UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Processament de Llenguatge Natural amb *Word2Vec* i Màquines de Suport Vectorial

Autor: Elena Blanco López

Directors:

Dr. Josep Vives i Santa-Eulàlia,
Departament de Matemàtica Econòmica, Financera i Actuarial.

Santiago Seguí Mesquida,
Departament de Matemàtiques i Informàtica.

Barcelona, 17 de gener de 2024

Abstract

In this project, the theoretical foundations of two machine learning methods are studied in detail to understand their methodology and learning process.

Word2Vec consists of a neural network capable of processing text data as numerical vectors that attempt to reflect the semantic and grammatical relationships between words.

Support Vector Machines construct a hyperplane or set of hyperplanes in a high-dimensional (or even infinite) space that can be used in classification or regression problems.

Finally, these two models will be combined in an algorithm that will allow us to carry out a multiclassification task on text data, which will involve detecting possible metastasis in cancer patients from clinical annotations.

Resum

En aquest projecte s'estudien amb detall les bases teòriques de dos mètodes d'aprenentatge automàtic per tal de compendre amb profunditat la seva metodologia i procès d'aprenentatge.

El primer model, *Word2Vec*, consisteix en una xarxa neuronal capaç de processar dades de text per transformar-les en vectors numèrics que tracten de reflectir les relacions semàntiques i gramaticals entre paraules.

En segon lloc, les màquines de suport vectorial constitueixen un seguit de mètodes que construeixen un hiperplà o conjunt d'hiperplans en un espai de dimensionalitat molt alta (o fins i tot infinita) que pot ser utilitzat en problemes de classificació o regressió.

Finalment, s'uneixen aquests dos models en un algoritme que ens permet dur a terme una tasca de multiclassificació de dades de text mitjançant un programa en *Python*. Aquest consistirà en detectar possibles metàstasis de pacients de càncer en anotacions clíniques.

2020 Mathematics Subject Classification. 68T50, 68T10, 30C40

Agraïments

Vull agraïr a en Josep i en Santi per acceptar la meva proposta de treball que tant em motivava i per proporcionar-me consells que han estat imprescindibles pel desenvolupament d'aquest treball.

També vull donar gràcies a en José Luis per proporcionar-me aquest projecte, i a tot el departament de l'Oficina Tècnica de l'ICO Hospitalet per acollir-me tan bé i pel suport diari.

Per últim, agraïr als meus pares, a l'Abel, a en Clemen i a la Cristina pel seu suport incondicional.

Índex

Introducció

El processament del llenguatge natural (PLN) és un camp de la intel·ligència artificial que permet als ordinadors entendre i comunicar-se en llenguatge natural tractant d'imitar l'èsser humà. Aquest combina tècniques i algoritmes d'aprenentatge automàtic (en anglés, *machine learning*) o d'aprenentatge profund (en anglés, *deep learning*) amb l'anàlisi de text i la lingüística computacional, amb la fi de dotar a les màquines amb la capacitat de compendre, manipular i generar llenguatge natural de la manera més semblant possible a la complexitat humana.

L'adopció del processament del llenguatge natural en l'àmbit sanitari no ha fet més que crèixer en els darrers anys pel seu potencial per analitzar i interpretar grans quantitats de dades de pacients. Aquesta tecnologia permet la identificació d'indicadors o mesures de la malaltia que s'especifiquen en les notes clíniques dels pacients i que es consideraven perdudes en format de text. El fet de poder estructurar aquest tipus de dades presenta un gran avenç envers la millora de tractaments de pacients.

Durant les meves pràctiques en l'Institut Català d'Oncologia (ICO) se'm va proposar un projecte d'extracció de l'estadi dels pacients de càncer de colon, una mesura descriptiva del tamany tumor i de la seva extensió a teixits o ganglis limfàtics adjacents. En el marc pràctic d'aquest treball s'utilitza una base de dades obtinguda a través de l'estudi *Estudi per a la creació i validació d'un mètode informàtic per a l'obtenció i estructuració de l'estadiatge en càncer a partir de les dades existents al curs clínic* desenvolupat a l'Oficina Tècnica de l'ICO Hospitalet, amb l'objectiu de disposar de dades estructurades necessàries per a l'avaluació de resultats i millora de qualitat a la pràctica clínica i assistencial. Per establir aquesta mesura és indispensable conèixer si el pacient té metàstasi o no. És per això que es va desenvolupar un programa informàtic d'extracció de frases, d'una llargada de tretze paraules, que contenen informació sobre la metàstasi dels pacients. L'objectiu del marc pràctic d'aquest treball és ser capaços de transformar aquestes frases en vectors numèrics per posteriorment emprar un classificador de màquines de suport vectorial que les etiqueti segons si el pacient té metàstasi, no en té o no s'ha pogut determinar.

Quan els ordinadors són emprats per solucionar problemes pràctics hi ha dues maneres d'enfocar la resolució, una és mitjançant la programació tradicional. Aquest és el cas en què el mètode per obtenir el resultat desitjat a partir d'un conjunt de dades d'entrada pot ser especificat pas a pas. La tasca del programador és traslladar aquest mètode a una seqüència d'instruccions de codi que l'ordinador pugui seguir. No obstant, a vegades no és possible dissenyar un sistema lògic que assoleixi el resultat desitjat a partir de les dades disponibles. Aquí és quan sorgeix una segona manera d'enfocar la qüestió la qual consisteix en què l'ordinador sigui capaç d'apendre una relació entre les dades disponibles i el resultat pertinent, a partir d'exemples que se li proporcionen. Així com un nen pot aprendre què és un elefant si se li assenyala dins d'un grup d'animals del zoo en comptes de donant una descripció precisa i detallada de l'animal. A aquest enfocament se l'anomena **metodologia d'aprenentatge**.

A continuació es proporcionarà el context teòric sobre el qual es desenvolupa aquest treball incloent conceptes bàsics de probabilitat, aprenentatge estadístic i mètodes d'aproximació necessàris per tal d'entendre la naturalesa del problema.

Donat un conjunt de dades que anomenarem **conjunt d'entrada** $X = \{\boldsymbol{x}_k\}_{k=1}^d$ es vol que l'ordinador sigui capaç de predir els resultats corresponents $Y = \{\boldsymbol{y}_k\}_{k=1}^d$ o **conjunt de sortida**. Anomenarem a la seqüència de parelles $S = \big((\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), ..., (\boldsymbol{x}_d, \boldsymbol{y}_d)\big)$,

conjunt d'entrenament. A aquest cas concret de metodologia d'aprenentatge on les dades d'entrenament són parelles del tipus entrada/sortida se l'anomena **aprenentatge supervisat**. Per simplificar, se suposarà durant el transcurs del treball que $Y \subseteq \mathbb{R}$. El conjunt S s'utilitza llavors per apendre una funció $f : X \longrightarrow \mathbb{R}$ tal que $f(\boldsymbol{x})$ sigui una bona aproximació de la possible resposta $\boldsymbol{y} \in Y$. Però, què significa que $f(\boldsymbol{x})$ sigui una bona aproximació? Per això es defineix una **funció de pèrdua** $L : X \times Y \times \mathbb{R} \longrightarrow [0, \infty)$ tal que el valor $L(\boldsymbol{x}, \boldsymbol{y}, f(\boldsymbol{x}))$ serà més petit com més bona sigui l'aproximació $f(\boldsymbol{x})$. És obvi que no és suficient amb conèixer el valor de $L(\boldsymbol{x}, \boldsymbol{y}, f(\boldsymbol{x}))$ per un element particular $(\boldsymbol{x}, \boldsymbol{y})$, en canvi, cal estudiar el seu valor per qualsevol parella del conjunt d'entrenament. En aprenentatge estadístic generalment es tracta de minimitzar la pèrdua esperada de f o **risc de** $\boldsymbol{f}$

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L(\boldsymbol{x}, \boldsymbol{y}, f(\boldsymbol{x}))dP(\boldsymbol{x}, \boldsymbol{y}) = \int_X \int_Y L(\boldsymbol{x}, \boldsymbol{y}, f(\boldsymbol{x}))dP(\boldsymbol{y}|\boldsymbol{x})dP_X(\boldsymbol{x}),$$

on P és la distribució de probabilitat en $X \times Y$ desconeguda que genera els parells $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ de S. Per tant, una aproximació f és millor com més petit sigui $\mathcal{R}_{L,P}(f)$ i és natural considerar el mínim risc possible

$$\mathcal{R}^*_{L,P} := \inf_{f:X \longrightarrow \mathbb{R}} \mathcal{R}_{L,P}(f).$$

Les màquines de suport vectorial conformen un model d'aprenentatge supervisat que resol problemes tant de regressió com de classificació. Per simplificar la notació denotarem $L(\boldsymbol{x}, \boldsymbol{y}, f(\boldsymbol{x})) := L(\boldsymbol{y}, f(\boldsymbol{x}))$, que es pot entendre com que la funció de pèrdua depèn de $\boldsymbol{x}$ només a través de $f(\boldsymbol{x})$. En el cas de la classificació binària el conjunt de sortida és $Y = \{-1, 1\}$ i una de les funcions de pèrdua emprades amb més freqüència és la L_{class}, la qual només penalitza les classificacions errònies ($L_{class}(\boldsymbol{y}, f(\boldsymbol{x})) = 1$ si $sgn(f(\boldsymbol{x})) \neq \boldsymbol{y}$ i és igual a 0 altrament). A partir d'ara considerarem el cas concret $L := L_{class}$.

Com que la distribució de probabilitat P és desconeguda i conseqüentment $\mathcal{R}_{L,P}(f)$ també ho és, aleshores és complicat trobar una funció que minimitzi aquest risc directament. Una solució a aquest problema consisteix en reemplaçar el risc de f per la seva equivalència empírica

$$\mathcal{R}_{L,P_S}(f) = \frac{1}{d} \sum_{i=1}^{d} L(\boldsymbol{y}_i, f(\boldsymbol{x}_i)),$$

on P_S és la distribució empírica* de les observacions (x_k, y_k), $k = 1, ..., d$. La llei dels grans nombres assegura que $\mathcal{R}_{L,P_S}(f)$ és una aproximació de $\mathcal{R}_{L,P}(f)$ per cada f, tot i així, trobar l'ínfim de $\mathcal{R}_{L,P_S}(f)$ per totes les funcions $X \longrightarrow \mathbb{R}$ no sempre dona lloc a una bona aproximació de $\mathcal{R}^*_{L,P}(f)$. Per exemple, si es pren una funció f que classifiqui tots els $\boldsymbol{x}_i$ correctament però que valgui 0 a la resta de valors, aleshores aquesta funció l'únic que fa és memoritzar S, conformant així una aproximació dolenta de $\mathcal{R}^*_{L,P}(f)$. Aquest és un exemple extrem del fenomen anomenat **sobreajustament**. L'objectiu de l'aprenentatge automàtic és obtenir models que actuïn bé no només sobre les dades d'entrenament sino sobretot sobre noves dades que aquest no hagi vist anteriorment. Una bona manera d'atacar aquesta qüestió és considerant un conjunt petit de funcions específiques $X \longrightarrow \mathbb{R}$, $\mathcal{F}$, per tal que el problema de minimització es simplifiqui en trobar la funció f^* solució de

*La corresponent funció de distribució empírica és $F_d(x, y) = \frac{1}{d} \sum_{i=1}^{d} \mathbb{1}_{\{X_i \geqslant x, Y_i \geqslant y\}}(x, y)$.

$$\inf_{f \in \mathcal{F}} \mathcal{R}_{L,P_S}(f), \tag{1}$$

on $\mathcal{F}$ és l'anomenat **conjunt de funcions hipòtesi**.

Al Capítol 1 es presentaran les funcions nucli (*kernel*, en anglès) que constitueixen un dels blocs principals per la construcció de màquines de suport vectorial. Es proporcionarà una caracterització rigurosa, tractant els dos casos X finit i infinit numerable, es provarà que defineixen íntegrament el conjunt de funcions hipòtesi mitjançant la representació dual d'aquestes i s'exposaran alguns dels nuclis més emprats a la pràctica. Posteriorment, al Capítol 2 s'explicarà amb detall com les màquines de suport vectorial troben una solució de (1) fent ús dels nuclis, per això s'inclourà un breu apartat de teoria d'optimització on s'introduiran els multiplicadors de Lagrange i el teorema de Kuhn-Tucker.

Per poder aplicar classificadors de suport vectorial a les nostres frases, abans s'ha de trobar una representació vectorial numèrica adient de les mateixes. Explorarem al Capítol 3 com dades de text poden ser transformades en vectors numèrics que puguin ser processats per mètodes d'aprenentatge automàtic. En concret, ens centrarem en explicar el mètode *Word2Vec* que transforma paraules en vectors numèrics intentant plasmar les possibles relacions semàntiques i gramaticals entre aquestes.

Els resultats dels capítols anteriors ens permetran al Capítol 4 estudiar com diferents models de classificadors de suport vectorial es comporten sobre les frases de la metàstasi. Abans però, s'iniciarà el capítol amb una explicació sobre la naturalesa de les dades d'entrenament per posteriorment determinar quines mètriques s'utilitzaran per mesurar l'eficàcia dels models. A continuació es farà un anàlisi sobre quin dels diferents models que *Word2Vec* ofereix s'adhereix millor a les nostres frases. I per últim, es farà un estudi sobre l'ús de diferents nuclis en els classificadors de suport vectorial i es trobaran els hiperparàmetres òptims en cada cas mitjançant el mètode de validació creuada.

Els objectius d'aquest projecte són per tant proporcionar una base teòrica el més detallada possible sobre els mètodes de les màquines de suport vectorial i *Word2Vec* per tal de finalment posar-los en pràctica sobre el nostre conjunt de dades, format per frases que parlen de la possible metàstasi de pacients de càncer de colon. Més encara, es farà un anàlisi dels resultats obtinguts amb el propòsit de poder visualitzar alguns aspectes del marc teòric.

1 Teoria del nucli

1.1 Un exemple concret: definició de l'espai característic

En l'aprenentatge supervisat l'algoritme és alimentat per un conjunt de dades d'entrada X que suposarem que consisteixen en vectors cada entrada dels quals correspon a un atribut, és a dir, $X \subseteq \mathbb{R}^n$, juntament amb les seves etiquetes associades, Y, que se suposarà que és un subconjunt de $\mathbb{R}$. Reprenent l'exemple dels elefants de la introducció, el conjunt X podria consistir en dos lleons, un ós formiguer, tres elefants i dues serps que tenen tres atributs: tamany, nombre de potes i si tenen una prolongació al nas o no. En aquest cas X és un subconjunt de $\mathbb{R}^3$ amb vuit elements, i $Y = \{-1, 1\}$ segons si és un elefant o no ho és.

