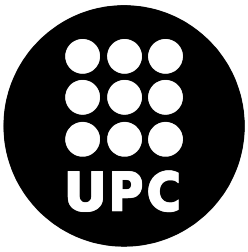
Universitat Politècnica de Barcelona

Facultat d'Informàtica

Grau en Ciència i Enginyeria de Dades



Implementació de SVM primal i dual

Optimització matemàtica

Noa Mediavilla

46998650J

Rebeca Torrecilla

46189534Z

Professor
Jordi Castro Pérez

Barcelona

Juliol de 2025

1 Introducció

Classificar dades resulta una tasca ordinària d'aprenentatge automàtic. Suposem que ens donen una sèrie de punts, cadascun pertanyent a una de dues classes disponibles amb l'objectiu de decidir a quina classe pertany cada nou punt que introduïm.

En el cas de les màquines de vectors de suport (*support vector machines, SVM*), cada punt es vist com un vector p-dimensional (una llista de p nombres) on volem saber si aquests seran separables per un hiperplà $(p-1)$-dimensional. Això representa l'anomenat classificador lineal.

Com pot haver-hi més d'un hiperplà possible que pugui separar les nostres dades, una manera lògica d'escollir el millor d'aquests és el que mantingui una major separació, o marge, entre les dues classes. Així, escollim un hiperplà de manera que la distància entre aquest i el punt més proper es maximitzi. Aquests punts que es situen justament al marge són els anomenats, vectors de suport.

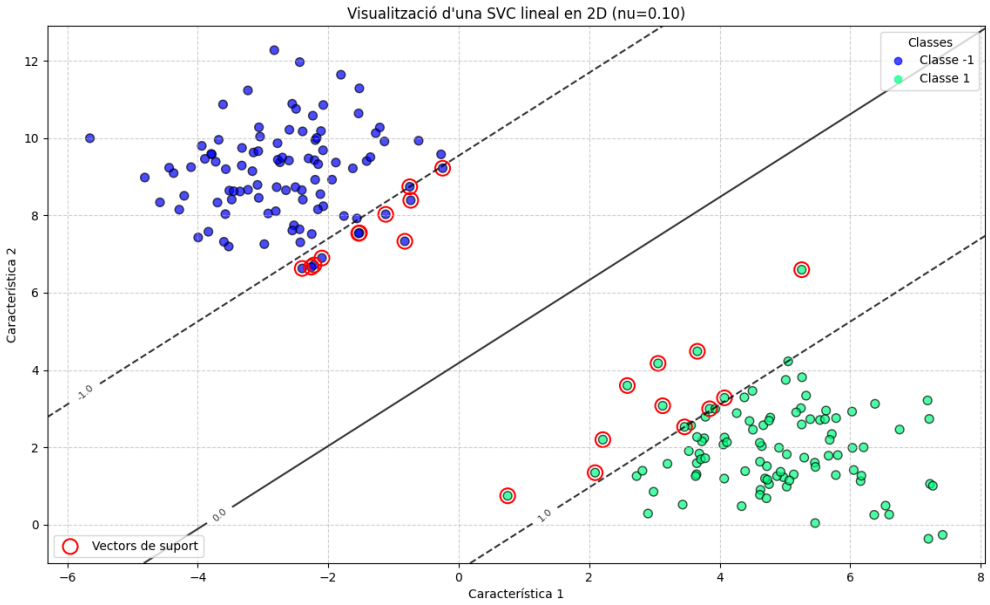


Figure 1: Gràfic d'un exemple de SVM en un problema de classificació lineal amb $\nu = 0.1$.

Formulant aquest problema matemàticament, podem dir que ens donen una base de dades d'entrenament que conté n punts de la forma $(x_1, y_1), ..., (x_n, y_n)$, on les y_i corresponen a les etiquetes dels punts, indicant amb un -1 o un 1 la classe a la que pertany el punt x_i. Volem doncs trobar l'hiperplà que maximitzi el marge entre el grup de punts x_i que comptin amb l'etiqueta $y_i = 1$ del grup de punts amb $y_i = -1$.

