant-lo en alguns casos particulars), per a valors de p reduïts i adequats. L'error de classificació per a la base de dades soyL, Figura 4.9(c), mostra una tendència a millorar lleugerament segons s'augmenta la quantitat de punts base i el cas de la mitjana té el valor d'error més elevat per a aquesta base de dades. Pot observar-se que la tendència general de la corba de l'error és anàleg al de les bases de dades soyS (Figura 4.7(a)) i iono (Figura 4.8(a)). En la base de dades chromo (Figura 4.8(d)), el cas de la mitjana resulta ser millor resultat de classificació i la tendència general dels mètodes es manté constant segons creix p, al igual que succeeix per a la base de dades spam (Figura 4.9(a)). El comportament per a les bases de dades balance (Figura 4.8(b)), nist16 (Figura 4.9(e)), Art100 (Figura 4.9(d)) és semblant al de satellite (Figura 4.9(b)), a on la mitjana mostra el millor resultat d'error i la tendència general és millorar l'error segons augmenta p però no aconseguint el valor d'error de la mitjana. Les bases de dades que no mostren cap tipus de millora ni empitjorament al fer variar la grandària de p són: wine (Figura 4.7(b)), malaysia (Figura 4.7(e)), breast (Figura 4.8(c)), mor (Figura 4.8(e)) i ecoli (Figura 4.7(d)). Aquestes es caracteritzen per conservar el valor d'error del MCML original i a l'hora mostrar un cost computacional inferior, fet que es posarà de manifest més endavant.

Tests estadístics

Les mitjanes mostrades no permeten establir un mètode òptim entre les diferents propostes (FIX, CAN i KMN) de manera objectiva . Això és en part degut a la variabilitat dels resultats. No obstant això, es pot estudiar si algun dels mètodes és sempre (és a dir, en cada experiment particular) significativament millor que un altre mitjançant tests estadístics basats en ordenacions [38].

Amb aquest propòsit s'ha emprat un test de comparacions múltiple de Friedman que estableix una ordenació. La Taula 4.2 mostra les mesures de les mitjanes de les posicions en les ordenacions resultants (de més efectiu a menys). A la vista

dels resultats podem establir que KMN és el millor en promig, seguit per CAN i FIX. Encara que les diferències entre CAN i FIX són menys clares.

Taula 4.2: Ordenació promig dels diferents mètodes comparats.

Algorisme	Classificació
KMN	1.67463
CAN	2.16304
FIX	2.16231

La Taula 4.3 mostra els $p-$valors ajustats corresponents al test post-hoc de Holm, comparant els mètodes dos a dos.

Taula 4.3: Resultats del test de Holm ($\alpha = 0.05$), la hipòtesi nul·la és rebutjada al nivell α i apareixen en negreta aquells valors menors que 0.05.

Algorismes	$p-$valor	$p-$valor ajustat
KMN vs. CAN	$\mathbf{1.17137}\times10^{-19}$	$\mathbf{3.51410}\times10^{-19}$
KMN vs. FIX	$\mathbf{1.32529}\times10^{-19}$	$\mathbf{3.51410}\times10^{-19}$
CAN vs. FIX	0.98926	0.98926

Pot concloure's que existeixen diferències significatives en quant a resultats de classificació entre els mètodes KMN i CAN, i també entre els mètodes KMN i FIX però no entre FIX i CAN. Amb la qual cosa, es pot establir com a significativament millor en l'experimentació de classificació al mètode KMN.

Diferències entre FIX i CAN

El mètode CAN va ser introduït amb l'esperança que fora més robust front a l'elecció dels punts base i que això es reflectirà en un millor comportament. En canvi, cap dels tests permet observar diferències significatives respecte a FIX. Per tal d'estudiar amb més detall els dos mètodes es realitza una comparació dels mètodes FIX i CAN, a on es mesura la diferència d'error entre el resultat de cada execució dels dos mètodes. La intenció és observar quin dels dos mètodes és més robust en cada repetició de l'experiment.

En particular, s'ha dissenyat l'experiment comparatiu de la manera següent: dels possibles valors de p (de grandària de Y) assignats, hem exclòs el cas del MCML ($Y = X$), 50% i el cas extrem de la mitjana. De manera que per a cadascun dels restants valors de p, es mesura la diferència d'error entre els dos mètodes. Els valors positius fan referència a un millor resultat per al mètode CAN i valors negatius ho fan per a FIX. Per a cada base de dades, es tria el major increment (en valor absolut) de les 10 repeticions que s'han fet a l'experiment i aquest valor $\Delta = \max_{i=1,\dots,10} \frac{\mathrm{CAN}_i - \mathrm{FIX}_i}{\sigma}$, a on σ és la desviació estàndard en els errors CAN i FIX per a tot i, s'il·lustra en la Figura 4.10. Així doncs, els valors que queden fora de la zona ombrejada és per que mostren un resultat significativament millor i en concret més de dues vegades la desviació estàndard total dels resultats de cada base de dades. Els resultats de la Figura 4.10, indiquen que el mètode

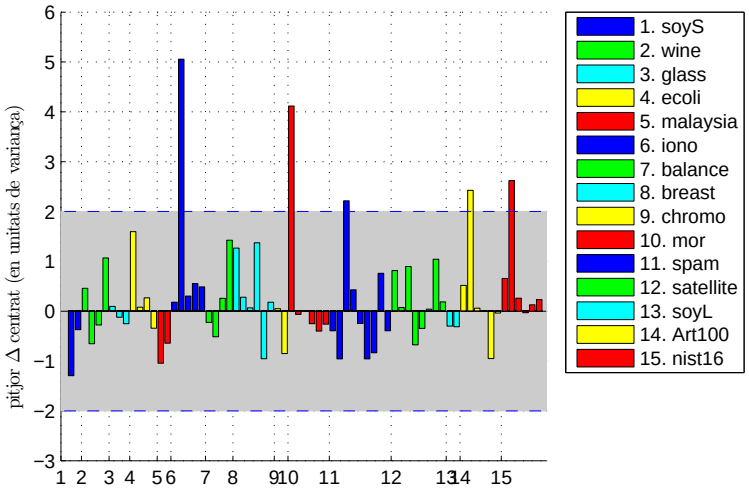


Figura 4.10: Representació gràfica de les diferències entre els errors de FIX i CAN sobre les 15 bases de dades.

CAN guanya amb més de 2σ en 5 de les 15 bases de dades de l'experimentació, mentre que el mètode FIX no ho fa en cap.

Es pot concloure que encara que en promig els dos mètodes són indistingibles, FIX pràcticament mai mostra una millora significativa respecte a CAN. Per a 5 de les bases de dades (iono, mor, spam, Art100 i nist16) el mètode CAN mostra una millora respecte a FIX al que supera en almenys 2σ i fins a 5σ per a iono o 4σ en la base de dades mor.

4.5.3 Temps d'execució

Els temps d'execució de cada un dels mètodes estudiats en aquest capítol han sigut recollits durant l'experimentació. A continuació, en les Figures 4.11 i 4.12 apareixen els temps calculats per iteració.

A la vista dels resultats es pot concloure que existeixen diferències significatives en els temps d'execució entre el MCML i els tres mètodes: FIX, CAN i KMN. En particular, a partir del primer valor de punts base 50% els tres mètodes són més ràpids que el MCML. Aquesta millora temporal va en augment segons es redueix el nombre de punts base. Aquest comportament se satisfà per a totes les bases de dades emprades en l'experimentació. També existeixen diferències significatives entre una versió (de FIX, CAN o KMN) i la immediatament major en quant a nombre de punts base.

D'aquesta manera pot concloure's que la utilització de punts base redueix el cost computacional del MCML de manera significativa i a l'hora preserva la

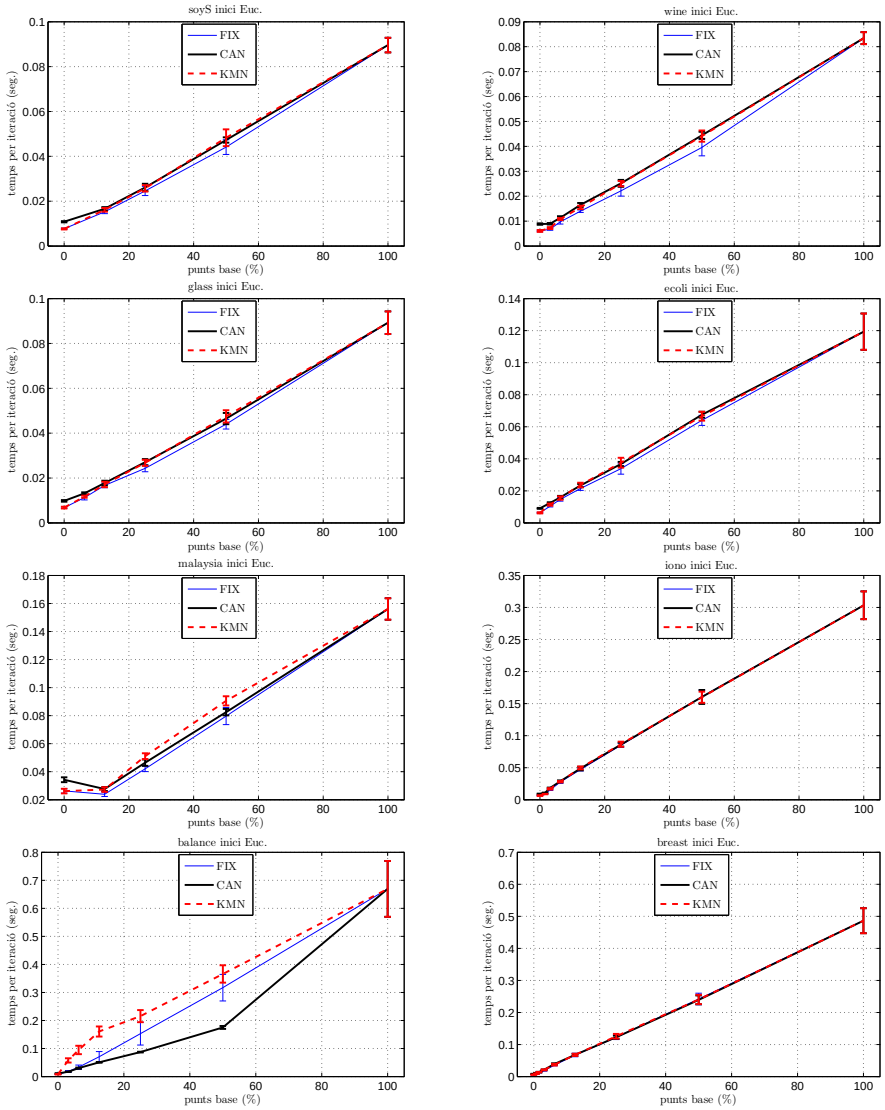


Figura 4.11: Temps d'execució dels mètodes FIX, CAN, KMN i MCML (representat amb el 100%), per a les bases de dades (de dalt a baix i d'equerra a dreta): soyS, wine, glass, ecoli, malaysia, iono, balance i breast.

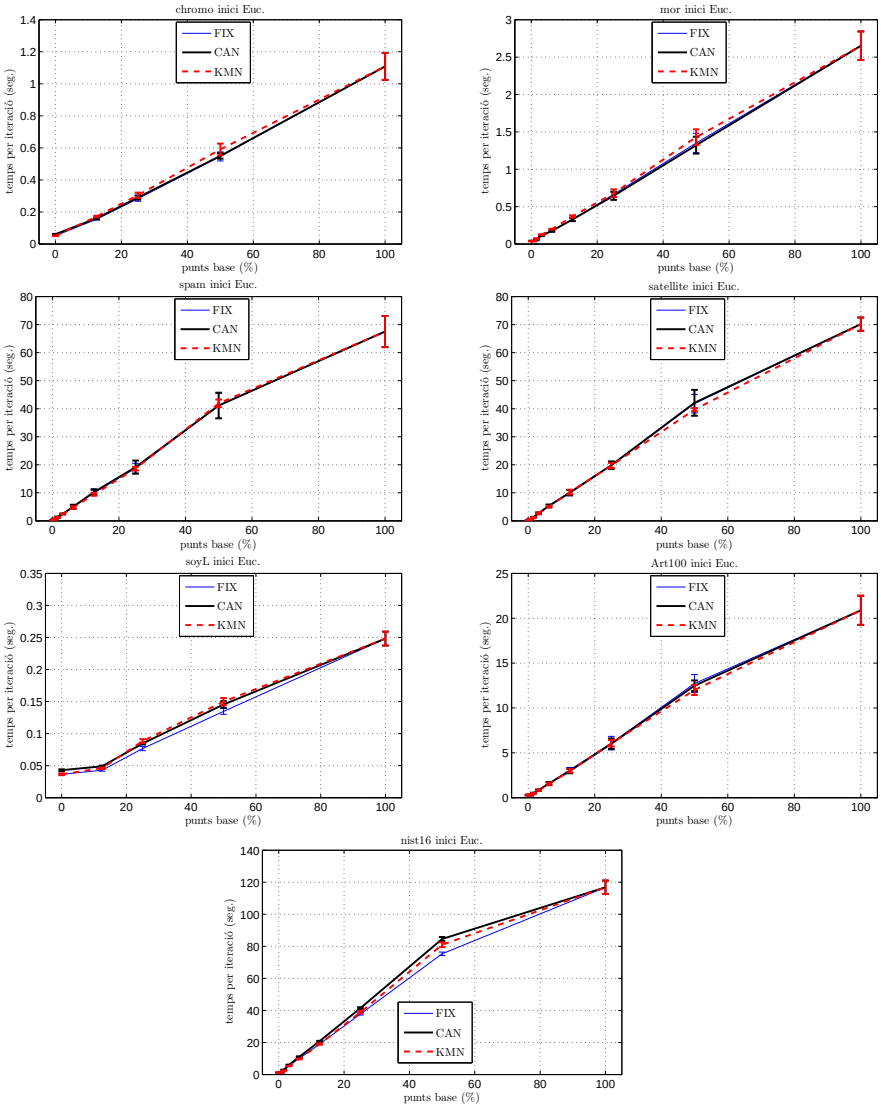


Figura 4.12: Temps d'execució dels mètodes FIX, CAN, KMN i MCML (representat amb el 100%), per a les bases de dades (de dalt a baix i d'equerra a dreta): chromo, mor, spam, satellite, soyL, Art100 i nist16.

efectivitat del mètode original.

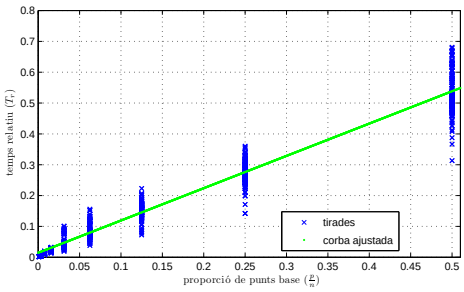
Anàlisi dels temps computacionals de FIX i CAN

A partir dels temps d'execució obtinguts en els experiments (il·lustrats amb anterioritat) es va a presentar un estudi sobre els temps relatius de les versions FIX i CAN respecte als temps del MCML.

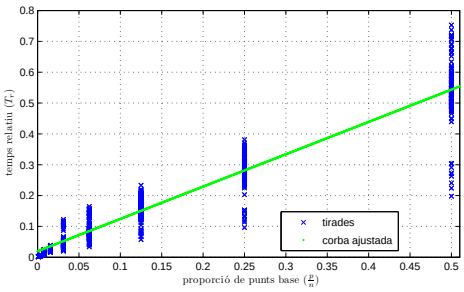
L'estudi teòric sobre la complexitat del MCML i les versions presentades concloïa amb un temps quadràtic i lineal respectivament (en n). Per tal de mostrar aquest comportament lineal del temps teòric de les versions amb punts base respecte al nombre de punts en el conjunt d'entrenament es mostraran els temps d'execució relatius al MCML dels experiments. El mètode FIX s'il·lustra en la Figura 4.13(a) i el mètode CAN en la Figura 4.13(b). El valor del temps relatiu s'ha calculat de la manera següent

$$T_r = \frac{T(n,d,p)}{T(n,d)} \in \frac{\mathcal{O}(pnd^2)}{\mathcal{O}(n^2d^2)} = \mathcal{O}\left(\frac{p}{n}\right) \tag{4.7}$$

Aquests temps només es mostren per als conjunts de punts base diferents a la mitjana i al MCML original.



(a) Temps relatius del mètode FIX.



(b) Temps relatius del mètode CAN.

A les figures s'han inclòs també les rectes ajustades per mínims quadrats. Aquestes rectes s'han calculat fent servir tots els valors de temps relatius obtinguts en la simulació amb cada base de dades emprada en l'experimentació.

Les equacions de les rectes expressades mitjançant els intervals de confiança al 95% són

$$y_{\text{FIX}} =]1.033, 1.062[\, x +]0.01078, 0.01791[\,, \tag{4.8}$$

$$y_{\text{CAN}} =]1.028, 1.069[\, x +]0.01436, 0.02439[\,, \tag{4.9}$$

els intervals que apareixen en les rectes estan solapats, però pot observar-se un poc més de variació en el valor del pendent de CAN, encara que aquest no pareix ser molt gran en relació a FIX. Les equacions de les rectes expressades mitjançant les mitjanes dels intervals de confiança són les següents

$$y_{\text{FIX}} = 1.048x + 0.01435, \tag{4.10}$$

$$y_{\text{CAN}} = 1.048x + 0.01937. \tag{4.11}$$

Les dos equacions de la recta (4.10), (4.11) tenen el mateix pendent de valor pròxim a 1. Açò representa una reducció del temps lineal respecte al nombre de punts base en els dos algorismes. El temps relatiu dels dos algorismes per al cas $x = 0.5$ es reflecteix amb aproximadament la meitat temps del MCML, de $x = 0.25$ amb un quart, etcètera. Es pot concloure que l'ús de punts base té una complexitat lineal respecte al MCML que és directament proporcional al nombre de punts base.

4.6 Discussió

En aquest capítol s'han analitzat diferents propostes fent servir punts base en el mètode MCML. Sobre l'experimentació duta a terme amb les tres versions FIX, CAN i KMN pot establir-se que totes mantenen l'efectivitat en l'error de classificació quan es redueix considerablement el nombre de punts base. El mètode KMN és en la mitjana el més efectiu entre els mètodes avaluats probablement per que els punts base representen millor el conjunt d'entrenament. Els mètodes que fan servir punts base aleatoris funcionen en general pitjor. Generar nous punts aleatoris en cada iteració fa que els resultats siguen més robusts però açò no compensa en promig. Tenint en compte totes les consideracions, es recomana la utilització del mètode KMN amb un nombre relativament reduït de punts base, al voltant del 10%. Aquesta configuració es correspon a l'experimentació amb un valor d'error de classificació comparable a l'obtingut amb el MCML i amb millores en torn al 90% del temps requerit per a obtindre una solució amb el MCML.