En un problema d'aprenentatge supervisat, el conjunt de funcions hipòtesi més simple a triar és el conjunt de funcions lineals

$$\{f : X \subseteq \mathbb{R}^n \longrightarrow \mathbb{R} \,|\, f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b,\ (\boldsymbol{w}, b) \in \mathbb{R}^n \times \mathbb{R}\}.$$

Notem que cadascuna d'aquestes funcions defineix un hiperplà en $\mathbb{R}^n$, una varietat lineal de dimensió n-1. Geomètricament parlant, en una tasca de classificació binària la funció solució definirà un hiperplà que separarà l'espai d'entrada en dues regions, una que representi la classe positiva i l'altra la negativa.

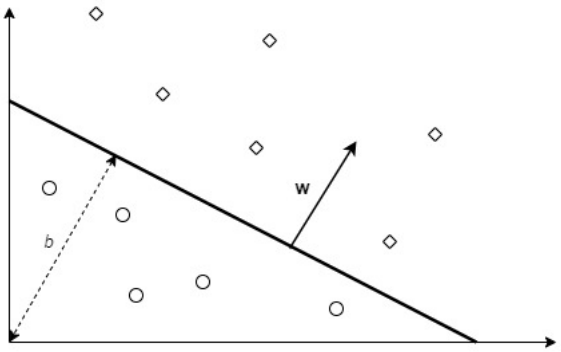

Figura 1 : Un hiperplà $(\boldsymbol{w}, b)$ que separa un conjunt de dades bidimensional

L'ús de funcions lineals com a conjunt d'hipòtesis al nostre model pot ser realment útil depenent del conjunt de dades d'entrada X i la tasca que es vulgui realitzar. No obstant, aquests models no presenten massa efectivitat davant d'alguns problemes del món real de més complexitat que requereixen la tria d'un conjunt de funcions hipòtesi que no siguin combinació lineal dels atributs donats.

Imaginem que comptem amb un conjunt de dades sobre malalts d'una certa malaltia. El nostre conjunt d'entrada consta de dues columnes: en una tenim els miligrams de medicació que se'ls ha receptat i a l'altra si el pacient s'ha curat o no. S'observa que la reacció dels pacients envers el medicament és notablement negativa si la dosi no és l'adequada, és a dir, si se li subministra una dosi massa baixa o massa alta. Si representem aquestes dades amb punts sobre la recta real positiva, observem que els punts més propers al 0 són de malalts que no s'han curat, a prop de la dosi recomanada són de malalts que sí s'han curat i, altra vegada, els punts de més a la dreta corresponen a no curats. Podem visualitzar aquesta representació en l'eix OX de la gràfica de la figura següent.

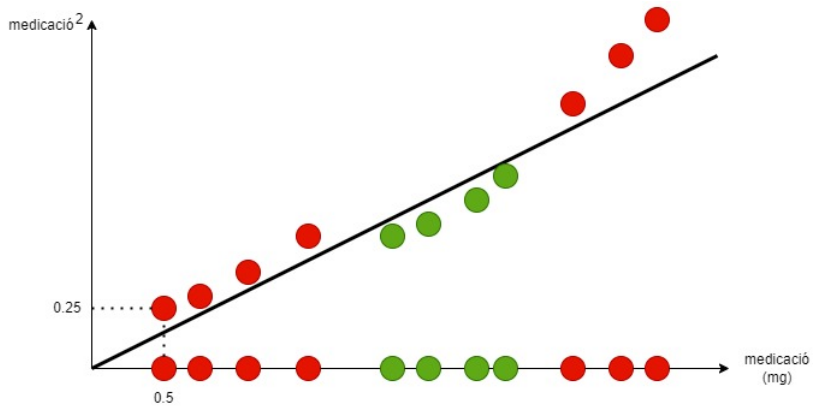

Figura 2 : Una assignació a un espai de major dimensió pot simplificar la tasca de classificació

Notem que afegint una dimensió més, proporcionada per la funció quadrada $\phi_2(x) = x^2$, essent x els miligrams de medicació subministrats, si definim

$$\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x})) := (\boldsymbol{x}, \boldsymbol{x}^2),$$

obtenim que sobre l'espai $\boldsymbol{\phi}(X) \subseteq \mathbb{R}^2$ es pot realitzar una separació de les dades mitjançant una funció lineal, és a dir, una recta. Aquest exemple dona lloc a poder definir, en general, l'espai vectorial resultat d'aplicar la següent aplicació:

$$\boldsymbol{x} = (x_1, ..., x_n) \longmapsto \boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), ..., \phi_N(\boldsymbol{x}))$$

amb $N \in \mathbb{N} \cup \{\infty\}$ no necessàriament igual a n. $F = \{\boldsymbol{\phi}(\boldsymbol{x}) \,|\, \boldsymbol{x} \in X\}$ és per tant l'**espai característic** i anomenarem **funció característica** a $\boldsymbol{\phi}$. En el context de l'aprenentatge automàtic per $i \in \{1, ..., N\}$ els elements x_i es solen denominar com **atributs** i els elements $\phi_i(\boldsymbol{x})$ com **característiques**.

1.2 El truc del nucli

Tal com il·lustra l'últim exemple de l'apartat anterior, per tal de trobar un hiperplà que separi les nostres dades, primer es reescriuran les dades d'entrada en una nova representació a l'espai característic F aplicant una funció no lineal a les nostres dades d'entrada, per posteriorment aplicar un classificador lineal a F. En aquests dos pasos és com es construeixen els models de classificació no lineal, com per exemple alguns models de màquines de suport vectorial. El conjunt de funcions hipòtesi estarà format aleshores per funcions del tipus

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} w_i \phi_i(\boldsymbol{x}) + b, \tag{1.1}$$

on $\boldsymbol{\phi} = (\phi_1, ..., \phi_N)$ és la funció característica i és no lineal. Recordem que aquesta pren valors a l'espai X i els transforma en elements de l'espai característic.

Una propietat molt important dels classificadors lineals és que les funcions hipòtesi es poden expressar com a combinació lineal de les dades d'entrenament de la manera

$$f(\boldsymbol{x}) = \sum_{i=1}^{d} \alpha_i y_i \langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}) \rangle_F + b, \tag{1.2}$$

on $d \in \mathbb{N} \cup \{\infty\}$, és el nombre d'elements de l'espai d'entrada X.

Notació 1.1. *Es denotarà $\langle \cdot, \cdot \rangle_V$ el producte escalar definit a l'espai vectorial V. Si no s'especifica res se suposarà que es parla del producte escalar euclidià real.*

No existeix una demostració general d'aquesta propietat ja que depèn completament de l'algoritme i del problema d'optimització plantejats pel mètode que s'estigui emprant. En el cas de les màquines de suport vectorial veurem com s'assoleix aquesta representació dual de les funcions hipòtesi al següent capítol amb més detall. No obstant, en el transcurs d'aquest capítol suposarem que tota funció hipòtesi té una representació primària (1.1) i una representació dual (1.2), i ens centrarem en les propietats que ens ofereixen les funcions nucli a més de com es construeixen.

Notem que si aconseguim calcular els productes escalars $\langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}) \rangle_F$ a l'espai característic es redueixen els dos pasos anteriors en un sol càlcul. Aquest fenomen és l'anomenat truc del nucli. En altres paraules, si es coneix la matriu de Gram a F $\mathbb{K} = \left(\langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle_F \right)_{i,j=1}^{d}$ també anomenada matriu nucli, aleshores s'evita el càlcul explícit dels punts $\boldsymbol{\phi}(\boldsymbol{x}) \in F$, és a dir, no és necessàri coneixer quin és l'espai F per poder entrenar el nostre model en aquest mateix espai.

1.3 Teorema de Mercer: caracterització dels nuclis

Iniciem aquest apartat definint el que és una funció nucli, l'eina que facilitarà el poder utilitzar una funció lineal per separar les diferents poblacions del nostre espai d'entrada. Per tal de fer-ho possible primer es crea un espai característic d'alta dimensionalitat per posteriorment establir quin és el producte escalar sobre aquest i, per últim, trobar un mètode directe per calcular el producte escalar de dos elements, en termes dels elements de l'espai d'entrada inicial. Veurem que el fet de trobar una funció nucli adequada ens permetrà definir implícitament l'espai característic sense haver de calcular els productes escalars manualment. En aquesta secció s'utilitzaran conceptes i resultats de teoria d'operadors i teoria espectral recollits als l'apèndixs A i B.

Per tal de generalitzar la definició de l'espai característic per a dimensió aribitrària, possiblement infinita, considerarem els espais de Hilbert els quals són una generalització dels espais euclidians. En concret, durant el treball considerarem espais de Hilbert que són una generalització de $\mathbb{R}^n$.

Definició 1.1 (Espai de Hilbert)**.** *Un espai de Hilbert $\mathcal{H}$ és un espai vectorial dotat d'un producte escalar (espai prehilbertià) que, a més, és complet respecte a la norma induïda pel producte escalar, $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$.*

Definició 1.2 (Funció nucli)**.** *Sigui $X \subseteq \mathbb{R}^n$ i sigui $\boldsymbol{K}$: $X \times X \longmapsto \mathbb{R}$. Diem que $\boldsymbol{K}$ és un nucli si existeix una funció $\boldsymbol{\phi} : X \longmapsto \mathcal{H}$, on $\mathcal{H}$ és un espai de Hilbert separable, tal que*

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z}) \rangle_{\mathcal{H}}. \tag{1.3}$$

Observació 1.1. *Notem que* $\boldsymbol{K(x,z)}$ *és simètrica ja que el producte escalar en* $\mathcal{H}$ *és simètric.*

Observació 1.2. *Se sap que el producte escalar és una mesura de similaritat basada en angles entre vectors o en termes de localització de punts respecte a l'origen de coordenades. Per tant, segons (1.3), un nucli pot ser interpretat com una mesura de similaritat entre punts de l'espai característic.*

A continuació s'exposarà un teorema que presenta condicions generals i suficients per tal que una funció contínua i simètrica $X \times X \longrightarrow \mathbb{R}$ sigui un nucli, admetent així una representació com (1.3). Aquest va ser introduit per James Mercer al 1909, però no va ser fins el 1964 que es va introduir la representació dels nuclis com productes escalars dels elements de l'espai característic en el món de l'aprenentatge automàtic, en el treball [4] d'Aizermann, Bravermann i Rozoener.

Primer s'exposarà la versió més general on X és un subconjunt compacte de $\mathbb{R}^n$ no necessàriament finit. Abans però, necessitem definir un parell de conceptes:

Definició 1.3 (Espai L^p)**.** *Considerem una funció* f *amb domini* X*. Per* $p > 0$ *definim la norma* L^p *com*

$$\|f\|_{L^p} := \left(\int_X |f(x)|^p dx\right)^{\frac{1}{p}}.$$

L'espai L^p *es el conjunt de funcions* $f : X \longmapsto \mathbb{R}$ *tals que la seva norma* L^p *és finita, és a dir,*

$$L^p(X) := \{f : X \longmapsto \mathbb{R} \,|\, \|f\|_p < \infty\}.$$

Definició 1.4 (Espai ℓ_p)**.** *Per* $0 < p < \infty$*,* ℓ_p *és l'espai de successions reals* $(x_n)_{n\in\mathbb{N}}$ *tals que*

$$\sum_{n\in\mathbb{N}} |x_n|^p < \infty.$$

Si $p \geqslant 1$ *aleshores* ℓ_p*, equipat amb la norma*

$$\|x_n\|_p = \Big(\sum_{n\in\mathbb{N}} |x_n|^p\Big)^{1/p},$$

és un espai mètric complet. Si $p = 2$ *aleshores* ℓ_2 *és també un espai de Hilbert amb producte escalar* $\big\langle (x_n)_{n\in\mathbb{N}}, (y_n)_{n\in\mathbb{N}} \big\rangle_{\ell_2} = \sum_{n\in\mathbb{N}} x_n y_n$*, això és el producte escalar euclidià.*

Teorema 1.1 (Teorema de Mercer)**.** *Sigui* X *un subconjunt compacte de* $\mathbb{R}^n$ *i sigui* $\boldsymbol{K} \in L^2(X \times X)$ *una funció continua i simètrica. Aleshores l'operador de Hilbert-Schmidt* $T_{\boldsymbol{K}} : L^2(X) \longmapsto L^2(X)$*, definit per*

$$(T_{\boldsymbol{K}} f)(\cdot) = \int_X \boldsymbol{K}(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}, \tag{1.4}$$

és positiu, és a dir,

$$\int_{X\times X} \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z} \geqslant 0$$

per tota $f \in L^2(X)$*, si i només si,* $\boldsymbol{K(x,z)}$ *és un nucli.*

Demostració. Aquesta demostració ha estat inspirada per les notes de John Thicksun que podem trobar a [5], sobre el teorema de Mercer.
($\Longrightarrow$) Primer notem que si $\boldsymbol{K}$ és simètrica, aleshores el seu operador de Hilbert-Schmidt $T_{\boldsymbol{K}}$ és autoadjunt, per $f, g \in L^2(X)$:

$$\langle f, T_{\boldsymbol{K}} g\rangle_{L^2} = \int_X f(\boldsymbol{x}) \int_X \boldsymbol{K}(\boldsymbol{x},\boldsymbol{z}) g(\boldsymbol{z})\, d\boldsymbol{z} d\boldsymbol{x} = \int_X \int_X \boldsymbol{K}(\boldsymbol{z},\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}\, g(\boldsymbol{z}) d\boldsymbol{z} = \langle T_{\boldsymbol{K}} f, g\rangle_{L^2}.$$

Ara, com $T_{\boldsymbol{K}}$ és un operador de Hilbert-Schmidt, és compacte (veure apèndix A), i pel teorema espectral per operadors autoadjunts i compactes (veure apèndix B), $T_{\boldsymbol{K}}$ té un nombre numerable de vectors propis (en el nostre cas funcions propies) ortonormals $\{\psi_i\}_{i\in\mathbb{N}}$ amb valors propis $\{\lambda_i\}_{i\in\mathbb{N}}$ tals que per a tot $f \in L^2(X)$,

$$T_{\boldsymbol{K}} f = \sum_{i=1}^{\infty} \lambda_i \langle f, \psi_i\rangle_{L^2} \psi_i.$$

Pel teorema de Fubini i pel fet que $T_{\boldsymbol{K}}$ sigui positiu, tenim que

$$\lambda_i = \langle \psi_i, T_{\boldsymbol{K}} \psi_i\rangle_{L^2} = \int_X \psi_i(\boldsymbol{x}) \int_X \boldsymbol{K}(\boldsymbol{x},\boldsymbol{z}) \psi_i(\boldsymbol{z})\, d\boldsymbol{z} d\boldsymbol{x} \geqslant 0,$$

del que deduïm que $\lambda_i \geqslant 0$. Definim $\boldsymbol{\phi} : X \longrightarrow \ell_2$ amb $\phi_i \in L^2(X)$, $\phi_i(\boldsymbol{x}) = \sqrt{\lambda_i}\psi_i(\boldsymbol{x})$.