Podem escriure qualsevol hiperplà com $w^T x - \gamma = 0$. El vector de pesos w determina l'orientació de l'hiperplà amb les seves components corresponents a les variables de les dades. El producte $w^T x_i$ mesura la projecció del punt x_i a w i, afegint un terme de desviació γ, canviem aquest valor per determinar la classe predita.

Si les dades d'entrenament donades resulten linealment separables, podem seleccionar dos hiperplans paral·lels que separin les dues classificacions de les dades per tal de maximitzar el marge (*hard-margin*). Aquests hiperplans es poden descriure amb les equacions $w^T x + \gamma \geq 1$ (qualsevol punt per sobre d'aquest límit pertany a una classe, d'etiqueta 1) i $w^T x + \gamma \leq -1$ (qualsevol punt per sota d'aquest límit pertany a l'altra classe, d'etiqueta -1). Els punts que compleixen la igualtat són els vectors de suport.

Per últim, computacionalment parlant, els classificadors de SVM (amb *soft-margin*) tracten de minimitzar la següent funció d'optimització:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i - b)\right)\right] + \lambda\|\mathbf{w}\|^2.$$

que combina els objectius de minimitzar l'error de classificació (primer terme) a la vegada que intentem maximitzar el marge (segon terme, regularització). Arreglant l'expressió com un problema d'optimització restringit amb una funció objectiu diferenciable s'obté l'anomenat problema primal.

$$\begin{aligned} \min_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} \quad & \tfrac{1}{2}w^T w + \nu e^T s \\ \text{s. a} \quad & -Y(Aw + \gamma e) - s + e \leq 0 \quad [\lambda \in \mathbb{R}^m] \\ & -s \leq 0 \quad [\mu \in \mathbb{R}^m] \end{aligned}$$

Al tractar de resoldre el primal Lagrangià es pot obtenir el problema dual, una formulació clau doncs resulta eficaçment resoluble per al nostre algorisme de programació quadràtica, especialment quan utilitzem kernels no lineals.

$$\begin{aligned} \max_{\lambda} \quad & \lambda^T e - \tfrac{1}{2}\lambda^T Y A A^T Y \lambda \\ \text{s. a} \quad & \lambda^T Y e = 0 \\ & 0 \leq \lambda \leq \nu \end{aligned}$$

Les funcions kernel fan possible l'aplicació de mètodes lineals a problemes no lineals a partir de la transformació de les dades a un espai de dimensions superiors sense portar a terme de manera explícita computacions en aquests espais. En el context de les SVM, aquesta funció s'encarrega de computar el producte escalar de parells de dades d'entrada a l'espai transformat, permetent la identificació d'hiperplans separadors en classificacions complexes.

Per poder tenir una intuïció de la seva formulació matemàtica, mostrem l'exemple d'un kernel gaussià:

$$K(x, y) = \varphi(x)^T \varphi(y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

expressió que resulta molt útil doncs, en hiperplans de la forma $w^T\varphi(x_j) + \gamma$ no sabem el vertader valor

de $\varphi(x_j)$ però l'aproximació del kernel ens permet fer el càlcul:

$$w^T\varphi(x_j) = \left(\sum_{i=1}^{m} \lambda_i y_i \varphi(x_i)\right)^T \varphi(x_j) = \sum_{i=1}^{m} \lambda_i y_i K(x_i, x_j)$$

En la formulació dual, el SVM utilitza els valors donats per les kernels en comptes del valor del producte escalar a l'espai original, el que fa millorar l'algorisme a l'hora d'avaluar problemes de classificació complexes i no lineals amb eficàcia.