Capítol 5

Aprenentatge en línia passiu-agressiu i extensió per mínims quadrats

Resum – L'aprenentatge per lots pot estar limitat per la grandària del conjunt d'entrenament. L'elevat requeriment computacional fa que determinats problemes siguen intractables degut a les limitacions físiques de les màquines. És per això que l'aprenentatge en línia es presenta com una alternativa eficient per a enfrontar-se a aquesta situació. En aquest capítol es descriu una família de mètodes d'aprenentatge de distàncies en línia i es desenvolupa un estudi al voltant de les diferents propostes de manera teòrica i pràctica mitjançant una experimentació exhaustiva.

Contingut

5.1 Introducció

Alguns dels mètodes d'aprenentatge de distàncies més coneguts s'han introduït en el Capítol 3. En la seua majoria es caracteritzen per realitzar un aprenentatge per lots. Aquest tipus d'aprenentatge requereix el conjunt $\mathcal{M}$ en la seua totalitat, de manera que quan no està disponible és necessari emprar un altre paradigma d'aprenentatge.

L'aprenentatge en línia s'enfronta a problemes a on les dades arriben de manera seqüencial o en línia, com per exemple en el processament de fluxos de

dades. Aquesta varietat d'aprenentatge també pot resultar útil quan l'objectiu és aprendre un model que canvia al llarg del temps.La utilització de mètodes d'aprenentatge en línia és també interessant quan el conjunt $\mathcal{M}$ està disponible però no pot ser processat (o emmagatzemat) degut a les limitacions físiques de la màquina en la qual es desenvolupa l'aprenentatge.

En un esquema d'aprenentatge en línia (com el de la Figura 3.9), l'única informació disponible en cada instant és la mostra més recent (juntament amb la seua etiqueta). Açò redueix les necessitats de memòria i demanda computacional al mínim.

En la Secció 5.2 d'aquest capítol es formula un mètode d'aprenentatge de distàncies per lots a través d'un problema d'optimització regularitzat amb restriccions. Les aportacions en línia que s'introdueixen es formulen a partir d'aquest algorisme. En la Secció 5.3, es plantegen diferents formulacions en línia sota un esquema Passiu-Aggressiu. Finalment, la Secció 5.4 recull tota l'experimentació relativa a la família de mètodes en línia presentats juntament amb les conclusions.

5.2 Aprenentatge de distàncies mitjançant maximització del marge

Una manera de formular l'aprenentatge de distàncies és mitjançant la maximització del marge entre els valors de les distàncies. Per a això, es consideren els valors de les distàncies com a quantitats escalars que seran relativament grans (o petites), depenent de si fan referència a parells d'elements de diferent classe (o mateixa classe). A més, aquests valors de distància estaran separats idealment quan la distància siga òptima per a discriminar-los.

La Figura 5.1 il·lustra un histograma en què s'han representat els valors de distància corresponents a parells d'elements similars i dissimilars. En aquest, els valors de distància entre elements similars són clarament inferiors al dels elements dissimilars.

Una possible formulació consisteix a maximitzar el marge de separació entre ambdós tipus de distàncies. Es pot pensar en una solució ideal com aquella que maximitza el marge de separació d'acord amb el principi de minimització del risc estructural [92]. Però en el cas (més habitual) en què les restriccions no es poden satisfer, és a dir, no pot establir-se un marge òptim entre els valors de les distàncies pot contemplar-se la utilització del que es coneix com un marge tou. Aquesta alternativa flexibilitza la condició de marge i permet intrusions dels valors de les distàncies respecte del marge. Habitualment es tracta d'optimitzar les intrusions i establir una mesura per controlar-les, amb la convicció que una bona distància és aquella en que tant el nombre com la intensitat de les intrusions siga mínima.

Denotem per $b \in \mathbb{R}$ el punt central del marge de separació òptim. Al valor b l'anomenarem llindar, ja que és el valor de distància que millor separa les mostres similars i dissimilars. A més, fixem un valor del marge de separació de 2 i establim

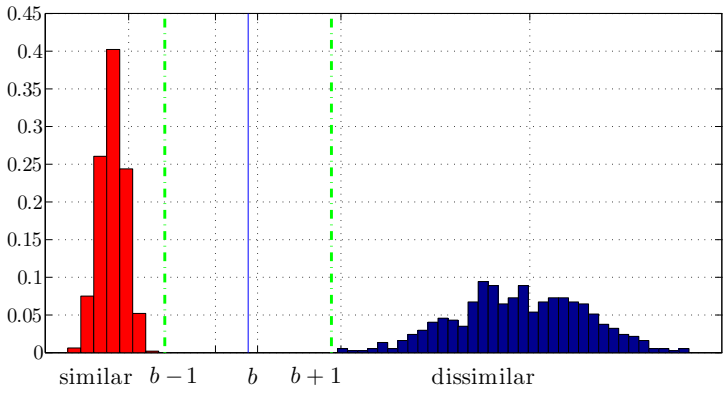


Figura 5.1: Histograma de distàncies d'una situació idealitzada de maximització del marge entre valors de distàncies.

les restriccions:

$$d_{ij}^M \leq b-1, \text{ si } y_{ij}=1,$$
$$d_{ij}^M \geq b+1, \text{ si } y_{ij}=-1,$$

o, en una forma més compacta

$$y_{ij}(b-d_{ij}^M) \geq 1. \tag{5.1}$$

Sota aquesta formulació, analitzem els casos separable i no separable.

5.2.1 Cas separable

El problema d'aprendre una distància que maximitze el marge es pot formular d'una manera similar a la manera en què es fa amb màquines de vectors suport [80]. En particular, s'estableix un valor fix per aquest marge i l'objectiu passa a ser la minimització d'un regularitzador sotmès a restriccions sobre aquest marge. En particular, s'utilitza com a regularitzador la norma de Frobenius que és el més habitual [54, 81].

Les restriccions de l'Equació (5.1), juntament amb el criteri de regularització i la restricció PSD, donen lloc a un problema d'optimització quadràtica que es pot resoldre de diverses maneres [98, 77]. La seua resolució ens dóna la matriu M i el valor b que separa de manera òptima ambdós tipus de valors de distància. Aquest problema pot formular-se de la manera següent

$$\begin{aligned} \min_{M,b,\xi} \quad & \frac{1}{2}\|M\|_{Fro}^2, \\ \text{tal que} \quad & y_{ij}(b-d_{ij}^M) \geq 1, \\ & M \succeq 0, \quad \forall 1 \leq i < j \leq N, \end{aligned} \tag{5.2}$$

on $N = \frac{n(n-1)}{2}$ són tots els possibles parells sense repeticions per a n elements.

5.2.2 Cas no separable

Independentment d'altres consideracions, les restriccions del problema (5.2) poden modificar-se per tenir en compte el cas no separable mitjançant la introducció d'una variable de folgança (no negativa) per a cada restricció. D'aquesta forma, les restriccions en (5.2) es modifiquen afegint aquesta variable de la següent manera

$$y_{ij}(b - d_{ij}^M) \geq 1 - \xi_{ij}. \tag{5.3}$$

Això s'ha de completar amb la inclusió d'un nou terme en el criteri d'optimització que penalitze valors alts de les variables de folgança. Altrament dit, s'estableix un marge tou per a separar els dos tipus de valors de distància. Això és equivalent a minimitzar la pèrdua bisagra associada amb tots els parells que violen la separació de marge (dur). Donat un model de distància (M, b) i un parell (i, j), la corresponent pèrdua bisagra ve donada com a

$$\ell_{ij}^{(M,b)} = \max\left\{0, p_{ij}^{(M,b)}\right\} = \max\left\{0, 1 - y_{ij}\left(b - d_{ij}^M\right)\right\}$$

on anomenarem $p_{ij}^{(M,b)}$ a la pèrdua amb signe predit per al parell (i, j), quan s'empra el model (M, b). Conseqüentment, la restricció en l'Equació (5.3) pot expressar-se com $p_{ij}^{(M,b)} \leq \xi_{ij}$ o equivalentment com $\ell_{ij}^{(M,b)} \leq \xi_{ij}$.

L'objectiu doncs, és separar les distàncies amb un marge màxim. Aquest es defineix formalment fent ús de la coneguda formulació de les màquines de vectors suport amb marge tou, definit mitjançant les variables de folgança ξ_{ij}.

$$\begin{aligned} \min_{M,b,\xi} \quad & \frac{1}{2}\|M\|_{Fro}^2 + C\sum_{ij}\xi_{ij}, \\ \text{tal que} \quad & \ell_{ij}^{(M,b)} \leq \xi_{ij}, \\ & \xi_{ij} \geq 0. \end{aligned} \tag{5.4}$$

5.2.3 Solució mitjançant programació quadràtica

Derivant el problema primal (5.4) i fent ús del mètode dels multiplicadors de Lagrange, que s'explica amb més detall a l'Apèndix D, s'obté el problema dual

$$\begin{aligned} \max_{\alpha} & \sum_{\ell=1}^{N}\alpha_\ell - \frac{1}{2}\sum_{\ell,m=1}^{N}\alpha_\ell\alpha_m y_\ell y_m \langle Z_\ell, Z_m\rangle, \\ \text{tal que} & \sum_{\ell=1}^{N}\alpha_\ell y_\ell = 0, \\ & 0 \leq \alpha_\ell \leq C, \quad 1 \leq \ell \leq N, \end{aligned} \tag{5.5}$$

a on $\langle Z_\ell, Z_m\rangle = \mathrm{Tr}(Z_\ell^\top Z_m)$, fa referència al producte escalar entre les matrius $Z_\ell = (x_{i_\ell} - x_{j_\ell})(x_{i_\ell} - x_{j_\ell})^\top$, i l'índex ℓ fa referència al parell (i_ℓ, j_ℓ). L'expressió

que s'obté per a construir la matriu M es detalla en l'Apèndix D, i finalment ve donada per:

$$M = -\sum_{\ell \in ij} \alpha_\ell y_\ell Z_\ell. \tag{5.6}$$

Anàlogament a les SVM, els vectors suport són els x_ℓ corresponents als $\alpha_\ell > 0$, que estan exactament sobre el marge i satisfan $y_\ell(b - d_\ell^M) = 1$. De manera que es pot deduir $b - d_\ell^M = \frac{1}{y_\ell} \Leftrightarrow b = d_\ell^M + y_\ell$ (fent servir que $\frac{1}{y_\ell} = y_\ell$), el que permet definir el valor b. En la pràctica, resulta més robust realitzar la mitjana sobre tots els vectors suport, definim SV com el conjunt que conté tots els índexs dels vectors suport, pot calcular-se b de la manera següent

$$b = \frac{1}{|\mathrm{SV}|} \sum_{\ell \in \mathrm{SV}} (d_\ell^M + y_\ell). \tag{5.7}$$

El problema (5.5) és un problema quadràtic i es pot resoldre amb qualsevol mètode estàndard. També es pot fer servir qualsevol mètode específic [12]. L'únic problema però, és que la matriu que finalment s'obté (5.6) no té per què complir la restricció $M \succeq 0$. La solució més natural consisteix a trobar una solució aproximada que satisfaça les restriccions. En particular, s'aproxima mitjançant la matriu semidefinida positiva més pròxima a la solució de l'Equació (5.6) en norma de Frobenius [75], que ve donada per

$$\sum_{i=1}^{d} \max\{0, \lambda_i\} v_i v_i^\top, \tag{5.8}$$

a on v_i és el $i-$èssim vector propi de la matriu M associat al valor propi λ_i.

En el cas particular d'aprenentatge de distàncies, és necessària altra restricció addicional en la definició del problema [83], el valor del llindar b ha d'estar per sobre d'1, és a dir $b \geq 1$.

En tots els plantejaments anteriors, aquestes restriccions no s'han considerat en la formulació del problema d'optimització.

5.3 Formulació en línia utilitzant maximització del marge

En lloc de considerar l'optimització amb totes les mostres disponibles, el problema d'aprenentatge de distàncies es pot resoldre d'una manera més convenient des del punt de vista de la computació mitjançant l'ús d'un enfocament d'aprenentatge en línia [83, 73]. A cada pas, un nou problema d'optimització es formula i es resol utilitzant només un parell etiquetat (i, j) que està a disposició del sistema. Com que a cada pas es processa un únic parell, ens referirem tant al parell com al pas mitjançant l'índex k. El problema empra el model (M^k, b^k) après en l'anterior pas per produir un nou model que té també la nova instància en compte.

El resultat final òbviament no serà una solució per al problema d'optimització sobre totes les mostres. Però, es poden establir certes garanties quant a la bondat de les solucions corresponents a les formulacions en línia.

Les següents formulacions , són conegudes com formulacions Passives-Agressives (PA) i van ser introduïdes en general en [21].

5.3.1 Formulació passiva-agressiva (PA)

Cas separable

En el cas separable es minimitza una mesura convenient de la distància entre l'anterior i el nou model, subjecte a la restricció que la nova instància ha de caure en el costat correcte del marge (dur) [83]. Més concretament, es pot escriure com

$$\min_{M,b} \quad \tfrac{1}{2}\left\|M - M^k\right\|_{\mathsf{Fro}}^2 + \tfrac{1}{2}\left(b - b^k\right)^2, \tag{5.9}$$

$$\text{tal que} \qquad \ell_k^{(M,b)} = 0, \tag{5.10}$$

on ℓ_k és la pèrdua frontissa comesa sobre el $k-$èssim parell, (x_i, x_j) i $\|\cdot\|_{\mathsf{Fro}}$ és la norma de Frobenius. Una conseqüència òbvia d'aquesta formulació és el fet que el model només necessita ser actualitzat si la nova instància origina una pèrdua estrictament positiva quan s'està emprant el model previ.

5.3.2 Cas no separable

En el cas no separable, es permet a la nova instància la violació de la condició de marge però és penalitzada emprant un paràmetre C per a donar-li més o menys pes relatiu a la restricció. Açò es pot fer com en el cas per lots, introduint una variable de folgança.

$$\min_{M,b,\xi} \quad \tfrac{1}{2}\left\|M - M^k\right\|_{\mathsf{Fro}}^2 + \tfrac{1}{2}\left(b - b^k\right)^2 + C\xi, \tag{5.11}$$

$$\text{tal que} \qquad \ell_k^{(M,b)} \leq \xi, \tag{5.12}$$

$$\xi \geq 0. \tag{5.13}$$

En aquesta formulació, la variable (no negativa) de folgança, ξ, permet a una instància particular la violació de la restricció a canvi d'un increment en el criteri (5.11) ponderat per C.

Un camí alternatiu per a plantejar el problema és eliminar la restricció de no negativitat de la variable de folgança (Equació (5.13)) i considerar un terme de penalització quadràtic en l'Equació (5.11).

$$\min_{M,b,\xi} \quad \tfrac{1}{2}\left\|M - M^k\right\|_{\mathsf{Fro}}^2 + \tfrac{1}{2}\left(b - b^k\right)^2 + C\xi^2, \tag{5.14}$$

$$\text{tal que} \qquad \ell_k^{(M,b)} \leq \xi, \tag{5.15}$$

Seguint la taxonomia presentada pels autors, es farà referència a aquestes dues formes de plantejar el problema com PAI (no-separable, penalització lineal) o

PAII (no-separable, penalització quadràtica). A més a més, ens referirem al cas separable com PA0.

Aquestes formulacions s'han aplicat en diversos problemes d'aprenentatge de distàncies o estan estretament relacionats amb aquests [83, 18, 73]. S'ha demostrat [83] que en tots els casos anteriors la solució d'aquests problemes té una expressió tancada que pot ser escrita com la següent regla d'actualització

$$M^{k+1} = M^k - \tau y_{ij}(x_i - x_j)(x_i - x_j)^\top, \tag{5.16}$$

$$b^{k+1} = b^k + \tau y_{ij}, \tag{5.17}$$

on el valor de τ s'anomena longitud de pas i el seu valor serà diferent per a cada una de les tres formulacions anteriors a les quals assignem els índexs $n = 0, 1$ i 2, respectivament. La actualització (5.16) resulta ser una correcció de rang 1 sobre la matriu anterior.

$$\text{PA0:} \quad \tau_0 = \frac{\ell_k}{1 + \|(x_i - x_j)(x_i - x_j)^\top\|^2_{\mathsf{Fro}}}, \tag{5.18}$$

$$\text{PAI:} \quad \tau_1 = \min\{C, \tau_0\}, \tag{5.19}$$

$$\text{PAII:} \quad \tau_2 = \frac{\ell_k}{1 + \frac{1}{2C} + \|(x_i - x_j)(x_i - x_j)^\top\|^2_{\mathsf{Fro}}}. \tag{5.20}$$

$$\tag{5.21}$$

Aquestes longituds de pas són dependents de la pèrdua bisagra comesa pel nou (k-èssim) parell, respecte a l'actual model predictiu, $\ell_k = \ell_k^{(M^k, b^k)}$. Si l'etiqueta del parell és consistent amb la predicció del model en la iteració anterior, tenim $\ell_k = 0$ i per tant $\tau_n = 0$. Això implica que no hi ha correcció en aquesta iteració (passivitat). D'altra banda, quan la predicció viola el marge (dur o bla), el model s'actualitza ja siga en sentit estricte (τ_0) o controlat pel paràmetre d'agressivitat C. Per tant, C controla la força amb que l'algorisme adapta el model a cada parell en cada pas.

5.3.3 Formulació basada en mínims quadrats

També és possible una formulació alternativa de l'anterior problema d'aprenentatge de distàncies [69] utilitzant mínims quadrats [86]. En lloc de forçar un marge suau al penalitzar la desviació de les condicions ideals, és possible obligar els valors de distància similars i dissimilars a caure prop dels valors representatius $b-1$ i $b+1$, respectivament. Amb aquesta finalitat, es pot minimitzar l'error quadràtic corresponent de manera seqüencial. Això correspon a la reformulació de l'anterior problema d'optimització (5.14), (5.15) (versió PAII) com:

$$\min_{M,b,\xi_k} \quad \tfrac{1}{2}\left\|M - M^k\right\|^2_{\mathsf{Fro}} + \tfrac{1}{2}\left(b - b^k\right)^2 + C\xi_k^2, \tag{5.22}$$

$$\text{tal que} \qquad p_{ij}^{(M,b)} = \xi_k. \tag{5.23}$$

El principal canvi en aquesta formulació és que la restricció de desigualtat en l'Equació (5.15) ha canviat a una igualtat que utilitza la funció de pèrdua amb signe, $p_{ij}^{(M,b)} = 1 - y_{ij}\left(b - d_{ij}^{M}\right)$. La violació d'aquesta restricció ara mesura com de lluny està el valor de distància del seu corresponent valor representatiu ($b-1$ o $b+1$).