$$\begin{aligned}\int_X \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z})\rangle_{\ell^2} f(\boldsymbol{z})\, d\boldsymbol{z} &= \int_X \sum_{i=0}^{\infty} \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{z}) f(\boldsymbol{z})\, d\boldsymbol{z} \\ = \sum_{i=0}^{\infty} \lambda_i \psi_i(\boldsymbol{x}) \int_X \psi_i(\boldsymbol{z}) f(\boldsymbol{z})\, d\boldsymbol{z} &= \sum_{i=0}^{\infty} \lambda_i \langle f, \psi_i\rangle_{L^2} \psi_i(\boldsymbol{x}) = T_{\boldsymbol{K}} f(\boldsymbol{x}).\end{aligned}$$

Com aquesta igualtat es compleix per qualsevol $f \in L^2(X)$, tenim que

$$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z})\rangle_{\ell^2}.$$

($\Longleftarrow$) Si $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{z})$ és un nucli aleshores

$$\begin{aligned}\int_{X\times X} \boldsymbol{K}(\boldsymbol{x},\boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z})\, d\boldsymbol{x} d\boldsymbol{z} &= \int_{X\times X} \langle f(\boldsymbol{x})\phi(\boldsymbol{x}), f(\boldsymbol{z})\phi(\boldsymbol{z})\rangle_{\mathcal{H}}\, d\boldsymbol{x} d\boldsymbol{z} \\ &= \left\langle \int_X f(\boldsymbol{x})\phi(\boldsymbol{x})\, d\boldsymbol{x}, \int_X f(\boldsymbol{z})\phi(\boldsymbol{z})\, d\boldsymbol{z} \right\rangle_{\mathcal{H}} \geqslant 0,\end{aligned}$$

per ser $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ un producte escalar. □

Observació 1.3. *La condició de positivitat*

$$\int_{X\times X} \boldsymbol{K}(\boldsymbol{x},\boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z})\, d\boldsymbol{x} d\boldsymbol{z} \geqslant 0$$

és equivalent a la condició de què $\forall l \in \mathbb{N}$, *siguin* $\alpha_1, ..., \alpha_l \in \mathbb{R}$ *i* $\boldsymbol{x}_1, ..., \boldsymbol{x}_l \in X$, *aleshores*

$$\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geqslant 0 \tag{1.5}$$

que, per últim, és equivalent a què la matriu de Gram o matriu nucli $\mathbb{K} = \big(\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)\big)_{i,j=1}^{l}$ *sigui definida positiva.*

Ara veurem que en el cas en què X és finit, la caracterització dels nuclis depèn només d'una condició matricial sobre la matriu nucli $\mathbb{K}$.

Proposició 1.1. *Sigui $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_d\}$ un subconjunt de $\mathbb{R}^n$ i sigui $\boldsymbol{K} : X \times X \longrightarrow \mathbb{R}$ una funció simètrica. Aleshores $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z})$ és un nucli si, i només si, la matriu nucli $\mathbb{K}$ és definida positiva.*

Demostració. ($\Longrightarrow$) Suposem que $\boldsymbol{K}$ és un nucli, aleshores $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z}) \rangle_{\mathcal{H}}$ per una funció $\boldsymbol{\phi} : X \longrightarrow \mathcal{H}$. Sigui $u \in \mathbb{R}^d$, volem veure que $u^T \mathbb{K} u \geqslant 0$, on u^T denota el vector transposat de u (o vector fila), que com hem vist a l'Observació 1.3 equival a (1.5).

$$\begin{aligned} u^T \mathbb{K} u &= \sum_{i=1}^{d} \sum_{j=1}^{d} u_i u_j \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \sum_{i=1}^{d} \sum_{j=1}^{d} u_i u_j \langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{d} u_i \boldsymbol{\phi}(\boldsymbol{x}_i) \,, \sum_{j=1}^{d} u_j \boldsymbol{\phi}(\boldsymbol{x}_j) \right\rangle_{\mathcal{H}} \geqslant 0, \end{aligned}$$

per ser $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ un producte escalar.

($\Longleftarrow$) Com $\mathbb{K}$ és simètrica aleshores existeix una matriu ortogonal $\mathbf{V}$ tal que la matriu nucli pot factoritzar com $\mathbb{K} = \mathbf{V} \Lambda \mathbf{V}^T$, on $\mathbf{V}^T$ és la matriu transposada de $\mathbf{V}$ i Λ és una matriu diagonal que conté els valors propis $\{\lambda_t\}_{t=1}^d$ de $\mathbb{K}$ de vectors propis $\{v_t\}_{t=1}^d$, amb $v_t = (v_{ti})_{i=1}^d \in \mathbb{R}^d$. Com $\mathbb{K}$ és definida positiva, tots els seus valors propis són positius i podem definir la funció següent:

$$\boldsymbol{\phi} \,:\, \boldsymbol{x}_i \mapsto (\sqrt{\lambda_t} v_{ti})_{t=1}^d \in \mathbb{R}^d, \quad i = 1, ..., d.$$

Ara, sigui $\boldsymbol{\phi}(X) := \mathcal{H}$ un espai de Hilbert d-dimensional dotat del producte escalar euclidià, tenim que

$$\langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle = \sum_{t=1}^{d} \lambda_t v_{ti} v_{tj} = (\mathbf{V} \Lambda \mathbf{V}^T)_{ij} = \mathbb{K}_{ij} = \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

□

En resum, si definim una funció característica de l'espai d'entrada X sobre l'espai de Hilbert de successions ℓ_2, de la forma

$$\begin{aligned} \boldsymbol{\phi} \colon X &\longrightarrow \ell^2 \\ \boldsymbol{x} &\longmapsto (\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), ...) := (\sqrt{\lambda_1}\psi_1(\boldsymbol{x}), \sqrt{\lambda_2}\psi_2(\boldsymbol{x}), ...), \end{aligned}$$

on $\{\psi_i\}_{i \in \mathbb{N}}$ són funcions propies de valors propis $\{\lambda_i\}_{i \in \mathbb{N}}$ de l'operador de Hilbert-Schmidt de $\boldsymbol{K}$ definit per (1.4), aleshores $\boldsymbol{K}$ admet la representació

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z}) \rangle_{\ell^2} = \sum_{i=1}^{\infty} \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{z})$$

i defineix un producte escalar sobre l'espai característic $F = \boldsymbol{\phi}(X) \subseteq \ell^2$, que és un espai de successions reals $(\sqrt{\lambda_i}\psi_i(\boldsymbol{x}))_{i\in\mathbb{N}}$ tals que

$$\sum_{i=1}^{\infty} \lambda_i \psi_i^2 < \infty.$$

Les característiques que genera aquesta funció $\boldsymbol{\phi}$ tenen una propietat especial i és que són funcions ortonormals de $L^2(X)$. Anomenem **nucli de Mercer** a la funció $\boldsymbol{K} \in L^2(X \times X)$ que satisfà les condicions del Teorema de Mercer. Més endavant veurem que aquesta tria concreta de la funció característica resulta en un espai característic dotat de certes propietats que facilitaran el nostre procès d'aprenentage.

1.3.1 Propietats bàsiques i alguns exemples de nuclis

Abans de presentar els nuclis més coneguts i emprats en alguns mètodes d'aprenentatge, com poden ser les màquines de suport vectorial, es proporcionaran les bases que ens permetran demostrar que aquests són exactament nuclis i no només funcions reals definides en $X \times X$. Recordem que fins ara hem vist que un nucli és una funció que defineix un producte escalar en l'espai característic, a més, sabem que una condició suficient i necessària perquè una funció real, contínua i simètrica en $X \times X$ sigui un nucli, és la condició de positivitat que s'exposa a la Observació 1.3.

Lemma 1.1. *Siguin $\boldsymbol{K}_1, \boldsymbol{K}_2$ nuclis en $X \times X$, $X \subseteq \mathbb{R}^n$, $a \in \mathbb{R}^+$ i $f(\cdot)$ una funció real en X. Aleshores les següents funcions són nuclis:*

i) $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) := \boldsymbol{K}_1(\boldsymbol{x}, \boldsymbol{z}) + \boldsymbol{K}_2(\boldsymbol{x}, \boldsymbol{z})$

ii) $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) := a\boldsymbol{K}_1(\boldsymbol{x}, \boldsymbol{z})$

iii) $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) := f(\boldsymbol{x})f(\boldsymbol{z})$

Demostració. Fixem un conjunt de punts de X, $\boldsymbol{x}_1, ..., \boldsymbol{x}_l$, $l \in \mathbb{N}$, i considerem les matrius nucli $\mathbb{K}_1$ i $\mathbb{K}_2$, resultat de restringir $\boldsymbol{K}_1$ i $\boldsymbol{K}_2$ a aquests punts. Considerem qualsevol vector $v \in \mathbb{R}^l$.

i) Tenim que

$$v'(\mathbb{K}_1 + \mathbb{K}_2)v = v'\mathbb{K}_1 v + v'\mathbb{K}_2 v \geqslant 0,$$

per tant, $\mathbb{K}_1 + \mathbb{K}_2$ és una matriu definida positiva i $\boldsymbol{K}_1 + \boldsymbol{K}_2$ és un nucli.

ii) $v'a\mathbb{K}_1 v = av'\mathbb{K}_1 v \geqslant 0$, que verifica que $a\boldsymbol{K}_1$ és un nucli.

iii) Notem que podem reescriure la condició de positivitat de (1.5) com

$$\begin{aligned}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j f(\boldsymbol{x}_i)f(\boldsymbol{x}_j) \\ &= \sum_{i=1}^{l} \alpha_i f(\boldsymbol{x}_i) \sum_{j=1}^{l} \alpha_j f(\boldsymbol{x}_j) \\ &= \left(\sum_{i=1}^{l} \alpha_i f(\boldsymbol{x}_i)\right)^2 \geqslant 0.\end{aligned}$$

□

Ara veiem un resultat recollit de [6] que ens permet identificar kernels que es poden expressar mitjançant una sèrie de Taylor.

Lemma 1.2. *Sigui $r \geqslant 0$ i $g : (-r, r) \longrightarrow \mathbb{R}$ una funció $\mathcal{C}^\infty$ tal que la seva expansió de Taylor és*

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \quad \forall x \in (-r, r).$$

Sigui $X := \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$. Si $a_n > 0$ per tot $n \geqslant 0$, aleshores $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) = f\big(\langle \boldsymbol{x}, \boldsymbol{z} \rangle\big)$ és un nucli en X.

Demostració. S'utilitzarà la fòrmula multinomial que diu que per $n \in \mathbb{N}$ i $z_1, ..., z_n \in \mathbb{C}$, es té

$$(z_1 + ... + z_d)^n = \sum_{\substack{j_1,...,j_d \geqslant 0 \\ j_1+...+j_d=n}} c_{j_1,...,j_d} \prod_{1 \leqslant i \leqslant d} z_i^{j_i}, \tag{1.6}$$

on $c_{j_1,...,j_d} = \binom{n}{j_1,...,j_d} = \frac{n!}{j_1!...j_d!}$.

Primer notem que $\boldsymbol{K}$ està ben definida ja que $|\langle \boldsymbol{x}, \boldsymbol{z} \rangle| = \|\boldsymbol{x}\|_2 \|\boldsymbol{z}\|_2 < r$ per tot $\boldsymbol{x}, \boldsymbol{z} \in X$. Aleshores,

$$\begin{aligned}
\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) &= \sum_{n=0}^{\infty} a_n \left(\sum_{k=1}^{d} x_k z_k \right)^n \\
&= \sum_{n=0}^{\infty} a_n \sum_{\substack{j_1,...,j_d \geqslant 0 \\ j_1+...+j_d=n}} c_{j_1,...,j_d} \prod_{1 \leqslant i \leqslant d} (x_i z_i)^{j_i}, \quad \text{per la fòrmula (1.6)} \\
&= \sum_{j_1,...,j_d \geqslant 0} a_{j_1+...+j_d} c_{j_1,...,j_d} \prod_{1 \leqslant i \leqslant d} x_i^{j_i} \prod_{1 \leqslant i \leqslant d} z_i^{j_i}.
\end{aligned}$$

Si definim $\boldsymbol{\phi} : X \longrightarrow \ell_2$ com

$$\boldsymbol{\phi}(\boldsymbol{x}) = \left(\sqrt{a_{j_1+...+j_d} c_{j_1,...,j_d}} \prod_{1 \leqslant i \leqslant d} x_i^{j_i} \right)_{j_1,...,j_d \geqslant 0},$$

aleshores tenim que $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{z}) \rangle_{\ell_2}$. □

Ara ja estem preparats per definir els nuclis més emprats en els models de màquines de suport vectorial:

Nucli polinòmic

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) := (\langle \boldsymbol{x}, \boldsymbol{z} \rangle + c)^m, \quad m, c \geqslant 0,$$

on m és un enter i c un nombre real. Clarament aquesta funció és un nucli si combinem les propietats *i)* i *ii)* del Lemma 1.1.

Els nuclis polinòmics amb $m = 1$ i $c = 0$ s'anomenen **nuclis lineals**. Notem que en aquest tipus de nuclis la funció característica queda determinada explícitament per $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$, per tant les dades no són assignades a cap espai característic diferent que el mateix espai d'entrada.

Nucli de funció de base radial (RBF) o nucli gaussià

$$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{z}) := exp\Big(-\gamma\|\boldsymbol{x}-\boldsymbol{z}\|_2^2\Big) = exp\Big(\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|_2^2}{2\sigma^2}\Big)$$

on $\gamma := 1/2\sigma^2$ i σ^2 és la variància del nucli. El paràmetre gamma controla la influència dels punts més propers a l'hiperplà definit per la funció solució, valors més alts d'aquest provoca que els punts més propers tinguin més influència. Per veure que aquesta funció és un nucli, l'hem de descomposar de la següent manera:

$$exp\Big(\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|_2^2}{2\sigma^2}\Big) = exp\Big(-\frac{\|\boldsymbol{x}\|_2^2}{2\sigma^2}\Big)exp\Big(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\Big)exp\Big(\frac{\langle\boldsymbol{x},\boldsymbol{z}\rangle}{\sigma^2}\Big).$$

Els primers dos factors formen un nucli per la propietat *iii)* del Lemma 1.1. L'últim factor és un nucli ja que la funció exponencial es pot aproximar per un polinomi de coeficients positius segons la sèrie de Taylor, per tant, podem aplicar el Lemma 1.2.