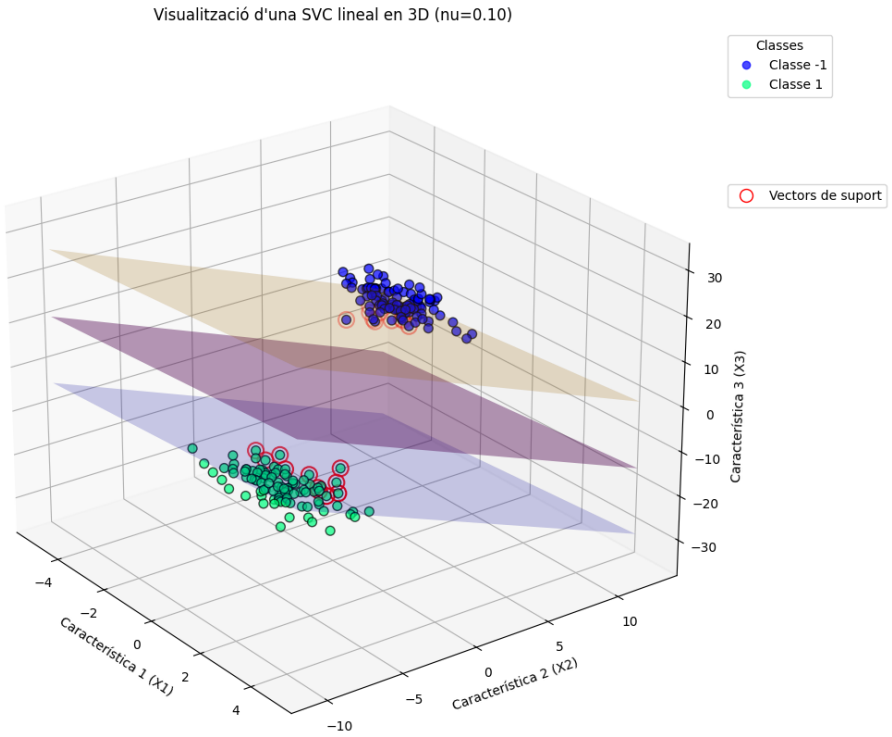


Figure 2: Visualització d'una SVM en un problema de classificació lineal tridimensional amb $\nu = 0.1$.

2 Anàlisis de dades separables

2.1 Dades separables generades artificialment

Per començar el nostre anàlisi, vam dividir les seccions en funció del tipus de dades utilitzades. Vam utilitzar un dataset de 1400 punts de train i 600 de test (2000 punts en total) amb 4 característiques que ens servirien més tard per dividir el dataset en grups. Fins i tot, per poder obtenir un model capaç de generalitzar correctament, vam assignar erròniament algunes dades a consciència. A més a més, això ens permetrà avaluar la robustesa de l'algorisme i entendre l'impacte dels paràmetres de regularització davant els errors.

2.2 Anàlisi dels resultats

Els resultats dels experiments realitzats han sigut els següents:

ν	Precisió	Valor de la funció objectiu	γ_{primal}	w	Temps (s)
0.0	0.495	0	0	0 0 0 0	0.017
0.1	0.94	60.16001045	-4.89836	2.42618 2.48193 2.41669 2.4697	0.025
0.5	0.94	234.2656996	-7.19041	3.62067 3.67822 3.53392 3.5671	0.043
0.75	0.938333	337.594947	-7.68619	3.88208 3.95582 3.78556 3.76602	0.030
2	0.938333	843.7753059	-8.84178	4.44839 4.57281 4.266 4.40826	0.045
5	0.938333	2046.715458	-9.43335	4.70445 4.88549 4.55287 4.72556	0.039

Table 1: Resultats obtinguts en la execució del problema primal amb dades linealment separables en funció del paràmetre ν (temps d'execució aproximats).

El Teorema de la Dualitat Forta demostra que, si un problema (primal) de programació lineal té una solució òptima, llavors el seu corresponent dual també té una solució òptima i els respectius valors de les funcions objectiu són idèntics. Per aquest motiu no ens sembla estrany que el valor de les dues funcions objectiu hagi coincidit en ambdues formulacions (3, 4).

A més a més, s'observa que aquest valor esdevé nul per a $\nu = 0$ (on no apliquem una penalització d'error) i augmenta monòtonament a mesura que ν s'incrementa. Aquest comportament es pot deure al creixent

ν	Precisió	Valor de la funció objectiu	γ_{dual}	w	Temps (s)
0.0	Error	-	-	-	-
0.1	0.94	60.16001042	-4.89836	2.42618 2.48193 2.41669 2.4697	16.211
0.5	0.94	234.2656995	-7.19041	3.62067 3.67822 3.53392 3.5671	11.515
0.75	0.938333	337.5949468	-7.68619	3.88208 3.95582 3.78556 3.76602	7.112
2	0.938333	843.7753052	-8.84178	4.44839 4.57281 4.266 4.40826	7.746
5	0.938333	2046.715458	-9.43335	4.70445 4.88549 4.55287 4.72556	10.758

Table 2: Resultats obtinguts en la execució del problema dual amb dades linealment separables en funció del paràmetre ν (temps d'execució aproximats).

esforç que porta a terme la SVM per minimitzar els errors de classificació, malgrat no poder fer res davant els punts intencionalment classificats de forma errònia. Aquests punts són els responsables de l'augment de la funció objectiu en un intent de ν per forçar la penalització d'aquests errors.