El problema d'optimització corresponent pot abordar-se de manera similar als anteriors. Açò també dóna lloc a una solució tancada (els detalls de la derivació es donen en l'Apèndix E), que consisteix en la mateixa regla d'actualització que l'expressada en les Equacions (5.16) i (5.17), però amb una longitud de pas diferent donada per

$$\tau_3 = \frac{p_k}{1 + \frac{1}{2C} + \|X_k\|_{\mathsf{Fro}}^2}, \tag{5.24}$$

on $p_k = p_{ij}^{(M^k,b^k)} = 1 - y_{ij}\left(b^k - d_{ij}^{M^k}\right)$ és la pèrdua amb signe predit fent ús del k-èssim model després i $X_k = (x_i - x_j)(x_i - x_j)^\top$. Cal notar que τ_3 ara pot prendre valors negatius i es compleix que $\tau_2 = \max\{0, \tau_3\}$. Conseqüentment, el corresponent algorisme pot considerar-se més agressiu perquè implica un major nombre d'actualitzacions del model. Com els diferents enfocaments comparteixen l'estructura i part dels objectius de la filosofia PA, ens referirem a la nova formulació com a PALS (de l'anglès Passive-Agressive Least Squares) en aquest treball. Només quan es compleix que $p_k = 0$, es realitza una etapa passiva. Això només es produeix quan el valor de distància té exactament el valor desitjat, la qual cosa està en contrast amb els enfocaments passius-agressius purs que realitzen un pas passiu quan $\ell_k = \max(0, p_k) = 0$, és a dir quan $p_k \leq 0$.

Es pot observar que tots els algoritmes d'aprenentatge anteriors i, en particular, les longituds de pas corresponents (τ_n), estan estretament relacionades entre si. En particular, podem escriure aquestes longituds de pas en funció de C i p_k per a presentar les seves interdependències de manera explícita.

$$\begin{array}{lrcll}
\text{PALS:} & \tau_3(C, p_k) & = & \dfrac{p_k}{1 + \frac{1}{2C} + \|X_k\|_{\mathsf{Fro}}^2} & \in \quad]\text{-}\infty, +\infty[\ , \\
\text{PAII:} & \tau_2(C, p_k) & = & \max(0, \tau_3(C, p_k)) & \in \quad [0, +\infty[\ , \\
\text{PA:} & \tau_0(p_k) & = & \lim\limits_{C \to \infty} \tau_2(C, p_k) & \in \quad [0, +\infty[\ , \\
\text{PAI:} & \tau_1(C, p_k) & = & \min(C, \tau_0(p_k)) & \in \quad [0, +\infty[\ .
\end{array}$$

En la Figura 5.2, es mostra una il·lustració gràfica per aclarir com les diferents longituds de pas es relacionen entre si. En la Figura 5.2a, els valors de τ_n es mostren com a funció de p_k per a un valor donat de C. Es pot veure clarament que τ_1 és una versió saturada de τ_o, mentre que τ_2 correspon a una versió més suau de la (insaturada) τ_1. D'altra banda, τ_3 és igual a τ_2 en el cas positiu però continua mantenint les correccions en el model també en el cas negatiu. En la Figura 5.2b, es mostren els diferents valors de τ_n per a dos valors diferents de p_k com a funció de C. Cal notar que aquests dos valors de p_k donen lloc a dos valors asimptòticament diferents de la forma $v_k = \frac{p_k}{1+\|(x_i-x_j)(x_i-x_j)^\top\|_{\mathsf{Fro}}^2}$, $k = 1, 2, \ldots$ Es pot observar que τ_3 és una aproximació de τ_0 per a valors grans de C. En

ambdues il·lustracions pot observar-se que τ_1 i τ_2 poden donar lloc a correccions més fortes o més suaus depenent dels valors de predicció i del paràmetre C.

5.3.4 Satisfacció de restriccions

El rendiment particular de les solucions anteriors pot dependre de dues consideracions pràctiques molt importants relacionades amb la satisfacció de restriccions PSD i ajustament de paràmetres. S'ha comentat en apartats anteriors que la restricció PSD s'aplica després de realitzar el pas d'optimització. En particular, com que en cada pas només es fa una correcció de rang 1, l'únic valor propi negatiu es fa zero tan aviat com apareix en cada iteració. Això condueix a una raonable aproximació al problema original [83]. No obstant això, també és possible ajornar les correccions PSD passat un nombre fix de passos, o fins i tot fins al final del procés iteratiu [18]. Aquesta estratègia té un benefici computacional obvi ja que fins i tot per als algorismes incrementals especialitzats, recalcular la descomposició en valors i vectors propis és una operació costosa. D'altra banda, la utilització de matrius indefinides durant el procés iteratiu d'optimització dóna lloc en la pràctica a un tipus de concentració de la informació en els valors propis negatius. Es consideren doncs dues possibilitats per a cada algorisme d'aprenentatge de distàncies en línia. Aquestes dues versions s'anomenen respectivament, positiva (marcada amb el símbol +) i negativa (denotada amb el símbol -). En el primer

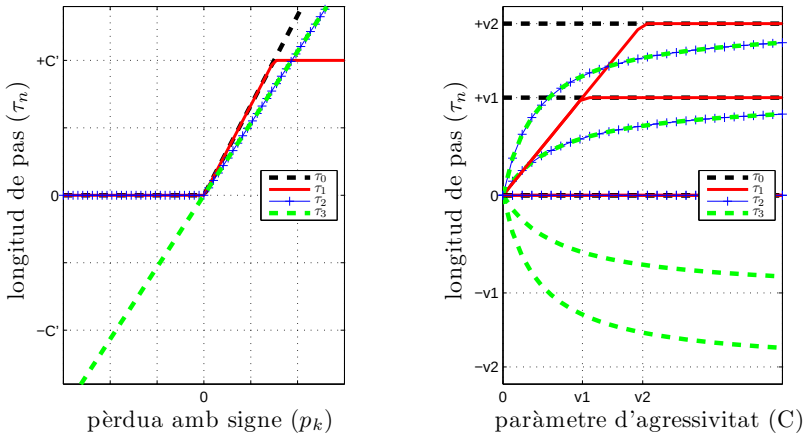


Figura 5.2: Diferents longituds de passos corresponents al PA separable, PAI, PAII i PALS. (a) Com a funció de la pèrdua amb signe per al parell actual, p_k, per a un valor donat de C. El valor C' mostrat correspon a $C \cdot (1 + \|X_k\|^2_{\mathsf{Fro}})$. (b) com a funció de C per a dos valors diferents de p_k que donen lloc a dos valors asimptòtics diferents (v_1 and v_2). Tant el cas positiu com el negatiu ($p_k > 0$ i $p_k \leq 0$) es mostren per a cada valor de p_k.

cas, la correcció PSD s'aplica a cada iteració. En el segon, les matrius indefinides es deixen com a tals fins al final del procés. Com es veurà en l'experimentació posterior, es produeixen matrius (molt) més disperses en les versions negatives, amb un poder de predicció lleugerament per sota de les corresponents positives.

L'altra qüestió important és establir un valor correcte del paràmetre d'agressivitat, C. El valor òptim de C pot dependre de diferents aspectes, com per exemple com de separable és el problema. En contrast amb el mateix concepte en classificació binaria, aquesta separabilitat en valors de distància és difícil d'analitzar i gestionar. Totes aquestes consideracions empitjoren perquè el valor òptim de C pot dependre fins i tot del parell en particular. Per tal de simplificar l'experimentació, el valor òptim del paràmetre C s'ha fixat per a totes les iteracions, però de manera diferent per a cada algorisme. Amb aquest fi, s'ha realitzat una ronda de validació, triant finalment el valor del paràmetre que dóna lloc a un millor resultat. Per establir aquest valor, s'han considerat dos criteris alternatius: a) el valor que minimitza la pèrdua acumulada i b) el que maximitza el rendiment de la matriu apresa quan s'utilitza juntament amb el classificador dels $k-$veïns, fent servir el millor k entre 1 i 25.

Val la pena mencionar la dependència del resultat amb la matriu inicial que es dóna a l'algorisme en línia. En general, es pot pensar en un inici amb un model aleatori, un model previ (per exemple obtingut fent servir un aprenentatge per lots sobre un conjunt reduït) o fer servir el model nul (o zero). L'ús d'un model prèviament aprés pot accelerar l'obtenció d'una bona solució però també pot condicionar el comportament de l'algorisme en línia. Un model aleatori pot forçar l'algorisme a començar lluny de les regions òptimes de l'espai de solucions però condicionant la dispersió del model. El model nul té l'avantatge de començar d'un model neutral, dispers i sense condicionar. En aquest cas, és possible fitar la pèrdua acumulada del model [83].

Finalment, es pot escriure un esquema algorísmic comú vàlid per als 3 mètodes considerats tant en les seves versions positives o negatives i que es mostra en l'Algorisme 2.

5.4 Experiments i resultats

Per tal de comparar i avaluar els beneficis i desavantatges dels algoritmes en línia considerats en aquest treball, s'ha realitzat una experimentació exhaustiva. Per aquest fi, una sèrie d'experiments mostren el seu comportament com a processos en línia. A més, els diferents indicadors (pèrdua acumulada i màxima taxa de classificació), s'han utilitzat en combinació amb el classificador dels $k-$veïns més pròxims per avaluar la bondat de l'espai de característiques en la tasca de classificació. A més a més, la càrrega computacional de cada mètode s'ha emmagatzemat per tal de mostrar i comparar la seva execució en un disseny equitatiu i neutral.

Algorisme 2 Aprenentatge en línia de distàncies passiu agressiu

Entrada: $\{x_i \mid i \in \mathbb{N}\}, P \subseteq \mathbb{N} \times \mathbb{N}$ //conjunt d'objectes i parells d'índexs
Entrada: $\{y_{ij} \mid (i,j) \in P\}$ //etiqueta de similitud
Entrada: M^0, b^0,C. //model inicial i paràmetre d'agressivitat
Entrada: versió, tolerància. //versió (+/−) i tolerància en l'agressivitat
Eixida: M, b. //model aprés

$k = 0$
repetir
  $(i,j) \leftarrow$ ObténouParellDe(P)
  Estableix τ d'acord amb les Equacions (5.19), (5.21) o (5.24).
  si $|\tau| \geq$ tolerància **aleshores**
    $M^{k+1} \leftarrow M^k - \tau y_{ij}(x_i - x_j)(x_i - x_j)^\top$
    $b^{k+1} \leftarrow b^k + \tau y_{ij}$
    $k \leftarrow k + 1$
    si versió = (+) **aleshores**
      (M^k, b^k) ←AplicaResctriccions(M^k, b^k)
    fi si
  sino
    $M^{k+1} \leftarrow M^k$
    $b^{k+1} \leftarrow b^k$
    $k \leftarrow k + 1$
  fi si
fins que convergència

si versió = (−) **aleshores**
  (M^k, b^k) ←AplicaRestriccions(M^k, b^k)
fi si

$(M, b) \leftarrow (M^k, b^k)$

5.4.1 Detalls dels experiments

La configuració de l'experimentació en aquest treball s'ha fixat com se suggereix en [24] i altres estudis preliminars anteriors [69]. A més a més, s'ha considerat el mètode ITML [24], com a punt de referència a efectes de comparació. L'algorisme ITML s'ha utilitzat com se suggereix en [24], utilitzant el programari posat a disposició pels autors que dóna el seu propi mètode per establir paràmetres.

S'ha utilitzat un conjunt P, composat per $|P| = 40c(c-1)$ parells sense repetició per a garantir que tots els mètodes considerats (incloent l'ITML) fan servir la mateixa quantitat d'informació. Aquest conjunt s'utilitza per a establir les relacions entre els elements de $\mathcal{M}$ mitjançant les seues etiquetes. El conjunt P s'ha seleccionat a l'atzar per a l'entrenament de tots els mètodes.

En particular, al conjunt P es barregen (aleatòriament) tots els parells i es proporcionen seqüencialment a l'algorisme. Aquest conjunt es proporciona com a entrada a l'algorisme almenys dues vegades i, a continuació aquest procés es repe-

teix fins que s'ha arribat a un nombre màxim d'iteracions. Aquest nombre màxim d'iteracions s'ha establit com el mínim entre el 20% de tots els possibles parells d'entrenament sobre les mostres (com en estudis previs [73]) i 50 vegades el nombre de parells en P (que és, amb molt, més que suficient per a les bases de dades més grans amb menor nombre de classes). Açò vol dir que almenys s'executen un total de t passos, amb $t = \min(\lfloor \frac{1}{5} \frac{\frac{n}{2}(\frac{n}{2}-1)}{40} \rfloor, 50|P|) = \min(\lfloor \frac{n(n-2)}{40} \rfloor, 50|P|)$. Aquest valor s'ha fixat com a un punt d'equilibri entre cost computacional i rendiment.

Per tal d'aconseguir un valor adequat de C de manera automàtica, cada conjunt fix de parells s'utilitza per entrenar un model per a cada versió de cada mètode en línia. Aquest model s'entrena fixant $t = |P|$, que correspon a donar-li a cada parell una única vegada. El model final s'ha validat sobre el conjunt d'entrenament complet.

Un rang apropiat de valors espaiats exponencialment en el rang $[10^{-4}, 10^{2}]$ s'ha considerat com l'espai de paràmetres. Tots els models en línia s'han inicialitzat amb el model buit, que és $b = 0$ i la matriu zero com se suggereix en [21]. Tots els resultats presentats són la mitjana de 10 execucions independents amb diferents inicialitzacions aleatòries per a obtindre conjunts diferents d'entrenament i test, però tenint en compte la utilització de la mateixa informació per a cadascun dels algorismes considerats.

5.4.2 Avaluació del rendiment

Comparativa de predicció en línia.

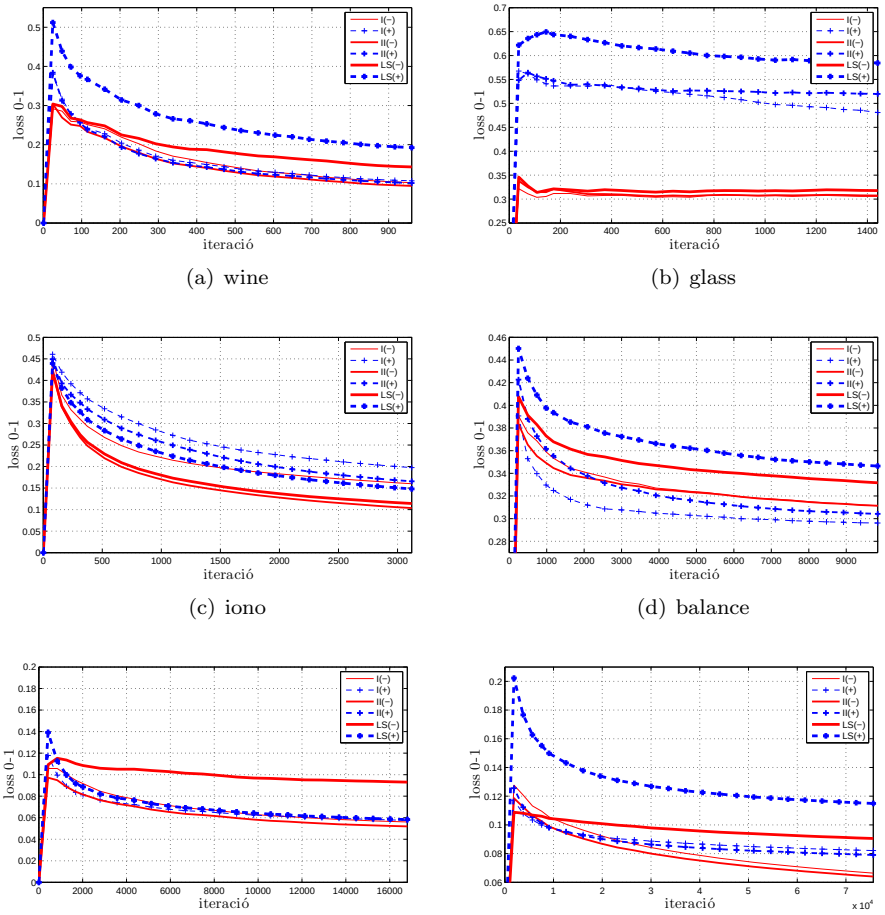
Els diferents algorismes s'han fet servir per a obtindre una matriu per a realitzar una classificació basada en distàncies. També s'ha avaluat el comportament com a mètodes en línia en els experiments. Amb aquesta finalitat, diferents pèrdues i mesures de rendiment s'han considerat al llarg del procés d'aprenentatge. Primer, la Figura 5.3 mostra la mitjana de la pèrdua predictiva 0-1 definida com

$$\ell = \frac{1}{t} \sum_{k=1}^{t} \text{sgn}(\ell_k) = \frac{1}{2t} \sum_{k=1}^{t} |y_k - \hat{y}_k| \,,$$

a on t és la longitud de la seqüència, y_k és la etiqueta real del parell subministrat al $k-$èssim pas, i $\hat{y}_k$ és la etiqueta predita emprant el $(k-1)-$èssim model. Aquesta mesura il·lustra el comportament dels diferents algorismes en línia al llarg del temps quan discriminen entre objectes similars i dissimilars. Només 6 de les 15 bases de dades es mostren en la Figura 5.3, sent aquestes representatives de tots els diferents comportaments observats en el conjunt complet d'experiments.

Els resultats en la resta de bases de dades es mostren en l'Apèndix C. En particular, el comportament de les bases de dades nist16 i spam són molt similars a la mostrada per a iono. Aquesta similitud també pot observar-se per a ecoli, malaysia, mor i satellite respecte a Art100. Les bases de dades chromo i breast mostren el mateix comportament que wine i glass respectivament. Finalment, soyS i soyL també exhibeixen un comportament molt similar.

Resumint, tots els algorismes d'aprenentatge en línia donen lloc a un comportament raonablement bo en els experiments, d'acord amb la mesura de la pèrdua.



(a) wine (b) glass

(c) iono (d) balance

(e) soyL (f) Art100

Figura 5.3: Pèrdua predictiva acumulada dels algorismes en línia.