Nucli sigmoide

$$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{z}) := tanh(\gamma\langle\boldsymbol{x},\boldsymbol{z}\rangle + c), \quad \gamma \in \mathbb{R}, \gamma > 0$$

Aquesta funció és simplement una funció de tangent hiperbòlica aplicada a un producte escalar. Notem que

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

per tant, si expandim la funció exponencial per Taylor tenim que pel Lemma 1.2, $\boldsymbol{K}$ és un nucli.

1.4 Representació de Mercer dels espais de Hilbert de nucli reproductor

Sigui $\boldsymbol{K}$ un nucli de Mercer, considerem la funció característica definida pels valors propis $\{\lambda_i\}_{i\in\mathbb{N}}$ i funcions propies $\{\psi_i\}_{i\in\mathbb{N}}$ de $\boldsymbol{K}$:

$$\boldsymbol{x} \longmapsto \boldsymbol{\phi}(\boldsymbol{x}) = (\sqrt{\lambda_1}\psi_1(\boldsymbol{x}), ..., \sqrt{\lambda_j}\psi_j(\boldsymbol{x}), ...),$$

on $\boldsymbol{\phi}(X) := F$ és un subespai de ℓ_2 dotat amb el producte escalar que defineix $\boldsymbol{K}$ (producte euclidià).

A continuació definim el conjunt

$$\mathscr{H} = \Big\{\sum_{i=1}^{\infty} a_i\sqrt{\lambda_i}\psi_i \ : \ (a_i)_{i\in\mathbb{N}} \in \ell_2\Big\}. \tag{1.7}$$

Notem que aquest és el conjunt de totes les funcions que obtenim aplicant una funció lineal a l'espai característic, és a dir, que $\mathscr{H}$ conté les nostres funcions hipòtesi. No obstant, si F té dimensió infinita $\mathscr{H}$ podria contenir massa funcions. En concret, veurem que l'anterior tria de l'espai característic assegurarà que el nostre conjunt $\mathscr{H}$ únicament contingui les funcions hipòtesi considerades, a més de ser un espai de Hilbert de nucli reproductor.

Definició 1.5. *Sigui $\mathcal{H}$ un espai de Hilbert de funcions reals. Diem que $\mathcal{H}$ és un* ***Espai de Hilbert de Nucli Reproductor*** *(en anglès, Reproducing Kernel Hibert Space, RKHS) si existeix $\boldsymbol{K} : X \times X \longmapsto \mathbb{R}$ tal que:*

1. $\forall \boldsymbol{x} \in X$ *la funció* $\boldsymbol{K}(\cdot, \boldsymbol{z}) \in \mathcal{H}$.
2. $\langle f(\cdot), \boldsymbol{K}(\cdot, \boldsymbol{z})\rangle = f(\boldsymbol{z}) \ \ \forall f \in \mathcal{H}$.

Si $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z})$ satisfà aquestes propietats diem que és un ***nucli reproductor*** *de l'espai $\mathcal{H}$.*

Un cop presentada la definició d'espai de Hilbert de nucli reproductor, l'objectiu d'aquest apartat serà demostrar que per a tot nucli de Mercer $\boldsymbol{K}$ existeix un RKHS tal que $\boldsymbol{K}$ és el seu nucli reproductor. El primer pas es definir un producte escalar en $\mathscr{H}$ per dues funcions $f, g \in \mathscr{H}$:

$$f(\cdot) := \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i}\psi_i(\cdot) \quad , \quad g(\cdot) := \sum_{i=1}^{\infty} \tilde{a}_i \sqrt{\lambda_i}\psi_i(\cdot)$$

$$\langle f, g\rangle_{\mathscr{H}} = \left\langle (a_i)_{i\in\mathbb{N}}, (\tilde{a}_i)_{i\in\mathbb{N}} \right\rangle_{\ell_2} = \sum_{i=1}^{\infty} a_i \tilde{a}_i.$$

Notem que com aquest coincideix amb el producte escalar euclidià en ℓ_2, per tant $\mathcal{H}$ és un espai de Hilbert. Primer veiem que se satisfà la primera condició de la Definició 1.5, tenim que per $\boldsymbol{z} \in X$

$$\boldsymbol{K}(\cdot, \boldsymbol{z}) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\psi_i(\boldsymbol{z})\sqrt{\lambda_i}\psi_i(\cdot),$$

on $\big(\sqrt{\lambda_i}\psi_i(\boldsymbol{z})\big)_{i\in\mathbb{N}} \in \ell_2$ ja que $\|\big(\sqrt{\lambda_i}\psi_i(\boldsymbol{z})\big)\|^2_{\ell_2} = \sum_{i=1}^{\infty} \lambda_i \psi_i^2(\boldsymbol{z}) = \boldsymbol{K}(\boldsymbol{z}, \boldsymbol{z}) < \infty$. Concloem llavors que $\boldsymbol{K}(\cdot, \boldsymbol{z}) \in \mathscr{H}$. Com a conseqüència, les funcions expressades en la forma dual

$$f(\cdot) = \sum_{i=1}^{l} \alpha_i \boldsymbol{K}(\boldsymbol{x}_i, \cdot) \quad \alpha_i \in \mathbb{R},$$

són també a l'espai $\mathscr{H}$. Per provar la segona condició només hem de prendre una funció general $f(\cdot) = \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i}\psi_i(\cdot)$ i $\boldsymbol{K}(\cdot, \boldsymbol{z})$, i fer el producte escalar entre ells. Efectivament

$$\langle f(\cdot), \boldsymbol{K}(\cdot, \boldsymbol{z})\rangle_{\mathscr{H}} = \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i}\psi_i(\boldsymbol{z}) = f(\boldsymbol{z}).$$

Fins ara hem provat que $\mathscr{H}$ és un espai de Hilbert de nucli reproductor, ens resta provar que $\mathscr{H}$ conté única i exactament les nostres funcions hipòtesi. Ho farem veient que el conjunt

$$\mathscr{K} := span\big(\{\boldsymbol{K}(\boldsymbol{x}_i, \cdot)\}_{i=1}^{l}\big) = \Big\{ \sum_{i=1}^{l} \alpha_i \boldsymbol{K}(\boldsymbol{x}_i, \cdot) \ : \ l \in \mathbb{N}, (\boldsymbol{x}_1, ..., , \boldsymbol{x}_l) \in X^l, \alpha_i \in \mathbb{R} \Big\}$$

és dens en $\mathscr{H}$. En primer lloc, sabem per la primera condició de la Definició 1.5 que $\mathscr{K} \subseteq \mathscr{H}$. Ara, sigui $f \in \mathscr{H}$ si projectem aquesta funció en $\mathscr{K}$ obtenim dues components

$f_\parallel$ i $f_\perp$ tal que $f = f_\parallel + f_\perp$. Com el nucli $\boldsymbol{K}$ té la propietat reproductora, aleshores per $\forall \boldsymbol{z} \in X$

$$f(\boldsymbol{z}) = \langle f(\cdot), \boldsymbol{K}(\cdot, \boldsymbol{z})\rangle_{\mathscr{H}} = \langle f_\parallel(\cdot), \boldsymbol{K}(\cdot, \boldsymbol{z})\rangle_{\mathscr{H}} + \langle f_\perp(\cdot), \boldsymbol{K}(\cdot, \boldsymbol{z})\rangle_{\mathscr{H}} = \langle f_\parallel(\cdot), \boldsymbol{K}(\cdot, \boldsymbol{z})\rangle_{\mathscr{H}} = f_\parallel(\boldsymbol{z}).$$

A més, si definim la norma

$$\|f\|_{\mathscr{H}} = \sqrt{\langle f, f\rangle_{\mathscr{H}}},$$

i afegim a $\mathscr{H}$ els límits de les seqüencies de funcions que convergeixen amb aquesta norma obtenim que $\bar{\mathscr{K}} = \mathscr{H}$. Això demostra, com volíem veure, que $\mathscr{H}$ no conté funcions que no puguin ser expressades de forma dual.

Així mateix, el producte escalar $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ no depèn de la representació de les funcions f i g (primària o dual). Siguin $f(\cdot) = \sum_{i=1}^{l} \alpha_i \boldsymbol{K}(\boldsymbol{x}_i, \cdot)$ i $g(\cdot) = \sum_{j=1}^{\tilde{l}} \tilde{\alpha}_j \boldsymbol{K}(\tilde{\boldsymbol{x}}_j, \cdot)$, aleshores

$$\begin{aligned}
\langle f, g\rangle_{\mathscr{H}} &= \left\langle \sum_{i=1}^{l} \alpha_i \boldsymbol{K}(\boldsymbol{x}_i, \cdot), \sum_{j=1}^{\tilde{l}} \tilde{\alpha}_j \boldsymbol{K}(\tilde{\boldsymbol{x}}_j, \cdot) \right\rangle_{\mathscr{H}} \\
&= \sum_{i=1}^{l} \alpha_i \sum_{j=1}^{\tilde{l}} \tilde{\alpha}_j \langle \boldsymbol{K}(\boldsymbol{x}_i, \cdot), \boldsymbol{K}(\tilde{\boldsymbol{x}}_j, \cdot)\rangle_{\mathscr{H}} \\
&= \sum_{i=1}^{l} \alpha_i \sum_{j=1}^{\tilde{l}} \tilde{\alpha}_j \boldsymbol{K}(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_j), \quad \text{per la propietat reproductora de } \boldsymbol{K} \\
&= \sum_{i=1}^{l} \alpha_i g(\boldsymbol{x}_i) = \sum_{j=1}^{\tilde{l}} \tilde{\alpha}_j f(\tilde{\boldsymbol{x}}_j).
\end{aligned}$$

Per últim, és important adonar-se de què assignar l'espai d'entrada, X, a l'espai característic

$$F = \{\boldsymbol{\phi}(\boldsymbol{x}) \,|\, \boldsymbol{x} \in X\} = \left\{(\sqrt{\lambda_1}\psi_1(\boldsymbol{x}), ..., \sqrt{\lambda_j}\psi_j(\boldsymbol{x}), ...) \,|\, \boldsymbol{x} \in X\right\} \subseteq \ell_2,$$

és equivalent a aplicar la funció característica següent

$$\begin{aligned}
\tilde{\boldsymbol{\phi}} : X &\longrightarrow \mathscr{H} \\
\boldsymbol{x} &\longmapsto \boldsymbol{K}(\cdot, \boldsymbol{x})
\end{aligned}$$

ja que el producte escalar d'elements de F definit per $\langle \cdot, \cdot \rangle_{\ell_2}$ equival al producte escalar d'elements de $\tilde{\boldsymbol{\phi}}(X)$ definit per $\langle \cdot, \cdot \rangle_{\mathscr{H}}$, i ambdós estan definits pel nucli $\boldsymbol{K}$, que és a més un nucli reproductor del RKHS $\mathscr{H}$:

$$\begin{aligned}
\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) &= \langle \boldsymbol{K}(\cdot, \boldsymbol{x}), \boldsymbol{K}(\cdot, \boldsymbol{z})\rangle_{\mathscr{H}} \\
&= \left\langle \sum_{i=1}^{\infty} \sqrt{\lambda_i}\psi_i(\boldsymbol{x})\sqrt{\lambda_i}\psi_i \,, \sum_{i=1}^{\infty} \sqrt{\lambda_i}\psi_i(\boldsymbol{z})\sqrt{\lambda_i}\psi_i \right\rangle_{\mathscr{H}} \\
&= \left\langle \left(\sqrt{\lambda_i}\psi_i(\boldsymbol{x})\right)_{i\in\mathbb{N}} , \left(\sqrt{\lambda_i}\psi_i(\boldsymbol{z})\right)_{i\in\mathbb{N}} \right\rangle_{\ell_2} = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{z})\rangle_{\ell_2}.
\end{aligned} \tag{1.8}$$

Per tant són equivalents $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ i $\langle \cdot, \cdot \rangle_{\ell_2} := \langle \cdot, \cdot \rangle_{\boldsymbol{K}}$, i és per això que per abús de llenguatge a l'espai característic se'l considera un espai de Hilbert de nucli reproductor.

Finalment, relacionarem els conceptes més importants explicats en aquest capítol per tal d'establir una idea clara de l'avantatge que té l'ús dels nuclis en un problema d'aprenentatge. L'assignació de les dades d'entrada a noves dades a l'espai característic es realitza mitjançant els nuclis, que no són res més que un producte escalar intern sobre un RKHS, segons (1.8). No obstant, l'expressió explícita d'aquestes noves dades o característiques no es coneix, el que es coneix és la similitud relativa (producte escalar) entre aquestes, és a dir, el nucli. Aquest fenomen és el que s'anomena com a truc del nucli, introduit a l'apartat 1.1.

En aprenentatge automàtic i ciència de dades, la "**nuclització**" vol dir fer una petita modificació en la formulació de l'algoritme per tal de què les dades del RKHS, o de l'espai característic, siguin utilitzades com a dades d'entrada en comptes de les dades originals. Per exemple, una tècnica de nuclització consisteix en reescriure les fòrmules de l'algoritme o el problema d'optimització de manera que les dades sempre apareguin com el producte escalar de dues instancies de dades i no com una instància de dades individual. En altres paraules, la formulació de l'algoritme només pot contenir $\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{K}}$, $\langle \boldsymbol{x}, X \rangle_{\boldsymbol{K}}$, $\langle X, X \rangle_{\boldsymbol{K}}$ i no només $\boldsymbol{x}, X$. D'aquesta manera el truc del nucli substitueix els $\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{K}}$ per $\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}) \rangle_{\boldsymbol{K}}$:

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{K}} \longmapsto \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}) \rangle_{\boldsymbol{K}} = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}).$$

.