Per últim fer una anotació respecte al valor de la precisió en el cas de $\nu = 0$ per al problema primal. Aquest resultat ha estat obtingut simplement pel fet que la màquina comença a assignar a totes les dades de test un valor constant ("-1") aconseguint endevinar per mera casualitat quasi la meitat de les vertaderes etiquetes. No obstant, això no es podria considerar pròpiament una actuació de SVM davant la falta de penalització d'error i probablement, paràmetres de $w = 0$ i $\gamma = 0$ (obtenint finalment $f(x) = 0$).

La precisió del model ha mostrat tenir un comportament característic. Al augmentar el valor de ν la precisió va pujar dràsticament al 94%, indicant que el model va aconseguir aprendre l'estructura de les dades. Curiosament, a mesura que el valor de ν s'incrementa, malgrat encara mantenir-se elevada, la precisió tendeix a decréixer. Aquest comportament, a despit del constant augment de la funció objectiu, reflexa el límit del model per millorar la classificació degut als punts erròniament classificats. En general però, el model sembla actuar de manera excel·lent donada la naturalesa inherentment separable de la major part de les dades, siguent la precisió limitada degut a aquest soroll manualment introduït.

El vector de pesos w també ha exhibit un augment en la magnitud de les seves components a mesura que ν creixia. Això és consistent amb un model que intenta ajustar-se cada cop més a les dades, potencialment reduint el marge de l'hiperplà (marge $= \frac{2}{||w||}$). El paràmetre del biaix, γ, també ha variat, mostrant ajustos en la posició de l'hiperplà separador.

2.3 Dades separables de bitllets de banc

En el segon experiment amb dades linealment separables, hem triat un dataset de classificació de bitllets bancaris, el qual consisteix en un recull de mostres obtingudes a partir d'imatges de bitllets genuïns i falsos. Es tracta de dades de tipus imatge, capturades amb una càmera industrial de gran precisió, la mateixa que s'utilitza habitualment en processos d'inspecció d'impressió, cosa que garanteix una qualitat constant i una resolució final de 400×400 píxels. A causa de la distància i l'òptica emprades, aquestes imatges en escala de gris presenten aproximadament 660 punts per polzada (dpi), cosa que assegura un grau de detall suficient per a detectar petites variacions tals com irregularitats en la trama del paper o en la tinta.

No obstant això, per tal de reduir la complexitat de les dades originals, s'ha aplicat sobre aquestes imatges la transformada de Wavelet, que descompon la informació espacial i freqüencial de la textura, permetent capturar propietats estadístiques rellevants per a la detecció d'anomalies. Aquesta informació descomposada ve donada per la variància, que mesura la dispersió de la intensitat dels píxels, reflectint com de heterogeni és el patró de brillantor d'un bitllet; la asimetria (o *skewness*), que avalua la desviació de la distribució de nivells de gris respecte a una simetria perfecta, cosa que pot indicar acumulacions irregulars d'ombra o de tinta; la curtosi, que determina si la distribució de valors té una forma més "punxeguda" o més "aplanada" que la d'una distribució gaussiana, ajudant a detectar variacions extremes degudes a forats, arrugues o restes de tinta; i l'entropia, que quantifica el grau d'aleatorietat o complexitat de la textura, ja que un valor alt sovint es relaciona amb patrons imprevisibles propis de falsificacions mal col·locades o impreses amb qualitat inferior.