En 12 de les 15 bases de dades, les versions negatives dels mètodes han donat lloc a millors resultats que les seues respectives versions positives. I en 5 d'elles la pitjor versió negativa és millor que totes les positives (glass, iono, breast, spam and nist16). D'altra banda, els mètodes LS han exhibit un comportament significativament pitjor en les bases de dades wine, chromo, balance, ecoli, mor, satellite, malaysia, soybean i Art100. Aquesta diferència en el comportament és encara més notòria en les versions positives dels algorismes.

Comportament de l'evolució de la dimensionalitat.

Per a comprendre millor el comportament dels algorismes, en la Figura 5.4 es mostra la mitjana de la dimensió efectiva de la matriu segons evoluciona l'optimització per al cas particular de la base de dades iono. Es pot observar que tots els algorismes en línia (que comencen amb una matriu zero) ràpidament incorporen noves dimensions i al mateix temps convergeixen també ràpidament a una dimensió particular. En la Figura 5.5, s'il·lustra el mateix experiment per a soyL mostrant el mateix comportament que iono.

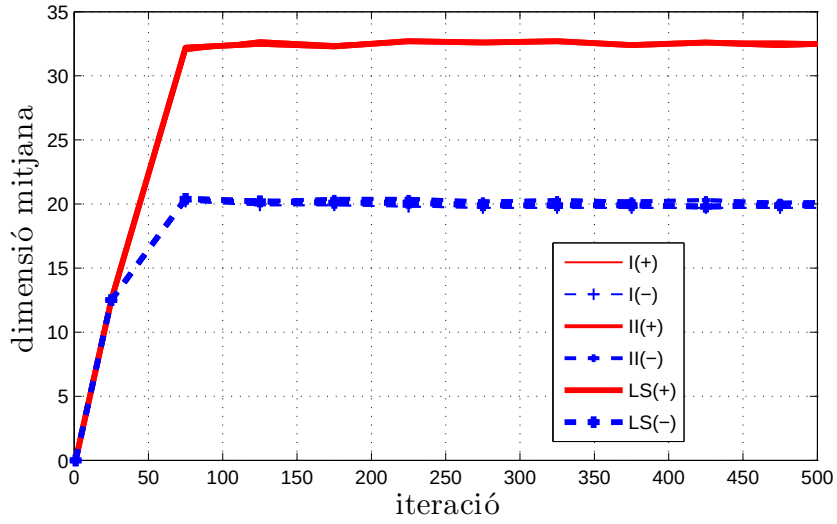


Figura 5.4: Mitjana de les dimensions efectives obtingudes emprant algorismes en línia en cada iteració en la base de dades iono.

La diferència entre les versions positives i negatives rau en el fet que les versions negatives convergeixen en una matriu molt més dispersa per que la informació es concentra en els vectors propis negatius de la corresponent matriu.

Les Figures 5.6 i 5.7, per a entrenament i test respectivament, il·lustren com es comporta la distància final quan s'utilitza en combinació amb un classificador dels $k-$veïns en funció de la dimensió de l'espai de característiques associat a la matriu M quan se seleccionen dimensions d'acord amb els valors propis. El

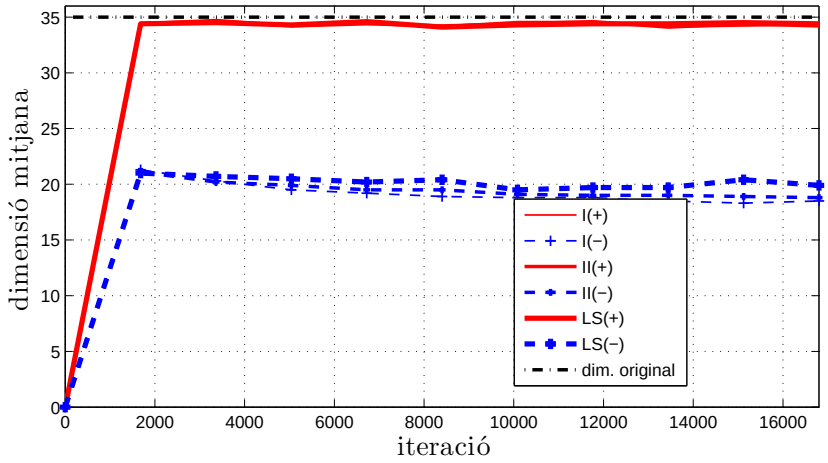


Figura 5.5: Mitjana de les dimensions efectives obtingudes emprant algorismes en línia en cada iteració en la base de dades soyL.

valor de k s'ha seleccionat com aquell que produeix millors resultats entre els 25 primers. Aquest valor resulta ser $k = 2$ per a iono. Aquestes dues figures il·lustren que el comportament en el conjunt d'entrenament és representatiu del que passa en el conjunt de test. En ambdós casos, les matrius obtingudes amb les versions positives i negatives encara contenen dimensions no informatives. Es pot veure que aproximadament els mateixos resultats es poden obtindre si les matrius es tallen als primers 10 vectors propis ordenats de major a menor. Les diferències entre els mètodes no són significatives i el mínim absolut mostrat en el conjunt d'entrenament no es correspon exactament amb el del conjunt de test.

En la Figura 5.8 (entrenament) i 5.9 (test) s'il·lustra el mateix experiment per a la base de dades soyL. El millor valor de k dins dels 25 primers valors és $k = 1$. En aquest cas no existeix un valor òptim de la dimensionalitat per baix de la dimensió original. La reducció de la dimensionalitat pareix no afectar al valor de l'error fins que no és relativament menuda (inferior a 10). Les corbes d'error són molt paregudes entre entrenament i test, i a més, el comportament de l'entrenament pot ser considerat com a un bon indicador sobre el valor de la dimensionalitat apropiada per aquest problema.

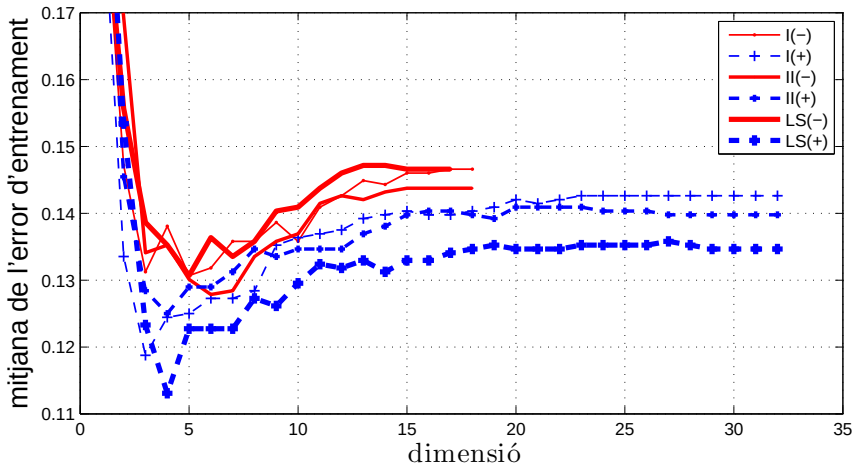


Figura 5.6: Mitjana dels errors sobre el conjunt d'entrenament obtinguts amb el classificador dels $k-$veïns considerant els vectors propis de les matrius obtingudes en els experiments en ordre d'importància en la Figura 5.4 (iono).

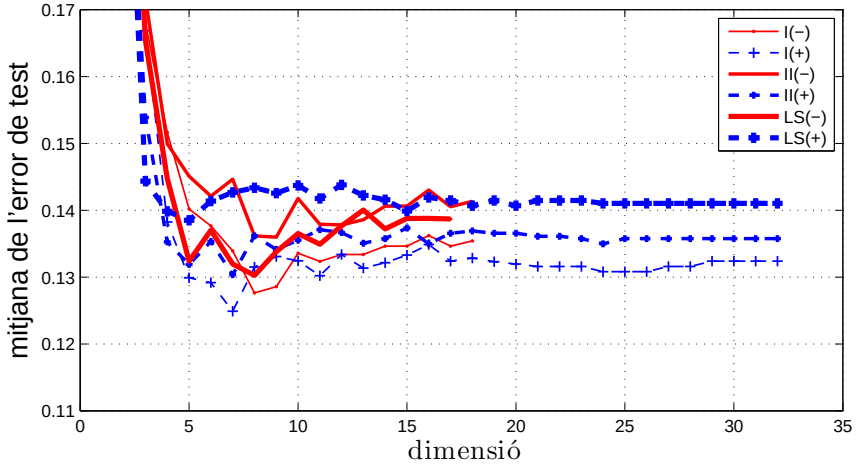


Figura 5.7: Mitjana dels errors sobre el conjunt de test obtinguts amb el classificador dels $k-$veïns considerant els vectors propis de les matrius obtingudes en els experiments en ordre d'importància en la Figura 5.4 (iono).

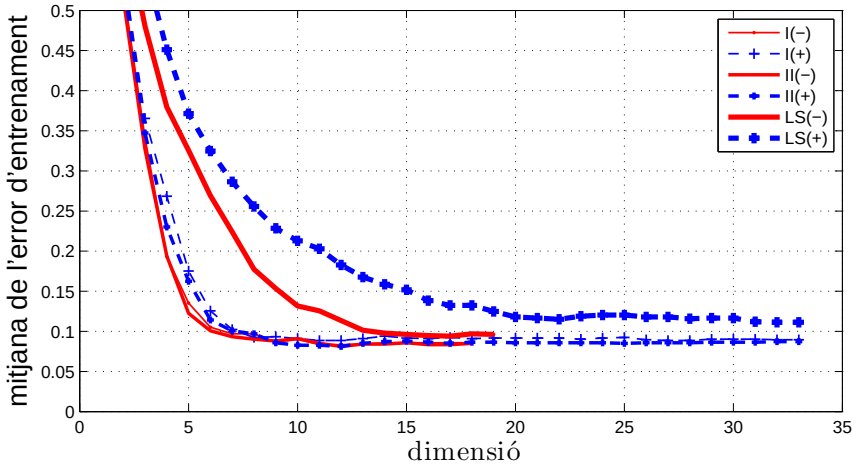


Figura 5.8: Mitjana dels errors sobre el conjunt d'entrenament obtinguts amb el classificador dels $k-$veïns considerant els vectors propis de les matrius obtingudes en els experiments amb soyL en ordre d'importància en la Figura 5.5.

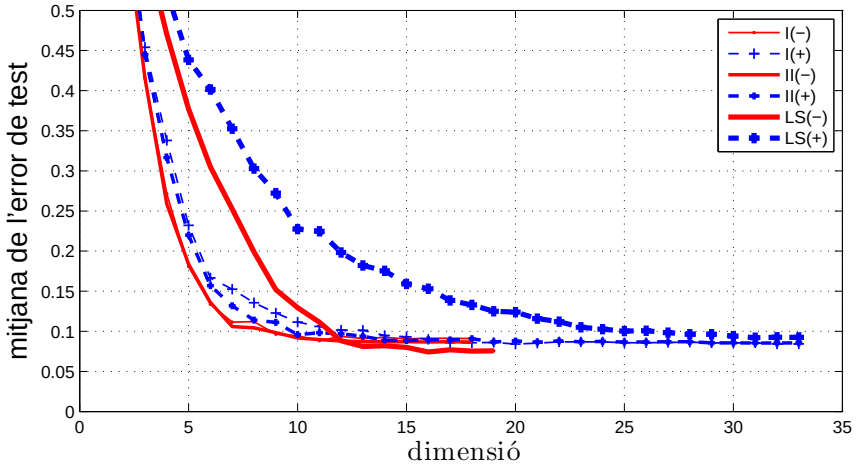


Figura 5.9: Mitjana dels errors sobre el conjunt de test obtinguts amb el classificador dels $k-$veïns considerant els vectors propis de les matrius obtingudes en els experiments amb soyL en ordre d'importància 5.5.

5.4.3 Mesures de càrrega computacional

Tots els experiments en totes les bases de dades s'han executat sobre la mateixa màquina. Les diferents execucions s'han restringit a utilitzar una única CPU per tal d'obtindre mesures més precises i objectives.

Les mitjanes dels temps de CPU en segons per a totes les bases de dades es mostren en la Taula 5.1, juntament amb les corresponents desviacions estàndard. Els mateixos temps d'execució apareixen en la Figura 5.10, fent servir una escala logarítmica en l'eix del temps per a il·lustrar gràficament els mèrits relatius de cada algorisme en cada base de dades.

Per completar, i per avaluar millor els temps de CPU dels diferents algorismes, s'ha realitzat un test de comparacions múltiple de Friedman. Per a això, s'han considerat els temps de totes les execucions sobre totes les bases de dades. Les mitjanes dels rangs resultants (de més ràpid a més lent) es mostren en la Taula 5.2. La Taula 5.3, mostra les comparacions per parells més importants sobre els mètodes i els seus p-valors ajustats respecte al test de Holm.

Taula 5.1: Temps de CPU en segons, es mostra la mitjana juntament amb la desviació estàndard (entre parèntesis).

	ITML		PAI+		PAII+		PALS+		PAI-		PAII-		PALS-	
soyS	10.56	(12.08)	0.30	(0.01)	0.30	(0.04)	0.49	(0.05)	0.11	(0.02)	**0.10**	(0.05)	0.11	(0.04)
wine	0.72	(1.17)	0.18	(0.03)	0.15	(0.04)	0.29	(0.02)	0.10	(0.00)	**0.10**	(0.00)	0.10	(0.00)
glass	0.53	(1.28)	0.26	(0.04)	0.25	(0.04)	0.32	(0.03)	**0.08**	(0.06)	0.09	(0.07)	0.11	(0.08)
ecoli	0.85	(1.08)	0.24	(0.07)	0.28	(0.08)	0.32	(0.03)	**0.05**	(0.01)	0.05	(0.01)	0.06	(0.01)
malaysia	4.02	(0.12)	1.25	(0.06)	1.24	(0.06)	2.06	(0.04)	1.15	(0.10)	1.10	(0.10)	**1.09**	(0.22)
iono	2.15	(0.69)	0.80	(0.06)	0.84	(0.13)	1.37	(0.11)	**0.28**	(0.03)	0.29	(0.04)	0.34	(0.04)
balance	2.11	(1.17)	0.46	(0.02)	0.46	(0.02)	0.58	(0.01)	**0.28**	(0.01)	0.28	(0.01)	0.30	(0.00)
breast	0.29	(0.45)	0.57	(0.12)	0.63	(0.14)	0.83	(0.12)	**0.26**	(0.10)	0.27	(0.12)	0.35	(0.09)
chromo	6.39	(0.30)	1.99	(0.11)	1.96	(0.09)	2.65	(0.28)	**1.71**	(0.07)	1.80	(0.07)	1.86	(0.15)
mor	4.08	(2.30)	4.77	(0.23)	4.72	(0.24)	8.00	(0.29)	**3.76**	(0.19)	3.81	(0.18)	4.23	(0.14)
spam	0.73	(0.57)	12.51	(1.08)	12.68	(1.31)	17.69	(1.17)	0.54	(0.07)	**0.50**	(0.07)	0.60	(0.01)
satellite	3.99	(7.65)	39.86	(2.40)	34.93	(5.20)	99.60	(0.42)	2.12	(0.09)	**2.07**	(0.05)	2.46	(0.06)
soyL	26.93	(0.97)	4.48	(1.03)	4.55	(1.01)	26.47	(2.37)	0.86	(0.12)	**0.85**	(0.09)	1.07	(0.04)
Art100	18.76	(0.12)	39.00	(1.30)	43.07	(3.20)	177.42	(18.85)	**4.32**	(0.06)	4.58	(0.32)	6.55	(0.06)
nist16	86.09	(46.08)	480.23	(17.15)	500.55	(19.73)	2007.20	(14.47)	**10.13**	(1.02)	10.81	(1.67)	34.06	(1.09)

Taula 5.2: Mitjana dels valors en les ordenacions dels mètodes d'acord amb el test de Friedman ($\alpha = 0.05$).

mètodes	posició promig
PAI-	1.49
PAII-	1.77
PALS-	2.87
PAII+	4.86
PAI+	4.93
ITML	5.67
PALS+	6.40

Taula 5.3: Resultats del test de Holm ($\alpha = 0.05$). Els casos en que es rebutja la hipòtesi nul·la es marquen amb negreta.

	ITML	PAI+	PAII+	PALS+
PAI-	$\mathbf{< 10^{-5}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-7}}$
PAII-	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-3}}$	**0.0011**	$\mathbf{< 10^{-7}}$
PALS-	**0.0046**	0.0967	0.1112	$\mathbf{< 10^{-3}}$

Dels resultats mostrats en la Taula 5.2, es pot concloure que tots els algorismes en línia són competitius a excepció del mètode PALS+ que queda al final de l'ordenació per sota de l'ITML. Com es natural, els menors temps d'execució s'han obtingut sempre utilitzant les versions negatives dels algorismes en línia. L'anàlisi de Friedman dóna diferències significatives entre les versions positives i negatives del mateix mètode. Açò ocorre fins i tot en el cas de PALS, que requereix aproximadament el doble d'actualitzacions que els altres algorismes en línia. També existeixen diferències significatives entre les versions negatives i l'ITML. Aquestes diferències augmenten amb la grandària de la base de dades. Es pot comentar que en tots els casos, el temps de CPU de la versió negativa està per sota del 40% de l'adoptat per l'algorisme ITML.

Finalment, la versió negativa de l'algorisme PALS és considerablement més lenta que les altres versions negatives dels algorismes en línia en les dues bases de dades més grans. En particular, és unes 3 vegades més lent per a nist16 i el doble de lent en el cas d'Art100. D'altra banda, els temps de CPU per a totes les versions negatives dels algorismes en línia són molt similars en totes les altres bases de dades.