2 Màquines de suport vectorial (SVM)

S'iniciarà el capítol amb una breu explicació sobre com les màquines de suport vectorial transformen el problema d'optimització (1) en un problema computacionalment factible. El primer pas és substituir la funció de pèrdua per classificació binària L_{class} per una altra que el converteixi en un problema d'optimització convex. Normalment es tria la funció de pèrdua de frontissa (en anglès, *hinge loss*), definida per

$$L_h(y,t) := max\{0, 1-yt\}, \quad y \in \{-1,+1\},\, t \in \mathbb{R},$$

que transforma el problema (1) en un problema d'optimització convex, ja que $\mathcal{R}_{L_h,S}(f)$ és convex en f. D'aquesta manera existeix una única funció f^* que minimitza el risc empíric

$$\inf_{f\in\mathcal{F}} \mathcal{R}_{L_h,P_S}(f). \tag{2.1}$$

El segon pas és considerar un conjunt de funcions hipòtesi específic que sigui un espai de Hilbert de nucli reproductor, $\mathcal{F} = \mathscr{H}$. Recordem que la propietat més important d'aquest tipus d'espais és que estan dotats d'una funció nucli $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{z})$ que es pot utilitzar per descriure totes les funcions contingudes en aquest. Si fixem aquest espai de Hilbert $\mathscr{H}$ i denotem per $\|\cdot\|_{\mathscr{H}}$ la seva norma, per un valor fix $\lambda > 0$, el problema d'optimització definit a (2.1) es transforma en

$$\inf_{f\in\mathscr{H}} \lambda\|f\|^2_{\mathscr{H}} + \mathcal{R}_{L_h,P_S}(f), \tag{2.2}$$

on el terme $\lambda\|f\|^2_{\mathscr{H}}$ s'anomena **terme de regularització** i s'utilitza per penalitzar les funcions f que tinguin una norma gran en $\mathscr{H}$. Això és degut a què funcions més complexes que modelen massa bé els valors de sortida del conjunt d'entrenament, és a dir, sobreajusten el model, tendeixen a tenir norma $\|\cdot\|_{\mathscr{H}}$ molt gran.

Arribats a aquest punt ja sabem que la funció solució de (2.2) serà de la forma

$$f^*_\lambda = \sum_{i=1}^{d} \alpha_i \boldsymbol{K}(\boldsymbol{x}_i,\cdot),\ \alpha_1,...,\alpha_d \in \mathbb{R}, \tag{2.3}$$

on $\boldsymbol{K}$ és el nucli reproductor de $\mathscr{H}$. En altres paraules, el minimitzador és una mitjana ponderada de com a molt d funcions $\boldsymbol{K}(\boldsymbol{x}_i,\cdot)$. A la pràctica és important trobar estratègies que sense conèixer P ens permetin trobar valors òptims de λ, a més de valors òptims pels hiperparàmetres del nucli que es triï.

2.1 Classificadors de suport vectorial

Sigui $X \subseteq \mathbb{R}^n$ finit, $Y = \{-1,1\}$ i S un conjunt d'entrenament

$$S = \big((\boldsymbol{x}_1,\boldsymbol{y}_1),...,(\boldsymbol{x}_d,\boldsymbol{y}_d)\big).$$

Suposarem que existeix un hiperplà que classifica correctament els punts de S, en aquest cas diem que el conjunt S és **linealment separable**. Notem que a partir d'aquest capítol suposarem que S té un nombre finit d'elements ja que és el cas que s'ajusta a un problema

real. Recordem que el problema de classificació binària es resol trobant una funció real $f : X \longrightarrow \mathbb{R}$ que per $\boldsymbol{x}_i = (x_1, ..., x_n) \in X$, $i \in \{1, ..., d\}$,

$$\begin{aligned} f(\boldsymbol{x}_i) &= \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \\ &= \sum_{i=1}^{d} w_i x_i + b. \end{aligned} \tag{2.4}$$

Ara, $f_D := sgn(f(\cdot))$ és l'anomenada **funció de decisió**, qui assigna a cada punt $\boldsymbol{x} \in X$ a la classe positiva (1) o, altrament, a la negativa (-1).

$$f_D(\boldsymbol{x}) := sgn(f(\boldsymbol{x})) = \begin{cases} 1, & \text{si } f(\boldsymbol{x}) \geqslant 0. \\ -1, & \text{si } f(\boldsymbol{x}) < 0. \end{cases} \tag{2.5}$$

La funció (2.4) defineix un hiperplà $(\boldsymbol{w}, b) := \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ que separa l'espai X en dues regions $\{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S \,|\, \boldsymbol{y}_i = +1\}$ i $\{(\boldsymbol{x}_i, \boldsymbol{y}_i) \,|\, \boldsymbol{y}_i = -1\}$. Ens referirem a aquest hiperplà com el **límit de decisió**, a $\boldsymbol{w}$ com el **vector de pesos** i a b com a **biaix**.

Tot i que geomètricament aquest algoritme sembla convincent, en realitat presenta dos inconvenients o defectes principals:

1. Una funció de decisió lineal potser que no s'adapti correctament a la tasca de classificació corresponent en casos en què el conjunt d'entrenament S no sigui linealment separable i, aleshores, no existeixin els paràmetres $\boldsymbol{w}$ i b descrits a dalt. Un exemple que ja hem vist és el de l'apartat 1.1.

2. En cas de presència de soroll a les dades, és possible que s'hagin de classificar erròneament alguns punts de l'espai d'entrada per tal d'evitar un sobreajustament del model. Amb soroll ens referim a valors atípics, inconsistencies en les dades o informació no rellevant. Generalment, si el nombre d'elements de X, d, és molt més gran que la dimensió de les mostres, n, el sobreajustament pot arribar a ser un problema greu.

L'objectiu de les màquines de suport vectorial per a la classificació és, per tant, trobar un hiperplà separador dins del conjunt d'hipòtesi

$$\mathcal{F} = \{f : X \subseteq \mathbb{R}^n \longrightarrow \mathbb{R} \,|\, f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b,\ (\boldsymbol{w}, b) \in \mathbb{R}^n \times \mathbb{R}\}$$

que doni solució a aquests problemes i que sigui computacionalment eficient. Per resoldre la primera qüestió ja hem vist al capítol anterior que es poden assignar les dades d'entrada $(\boldsymbol{x}_1, ..., \boldsymbol{x}_d)$ a un espai de Hilbert $\mathcal{H}$ mitjançant una funció no lineal $\boldsymbol{\phi} : X \longrightarrow \mathcal{H}$. Aleshores l'algoritme descrit a dalt s'aplica al conjunt de dades

$$S' = \big((\boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{y}_1), ..., (\boldsymbol{\phi}(\boldsymbol{x}_d), \boldsymbol{y}_d)\big),$$

és a dir, es buscarà un hiperplà separador en l'espai característic $\boldsymbol{\phi}(X)$ en comptes de en X, i el nostre conjunt de funcions hipòtesi serà per tant

$$\mathcal{F} = \{f : \boldsymbol{\phi}(X) \subseteq \mathcal{H} \longrightarrow \mathbb{R} \,|\, f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle + b,\ (\boldsymbol{w}, b) \in \mathcal{H} \times \mathbb{R}\},$$

No obstant, durant els següents apartats es considerarà que es treballa sobre el conjunt d'entrenament S, per simplificar la notació. Únicament cal tenir en compte que en cas de

voler treballar sobre un espai característic s'han de substituir les $\boldsymbol{x}_i$ per $\boldsymbol{\phi}(\boldsymbol{x}_i)$, i buscar un vector de pesos $\boldsymbol{w} \in \mathcal{H}$.

A continuació, definim un seguit de mesures a les quals es farà referència als apartats d'aquest capítol i que jugaran un rol molt important als problemes d'optimització que es presentaran.

Definició 2.1. *Definim el* ***marge (funcional) d'un punt*** *d'entrada* $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ *respecte a un hiperplà* $(\boldsymbol{w}, b)$ *com la quantitat*

$$\gamma_i = \boldsymbol{y}_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b).$$

Notem que $\gamma_i \geqslant 0$ *implica una correcte classficació de* $(\boldsymbol{x}_i, \boldsymbol{y}_i)$*. Anomenem* ***distribució (funcional) dels marges d'un hiperplà*** $(\boldsymbol{w}, b)$ *respecte a* S *com la distribució dels marges (funcionals) dels punts de* S*. El marge mínim d'aquesta distribució és el que sovint s'anomena marge (funcional) de* $(\boldsymbol{w}, b)$ *respecte a* S *i es denota* γ*. Ara, els marges funcionals dels punts de* S *respecte l'hiperplà* $\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \frac{b}{\|\boldsymbol{w}\|}\right)$ *s'anomenen* ***marges geomètrics*** *i mesuren la distància euclidiana dels punts* $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ *a l'hiperplà.*

Definició 2.2. *El* ***marge d'un conjunt d'entrenament*** S *és el màxim dels marges geomètrics de tots els hiperplans separadors de* S*. L'hiperplà que assoleix aquest marge màxim és l'anomenat* ***hiperplà de marge màxim****.*

Per donar amb un hiperplà separador adient dins del conjunt d'hipòtesi caldrà optimitzar un seguit de mesures dels seus marges, per exemple un pot optimitzar el marge geomètric màxim, la distribució dels marges, el nombre de vectors de suport (veurem més endavant el que són), etc. Cada tasca d'optimització motiva un algoritme diferent. En el nostre cas ens centrarem en aquells algoritmes que minimitzen la norma del vector de pesos.

L'organització d'aquest capítol és la següent. Es tractaran alguns aspectes de la teoria d'optimització de Lagrange per tal de desenvolupar eines per convertir un problema d'optimització primari en el seu corresponent dual, fet imprescindible per l'ús dels nuclis. Posteriorment, s'exposaran els dos classificadors més bàsics de SVM i s'explicarà com obtenir els duals dels problemes d'optimització que presenten. Veurem al final que els problemes que proposen aquests dos classificadors són equivalents a trobar una solució de (2.2).

Generalització cas multiclassificació

El problema de classificació binària plantejat abans es pot resoldre també si definim un vector de pesos $\boldsymbol{w}_1, \boldsymbol{w}_{-1}$ i un biaix b_1, b_{-1} per cada classe. Un punt $\boldsymbol{x} \in X$ serà assignat a la classe positiva 1 si $\langle \boldsymbol{w}_1, \boldsymbol{x} \rangle + b_1 \geqslant \langle \boldsymbol{w}_{-1}, \boldsymbol{x} \rangle + b_{-1}$, i a la classe negativa -1 altrament. Això és equivalent a (2.5) emprant $\boldsymbol{w} = \boldsymbol{w}_1 - \boldsymbol{w}_{-1}$ i $b = b_1 - b_{-1}$.

Pel cas de multiclassificació el conjunt de sortida és $Y = \{1, 2, ..., m\}$ i, per cada classe, s'assigna un vector de pesos i un biaix $(\boldsymbol{w}_i, b_i)$, $i \in \{1, 2, ..., m\}$. La funció decisió és llavors

$$c(x) = arg \max_{1 \leqslant i \leqslant m} \big(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle + b_i\big).$$

Geomètricament és equivalent a associar un hiperplà separador a cada classe i assignar a cada punt $\boldsymbol{x} \in X$ la classe amb l'hiperplà més llunyà a aquest. És per això que als següents apartats ens centrarem en el cas en què només tenim dues classes ja que un problema de multiclassificació es resol simplement entrenant un classificador binàri per cada classe.

2.1.1 Dualitat: mètode general de Lagrange

La teoria de l'optimització es remunta al treball de Fermat, qui va formular el resultat sobre l'estacionarietat per a problemes sense restriccions al segle XVII. L'extensió al cas amb restriccions va ser feta per Lagrange el 1788, en el cas de restriccions d'igualtat. No va ser fins el 1951 que la teoria es va generalitzar al cas de restriccions de desigualtat per Harold W. Kuhn i Albert W. Tucker. En aquest apartat s'exposarà el que és un problema d'optimització primari i es donaran eines per obtenir el seu equivalent en forma dual, una reformulació del problema en funció de les anomenades funcions generals de Lagrange. Posteriorment es donaran condicions necessàries i suficients perquè un punt sigui una solució òptima del problema d'optimització primari en termes de les funcions generals de Lagrange.

Definició 2.3 (Problema d'optimització primari)**.** *Donades funcions f, g_i, $i = 1, ..., k$, i h_i, $i = 1, ..., m$, definides sobre un domini $\Omega \subseteq \mathbb{R}^n$,*

$$\begin{aligned} &\textit{minimitzar} \;\; f(\boldsymbol{w}), && \boldsymbol{w} \in \Omega, \\ &\textit{tal que} \;\; g_i(\boldsymbol{w}) \leqslant 0, && i = 1, ..., k, \\ &\qquad\qquad h_i(\boldsymbol{w}) = 0, && i = 1, ..., m, \end{aligned}$$

*és un problema d'optimització primari, on $f(\cdot)$ s'anomena **funció objectiu**, i les altres relacions s'anomenen respectivament **restriccions de desigualtat** i **d'igualtat**.*

La regió del domini Ω on la funció objectiu està definida i on totes les restriccions se satisfan s'anomena **regió de possibilitat**

$$R := \{\boldsymbol{w} \in \Omega \,:\, \boldsymbol{g}(\boldsymbol{w}) \leqslant 0,\, \boldsymbol{h}(\boldsymbol{w}) = 0\}$$

Una solució del problema d'optimització primari és un punt $\boldsymbol{w}^* \in R$ tal que no existeix un altre punt $\boldsymbol{w} \in R$ pel qual $f(\boldsymbol{w}) \leqslant f(\boldsymbol{w}^*)$. Un punt que satisfà aquesta propietat també es denomina mínim global.

Una restricció de desigualtat $g_i(\boldsymbol{w}) \leqslant 0$ es diu que és **activa** si la solució $\boldsymbol{w^*}$ satisfà $g_i(\boldsymbol{w^*}) = 0$, altrament es diu que és **inactiva**.

Definició 2.4. *Donat un problema d'optimització primari amb domini $\Omega \subseteq \mathbb{R}^n$, definim la **funció general de Lagrange** com*

$$\begin{aligned} L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &:= f(\boldsymbol{w}) + \sum_{i=1}^{k} \alpha_i g_i(\boldsymbol{w}) + \sum_{i=1}^{m} \beta_i h_i(\boldsymbol{w}) \\ &:= f(\boldsymbol{w}) + \boldsymbol{\alpha'} \boldsymbol{g}(\boldsymbol{w}) + \boldsymbol{\beta'} \boldsymbol{h}(\boldsymbol{w}). \end{aligned}$$

*Anomenem als punts α_i, β_i **multiplicadors de Lagrange**.*

Definició 2.5 (Problema d'optimització dual de Lagrange)**.** *El problema d'optimització dual de Lagrange del problema primari de la Definició 2.3 és el següent:*

$$\begin{aligned} \textit{maximitzar} \quad & \theta(\boldsymbol{\alpha}, \boldsymbol{\beta}), \\ \textit{tal que} \quad & \boldsymbol{\alpha} \geqslant 0, \end{aligned}$$

on $\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \inf_{\boldsymbol{w} \in \Omega} L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

Per tal de transformar el problema primari en un problema dual simplement cal igualar a zero les derivades de la funció de Lagrange respecte les variables primàries i substituir les relacions obtingudes a la funció de Lagrange. D'aquesta manera s'elimina la dependència sobre les variables primàries a més de simplificar-se les restriccions d'igualtat i desigualtat. Això correspon a calcular explícitament $\theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Aquest procediment és elemental en les màquines de suport vectorial ja que com hem vist tenir una representació dual de la problemàtica i, per tant, de les funcions hipòtesi, és imprescindible per poder treballar eficientment en espais de dimensió molt alta.