Aquestes quatre dimensions ofereixen una representació estadística de cada bitllet que, en molts casos, és suficient perquè classes com "genuí" i "fals" resultin aproximadament linealment separables a l'espai de característiques. Abans d'entrenar qualsevol model, cal normalitzar o escalar adequadament aquestes mètriques per garantir que cap característica domini les distàncies de manera desproporcionada i que el procés d'optimització convergeixi amb estabilitat numèrica.

2.3.1 Problema Primal

A l'hora d'analitzar l'evolució del paràmetre ν, observem un comportament dividit en la capacitat del SVM per aprendre i generalitzar sobre el dataset de bitllets.

Primerament, quan $\nu = 0$, el model es redueix a una solució trivial: el vector de pesos és nul i la funció de decisió no discrimina gens entre bitllets genuïns i falsificats, resultant en una precisió del 57,04%, mesura que s'acosta a la classificació aleatòria. A efectes pràctics, això equival a no aplicar cap procés d'aprenentatge, ja que ni tan sols s'estableix un hiperplà separador. En aquest punt, no existeix marge ($\gamma = 0$) i la funció objectiu és zero, reflectint l'absència tant del terme de norma com del terme de penalització per errors. Es tracta, doncs, d'un subajust extrem, on el model no aprofita la informació estadística de les característiques extretes per la transformada Wavelet.

ν	Precisió	Valor de la funció objectiu	γ_{primal}	w	Temps (s)
0.0	0.5704	0.0000	0.0000	0 0 0 0	0.0118
0.1	0.9879	3.9272	1.3993	-0.9574 -0.6216 -0.7091 -0.0145	0.0612
0.5	0.9927	13.8217	1.8061	-1.4606 -0.9549 -1.0991 -0.0980	0.033
0.75	0.9927	19.5345	1.9048	-1.6362 -1.0327 -1.1980 -0.1114	0.029
2.0	0.9879	45.1841	2.4144	-2.4877 -1.4230 -1.7156 -0.1791	0.028
5	0.9879	104.0490	2.7967	-3.0011 -1.6938 -2.0669 -0.2957	0.032

Table 3: Resultats del problema *primal* per diferents valors de ν sobre el dataset linealment separable (temps d'execució aproximats).

Per una altra banda, quan fem incrementar ν a 0,5 i 0,75, veiem una millora significativa del comportament del model: la precisió assolida arriba al 99,27% i es manté estable, mentre que la magnitud del vector de pesos s'incrementa moderadament. Aquest augment progressiu de $\|\mathbf{w}\|$ implica una reducció lenta del marge, però manté γ per sobre de 1,8, cosa que permet separar les dues classes de forma diferenciada sense sobreajustar excessivament. Aquests ajustaments de ν reflecteixen la capacitat del SVM per trobar un punt de compromís en què l'hiperplà s'apropa més a la distribució dels exemples, aprofitant la informació estadística de cada característica per maximitzar la precisió sense perdre la capacitat de generalització. Les tendències en el valor òptim de la funció (que puja fins a 19,53 per $\nu = 0,75$) destaquen que la complexitat del model augmenta de manera controlada i progressiva.

Sembla que, a partir de $\nu = 2$, podem concloure que hi ha cert sobreajust de l'hiperplà divisor.

Tot i que el marge γ arriba a 2,41 i la norma de $\mathbf{w}$ augmenta substancialment (amb components que superen 2 en magnitud), la precisió cau una mica fins al 98,79%. Això indica que la frontera de decisió s'ha tornat excessivament rígida, ajustant-se massa de prop als punts d'entrenament i capturant sorolls o particularitats de mostres concretes. El valor òptim de la funció (45,18) revela una dominància gairebé total del terme quadràtic, fet que desaprofita la penalització dels errors en favor d'una complexitat creixent. Aquest patró corrobora que un ν massa elevat desequilibra el balanç entre marge i penalització, reduint així la capacitat de generalització en noves dades.

2.3.2 Problema Dual

ν	Precisió	Valor de la funció objectiu	γ_{dual}	w	Temps (s)
0.0	Error	–	–	–	2.1734
0.1	0.9879	3.9272	1.3993	-0.9574 -0.6216 -0.7091 -0.0145	4.7066
0.5	0.9927	13.8217	1.8061	-1.4606 -0.9549 -1.0991 -0.0980	6.7896
0.75	0.9927	19.5345	1.9048	-1.6362 -1.0327 -1.1980 -0.1114	3.9223
2.0	0.9879	45.1841	2.4144	-2.4877 -1.4230 -1.7156 -0.1791	5.5822
5	0.9879	104.0490	2.7967	-3.0011 -1.6938 -2.0669 -0.2957	15.2014

Table 4: Resultats del problema *dual* per diferents valors de ν sobre el mateix dataset (temps d'execució aproximats).