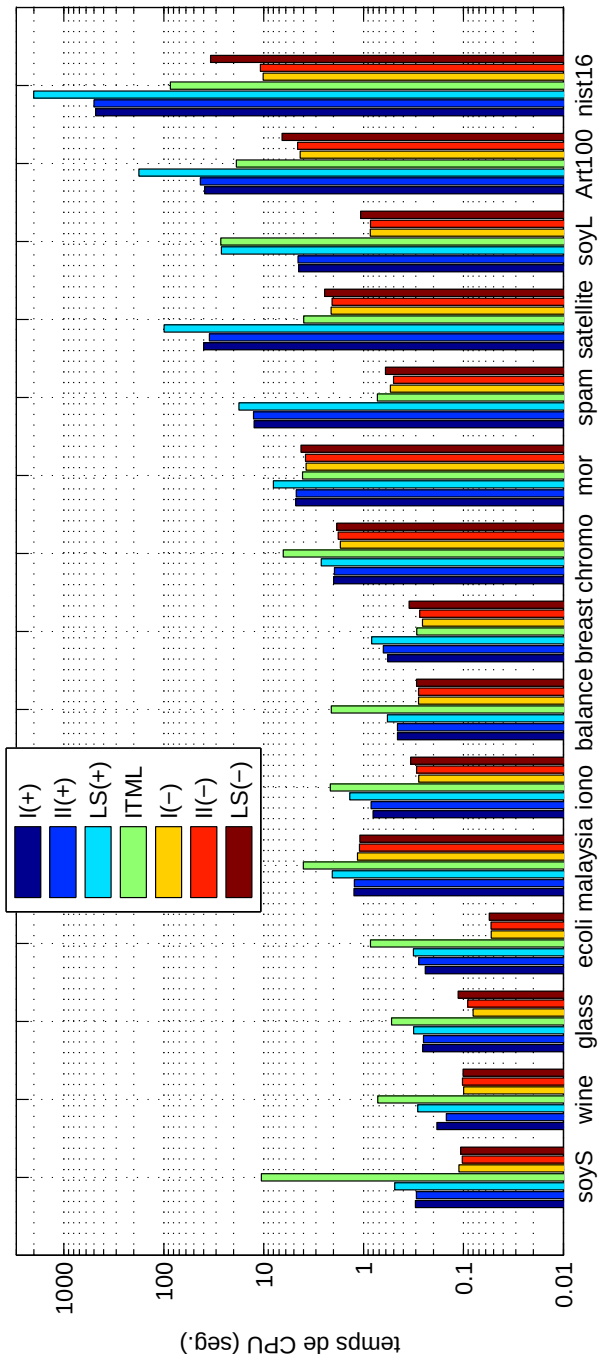


Figura 5.10: Mitjana dels temps de CPU per a cada algorisme en totes les bases de dades

Tasca de classificació

Per tal de mesurar la qualitat del resultat final dels diferents algorismes, s'han utilitzat les corresponents matrius, M, per a construir un classificador dels $k-$veïns més pròxims. S'ha calculat l'error de classificació utilitzant els 25 primers veïns i els millors resultats per a cada base de dades i cada mètode es mostren en la Taula 5.4.

Els errors de classificació obtinguts amb tots els algorismes, incloent l'ITML, indiquen un bon rendiment en la classificació en el context de l'experimentació duta a terme en aquest treball.

Un test de comparació multiple de Friedman [38] no ha revelat cap diferència significativa entre els resultats de classificació obtinguts amb l'ITML i qualsevol de les sis alternatives proposades. Es pot concloure doncs, que tots els algorismes donen lloc a resultats molt similars. També val la pena mencionar el fet que la combinació de l'enfocament LS amb matrius indefinides ha donat lloc als millors resultats en 3 de les 15 bases de dades. Aquestes són precisament les tres de major grandària, tant en nombre d'elements com en dimensionalitat. També podem comentar que, emprar la distància Euclidiana és la millor opció per a una de les bases de dades, ecoli, que ofereix un exemple on cap dels mètodes estudiats no han pogut ajudar a millorar els resultats de classificació de la distància Euclidana.

Taula 5.4: Promig dels errors de classificació i millor nombre de veïns (entre parèntesis). El millor resultat per a cada base de dades es mostra en negreta.

	Euclidean	ITML	PAI+	PAI-	PAII+	PAII-	PALS+	PALS-
soyS	0.158 (1)	**0.119** (2)	0.133 (14)	0.125 (13)	0.136 (12)	0.121 (15)	0.125 (9)	0.126 (12)
wine	0.027 (13)	0.033 (3)	0.018 (5)	0.017 (6)	**0.016** (3)	0.017 (9)	0.024 (7)	0.019 (3)
glass	0.510 (1)	0.513 (1)	0.523 (1)	**0.506** (1)	0.518 (1)	0.527 (1)	0.507 (1)	0.518 (1)
ecoli	**0.064** (3)	0.077 (5)	0.075 (3)	0.075 (17)	0.076 (15)	0.075 (7)	0.076 (17)	0.070 (9)
malaysia	0.301 (1)	0.298 (1)	0.281 (1)	0.289 (1)	**0.281** (1)	0.289 (1)	0.327 (1)	0.318 (1)
iono	0.153 (2)	0.155 (2)	**0.129** (2)	0.136 (2)	0.14 (2)	0.139 (2)	0.143 (2)	0.138 (2)
balance	0.335 (6)	0.263 (3)	**0.219** (3)	0.219 (2)	0.258 (3)	0.234 (2)	0.312 (6)	0.258 (6)
breast	0.033 (3)	0.037 (3)	**0.025** (9)	0.026 (11)	0.027 (17)	**0.025** (11)	0.028 (9)	0.030 (9)
chromo	**0.467** (1)	0.489 (1)	0.470 (1)	0.489 (1)	**0.467** (1)	0.488 (1)	0.476 (1)	0.481 (1)
mor	0.277 (9)	0.273 (12)	0.296 (12)	0.299 (10)	0.294 (10)	0.297 (8)	**0.269** (11)	0.279 (10)
spam	0.115 (1)	**0.109** (3)	0.118 (5)	0.12 (3)	0.119 (5)	0.121 (3)	0.12 (3)	0.121 (1)
satellite	0.119 (3)	**0.118** (3)	0.137 (4)	0.132 (3)	0.132 (3)	0.127 (3)	0.132 (3)	0.127 (3)
soyL	0.136 (1)	0.086 (1)	0.107 (1)	0.1 (1)	0.099 (1)	0.097 (1)	0.098 (1)	**0.083** (1)
Art100	0.219 (6)	0.203 (6)	0.209 (7)	0.217 (6)	0.207 (7)	0.214 (7)	0.203 (9)	**0.202** (9)
nist16	0.057 (1)	0.057 (1)	0.057 (1)	0.059 (1)	0.057 (1)	0.060 (1)	0.066 (1)	**0.056** (1)

5.5 Discussió

En aquest capítol, s'han analitzat diferents propostes d'aprenentatge de distàncies en línia que utilitzen un esquema passiu-agressiu. En particular s'ha aplicat aquest tipus d'aprenentatge fent servir diferents restriccions sota una mateixa formulació. A més, les diferents formulacions s'han considerat com versions que satisfan les restriccions necessàries en cada pas de l'aprenentatge i alternativament al final d'aquest (definides com versions positives (+) i negatives (-), respectivament). Sobre l'experimentació duta a terme, totes les propostes han sigut comparades amb un mètode de referència (l'ITML). Cal destacar l'eficiència computacional respecte a l'ITML que ofereixen totes les versions negatives.

La proposta PALS, està basada en mínims quadrats i es caracteritza per ser una versió més agressiva. En particular, la versió negativa (PALS-) obté 3 dels millors resultats de classificació. Aquesta posició està compartida juntament amb PAI+, de manera que entre aquests mètodes, l'elecció del mètode PALS-, resulta ser una elecció més interessant per assegurar l'estalvi en el temps computacional i un valor d'error de classificació competitiu.

Capítol 6

Conclusions i treballs futurs

Resum – Aquest capítol estableix les conclusions de la tesi i resumeix de manera global el treball dut a terme. També tracta de definir les direccions de recerca que donen continuïtat a l'estudi desenvolupat. Finalment, es presenten les publicacions tant nacionals com internacionals que recolzen tota aquesta investigació.

Contingut

6.1 Introducció

En aquesta tesi s'han presentat els fonaments teòrics necessaris per a introduir l'aprenentatge de distàncies així com per al desenvolupament i implementació de les corresponents contribucions originals. També s'han revisat alguns dels mètodes de major impacte dins de la literatura, i s'han seleccionat aquells que s'aproximen més a la investigació desenvolupada en aquest treball. Tot plegat, s'han realitzat dues aportacions originals en les vessants de l'aprenentatge per lots i en línia, respectivament.

El present capítol s'ha organitzat en relació a aquestes contribucions. Per cada idea desenvolupada s'introdueix l'estudi dut a terme i s'exposen les millores i conclusions derivades de les aportacions juntament amb alguns comentaris sobre diferents possibilitats de continuació d'aquestes.

6.2 Millora computacional del MCML a través d'un subconjunt d'elements

El MCML és un mètode d'aprenentatge de distàncies basat en la cerca d'una representació alternativa en què les dades d'entrenament es distribueixen de manera que els elements de la mateixa classe estan a distància mínima i els de classe diferent a distància màxima. En el context d'aquesta tècnica, s'han proposat diferents formes d'elegir un conjunt d'elements representatius anomenats punts base. Els integrants d'aquest conjunt es fan servir com referència per a prendre distàncies a la resta d'elements a l'hora d'implementar mètodes de reconeixement automàtic basats en distàncies.

Encara que el MCML ofereix un comportament competitiu quant a la bondat de les seues solucions, la formulació original limita la seua capacitat d'enfrontar-se a problemes a on la quantitat d'elements en el conjunt d'entrenament és relativament gran. La introducció dels punts base com a elements de referència resulta en una reducció significativa del seu requeriment computacional.

En l'estudi s'han considerat diversos criteris d'elecció dels punts base: selecció inicial aleatòria fixa, selecció inicial amb l'algorisme de les $k-$mitjanes fixa i selecció aleatòria adaptativa durant el procés d'optimització. La intenció de les diferents alternatives presentades és construir conjunts representatius amb cardinal relativament menut però que es distribuesquen estratègicament en l'espai original de manera que s'observe un millor comportament dels mètodes de reconeixement que s'hi deriven.

Per avaluar el rendiment de les diferents propostes s'han dissenyat experiments seleccionant conjunts de punts base amb diferent grandària. S'ha emprat com a mesura comparativa l'error de classificació amb el veí més proper i també els temps emprats per cada mètode per a arribar a la seua solució. Totes les versions presentades han mostrat una efectivitat en el valor d'error equiparable a la del mètode original. En línies generals, sempre pot trobar-se un valor òptim de la grandària del conjunt de punts base que almenys iguala la seua efectivitat. En els experiments s'ha arribat a conclusions positives quant a la classificació. Com a resultat, s'ha comprovat que el mètode més eficaç en termes d'error és el basat en la selecció inicial mitjançant les $k-$mitjanes. Com a resum, l'ús de punts base resulta ser una aportació que redueix el cost computacional del mètode original sense degradar el comportament dels corresponents mètodes de reconeixement. La utilització dels punts base en un nombre adequat permet arribar a un compromís adequat entre cost computacional de l'aprenentatge i efectivitat dels reconeixedors.

Una de les possibles direccions de recerca futura és definir el criteri del MCML considerant els punts base com variables desconegudes, amb el propòsit d'establir els valors òptims d'aquests per a les dades d'entrenament. D'aquesta manera no només s'aprendria una distància si no que a més a més s'obtindria una representació millorada del problema formada pels punts base resultants. Aquesta idea està a hores d'ara en desenvolupament, encara que ja s'han publicat resultats preliminars d'aquest estudi en un congrés de reconeguda rellevància [70].

6.3 Aprenentatge passiu-agressiu de distàncies i extensió per mínims quadrats

La principal aportació sobre l'aprenentatge en línia de distàncies és la presentació unificada d'una família d'algorismes basats en un esquema passiu-agressiu. Aquest tipus d'aprenentatge es fonamenta en la construcció progressiva d'un model que s'actualitza cada vegada que es considera un nou exemple sobre el qual es realitza una predicció incorrecta (cas agressiu), i no s'actualitza en cas d'encertar en aquesta predicció (cas passiu).Entre aquests, s'introdueix una vessant inspirada en mínims quadrats. S'ha realitzat un estudi teòric conjunt per a les diferents formulacions que es recolza sobre una exhaustiva experimentació. El principal resultat de l'experimentació és que la família d'algorismes en línia (inclosa la nova aportació) dóna lloc a resultats molt competitius tot i que els algorismes per lots disposen de cert avantatge en tindre accés des del principi a tota la informació disponible. A més a més, s'introdueix una vessant inspirada en mínims quadrats, que s'escapa de la filosofia del model passiu-agressiu ja que sempre (o quasi sempre) realitza una modificació.

Una possible línia de continuïtat d'aquesta investigació seria l'extensió al cas no lineal mitjançant kernels de les diferents formulacions presentades. D'altra banda, s'han realitzat estudis en el projecte d'investigació Consolider MIPRCV en la línia de l'aprenentatge actiu, que tracta de minimitzar la quantitat d'informació etiquetada durant l'entrenament. Aquest tipus d'aprenentatge considera els problemes a on la informació d'etiquetat resulta per algun motiu costosa d'obtindre. Els mètodes en línia presentats en aquesta tesi poden ser adaptats d'una manera teòrica i pràctica en tasques d'aquest àmbit.

Una de les idees que està en desenvolupament és l'extensió natural entre l'aprenentatge mostra a mostra (en línia) i l'aprenentatge amb totes les mostres (per lots). Suposant que a cada instant estan disponibles una quantitat relativament menuda d'elements (en principi major que 1), es plantejaria una formulació per a realitzar un aprenentatge passiu agressiu amb diverses mostres. La intenció seria trobar el model més pròxim que satisfaça les restriccions definides. Aquest tipus d'aprenentatge podria anomenar-se com aprenentatge mitjançant "mini-lots".

6.4 Publicacions resultants

Els diferents resultats de la investigació realitzada en aquesta tesi han donat lloc a diverses publicacions en revistes, congressos nacionals i internacionals de les àrees de reconeixement de patrons i aprenentatge automàtic.

- [71], Adrián Pérez-Suay and Francesc J. Ferri. Scaling up a metric learning algorithm for image recognition and representation. In *Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II*, ISVC '08, pages 592–601, Berlin, Heidelberg, 2008. Springer-Verlag.
- [72], Adrián Pérez-Suay, Francesc J. Ferri, and Jesús V. Albert. A random extension for discriminative dimensionality reduction and metric learning.

In *Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis*, (IbPRIA '09), pages 370–377, Berlin, Heidelberg, 2009. Springer-Verlag.

- [77], Adrián Pérez-Suay and Francesc J. Ferri. Algunas consideraciones sobre aprendizaje de distancias mediante maximización del margen. In *Actas del V Taller de Minería de Datos y Aprendizaje (TAMIDA'2010)*, pages 83–91. CEDI.

- [73], Adrián Pérez-Suay, Francesc J. Ferri, and Jesús V. Albert. An online metric learning approach through margin maximization. Jordi Vitrià, João Miguel Raposo Sanches, and Mario Hernández, editors, In *Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis*, (IbPRIA '11), pages 500–507, Berlin, Heidelberg, 2011. Springer-Verlag.

- [69], Adrián Pérez-Suay and F.J. Ferri. Online metric learning methods using soft margins and least squares formulations. In G.L. Gimel'farb et al., editor, *Structural, Syntactic and Statistical Pattern Recognition. Lecture Notes in Computer Science. Vol. 7626*, pages 373–381. Springer-Verlag, 2012.

- [75], Adrián Pérez-Suay, Francesc J. Ferri, Miguel Arevalillo-Herráez, and Jesús V. Albert. Comparative evaluation of batch and online distance metric learning approaches based on margin maximization. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3511–3515, Manchester, UK, 2013.

- [74], Adrián Pérez-Suay, Francesc J. Ferri, and Miguel Arevalillo-Herráez. Passive-aggressive online distance metric learning and extensions. *Progress in AI*, 2(1):85–96, 2013.

- [70] Adrián Pérez-Suay, F.J. Ferri, M. Arevalillo-Herráez, and J.V. Albert. About combining metric learning and prototype generation. In P. Fränti et al., editor, *Structural, Syntactic and Statistical Pattern Recognition. Lecture Notes in Computer Science*, volume 8621, pages 323–332. 2014.

Apèndix A
Bases de dades

A.1 Descripció de les bases de dades

Taula A.1: Característiques de les bases de dades: grandària (n), dimensionalitat (d), nombre de classes (c) i nombre de parells d'entrenament emprats en l'entrenament en línia (r).

	n	d	c	r
soyS	136	35	4	480
wine	178	13	3	240
glass	214	9	4	480
ecoli	272	7	3	240
malaysia	291	8	20	15200
iono	351	34	2	80
balance	625	4	3	240
breast	683	9	2	80
chromo	1143	8	24	22080
mor	2000	6	10	3600
spam	4601	57	2	80
satellite	6435	36	6	1200
soyL	266	35	15	8400
Art100	1710	104	10	3600
nist16	2000	256	10	3600

Per dur a terme l'avaluació experimental, s'han fet servir un total de 15 bases de dades. En particular, 12 bases de dades diferents de [29]; 2 bases de dades (spam, balance) del repositori de l'UCI [33]; i una més realista prèviament utilitzada en tasques de CBIR [2] s'han utilitzat com a conjunt estàndard per a experimentar en aquesta tesi.

Per aclarir les abreviacions, les bases de dades soybean es representen com soyS i soyL. Aquestes es refereixen a soybean Small i soybean Large, respectivament. També les característiques morfològiques en el conjunt de dades de dígits serà

abreviat com mor. Art100 es refereix a una col·lecció comercial anomenada "Art Explosion", distribuïda per la companyia Nova Development[1].

Les imatges en aquesta col·lecció s'han classificat manualment d'una manera subjectiva, d'acord a les resolucions emeses per usuaris reals. Per dur a terme experiments de classificació més significatius, només les classes amb més de 100 elements s'han considerat en aquesta col·lecció.

En totes les bases de dades, els objectes es consideren similars si comparteixen la mateixa etiqueta de classe. Un resum de les característiques particulars de cada conjunt de dades utilitzats en l'estudi comparatiu es mostra a la taula A.1. Els repositoris han estat ordenats aproximadament per ordre de complexitat relativa, tenint en compte tant el nombre de mostres com la dimensionalitat. Per raons de coherència, aquest ordre ha estat respectat en totes les figures i taules que fan referència a un mateix experiment en diferents bases de dades.

En els experiments, totes les bases de dades s'han dividit aleatòriament en dos subconjunts disjunts de la mateixa mida. Aquests s'utilitzen per a l'entrenament i el test, respectivament.