Ara, donarem un seguit de resultats que ens ajudaran a caracteritzar les solucions dels dos tipus de problema i a relacionar-les entre elles.

Teorema 2.1. *Sigui* $\boldsymbol{w} \in \Omega$ *una possible solució del problema d'optimització primari i* $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *una possible solució del corresponent problema dual. Aleshores* $f(\boldsymbol{w}) \geqslant \theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Demostració. Sigui $\boldsymbol{w} \in \Omega$ una possible solució

$$\begin{aligned} \theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \inf_{u \in \Omega} L(u, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\leqslant L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= f(\boldsymbol{w}) + \boldsymbol{\alpha}' \boldsymbol{g}(\boldsymbol{w}) + \boldsymbol{\beta}' \boldsymbol{h}(\boldsymbol{w}) \leqslant f(\boldsymbol{w}), \end{aligned}$$

ja que $\boldsymbol{g}(\boldsymbol{w}) \leqslant 0$ i $\boldsymbol{h}(\boldsymbol{w}) = 0$ per ser $\boldsymbol{w}$ una possible solució del problema primari, i $\boldsymbol{\alpha} \geqslant 0$ per ser $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ una possible solució del dual. □

Si $\boldsymbol{w}^*$ i $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ resolen els problemes primari i dual respectivament, aleshores a la diferència $f(\boldsymbol{w}^*) - \theta(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ se l'anomena **bretxa dual**. Si aquesta quantitat és exactament zero, aleshores es diu que $\boldsymbol{w}^*$ i $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ són solucions **òptimes**. Per últim, citarem el teorema de Kuhn-Tucker el qual presenta condicions necessàries i suficients perquè les solucions del problema primari i dual d'optimització siguin òptimes. No es proporcionarà la demostració d'aquest, que es pot trobar a [8], en canvi, es presentaran algunes observacions sobre com interpretar les condicions del teorema.

Teorema 2.2 (Kuhn-Tucker)**.** *Donat un problema d'optimització com el de la Definició 2.3 amb domini convex* $\Omega \subseteq \mathbb{R}^n$*,* $f \in \mathcal{C}^1$ *convexa i* g_i, h_i *funcions afins. Per un punt* $\boldsymbol{w}^* \in \mathbb{R}^n$ *si existeixen* $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ *tals que*

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \boldsymbol{w}} &= 0, \\ \frac{\partial L(\boldsymbol{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}} &= 0, \\ \alpha_i^* g_i(\boldsymbol{w}^*) &= 0,\ i = 1, ..., k, \\ g_i(\boldsymbol{w}^*) &\leqslant 0,\ i = 1, ..., k, \\ \alpha_i^* &\geqslant 0,\ i = 1, ..., k, \end{aligned}$$

aleshores $\boldsymbol{w}^*$ és una solució òptima del problema d'optimització primari.

La tercera condició és l'anomenada condició complementària de Karush-Kuhn-Tucker. Aquesta implica que per restriccions actives aleshores $\alpha_i^* > 0$ i $\boldsymbol{w}^*$ pertany a la frontera de la regió de possibilitat R, mentre que per restriccions inactives $\alpha_i^* = 0$ i la solució òptima pertany a la l'interior de R.

2.1.2 Classificador de marge màxim

El model més simple de classificadors de suport vectorial que representa un punt de partida per la construcció d'algoritmes de SVM més sotisficats, va ser introduït per Vapnik i Chervonenkis el 1963 i és el primer algoritme que es va proposar com a solució del problema de classificació binària.

Aquest classificador tracta de trobar l'hiperplà de marge màxim de la Definició 2.2 suposant que les dades d'entrenament són linealment separables. Aquesta suposició no és realista per tant el seu ús no és massa freqüent davant de problemes reals. Tot i així, és la base de molts altres models de SVM. Comencem veient que buscar aquest hiperplà equival a minimitzar la norma de $\boldsymbol{w} \in \mathbb{R}^n$ sota un conjunt de restriccions específic.

Proposició 2.1. *Donat un conjunt d'entrenament linealment separable*

$$S = \big((\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_d, \boldsymbol{y}_d)\big)$$

l'hiperplà ($\boldsymbol{w}$,b) que resol el problema d'optimització

$$\begin{aligned} &\textit{minimitzar}_{\boldsymbol{w},b} \quad \|\boldsymbol{w}\|_2^2 \\ &\textit{tal que} \quad \boldsymbol{y}_i\big(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\big) \geqslant 1, \quad i = 1, ..., d \end{aligned} \tag{2.6}$$

és exactament l'hiperplà de marge màxim amb marge geomètric $\gamma = 1/\|\boldsymbol{w}\|_2$.

Demostració. Sigui ($\boldsymbol{w}$,b) un hiperplà separador del conjunt S, és a dir, un hiperplà que satisfà $\boldsymbol{y}_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geqslant 1 \ \forall i \in \{1, ..., d\}$. Sigui $\boldsymbol{x}^+$ el punt de la classe positiva que té marge funcional menor respecte a $(\boldsymbol{w}, b)$, i $\boldsymbol{x}^-$ de manera equivalent per la classe negativa. Els marges geomètrics d'aquests punts són els següents

$$\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \boldsymbol{x}^+ \right\rangle + b, \quad \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \boldsymbol{x}^- \right\rangle + b.$$

Com que la funció associada a l'hiperplà $(\boldsymbol{w}, b)$ no varia encara que reescalem l'hiperplà com $(\lambda\boldsymbol{w}, \lambda b)$ amb $\lambda \in \mathbb{R}^+$, aleshores podem optimitzar el marge geomètric de l'hiperplà igualment fixant que el marge funcional d'aquest sigui 1, notem que això implica que

$$\begin{aligned} \langle \boldsymbol{w}, \boldsymbol{x}^+\rangle + b &= 1 \\ \langle \boldsymbol{w}, \boldsymbol{x}^-\rangle + b &= -1. \end{aligned}$$

Podem calcular el marge geomètric de $(\boldsymbol{w}, b)$ com:

$$\begin{aligned} \gamma &= \frac{1}{2}\left(\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \boldsymbol{x}^+ \right\rangle + b - \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \boldsymbol{x}^- \right\rangle - b\right) \\ &= \frac{1}{2\|\boldsymbol{w}\|_2}\big(\langle \boldsymbol{w}, \boldsymbol{x}^+\rangle - \langle \boldsymbol{w}, \boldsymbol{x}^-\rangle\big) \\ &= \frac{1}{\|\boldsymbol{w}\|_2}. \end{aligned}$$

És clar que maximitzar γ és equivalent a minimitzar $\|\boldsymbol{w}\|_2$. La restricció de desigualtat $\boldsymbol{y}_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geqslant 1$, $i \in \{1, ..., d\}$, és necessària per assegurar que cada punt de l'entrenament estigui ben classificat i estigui fora del marge. □

Ara, veurem com transformar aquest problema d'optimització en el seu corresponent problema dual, mitjançant el procediment explicat a l'apartat anterior. La funció general de Lagrange és

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle - \sum_{i=1}^{d} \alpha_i \big[\boldsymbol{y}_i\big(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\big) - 1\big], \tag{2.7}$$

on $\alpha_i \geqslant 0$ són els multiplicadors de Lagrange. La presència del terme $1/2$ multiplicant la norma al quadrat de $\boldsymbol{w}$ és una elecció de conveniència matemàtica que no afecta la solució òptima del problema d'optimització. A continuació, calculem les derivades parcials respecte a les variables primàries $\boldsymbol{w}$ i b:

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_{i=1}^{d} \boldsymbol{y}_i \alpha_i \boldsymbol{x}_i = 0 \iff \boldsymbol{w} = \sum_{i=1}^{d} \boldsymbol{y}_i \alpha_i \boldsymbol{x}_i, \\ \frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial b} &= \sum_{i=1}^{d} \boldsymbol{y}_i \alpha_i = 0, \end{aligned} \tag{2.8}$$

i resubstituint a (2.7)

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\sum_{i,j=1}^{d} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle - \sum_{i,j=1}^{d} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle + \sum_{i=1}^{d} \alpha_i \\ &= \sum_{i=1}^{d} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{d} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle. \end{aligned}$$

Podem reformular llavors la Proposició 2.1 com

Proposició 2.2. *Donat un conjunt d'entrenament linealment separable*

$$S = \big((\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_d, \boldsymbol{y}_d)\big),$$

i suposem que els paràmetres $\boldsymbol{\alpha}^*$ resolen el següent problema d'optimització

$$\begin{aligned} \textit{maximitzar } & W(\boldsymbol{\alpha}) = \sum_{i=1}^{d} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{d} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \\ \textit{tal que } & \sum_{i=1}^{d} \boldsymbol{y}_i \alpha_i = 0, \\ & \alpha_i \geqslant 0,\ i = 1, ..., d. \end{aligned}$$

Aleshores el vector de pesos $\boldsymbol{w}^* = \sum_{i=1}^{d} \boldsymbol{y}_i \alpha_i^* \boldsymbol{x}_i$ assoleix l'hiperplà de marge màxim amb marge geomètric

$$\gamma = 1/\|\boldsymbol{w}^*\|_2.$$

Observació 2.1. *El valor de b no apareix al problema dual, per tant, b^* es troba fent ús de la restricció primària*

$$b^* = -\frac{\max_{\boldsymbol{y}_i=-1}\big(\langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle\big) + \min_{\boldsymbol{y}_i=1}\big(\langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle\big)}{2}.$$

Com volem que la solució $\boldsymbol{w}^*$ sigui òptima hem d'imposar la condició complementària de Karush-Kuhn-Tucker (ja que les altres ja es compleixen) aleshores $\boldsymbol{\alpha}^*$, $(\boldsymbol{w}^*, b^*)$ han de satisfer

$$\alpha_i^*\big[\boldsymbol{y}_i\big(\langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle + b^*\big) - 1\big] = 0, \quad i = 1, ..., d. \tag{2.9}$$

Per tant, només quan el marge funcional és 1 els corresponents α_i^* seran diferents de zero, la resta s'anul·laran. Notem que aquests punts $\boldsymbol{x}_i$ d'entrada pels quals el marge funcional és 1, són aquells posicionats més a prop de l'hiperplà separador. S'anomenen **vectors de suport** ja que en la expressió del vector de pesos de (2.8) són els únics punts involucrats. Denotem sv el conjunt d'índexs dels vectors de suport.

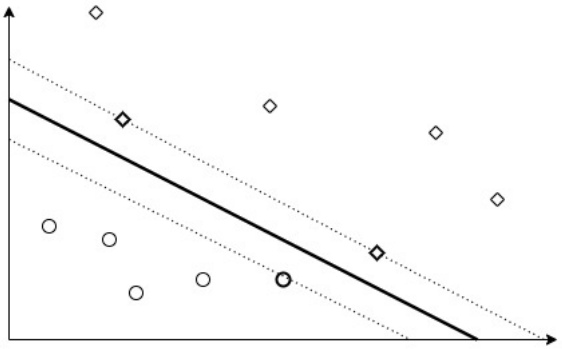

Figura 3 : Un hiperplà de marge màxim amb els seus vectors de suport destacats

Podem escriure l'hiperplà òptim de marge màxim en representació dual de la següent forma

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\alpha}^*, b^*) &= \langle \boldsymbol{w}^*, \boldsymbol{x}\rangle + b^* \\ &= \sum_{i=1}^{d} \boldsymbol{y}_i \alpha_i^* \langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + b^* \\ &= \sum_{i\in sv} \boldsymbol{y}_i \alpha_i^* \langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + b^* \end{aligned}$$

Els multiplicadors de Lagrange associats a cada punt es converteixen en variables duals que quantifiquen com d'important és un punt del conjunt d'entrenament envers la solució final. Els punts que no siguin vectors de suport no tenen cap tipus d'influència.

Una altra conseqüència de la condició complementària de Karush-Kuhn-Tucker és que per $j \in sv$

$$\boldsymbol{y}_j f(\boldsymbol{x}_j, \boldsymbol{\alpha}^*, b^*) = \boldsymbol{y}_j \left(\sum_{i\in sv} \boldsymbol{y}_i \alpha_i^* \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle + b^* \right) = 1,$$

ja que quan $j \in sv$, $\alpha_j^* \neq 0$ en (2.9) que implica que

$$\boldsymbol{y}_j\big(\langle \boldsymbol{w}^*, \boldsymbol{x}_j\rangle + b^*\big) - 1 = \boldsymbol{y}_j f(\boldsymbol{x}_j, \boldsymbol{w}^*, b^*) - 1 = \boldsymbol{y}_j f(\boldsymbol{x}_j, \boldsymbol{\alpha}^*, b^*) - 1 = 0,$$

i per tant,

$$\begin{aligned}\langle \boldsymbol{w}^*, \boldsymbol{w}^* \rangle &= \sum_{i,j=1}^{d} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i^* \alpha_j^* \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ &= \sum_{j \in sv} \alpha_j^* \boldsymbol{y}_j \sum_{i \in sv} \boldsymbol{y}_i \alpha_i^* \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ &= \sum_{j \in sv} \alpha_j^* (1 - \boldsymbol{y}_j b^*) \\ &= \sum_{i \in sv} \alpha_i^*.\end{aligned}$$

Si la solució $\boldsymbol{w}^*$ del problema primari de (2.6) és òptima, hem trobat una altra forma de calcular el marge geomètric de l'hiperplà òptim, en funció dels multiplicadors de Lagrange α_i^* dels vectors de suport amb $i \in sv$:

$$\gamma = 1/\|\boldsymbol{w}^*\|_2 = \left(\sum_{i \in sv} \alpha_i^* \right)^{-1/2}.$$

Tant el problema d'optimització dual com la funció objectiu f es poden expressar en funció de productes escalars dels punts d'entrenament. Això fa possible el poder trobar un hiperplà òptim a l'espai característic mitjançant l'ús de nuclis com mostrem a la següent proposició.