En la formulació dual del SVM, el paràmetre ν juga un paper anàleg al de la formulació primal però canalitza directament la influència dels vectors de suport mitjançant les variables duals α_i, cosa que es tradueix en un marge γ_{dual} i un valor òptim de la funció objectiu que incorpora tant la norma de $\mathbf{w}$ com la suma dels α_i associats a errors. Quan $\nu = 0$, la resolució de la formulació dual no pot trobar cap α_i significatiu, per la qual cosa es produeix un error i no es construeix cap hiperplà separador. Aquest cas reflecteix un aprenentatge nul per part del model, igual que en la formulació primal.

En fixar $\nu = 0{,}1$, s'activa un salt dramàtic en el comportament del dual: la precisió arriba al 98,79 %, la funció objectiu s'estabilitza en 3,927 i es genera un marge $\gamma_{\text{dual}} \approx 1{,}399$. Des del punt de vista de les α_i, això significa que només un petit subconjunt de mostres es converteixen en vectors de suport amb $\alpha_i > 0$, mentre que la resta tenen $\alpha_i = 0$, tot contribuint a una frontera nítida. El vector de pesos $\mathbf{w}$, calculat com a combinació lineal de les α_i i dels productes interns amb les mostres, queda definit amb components aproximadament $(-0{,}96, -0{,}62, -0{,}71, -0{,}01)$, que reflecteixen com de rellevants resulten cada una de les quatre característiques de textura per separar bitllets reals i falsos.

A mesura que augmentem ν cap a 0,5 i 0,75, apreciem com la formulació dual propicia una selecció més àmplia de vectors de suport (les α_i menys restrictives), fet que eleva la norma de $\mathbf{w}$ i, paradoxalment, amplia encara més el marge ($\gamma_{\text{dual}} = 1{,}806$ per a $\nu = 0{,}5$ i 1,905 per a $\nu = 0{,}75$). La precisió creix lleugerament fins al 99,27 %. El valro òptim de la funció passa de 13,82 a 19,53, mostrant una contribució més gran del terme quadràtic i dels termes duals. En aquests punts, el model aconsegueix aprofitar millor la informació distribuïda en la variància, skewness, curtosi i entropia, tot mantenint un equilibri entre

nombre de vectors de suport i robustesa del marge.

Quan ν arriba a 2, la formulació dual reflecteix característiques de sobreajust similars als observats en el primal: la precisió cau una mica fins al 98,79 %, γ_{dual} es dispara fins a 2,414 i el valor de la funció objectiu puja fins a 45,18. Això significa que nombrosos α_i adopten valors intermedis, incorporant massa punts com a suports, la qual cosa augmenta la complexitat del model i el fa excessivament sensible al soroll de les dades d'entrenament. El vector de pesos resultant, amb components fins i tot superiors a -2, destaca com de crucial esdevé cada característica i com la frontera s'ajusta excessivament al conjunt inicial.

En comparar els resultats obtinguts pels problemes primal i dual, el primer element a subratllar és que, tal com prediu el Teorema de la Dualitat Forta, els valors òptims de la funció objectiu coincideixen en ambdues formulacions per a cada valor de ν. Això confirma que, malgrat la diferent manera de formular el problema, s'està trobant la mateixa solució òptima en termes de marge i errors penalitzats. De fet, els valors de γ (primal) i γ_{dual} (dual) són pràcticament idèntics en cada cas.

Pel que fa als pesos $\mathbf{w}$, en ambdós enfocaments observem que tots els components són negatius i que el quart—l'entropia de la imatge—presenta magnituds molt menors que la resta. Això ens indica que, en el context de la nostra classificació, valors més alts de variància, skewness i curtosi són molt més discriminatius que l'entropia, la qual aporta poc valor addicional. El signe negatiu invariat de tots els pesos suggereix, a més, que un increment en qualsevol d'aquestes característiques tendeix a predir la classe "genuí" (associada a -1), especialment per la variància, on les magnituds més grans de $\mathbf{w}$ apunten que una elevada dispersió de la intensitat dels píxels és senyal d'autenticitat.