[1] `http://www.novadevelopment.com`

Apèndix B
Gràfiques dels experiments del MCML amb punts base

B.1 Evolució dels criteris amb diferents inicialitzacions

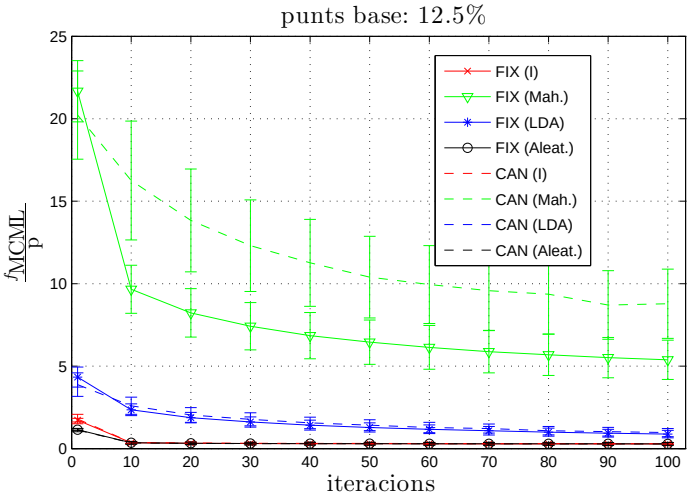


Figura B.1: Valors de criteri amb les diferents inicialitzacions en la base de dades soyS.

B. Gràfiques dels experiments del MCML amb punts base

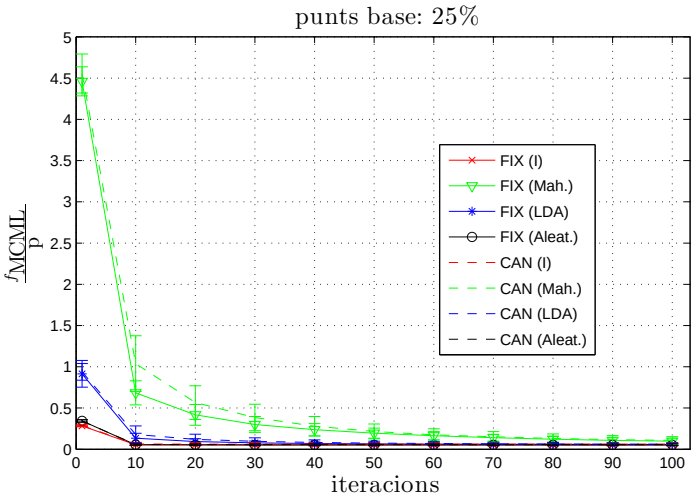


Figura B.2: Valors de criteri amb les diferents inicialitzacions en la base de dades wine.

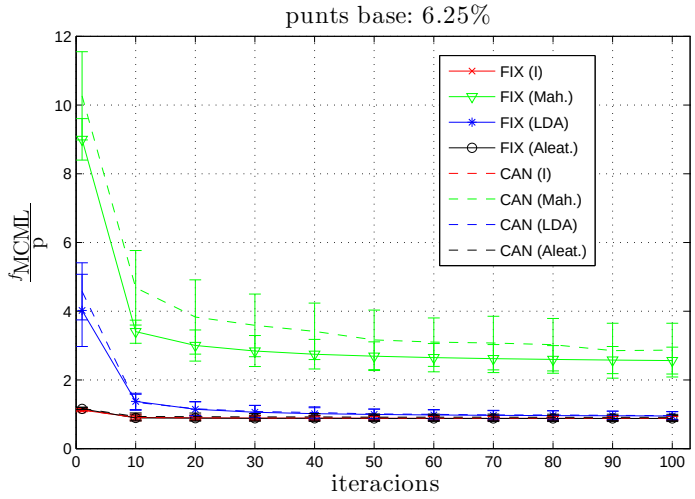


Figura B.3: Valors de criteri amb les diferents inicialitzacions en la base de dades glass.

B. Gràfiques dels experiments del MCML amb punts base

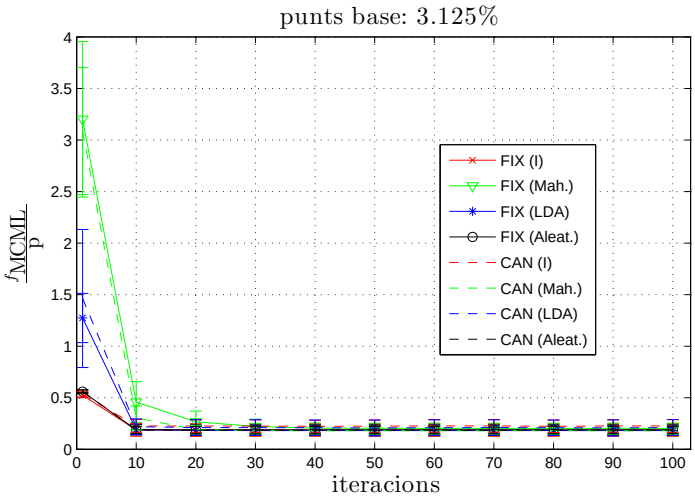


Figura B.4: Valors de criteri amb les diferents inicialitzacions en la base de dades ecoli.

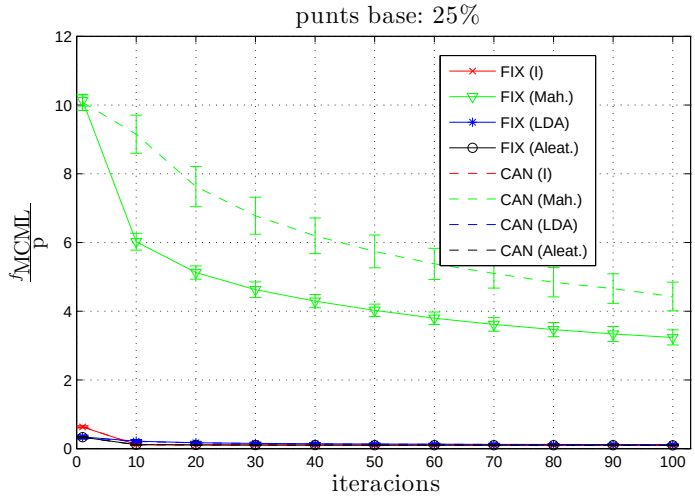


Figura B.5: Valors de criteri amb les diferents inicialitzacions en la base de dades ionosphere.

B. Gràfiques dels experiments del MCML amb punts base

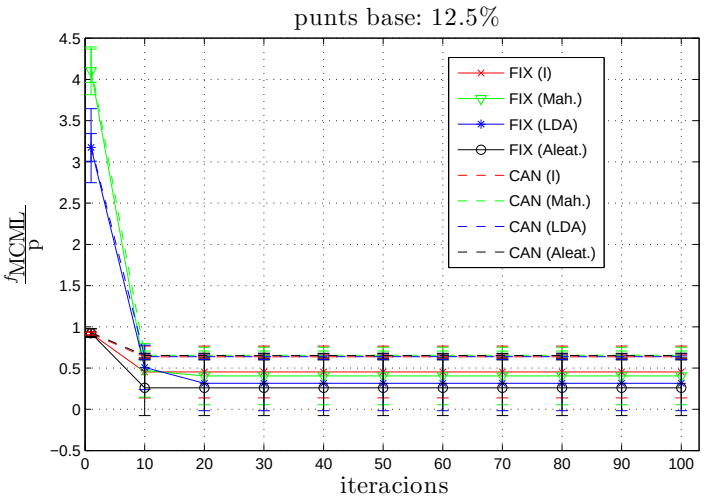


Figura B.6: Valors de criteri amb les diferents inicialitzacions en la base de dades balance.

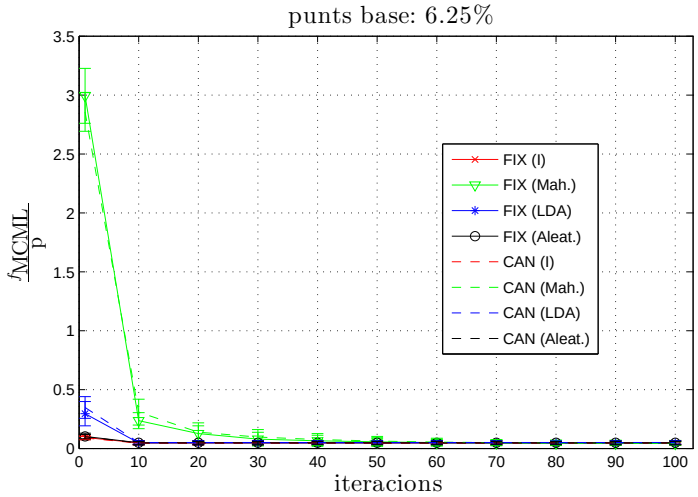


Figura B.7: Valors de criteri amb les diferents inicialitzacions en la base de dades breast.

B. Gràfiques dels experiments del MCML amb punts base

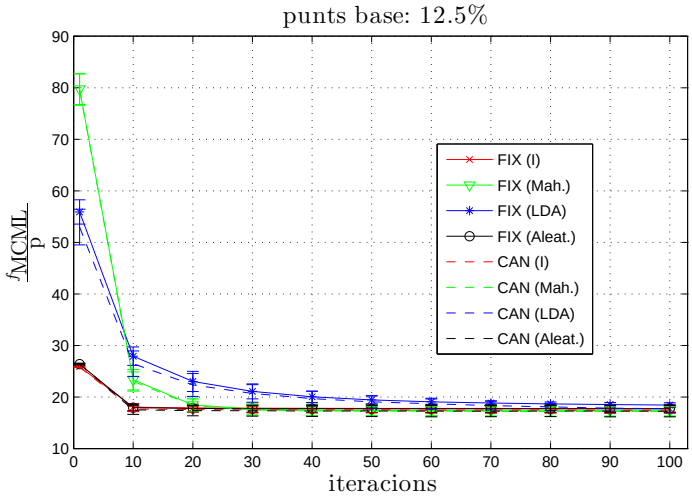


Figura B.8: Valors de criteri amb les diferents inicialitzacions en la base de dades chromo.

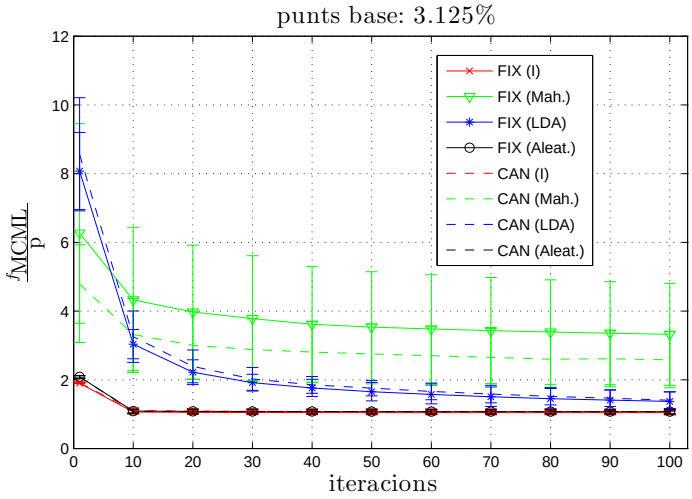


Figura B.9: Valors de criteri amb les diferents inicialitzacions en la base de dades mor.

B. Gràfiques dels experiments del MCML amb punts base

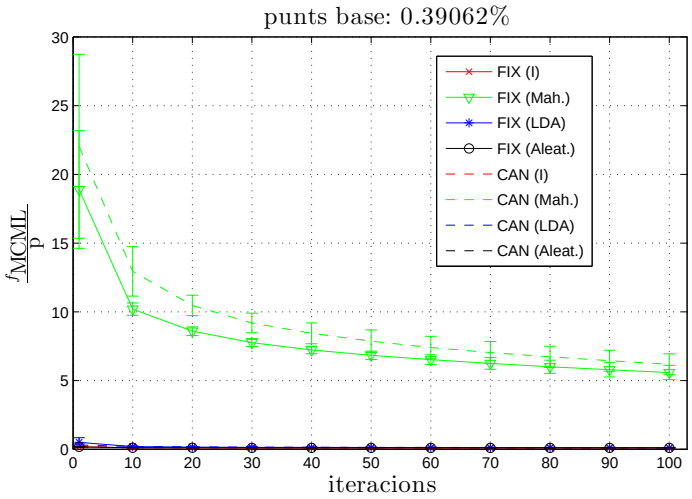


Figura B.10: Valors de criteri amb les diferents inicialitzacions en la base de dades spam.

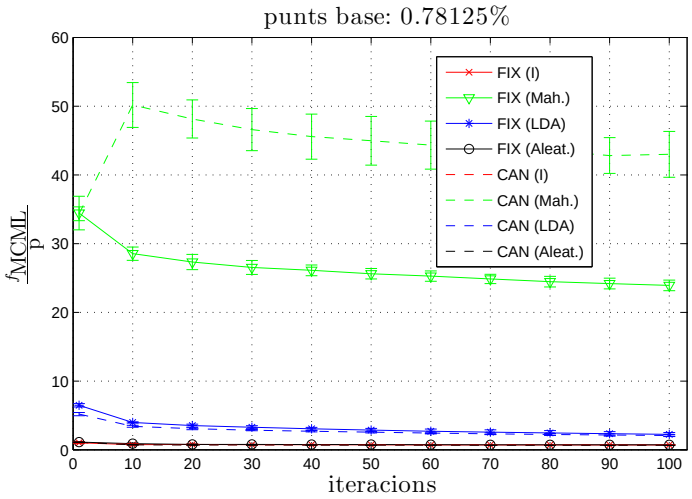


Figura B.11: Valors de criteri amb les diferents inicialitzacions en la base de dades satellite.

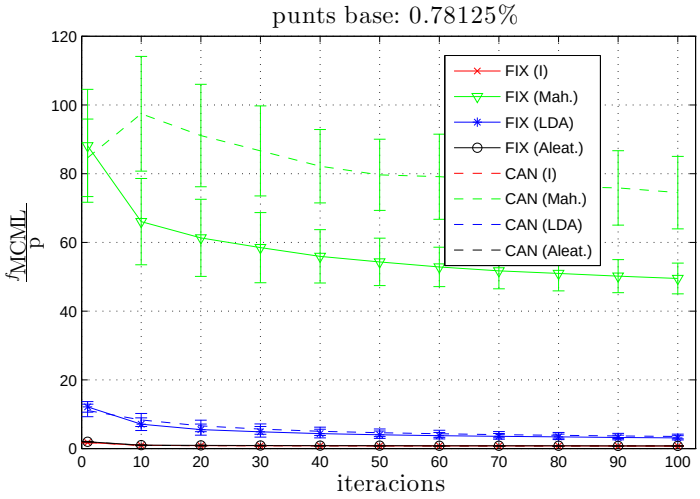


Figura B.12: Valors de criteri amb les diferents inicialitzacions en la base de dades Art100.

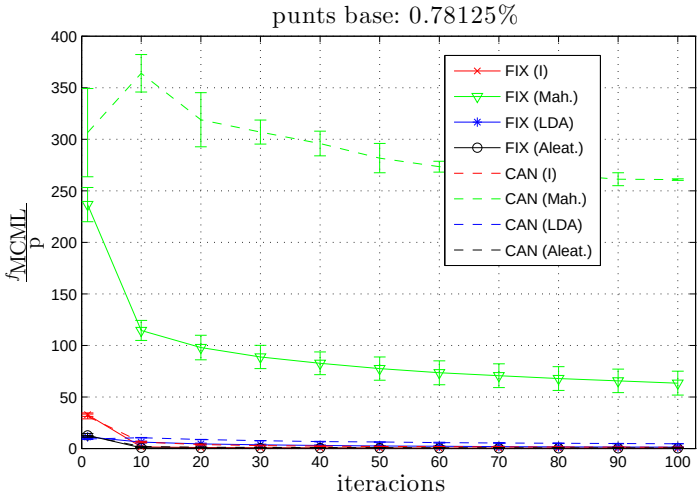
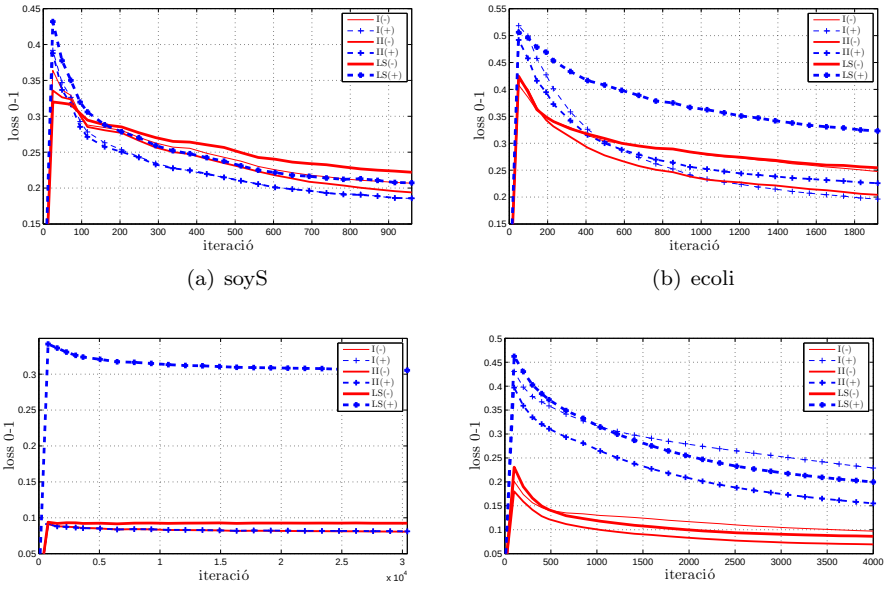


Figura B.13: Valors de criteri amb les diferents inicialitzacions en la base de dades nist16.

Apèndix C
Gràfiques dels experiments de la família PA

C.1 Pèrdua 0-1

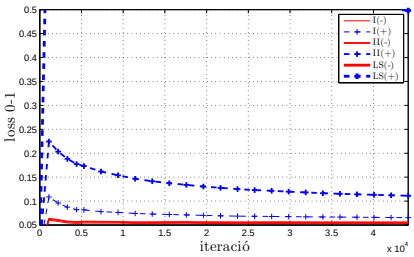


(a) soyS (b) ecoli

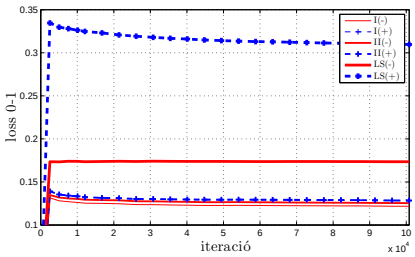
(c) malaysia (d) breast

Figura C.1: Error predictiu dels algorismes en línia.