Proposició 2.3. *Donat un conjunt d'entrenament*

$$S = \big((\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_d, \boldsymbol{y}_d) \big),$$

linealment separable a l'espai característic definit implícitament pel nucli $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z})$, i suposem que els paràmetres $\boldsymbol{\alpha}^*$ i b^* resolen el següent problema d'optimització:

$$\begin{aligned} &\textit{maximitzar } W(\boldsymbol{\alpha}) = \sum_{i=1}^{d} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{d} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i \alpha_j \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j), \\ &\textit{tal que } \sum_{i=1}^{d} \boldsymbol{y}_i \alpha_i = 0, \\ &\qquad \alpha_i \geqslant 0,\ i = 1, ..., d. \end{aligned} \tag{2.10}$$

Aleshores la funció decisió $f_D(\boldsymbol{x}) = sgn(f(\boldsymbol{x}))$, on $f(\boldsymbol{x}) = \sum_{i=1}^{d} \boldsymbol{y}_i \alpha_i^* \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b^*$, és equivalent a l'hiperplà òptim de marge màxim a l'espai característic, i aquest hiperplà té marge geomètric

$$\gamma = \left(\sum_{i \in sv} \alpha_i^* \right)^{-1/2}.$$

Notem que aquí l'espai característic definit implícitament pel nucli $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z})$ és $F = \boldsymbol{\phi}(X)$, on $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{K}(\cdot, \boldsymbol{x})$. A més, la condició de què la matriu $\mathbb{K} = \big(\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \big)_{i,j=1}^{d}$ sigui definida positiva per tots els punts de l'entrenament implica que la matriu $\big(\boldsymbol{y}_i \boldsymbol{y}_j \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \big)_{i,j=1}^{d}$

és també definida positiva, i de forma contrària, $\left(-\boldsymbol{y}_i\boldsymbol{y}_j\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)_{i,j=1}^{d}$ és definida negativa. Veiem que aquesta última matriu coincideix exactament amb la hessiana de W, $(H_W)_{i,j=1}^{d} = \frac{\partial^2 W}{\partial\alpha_i\alpha_j}$.

$$\begin{aligned}
\frac{\partial W(\boldsymbol{\alpha})}{\partial\alpha_k} &= 1 - \sum_{j=1,j\neq k}^{d} \boldsymbol{y}_k\boldsymbol{y}_j\alpha_j\boldsymbol{K}(\boldsymbol{x}_k,\boldsymbol{x}_j) - \boldsymbol{y}_k\boldsymbol{y}_k\alpha_k\boldsymbol{K}(\boldsymbol{x}_k,\boldsymbol{x}_k)\\
\frac{\partial^2 W(\boldsymbol{\alpha})}{\partial^2\alpha_k} &= -\boldsymbol{y}_k\boldsymbol{y}_k\boldsymbol{K}(\boldsymbol{x}_k,\boldsymbol{x}_k)\\
\frac{\partial^2 W(\boldsymbol{\alpha})}{\partial\alpha_k\alpha_h} &= -\boldsymbol{y}_k\boldsymbol{y}_h\boldsymbol{K}(\boldsymbol{x}_k,\boldsymbol{x}_h),\ h\neq k.
\end{aligned}$$

Com la hessiana és definida negativa aleshores W és còncava i, com a conseqüència, un màxim local $\boldsymbol{w}^*$ és en concret un màxim global ja que per $\boldsymbol{u}\in\mathbb{R}^d$, $\boldsymbol{u}\neq\boldsymbol{w}^*$, per la definició de màxim local existeix θ suficientment proper a 1 tal que

$$\begin{aligned}
f(\boldsymbol{w}^*) &\geqslant f(\theta\boldsymbol{w}^* + (1-\theta)\boldsymbol{u})\\
&\geqslant \theta f(\boldsymbol{w}^*) + (1-\theta)f(\boldsymbol{u}) \Longleftrightarrow f(\boldsymbol{w}^*)\geqslant f(\boldsymbol{u}).
\end{aligned}$$

Concloem llavors que la propietat necessària perquè una funció sigui un nucli, i per tant defineixi un espai característic, també asegura que el problema d'optimització dual del classificador de marge màxim (2.10) tingui una única solució òptima.

2.1.3 Classificador de marge dèbil

Aquest classificador proposat per Corinna Cortes i Vladimir Vapnik el 1995 tracta de donar solució al problema del sobreajustament. Recordem que al problema d'optimització primària del classificador de marge màxim plantejat a (2.6), la restricció de desigualtat $\boldsymbol{y}_i(\langle w,\boldsymbol{x}_i\rangle + b)\geqslant 1$ forçava als hiperplans a no cometre cap error de classificació sobre el conjunt d'entrenament. El classificador de marge dèbil proposa afegir un seguit de variables ξ_i, anomenades *slack variables*, però que nosaltres denominarem com **variables dèbils**, amb l'objectiu de suavitzar les restriccions. El problema d'optimització primari que proposa el classificador de marge dèbil és el següent:

$$\begin{aligned}
\mathit{minimitzar}_{\boldsymbol{\xi},\boldsymbol{w},b}\quad & \langle\boldsymbol{w},\boldsymbol{w}\rangle + C\sum_{i=1}^{d}\xi_i, && \boldsymbol{\xi}\in\mathbb{R}^d\\
\mathit{tal\ que}\quad & \boldsymbol{y}_i\big(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle + b\big)\geqslant 1-\xi_i, && i=1,...,d,\\
& \xi_i\geqslant 0, && i=1,...,d,
\end{aligned}\tag{2.11}$$

on $C>0$ és un paràmetre de regularització que controla l'equilibri entre maximitzar el marge i minimitzar les violacions del marge o errors de classificació. Com més alt sigui aquest valor menys classificacions errònies permet el model però més risc de sobreajustament. Notem que si substituïm les variables dèbils de la funció objectiu per els seus quadrats podem prescindir de les restriccions de positivitat sobre les ξ_i, i el problema d'optimització es converteix en

$$\begin{aligned}
\mathit{minimitzar}_{\boldsymbol{\xi},\boldsymbol{w},b}\quad & \langle\boldsymbol{w},\boldsymbol{w}\rangle + C\sum_{i=1}^{d}\xi_i^2, && \boldsymbol{\xi}\in\mathbb{R}^d\\
\mathit{tal\ que}\quad & \boldsymbol{y}_i\big(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle + b\big)\geqslant 1-\xi_i, && i=1,...,d.
\end{aligned}\tag{2.12}$$

L'equivalència entre els problemes (2.11) i (2.12) es deu a què si $\xi_i < 0$ aleshores la primera restricció de (2.11) també se satisfà si s'estableix que $\xi_i = 0$. Per tant, la solució òptima de (2.11) coincidirà amb la solució òptima del mateix problema eliminant la restricció de positivitat sobre les ξ_i.

A la pràctica, el paràmetre C varia dins d'un ampli rang de valors i es troba mitjançant el mètode de validació creuada utilitzant només el conjunt d'entrenament. Realment trobar C en (2.12) correspon a fixar un valor per $\|\boldsymbol{w}\|_2$ i minimitzar $\|\boldsymbol{\xi}\|_2$ per aquella mida de $\boldsymbol{w}$. De la mateixa manera, trobar C en (2.11) correspon a trobar un mínim de $\|\boldsymbol{\xi}\|_1$.

Marge dèbil amb norma $\|\cdot\|_2$

La funció general de Lagrange del problema (2.12) és

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + \frac{C}{2}\sum_{i=1}^{d}\xi_i^2 - \sum_{i=1}^{d}\alpha_i\big[\boldsymbol{y}_i\big(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b\big) - 1 + \xi_i\big].$$

Si imposem estacionarietat sobre les derivades parcials respecte de $\boldsymbol{w}$, $\boldsymbol{\xi}$ i b

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{d}\boldsymbol{y}_i\alpha_i\boldsymbol{x}_i = 0, \tag{2.13}$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha})}{\partial \boldsymbol{\xi}} = C\boldsymbol{\xi} - \boldsymbol{\alpha} = 0, \tag{2.14}$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha})}{\partial b} = \sum_{i=1}^{d}\boldsymbol{y}_i\alpha_i = 0,$$

i substituïm les relacions obtingudes a la funció general de Lagrange, tenim que

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = {} & \frac{1}{2}\sum_{i,j=1}^{d}\boldsymbol{y}_i\boldsymbol{y}_j\alpha_i\alpha_j\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle + \frac{C}{2}\sum_{i=1}^{d}\frac{\alpha_i^2}{C^2} - \sum_{i,j=1}^{d}\boldsymbol{y}_i\boldsymbol{y}_j\alpha_i\alpha_j\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \\ & - b\sum_{i=1}^{d}\boldsymbol{y}_i\alpha_i - \sum_{i=1}^{d}\alpha_i + \sum_{i=1}^{d}\frac{\alpha_i^2}{C} = \sum_{i=1}^{d}\alpha_i - \frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_j\alpha_i\alpha_j\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle - \frac{1}{2C}\langle \boldsymbol{\alpha}, \boldsymbol{\alpha}\rangle. \end{aligned}$$

Per tant, minimitzar la funció objectiu de dalt correspon a maximitzar

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{d}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{d}\boldsymbol{y}_i\boldsymbol{y}_j\alpha_i\alpha_j\left(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle + \frac{1}{C}\delta_{ij}\right),$$

on δ_{ij} és l'anomenada delta de Kronecker que val 1 si $i = j$, i 0 altrament. Ara obtenim el següent resultat on hem introduït directament l'ús nucli.

Proposició 2.4. *Considerem un conjunt d'entrenament*

$$S = \big((\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_d, \boldsymbol{y}_d)\big),$$

utilitzant l'espai característic definit implícitament pel nucli $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{z})$. *Suposem que* $\boldsymbol{\alpha^*}$ *resol el problema d'optimització:*

$$\begin{aligned} \textit{maximitzar} \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{d} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{d} \boldsymbol{y}_i\boldsymbol{y}_j\alpha_i\alpha_j\left(\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) + \frac{1}{C}\delta_{ij}\right), \\ \textit{tal que} \quad & \sum_{i=1}^{d} \boldsymbol{y}_i\alpha_i = 0, \\ & \alpha_i \geqslant 0,\ i = 1,...,d. \end{aligned}$$

Sigui $f(\boldsymbol{x}) = \sum_{i=1}^{d} \boldsymbol{y}_i\alpha_i^*\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}) + b^*$, *on* $b^* \in \mathbb{R}$ *es triat tal que* $\boldsymbol{y}_i f(\boldsymbol{x}_i) = 1 - \alpha_i^*/C$ *per qualsevol* i *amb* $\alpha_i^* \neq 0$. *Aleshores la funció de decisió* $f_D(\boldsymbol{x}) = sgn(f(\boldsymbol{x}))$ *és equivalent a l'hiperplà a l'espai característic que resol el problema d'optimització de (2.12), on les variables dèbils estan definides relatives al marge geomètric*

$$\gamma = \left(\sum_{i\in sv} \alpha_i^* - \frac{1}{C}\langle\boldsymbol{\alpha^*},\boldsymbol{\alpha^*}\rangle\right)^{-1/2}.$$

Demostració. Recordem que la condició de complementarietat de Karush-Kuhn-Tucker per $(\boldsymbol{w^*}, b^*)$ amb $\boldsymbol{w^*}$ definit per la relació (2.13) ens diu que

$$\alpha_i^*\left[\boldsymbol{y}_i\big(\langle\boldsymbol{w^*},\boldsymbol{x}_i\rangle + b^*\big) - 1 + \xi_i\right] = 0,\ i = 1,...,d.$$

Si $\alpha_i^* \neq 0$ aleshores s'ha de complir

$$\boldsymbol{y}_i f(\boldsymbol{x}) - 1 + \xi_i = 0 \iff \boldsymbol{y}_i f(\boldsymbol{x}) = 1 - \xi_i = 1 - \frac{\alpha_i^*}{C},$$

on hem utilitzat la relació (2.14).

Ara, calculem la norma de $\boldsymbol{w^*}$ que és la que defineix el marge geomètric.

$$\begin{aligned} \langle\boldsymbol{w^*},\boldsymbol{w^*}\rangle &= \sum_{i,j=1}^{d} \boldsymbol{y}_i\boldsymbol{y}_j\alpha_i^*\alpha_j^*\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ &= \sum_{j\in sv} \alpha_j^*\boldsymbol{y}_j \sum_{i\in sv} \boldsymbol{y}_i\alpha_i^*\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ &= \sum_{j\in sv} \alpha_j^*(1 - \xi_j^* - \boldsymbol{y}_j b^*) \\ &= \sum_{i\in sv} \alpha_i^* - \sum_{i\in sv} \alpha_i^*\xi_i^* \\ &= \sum_{i\in sv} \alpha_i^* - \frac{1}{C}\langle\boldsymbol{\alpha^*},\boldsymbol{\alpha^*}\rangle. \end{aligned}$$

□

L'única diferència amb el model del classificador de marge màxim és que la matriu nucli $\mathbb{K} = \big(\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)\big)_{i,j=1}^{d}$ és la mateixa però amb els elements de la diagonal multiplicats per 1/C, és a dir, que els valors propis d'aquesta també estaran multiplicats per 1/C. Per tant, es pot plantejar el problema com un canvi en el nucli

$$\boldsymbol{K'}(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{K}(\boldsymbol{x},\boldsymbol{z}) + \frac{1}{C}\delta_{\boldsymbol{x}}(\boldsymbol{z}),$$

on $\delta_{\boldsymbol{x}}(\boldsymbol{z}) = 1$ si $\boldsymbol{x} = \boldsymbol{z}$, i 0 altrament.

Marge dèbil amb norma $\|\cdot\|_1$

En aquest cas la funció general de Lagrange pel probema d'optimització per la norma $\|\cdot\|_1$ és

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r}) = \frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle + C\sum_{i=1}^{d}\xi_i - \sum_{i=1}^{d}\alpha_i\big[\boldsymbol{y}_i\big(\langle \boldsymbol{x}_i, \boldsymbol{w}\rangle + b\big) - 1 + \xi_i\big] - \sum_{i=1}^{d} r_i\xi_i,$$

amb $\alpha_i \geqslant 0$ i $r_i \geqslant 0$. Ara derivant respecte a $\boldsymbol{w}, \boldsymbol{\xi}$ i b, i imposant estacionarietat

$$\begin{aligned}
\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r})}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_{i=1}^{d}\boldsymbol{y}_i\alpha_i\boldsymbol{x}_i = 0,\\
\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r})}{\partial \xi_i} &= C - \alpha_i - r_i = 0,\\
\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r})}{\partial b} &= \sum_{i=1}^{d}\boldsymbol{y}_i\alpha_i = 0.
\end{aligned}$$

Resubstituint a la funció general de Lagrange obtenim

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r}) = \sum_{i=1}^{d}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{d}\boldsymbol{y}_i\boldsymbol{y}_j\alpha_i\alpha_j\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle.$$

Notem que la funció de Lagrange per aquest problema coincideix amb la del marge màxim. L'única diferència és que la condició $C - \alpha_i - \xi_i = 0$, juntament amb $r_i \geqslant 0$, força a què $\alpha_i \leqslant C$. Les condicions complementàries de Karush-Kuhn-Tucker són

$$\begin{aligned}
&\alpha_i\big[\boldsymbol{y}_i\big(\langle \boldsymbol{x}_i, \boldsymbol{w}\rangle + b\big) - 1 + \xi_i\big] = 0, \quad && i = 1, ..., d,\\
&\xi_i(\alpha_i - C) = 0, && i = 1, ..., d.
\end{aligned}$$

Per tant, les variables dèbils són diferents de zero només quan $\alpha_i = C$ i aleshores $r_i = 0$. Els punts $\boldsymbol{x}_i$ pels quals $\xi_i \neq 0$ tenen marge geomètric menor a $1/\|\boldsymbol{w}\|$, d'altra banda els $\boldsymbol{x}_i$ pels quals $0 < \alpha_i < C$ estan a distància $1/\|\boldsymbol{w}\|$ de l'hiperplà.