Una diferència operativa, però, rau en la complexitat computacional: hem observat que el primal convergia en unes 26 iteracions mentre que el dual n'exigia unes 32, tot i que el temps de càlcul resultant va ser pràcticament idèntic. Això es deu a la naturalesa de la matriu amb la qual treballa cada enfocament: la formulació dual maneja una matriu quadrada de mida igual al nombre de mostres (que en el nostre cas no és molt gran, però sí suficient per notar-ho en el nombre d'iteracions), mentre que el primal opera directament sobre el nombre de variables, que és molt menor.

3 Anàlisis de dades no separables

En tractar un dataset linealment no separable com el *Swiss Roll*, la incapacitat de la formulació primal-dual clàssica d'ajustar una frontera lineal revela ràpidament els seus límits: a mesura que incrementa ν, el valor de la funció objectiu creix de manera quasi lineal i la precisió es manté constant en un valor moderat, perquè simplement no existeix cap hiperplà lineal que pugui discriminar adequadament les dues classes. L'augment de ν, que estipula la importància relativa de les penalitzacions per errors en contraposició amb l'amplitud del marge, fa que el model admeti un nombre creixent de punts dins del marge o directament mal classificats, incrementant la suma dels valors "d'slack" i, per tant, fent créixer la funció objectiu encara que el marge es vagi fent més estret. Aquesta evolució quasi lineal del cost reflecteix la pròpia naturalesa de l'enfocament: sense transformar l'espai de característiques, cada increment de penalització es veu traduït en més punts "tolerats" dins del marge, però sense millorar la separabilitat real de les dades.

Per contra, quan incorporem un nucli gaussià, l'estratègia canvia radicalment: en projectar implícitament les mostres a un espai d'alta dimensió on la superfície de separació ja no és necessàriament un hiperplà en el sentit clàssic, el SVM amb kernel fa possible que les classes que en l'espai original s'enrotllen en una espiral esdevinguin clarament separables. Així doncs, la funció objectiu ja no creix indefinidament amb ν sinó que, un cop apresa una frontera complexa que ajusta bé les regions de cada classe, es tendeix a estabilitzar en un valor: els slacks addicionals que permet un ν més alt tenen poc impacte, perquè la majoria dels punts ja queden perfectament classificats fora del marge. Aquesta mesura de "creixement lent i estabilització" de la funció de cost reflecteix la capacitat del nucli per capturar la geometria intrínseca de les dades.

En termes de precisió, l'avantatge és encara més evident: mentre que la formulació primal-dual obté una *accuracy* modest i invariable (fruit de la frontera lineal immòbil), el SVM amb nucli gaussià supera ràpidament aquests valors i arriba a percentatges d'encerts de fins al 99 %. D'entrada, per a valors petits de ν, el kernel pot mostrar una precisió inicial menor per l'efecte de la regularització i d'un marge encara massa ampli, però un cop s'assegura una correcta penalització dels errors, s'assoleix una classificació pràcticament perfecta. Aquesta progressió posa de manifest el caràcter no lineal de la frontera necessària per segmentar el *Swiss Roll*, ja que només un nucli capaç de generar superfícies corbes pot adaptar-se a la forma espiralada de les classes.

Cal subratllar que, més enllà d'un cert llindar (al voltant de $\nu = 0{,}5$), la precisió del kernel ja no varia: la frontera s'ha adaptat tan bé a la distribució complexa que la variació posterior de la penalització ja no modifica els α_i essencials ni la geometria de la solució.

Aquestes conclusions mostren com, per a dades linealment no separables, la introducció d'un nucli adequat (en aquest cas, el gaussià) no només millora dramàticament l'ajust de la frontera, sinó que també permet contenir el creixement de la funció objectiu i aconseguir una precisió molt superior, tot mantenint un marge ben calibrat i uns slacks sota control fins i tot per a valors alts de ν.