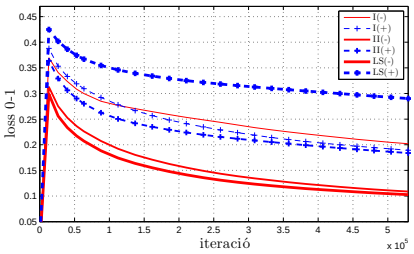
C. Gràfiques dels experiments de la família PA



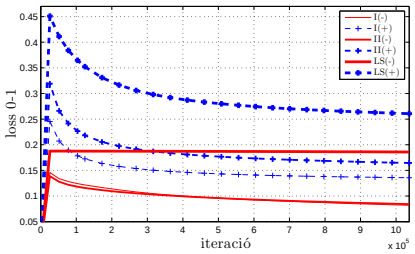
(a) chromo



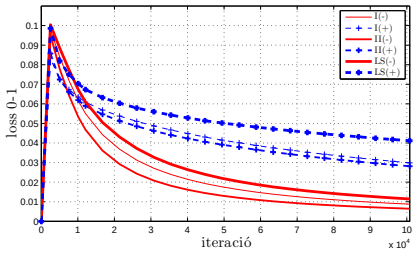
(b) mor



(c) spam



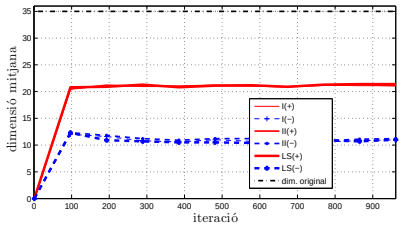
(d) satellite



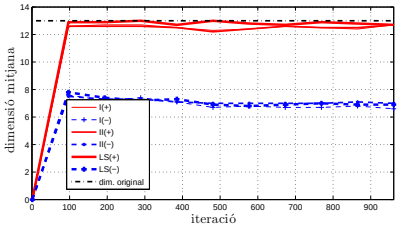
(e) nist16

Figura C.2: Error predictiu dels algorismes en línia.

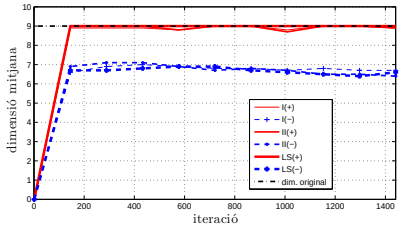
C.2 Estudi de la dimensió al procés iteratiu



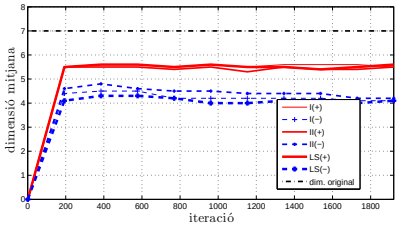
(a) soyS



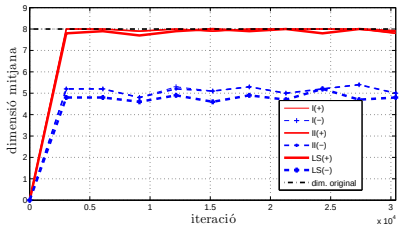
(b) wine



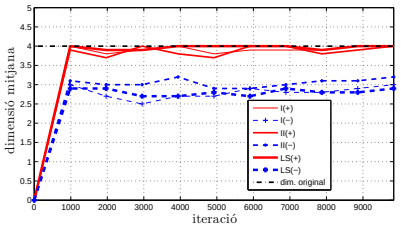
(c) glass



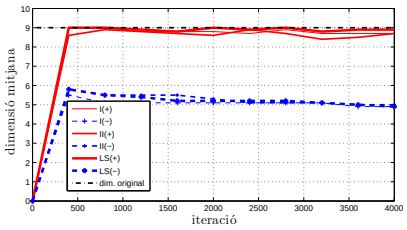
(d) ecoli



(e) malaysia



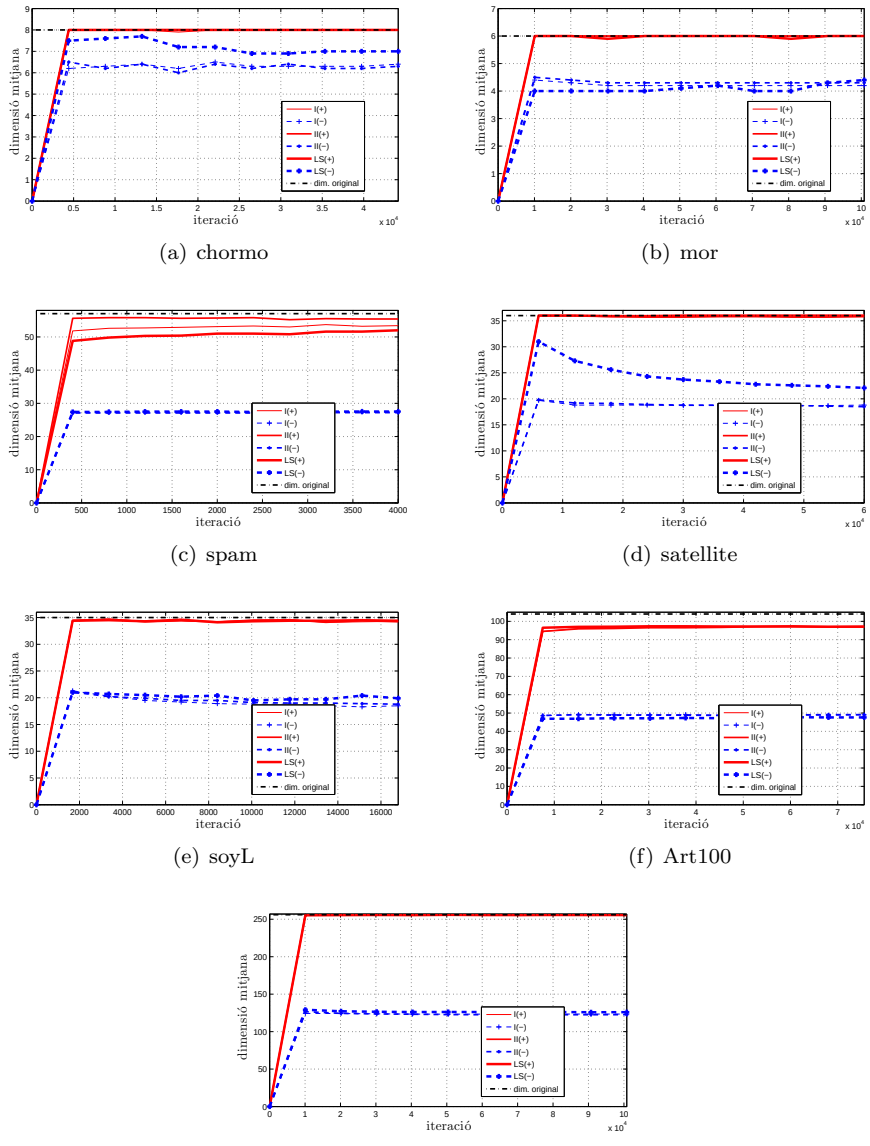
(f) balance



(g) breast

Figura C.3: Evolució de la dimensió al llarg del procés d'aprenentatge en línia.

C. Gràfiques dels experiments de la família PA



(a) chormo (b) mor

(c) spam (d) satellite

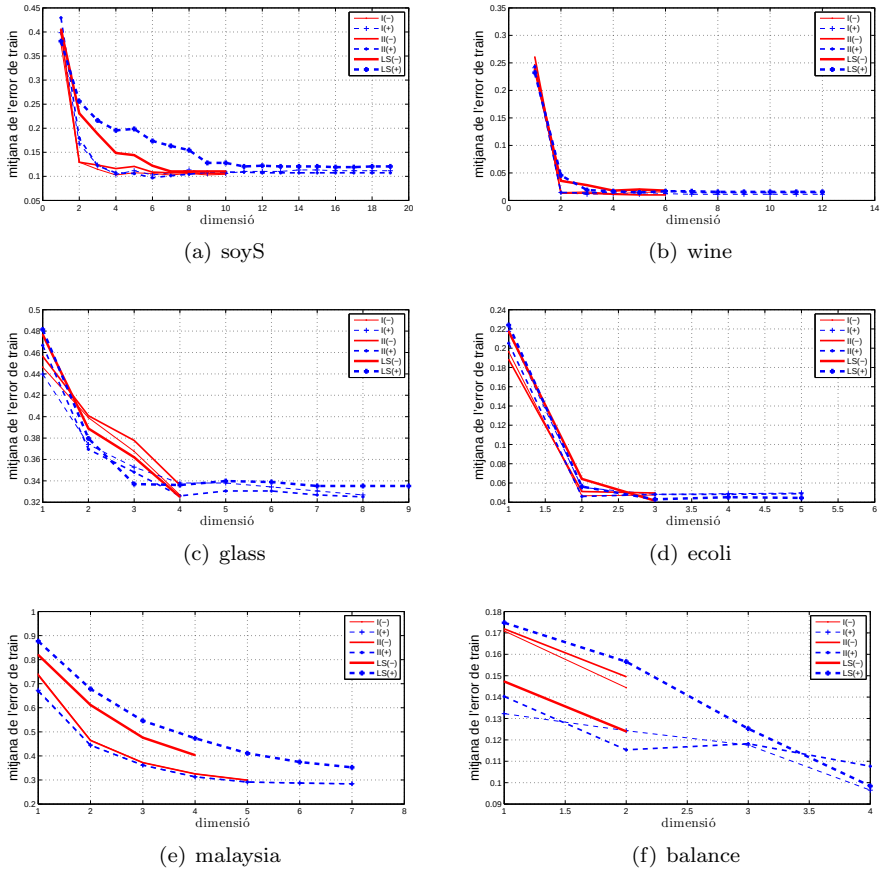
(e) soyL (f) Art100

(g) nist16

Figura C.4: Evolució de la dimensió al llarg del procés d'aprenentatge en línia.

C.3 Estudi de l'error amb la matriu final en funció de la dimensió

C.3.1 Error sobre el conjunt d'entrenament



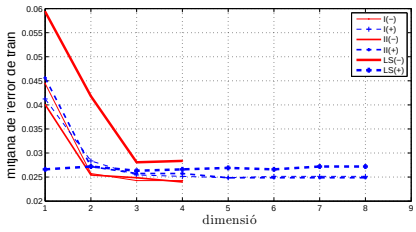
(a) soyS (b) wine

(c) glass (d) ecoli

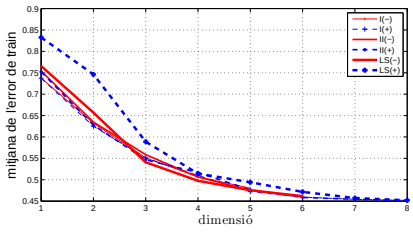
(e) malaysia (f) balance

Figura C.5: Mitjana de l'error sobre el conjunt d'entrenament amb la matriu final del procés d'aprenentatge en línia.

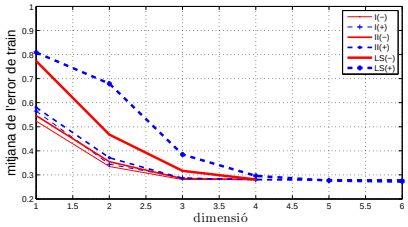
C. Gràfiques dels experiments de la família PA



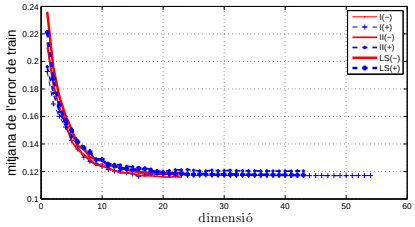
(a) breast



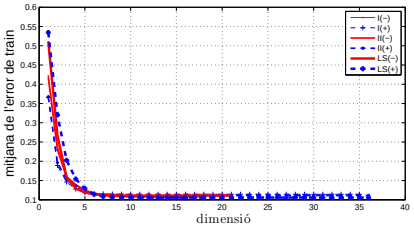
(b) chormo



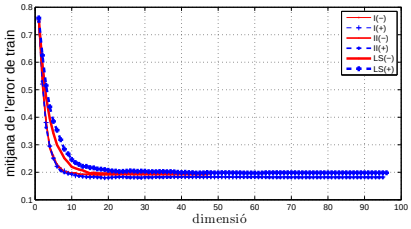
(c) mor



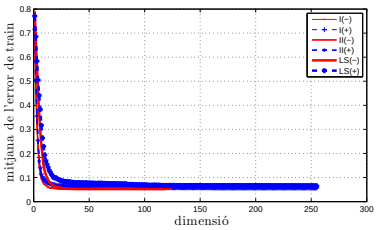
(d) spam



(e) satellite



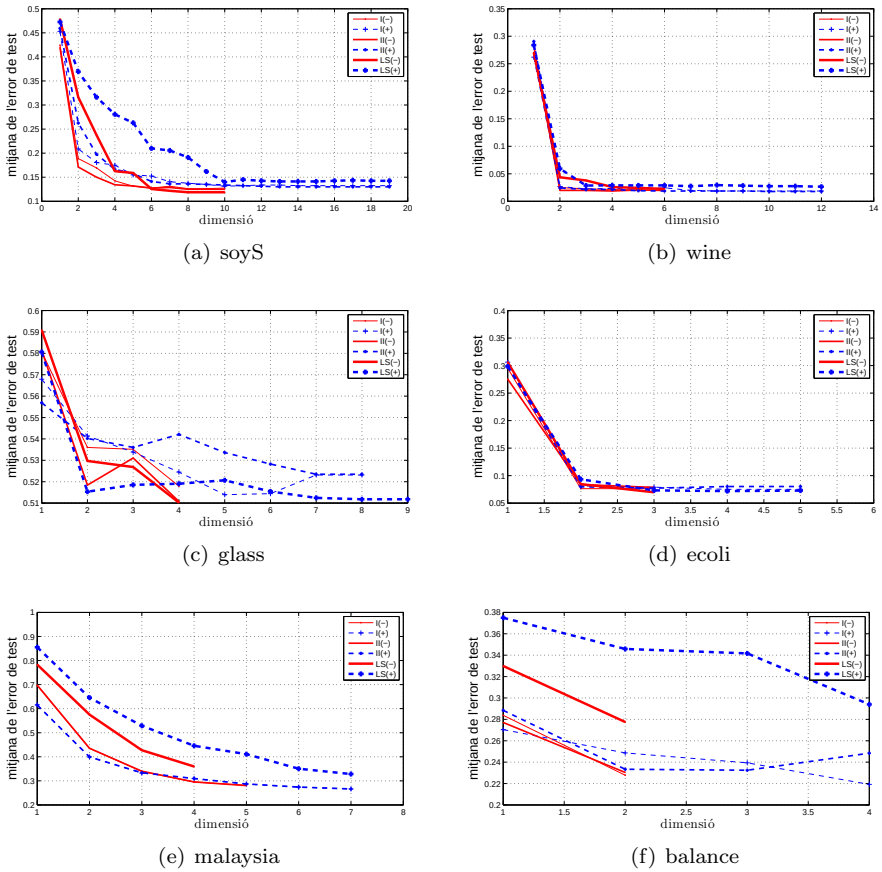
(f) Art100



(g) nist16

Figura C.6: Mitjana de l'error sobre el conjunt d'entrenament amb la matriu final del procés d'aprenentatge en línia.

C.3.2 Error sobre el conjunt de test



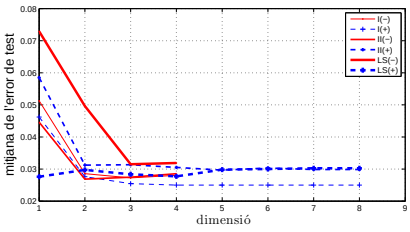
(a) soyS (b) wine

(c) glass (d) ecoli

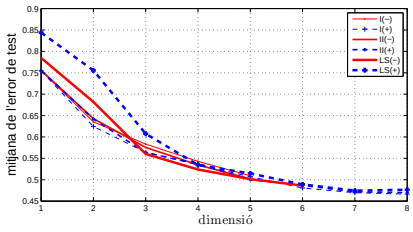
(e) malaysia (f) balance

Figura C.7: Mitjana de l'error sobre el conjunt de test amb la matriu final del procés d'aprenentatge en línia.

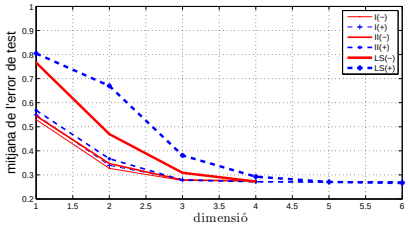
C. Gràfiques dels experiments de la família PA



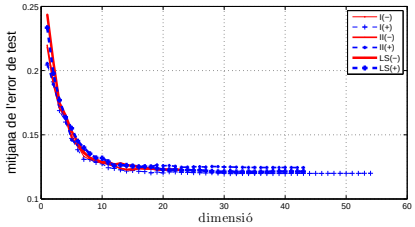
(a) breast



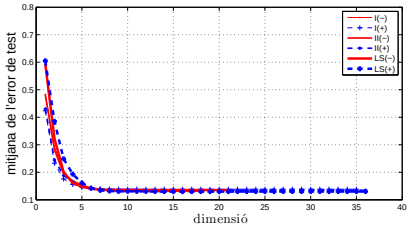
(b) chormo



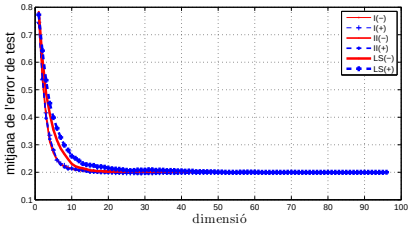
(c) mor



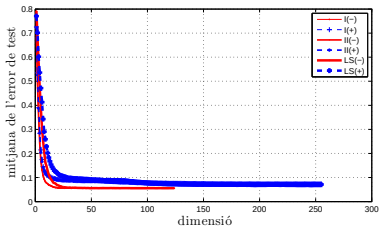
(d) spam



(e) satellite



(f) Art100



(g) nist16

Figura C.8: Mitjana de l'error sobre el conjunt de test amb la matriu final del procés d'aprenentatge en línia.

Apèndix D

Problema dual

L'objectiu d'aquest apèndix és obtindre el problema dual (5.5) a partir del problema d'aprenentatge de distàncies (5.4).

Cada una de les restriccions del problema (5.4) està definida per un parell d'índexs (i, j), sobre el conjunt d'entrenament. Primer anem a introduir un nou índex k, que farà referència al parell (i, j). I també, per simplificar la derivació, s'introdueix $d_k^M = z_k^\top M z_k$, on $z_k = (x_i - x_j)$. De manera que (5.4) pot reescriure's com

$$\begin{aligned} \min_{M,b,\xi} \quad & \frac{1}{2}\|M\|_{Fro}^2 + C\sum_k \xi_k, \\ \text{tal que} \quad & y_k(b - d_k^M) \geq 1 - \xi_k, \\ & \xi_k \geq 0. \end{aligned} \tag{D.1}$$

L'índex k varia entre $1 \leq k \leq N = \frac{n(n-1)}{2}$. Introduint els multiplicadors no negatius α_k i λ_k, definim $\mathcal{L} = \mathcal{L}(M, b, \xi, \alpha, \lambda)$

$$\mathcal{L} = \frac{1}{2}\|M\|_{Fro}^2 + C\sum_{k=1}^{N}\xi_k + \sum_{k=1}^{N}\alpha_k\left(1 - y_k\left(b - z_k^\top M z_k\right) - \xi_k\right) - \sum_{k=1}^{N}\lambda_k\xi_k. \tag{D.2}$$

Calculant les derivades parcials de $\mathcal{L}$ (Equació (D.2)), respecte a M, b, ξ_k i igualant-les a zero

$$\frac{\partial \mathcal{L}}{\partial M} = 0 \leftrightarrow M = -\sum_{k=1}^{N}\alpha_k y_k z_k z_k^\top, \tag{D.3}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \leftrightarrow \sum_{k=1}^{N}\alpha_k y_k = 0, \tag{D.4}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_k} = 0 \leftrightarrow \alpha_k = C - \lambda_k \leq C. \tag{D.5}$$

Les dues primeres Equacions (D.3), (D.4) són immediates. La tercera (D.5) s'obté tenint en compte que $\frac{\partial \mathcal{L}}{\partial \xi_k} = 0 \leftrightarrow C - \alpha_k - \lambda_k = 0$, i com $\lambda_k \geq 0$, necessàriament se satisfà la condició $\alpha_k \leq 0$, de l'Equació (D.5).

D. Problema dual

Substituint les relacions (D.3), (D.4) i (D.5) en (D.2)

$$\mathcal{L}(\alpha,\lambda)=\frac{1}{2}\left\|-\sum_{k=1}^{N}\alpha_k y_k z_k z_k^{\top}\right\|_{Fro}^{2}+\sum_{k=1}^{N}\alpha_k(1-y_k(-z_k^{\top}\left(-\sum_{l=1}^{N}\alpha_l y_l z_l z_l^{\top}\right)z_k)). \tag{D.6}$$

Tenint en compte la relació $\|M\|_{Fro}^2=\mathrm{Tr}(M^{\top}M)$ i algunes propietats fonamentals de la traça, es pot reescriure l'Equació (D.6)

$$\mathcal{L}(\alpha,\lambda)=\sum_{k=1}^{N}\alpha_k-\frac{1}{2}\sum_{k,l=1}^{N}\alpha_k\alpha_l y_k y_l\,\mathrm{Tr}(z_k z_k^{\top}z_l z_l^{\top}). \tag{D.7}$$

I l'Equació (D.7), amb les restriccions (D.4) i (D.5), formen el problema dual (5.5) que tornem a escriure a continuació

$$\begin{aligned}&\max_{\alpha}\sum_{k=1}^{N}\alpha_k-\frac{1}{2}\sum_{k,l=1}^{N}\alpha_k\alpha_l y_k y_l\left\langle Z_k,Z_l\right\rangle,\\ \text{tal que}\quad&\sum_{k=1}^{N}\alpha_k y_k=0,\\ &0\le\alpha_k\le C.\end{aligned}$$

Apèndix E

Derivació de l'actualització de PA-LS

Com s'indica en les Equacions (5.22) i (5.23), la formulació per mínims quadrats del problema original PA porta a un problema d'optimització amb només una restricció d'igualtat.

$$\begin{aligned} \min_{M,b,\xi} \quad & \tfrac{1}{2}\left\|M - M^k\right\|_{\mathsf{Fro}}^2 + \tfrac{1}{2}\left(b - b^k\right)^2 + C\xi^2, \\ s.t. \quad & 1 - y_{ij}\left(b - d_{ij}^M\right) = \xi. \end{aligned}$$

El corresponent problema de minimització sense restriccions s'obté mitjançant la introducció d'un multiplicador de Lagrange, τ, per arribar al següent Lagrangià

$$\begin{aligned} \mathcal{L}(M, b, \xi, \tau) \quad = \quad & \tfrac{1}{2}\|M - M^k\|_{\mathsf{Fro}}^2 + \tfrac{1}{2}(b - b^k)^2 + \\ & + C\xi^2 + \tau\left(1 - y_{ij}(b - d_{ij}^M) - \xi\right). \end{aligned} \tag{E.1}$$

Diferenciant la funció de Lagrange anterior respecte a M, b, ξ i igualant aquestes derivades parcials a zero condueix a

$$M = M^k - \tau y_{ij}(x_i - x_j)(x_i - x_j)^\top, \tag{E.2}$$

$$b = b^k + \tau y_{ij}, \tag{E.3}$$

$$2C\xi = \tau \Longrightarrow \xi = \frac{\tau}{2C}. \tag{E.4}$$

Ara podem emprar les Equacions (E.2), (E.3) i (E.4) en el Lagrangià (E.1) per a obtindre

$$\begin{aligned} \mathcal{L}(\tau) \quad = \quad & -\tfrac{\tau^2}{2}\left\|(x_i - x_j)(x_i - x_j)^\top\right\|_{\mathsf{Fro}}^2 - \\ & -\tfrac{\tau^2}{2} - \tfrac{\tau^2}{4C} + \tau\left(1 - y_{ij}\left(b^k - d_{ij}^{M^k}\right)\right). \end{aligned} \tag{E.5}$$

Prenent la derivada de $\mathcal{L}(\tau)$ respecte τ i establint-la a zero ens donarà el seu únic punt crític que correspon al mínim dels problemes anteriors.

E. Derivació de l'actualització de PA-LS

$$\tau = \frac{1 - y_{ij}\left(b^k - d_{ij}^{M^k}\right)}{1 + \frac{1}{2C} + \|(x_i - x_j)(x_i - x_j)^\top\|_{\mathsf{Fro}}^2}.$$

Cal tindre en compte que aquesta expressió s'ha introduït en l'Equació (5.24) com τ_3. Les Equacions (E.2) i (E.3) amb aquest valor de τ condueixen a la solució del problema PALS.

Bibliografia

[1] *Adaptive Distance Metric Learning for Clustering.* IEEE Computer Society, 2007.

[2] Miguel Arevalillo-Herráez, Francesc J. Ferri, and Juan Domingo. A naive relevance feedback model for content-based image retrieval using multiple similarity measures. *Pattern Recognition*, 43(3):619 – 629, 2010.

[3] K.C. Assi, H. Labelle, and F. Cheriet. Modified large margin nearest neighbor metric learning for regression. *Signal Processing Letters, IEEE*, 21(3):292–296, March 2014.

[4] Jonathan Baxter and Peter L. Bartlett. The canonical distortion measure in feature space and 1-nn classification. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *NIPS*. The MIT Press, 1997.

[5] Shai Ben-David, Nadav Eiron, and Philip M. Long. On the difficulty of approximately maximizing agreements. *JCSS*, 66(3):496–514, May 2003.

[6] Kristin P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1):23–34, January 1992.

[7] Mikhail Bilenko, Sugato Basu, and Raymond J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of 21st International Conference on Machine Learning (ICML)*, pages 81–88, Banff, Canada, July 2004.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[9] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer, 2003.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

BIBLIOGRAFIA

[11] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, June 1998.

[12] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[13] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., USA,, 2006.

[14] Ratthachat Chatpatanasiri, Teesid Korsrilabutr, Pasakorn Tangchanachaianan, and Boonserm Kijsirikul. On kernelization of supervised mahalanobis distance learners. *CoRR*, abs/0804.1441, 2008.

[15] Ratthachat Chatpatanasiri, Teesid Korsrilabutr, Pasakorn Tangchanachaianan, and Boonserm Kijsirikul. A new kernelization framework for mahalanobis distance learning algorithms. *Neurocomputing*, 73(10-12):1570–1579, 2010.

[16] Gal Chechik, Uri Shalit, Varun Sharma, and Samy Bengio. An online algorithm for large scale image similarity learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 306–314. 2009.

[17] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. In Helder Araújo, Ana Maria Mendonça, Armando J. Pinho, and M. Inés Torres, editors, *IbPRIA*, volume 5524 of *Lecture Notes in Computer Science*, pages 11–14. Springer, 2009.

[18] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, 11:1109–1135, March 2010.

[19] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[20] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, 13(1):21–27, September 2006.

[21] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

[22] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, UK, March 2000.

BIBLIOGRAFIA

[23] Jason V. Davis and Inderjit Dhillon. Differential entropic clustering of multivariate gaussians. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 337–344. MIT Press, Cambridge, MA, 2007.

[24] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML, pages 209–216, New York, NY, USA, 2007. ACM.

[25] Jia Deng, A.C. Berg, and Li Fei-Fei. Hierarchical semantic indexing for large scale image retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE*, pages 785–792, 2011.

[26] Huyen Do, Alexandros Kalousis, Jun Wang, and Adam Woznica. A metric learning perspective of svm: on the relation of lmnn and svm. *Journal of Machine Learning Research - Proceedings Track*, 22:308–317, 2012.

[27] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1 edition, June 1973.

[28] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[29] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, and D.M.J. Tax. Prtools4, a matlab toolbox for pattern recognition, 2004.

[30] R. Escalante and M. Raydan. *Alternating Projection Methods*. Fundamentals of Algorithms. Society for Industrial and Applied Mathematics, 2011.

[31] Ronen Feldman and James Sanger. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA, 2006.

[32] E. Fix and J. L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. *US Air Force School of Aviation Medicine*, Technical Report 4(3):477+, January 1951.

[33] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[34] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, EuroCOLT '95, pages 23–37, London, UK, UK, 1995. Springer-Verlag.

[35] Jerome H. Friedman. Flexible metric nearest neighbor classification. Technical report, Department of Statistics, Stanford University, 1994.

[36] Keinosuke Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

BIBLIOGRAFIA

[37] Keinosuke Fukunaga and Thomas E. Flick. An Optimal Global Nearest Neighbor Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(3):314–318, May 1984.

[38] Salvador García and Francisco Herrera. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2008.

[39] Zoubin Ghahramani. Unsupervised learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 72–112. Springer, 2003.

[40] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 451–458. MIT Press, Cambridge, MA, 2006.

[41] Jacob Goldberger, Sam Roweis, Geoffrey Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, Cambridge, MA, 2005.

[42] Gene H. Golub and Charles F. Van Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[43] Trevor Hastie and Robert Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE TPAMI*, 18(6):607–616, 1996.

[44] H. Hotelling. Analysis of a complex of statistical variables into principal components. *JEP*, 24, 1933.

[45] P.J. Huber, J. Wiley, and W. InterScience. *Robust statistics*. Wiley New York, 1981.

[46] Prateek Jain, Brian Kulis, Inderjit S. Dhillon, and Kristen Grauman. Online metric learning and fast similarity search. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 761–768. 2009.

[47] Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning:theory and algorithm. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 862–870. Curran Associates, Inc., 2009.

[48] Dor Kedem, Stephen Tyree, Kilian Weinberger, Fei Sha, and Gert Lanckriet. Non-linear metric learning. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2582–2590. 2012.

BIBLIOGRAFIA

[49] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, January 1997.

[50] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition*.

[51] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[52] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[53] Gautam Kunapuli and Jude W. Shavlik. Mirror descent for metric learning: A unified approach. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *ECML/PKDD (1)*, volume 7523 of *Lecture Notes in Computer Science*, pages 859–874. Springer, 2012.

[54] J.T. Kwok and I.W. Tsang. Learning with idealized kernels. *Proceedings of the Twentieth International Conference on Machine Learning*, pages 400–407, 2003.

[55] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285–318, April 1988.

[56] Bo Liu, Meng Wang, Richang Hong, Zhengjun Zha, and Xian-Sheng Hua. Joint learning of labels and distance metric. *Trans. Sys. Man Cyber. Part B*, 40(3):973–978, June 2010.

[57] Yangjing Long. Human age estimation by metric learning for regression problems. In *Computer Graphics, Imaging and Visualization, 2009. CGIV '09. Sixth International Conference on*, pages 343–348, Aug.

[58] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[59] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, April 1936.

[60] O. L. Mangasarian and David R. Musicant. Lagrangian support vector machines. *J. Mach. Learn. Res.*, 1:161–177, September 2001.

[61] María Luisa Micó, José Oncina, and Enrique Vidal. A new version of the nearest-neighbour approximating and eliminating search algorithm (aesa) with linear preprocessing time and memory requirements. *Pattern Recogn. Lett.*, 15(1):9–17, 1994.

[62] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1st edition, 1997.

BIBLIOGRAFIA

[63] Yvonne Moh and Joachim M. Buhmann. Regularized online learning of pseudometrics. In *ICASSP*, pages 1990–1993. IEEE, 2010.

[64] Nam Nguyen and Yunsong Guo. Metric learning: A support vector approach. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, ECML PKDD '08, pages 125–136, Berlin, Heidelberg, 2008. Springer-Verlag.

[65] Shibin Parameswaran and Kilian Weinberger. Large margin multi-task metric learning. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1867–1875. 2010.

[66] Roberto Paredes and Enrique Vidal. Learning weighted metrics to minimize nearest-neighbor classification error. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(7):1100–1110, 2006.

[67] Kyoungup Park, Chunhua Shen, Zhihui Hao, and Junae Kim. Efficiently learning a distance metric for large margin nearest neighbor classification. In Wolfram Burgard and Dan Roth, editors, *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*. AAAI Press, 2011.

[68] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[69] Adrián Pérez-Suay and F.J. Ferri. Online metric learning methods using soft margins and least squares formulations. In G.L. Gimel'farb et al., editor, *Structural, Syntactic and Statistical Pattern Recognition. Lecture Notes in Computer Science. Vol. 7626*, pages 373–381. Springer-Verlag, 2012.

[70] Adrián Pérez-Suay, F.J. Ferri, M. Arevalillo-Herráez, and J.V. Albert. About combining metric learning and prototype generation. In P. Fränti et al., editor, *Structural, Syntactic and Statistical Pattern Recognition. Lecture Notes in Computer Science.*, volume 8621, pages 323–332. 2014.

[71] Adrián Pérez-Suay and Francesc J. Ferri. Scaling up a metric learning algorithm for image recognition and representation. In *Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II*, ISVC '08, pages 592–601, Berlin, Heidelberg, 2008. Springer-Verlag.

[72] Adrián Pérez-Suay, Francesc J. Ferri, and Jesús V. Albert. A random extension for discriminative dimensionality reduction and metric learning. In *Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis*, IbPRIA '09, pages 370–377, Berlin, Heidelberg, 2009. Springer-Verlag.

[73] Adrián Pérez-Suay, Francesc J. Ferri, and Jesús V. Albert. An online metric learning approach through margin maximization. In Jordi Vitrià, João Miguel Raposo Sanches, and Mario Hernández, editors, *IbPRIA*, volume 6669 of *Lecture Notes in Computer Science*, pages 500–507. Springer, 2011.

BIBLIOGRAFIA

[74] Adrián Pérez-Suay, Francesc J. Ferri, and Miguel Arevalillo-Herráez. Passive-aggressive online distance metric learning and extensions. *Progress in AI*, 2(1):85–96, 2013.

[75] Adrián Pérez-Suay, Francesc J. Ferri, Miguel Arevalillo-Herráez, and Jesús V. Albert. Comparative evaluation of batch and online distance metric learning approaches based on margin maximization. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3511–3515, Manchester, UK, 2013.

[76] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007.

[77] Adrián Pérez-Suay and Francesc J. Ferri. Algunas consideraciones sobre aprendizaje de distancias mediante maximización del margen. In *Actas del V Taller de Minería de Datos y Aprendizaje (TAMIDA'2010)*, pages 83–91. CEDI 2010, 2010.

[78] Guo-Jun Qi, Jinhui Tang, Zheng-Jun Zha, Tat-Seng Chua, and Hong-Jiang Zhang. An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML, pages 841–848, New York, NY, USA, 2009. ACM.

[79] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000.

[80] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[81] Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[82] Shai Shalev-Shwartz, Koby Crammer, Ofer Dekel, and Yoram Singer. Online passive-aggressive algorithms. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[83] Shai Shalev-Shwartz, Yoram Singer, and Andrew Y. Ng. Online and batch learning of pseudo-metrics. In *Proceedings of the twenty-first International Conference on Machine Learning*, ICML, pages 743–750, New York, NY, USA, 2004. ACM.

[84] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine Learning*, ICML, pages 807–814, New York, NY, USA, 2007. ACM.

BIBLIOGRAFIA

[85] R. Short, II and K. Fukunaga. The optimal distance measure for nearest neighbor classification. *IEEE Trans. Inf. Theor.*, 27(5):622–627, September 1981.

[86] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Process. Lett.*, 9:293–300, June 1999.

[87] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

[88] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition, Fourth Edition*. Academic Press, 4th edition, 2008.

[89] Ivor W. Tsang, James T. Kwok, and Clear Water Bay. Distance metric learning with kernels. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 126–129, 2003.

[90] L.J.P. van der Maaten, E.O. Postma, , and H.J. van den Herik. Dimensionality reduction: A comparative review. Technical report, Tilburg University, TiCC-TR 2009-005, 2009.

[91] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[92] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[93] Vladimir N. Vapnik. *Statistical learning theory*. Wiley, 1 edition, September 1998.

[94] Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, and Francesca Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research (JMLR)*, 6:883–904, December 2005.

[95] Fei Wang and Jimeng Sun. Survey on distance metric learning and dimensionality reduction. *Data Mining and Knowledge Discovery*, 2014.

[96] Kilian Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1473–1480. MIT Press, Cambridge, MA, 2006.

[97] Kilian Q. Weinberger and Lawrence K. Saul. Fast solvers and efficient implementations for distance metric learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML, pages 1160–1167, New York, NY, USA, 2008. ACM.

BIBLIOGRAFIA

[98] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, June 2009.

[99] Patrick Henry Winston. *Artificial Intelligence (3rd Ed.)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1992.

[100] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[101] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, Cambridge, MA, 2003.

[102] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2, 2006.

[103] Yiming Ying, Kaizhu Huang, and Colin Campbell. Sparse metric learning via smooth optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2214–2222. 2009.

[104] Yiming Ying and Peng Li. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 13:1–26, March 2012.

BIBLIOGRAFIA

BIBLIOGRAFIA