Proposició 2.5. *Considerem el conjunt d'entrenament*

$$S = \big((\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_d, \boldsymbol{y}_d)\big),$$

utilitzant l'espai característic definit implícitament pel nucli $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z})$. Suposem que $\boldsymbol{\alpha}^*$ resol el problema d'optimització:

$$\begin{aligned}
\textit{maximitzar} \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{d}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{d}\boldsymbol{y}_i\boldsymbol{y}_j\alpha_i\alpha_j\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j),\\
\textit{tal que} \quad & \sum_{i=1}^{d}\boldsymbol{y}_i\alpha_i = 0,\\
& 0 \leqslant \alpha_i \leqslant C,\ i = 1, ..., d.
\end{aligned}$$

Sigui $f(\boldsymbol{x}) = \sum_{i=1}^{d}\boldsymbol{y}_i\alpha_i^*\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b^*$, on b^* es triat tal que $\boldsymbol{y}_i f(\boldsymbol{x}_i) = 1$ per qualsevol i amb $0 < \alpha_i^* < C$. Aleshores $f_D(\boldsymbol{x}) = sgn(f(\boldsymbol{x}))$ és equivalent a l'hiperplà a l'espai

característic que resol el problema d'optimització de (2.11), on les variables dèbils estan definides relatives al marge geomètric

$$\gamma = \left(\sum_{i,j \in sv} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i^* \alpha_j^* \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)^{-1/2}.$$

Demostració. El valor de b^* és triat segons les condicions de Karush-Kuhn-Tucker que impliquen que per $0 < \alpha_i^* < C$

$$\begin{aligned} &\alpha_i^* \left[f(\boldsymbol{x}_i) - 1 + \xi_i^* \right] = 0, \text{ i} \\ &\xi_i^* = 0 \Longrightarrow \boldsymbol{y}_i f(\boldsymbol{x}_i) = 1. \end{aligned}$$

La norma de $\boldsymbol{w^*} = \sum_{i=1}^d \boldsymbol{y}_i \alpha_i \boldsymbol{x}_i$ està donada per l'expressió

$$\begin{aligned} \langle \boldsymbol{w^*}, \boldsymbol{w^*} \rangle &= \sum_{i,j=1}^{d} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i^* \alpha_j^* \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \sum_{j \in sv} \sum_{i \in sv} \boldsymbol{y}_i \boldsymbol{y}_j \alpha_i^* \alpha_j^* \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j). \end{aligned}$$

□

Per últim, relacionarem el problema d'optimització amb norma $\| \cdot \|_1$ sobre l'espai característic $\boldsymbol{\phi}(X) \subseteq \mathcal{H}$, amb el problema d'optimització (2.2). Primer notem que el primer conjunt de restriccions de (2.11) es pot reescriure com $\xi_i \geqslant 1 - \boldsymbol{y}_i(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_i) \rangle + b)$. Si ho combinem amb el segon conjunt de restriccions $\xi_i \geqslant 0$ obtenim que les variables dèbils han de satisfer

$$\xi_i \geqslant max\{0, 1 - \boldsymbol{y}_i \big(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_i) \rangle + b \big) \} = L_h(\boldsymbol{y}_i, \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_i) \rangle + b).$$

És clar que si aquesta desigualtat es transforma en una igualtat, aleshores la funció objectiu continua essent mínima en ξ_i. Donat $(\boldsymbol{w}, b) \in \mathcal{H} \times \mathbb{R}$ definim la funció $f_{(\boldsymbol{w},b)}(\boldsymbol{x}) := \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_i) \rangle + b$. Ara, si multipliquem la funció objectiu de (2.11) per $2\lambda = 1/dC$ podem reescriure-la com

$$\min_{(\boldsymbol{w},b) \in \mathcal{H} \times \mathbb{R}} \lambda \langle \boldsymbol{w}, \boldsymbol{w} \rangle + \frac{1}{d} \sum_{i=1}^{d} L_h(\boldsymbol{y}_i, f_{(\boldsymbol{w},b)}(\boldsymbol{x}_i)), \tag{2.15}$$

Recordem que el problema d'optimització de (2.2) era

$$\inf_{f \in \mathscr{H}} \lambda \|f\|_{\mathscr{H}} + \frac{1}{d} \sum_{i=1}^{d} L_h(\boldsymbol{y}_i, f(\boldsymbol{x}_i)). \tag{2.16}$$

La major diferència entre aquests dos plantejaments és que el primer consisteix en considerar un espai de Hilbert general $\mathcal{H}$ i definir una funció $f_{(\boldsymbol{w},b)}$ en aquest espai, mentre que, en el segon, directament es consideren les funcions contingudes en el RKHS $\mathscr{H}$. Ara considerem la representació de Mercer de $\mathscr{H}$ definida a (1.7). Recordem que $f \in \mathscr{H}$, s'escriu com

$$f(\boldsymbol{x}) = \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i} \psi_i(\boldsymbol{x}) = \left\langle (a_i)_{i \in \mathbb{N}}, (\sqrt{\lambda_i} \psi_i(\boldsymbol{x}))_{i \in \mathbb{N}} \right\rangle_{\mathscr{H}} = \left\langle (a_i)_{i \in \mathbb{N}}, \boldsymbol{\phi}(\boldsymbol{x}) \right\rangle_{\mathscr{H}}.$$

Geomètricament, es pot entendre com que $(a_i)_{i\in\mathbb{N}}$ és el vector de pesos, $\boldsymbol{w}$, a l'espai característic ℓ_2 i, per tant, podem escriure la norma de f com

$$\|f\|^2_{\mathscr{H}} = \langle f, f\rangle_{\mathscr{H}} = \sum_{i=1}^{\infty} a_i^2 = \langle \boldsymbol{w}, \boldsymbol{w}\rangle_{\ell_2},$$

aleshores és directe concloure que (2.16) equival al problema

$$\inf_{(f,b)\in\mathscr{H}\times\mathbb{R}} \lambda\langle \boldsymbol{w}, \boldsymbol{w}\rangle_{\ell_2} + \frac{1}{d}\sum_{i=1}^{d} L(\boldsymbol{y}_i, f(\boldsymbol{x}_i) + b),$$

que alhora és equivalent a (2.15).

3 Processament del llenguatge natural: com podem representar dades de text?

Com a procès previ a la implementació de qualsevol mètode d'aprenentatge automàtic és necessari tractar les dades per tal que tinguin una representació numèrica, degut a què generalment aquests només processen vectors numèrics, per tant, no poden rebre text com a dades d'entrada. El preprocessament de dades és un procès que consisteix en netejar, transformar, reduir i normalitzar un conjunt de dades per tal d'obtenir una font de dades consistent i adequada pel cumpliment de l'objectiu del model. En el cas de dades de text aquest procès consistirà en trobar una representació que permeti al model d'aprenentatge retornar una predicció útil per la tasca concreta que es vol dur a terme. Per exemple, sobre un mateix conjunt de dades no és el mateix preguntar-se *De quin tema està parlant?* (tasca de classificació), que preguntar-se *Aquest text té un to positiu o negatiu?* (tasca d'anàlisi de sentiments); és possible que aquestes dues tasques necessitin una representació del llenguatge diferent.

D'altra banda, és important entendre que el llenguatge natural va ser creat a través d'un procès d'evolució, és a dir, les seves regles van ser formalitzades un cop els usuaris van fer ús d'ell, això implica que aquests mateixos usuaris ignorin o trenquin les normes amb freqüència i, conseqüentment, trobar un algoritme per representar el llenguatge natural a la perfecció sigui complicat. Per tant, s'ha de tenir en compte que els models de processament de text que es tractaran en aquesta secció no posseeixen un coneixement humà del llenguatge sino que simplement busquen regularitats estadístiques en les seves dades d'entrada.

Tot procès de preprocessament de dades de text consta obligatòriament de tres fases: **estandarditzación**, **tokenització** i **vectorització**.

- Primer el text s'estandarditza per tal de fer-lo més fàcil de processar pel model, per exemple es converteix tot en lletra petita i s'eliminen els caràcters especials com signes de puntuació.
- En segon lloc es divideix el text en unitats o ***tokens*** (poden ser paraules o grups de paraules).
- Per últim es converteix cada *token* en un vector numèric. Existeixen molts mètodes per dur a terme aquesta tasca, posteriorment es presentaran alguns dels més comuns.

Prèviament a l'últim procès de vectorització cal donar un valor numèric a cada *token*, és a dir, cal assignar un enter únic a cada unitat del nostre vocabulari el qual constarà de totes les entrades del conjunt d'entrenament del mètode d'aprenentatge que es vulgui emprar posteriorment. A continuació, ens centrarem en presentar alguns dels mètodes de vectorització de text més comuns.

3.1 Codificació *One-hot*

Definició 3.1 (Representació one-hot)**.** *Sigui $V = \{1, ..., N\}$ el nostre vocabulari indexat i sigui $n \in V$ un token indexat, la representació one-hot de n es defineix com el vector binàri $v=(v_i)_{i=1}^N$, amb:*

$$v_i = \begin{cases} 1, & \textit{si } i = n. \\ 0, & \textit{altrament.} \end{cases}$$

és a dir, $v_i = \mathbb{1}_{\{n\}}(i)$.

Un avantatge que presenta aquesta representació és la senzillesa de la seva implementació i comprensió, a més de què ens permet preservar l'ordre de les paraules al text. En canvi, una de les raons per les quals no se sol implementar en models d'aprenentatge profund és per l'alta dimensionalitat que comporta. Suposem que un element del nostre conjunt d'entrada té 10 paraules i la mida del vocabulari és de 10.000 paraules úniques, si tokenitzem per paraules, el nostre model estarà passant d'una mostra de dimensió 10 a una mostra multidimensional de dimensió 10×10.000. Això pot afectar negativament en la complexitat computacional i la memòria requerida.

Un altre desaventatge és que aquest tipus de codificació no captura la informació semàntica i gramatical de les paraules ni les relacions sintàctiques entre elles. Observem que si considerem $\boldsymbol{\beta}_e=\{v_1, ..., v_N\}$ la base canònica de $\mathbb{R}^N$, aquestes són les representacions *one-hot* de cada *token* del vocabulari V i, conseqüentment, tots els vectors v_i són ortogonals. Això dona lloc a què la similitud entre els vectors sigui nul·la, el que significa que no es pot mesurar la similitud semàntica entre paraules, com per exemple la sinonímia.

Tal com es desenvolupa en [10], per $v, u \in \boldsymbol{\beta}_e$ amb $v \neq u$ i per $i, j \in \{0, 1\}$, definim $S_{i,j} := S_{i,j}(v, u) = \sum_{h=1}^{N} \mathbb{1}_{\{i\}}(v_h)\mathbb{1}_{\{j\}}(u_h)$. De manera intuïtiva, $S_{0,1}$ és la quantitat de vegades en què v té un 0 i u té un 1 a la mateixa posició $h \in \{1, ..., N\}$, i idènticament pels altres casos variant i i j. Observem que $S_{1,1}$ no és res més que el producte escalar euclidià entre v i u. Per tant, en el cas de les representacions one-hot, sempre tindrem: $S_{0,0} = N - 2$, $S_{0,1} = 1$, $S_{1,0} = 1$ i $S_{1,1} = 0$ per qualsevol parell de vectors $v, u \in \boldsymbol{\beta}_e$, $v \neq u$. A continuació, observem una taula que mostra un seguit de mesures de similitud binària que, com el seu nom indica, mesuren com de semblants són dos objectes. Com hem vist abans tots els vectors one-hot són ortogonals i, conseqüentment, donen una puntuació de 0 en aquestes mesures.

Jaccard	Cosine	Dice	Russel-Rao
$\frac{S_{1,1}}{S_{1,1}+S_{0,1}+S_{1,0}}$	$\frac{S_{1,1}}{\sqrt{(S_{1,1}+S_{0,1})(S_{1,1}+S_{1,0})^2}}$	$\frac{2S_{1,1}}{2S_{1,1}+S_{0,1}+S_{1,0}}$	$\frac{S_{1,1}}{S_{1,1}+S_{0,1}+S_{1,0}+S_{0,0}}$

3.2 *Word2Vec*

Word2Vec és un model de representació vectorial de paraules en un espai vectorial de dimensió relativament petita, comparat amb el cas de la codificació *One-hot*, el qual tracta de reflexar la relació semàntica de paraules mitjançant relacions geomètriques, per exemple assignant vectors propers a paraules sinònimes. Aquests vectors s'anomenen *word embeddings* i el que fan és conformar un espai geomètric estructurat que representa el llenguatge humà (espai *embedding*). Notem que només tenint en compte aquestes característiques *Word2Vec* ja presenta notables millores respecte el model *One-hot*.

Aquest model, desenvolupat per un grup d'investigadors de Google liderat per Thomas Mikolov el 2013, és actualment un dels més populars i emprats. El model aprèn els *word embeddings* a través de la implementació d'una xarxa neuronal que consta d'una capa d'entrada, una capa oculta i una capa de sortida i, durant el seu aprenentatge, calcula les probabilitats condicionades de les paraules respecte el seu context utilitzant la funció exponencial *softmax* $\sigma : \mathbb{R}^n \longmapsto (0,1)^n$, definida per:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \quad i \in \{1, ..., n\}.$$

Observem que aquesta pren un vector z de dimensió n i retorna una distribució de probabilitat.

Existeixen dues versions d'aquest model que tenen diferents enfocaments:

3.2.1 Model *Continuous-bag-of-words* (CBOW)

Sigui w_n una paraula del vocabulari i sigui $\mathcal{C} = \{w^{(n-m)}, ..., w^{(n-1)}, w^{(n+1)}, ..., w^{(n+m)}\}$ amb $m > 0$, el conjunt de paraules que es troben properes a w_n dins del text, que anomenarem el **context** de w_n. El model *Continuous-bag-of-words* té com a objectiu predir o generar la paraula w_n, o **paraula central**, segons el seu context $\mathcal{C}$.

L'algoritme consisteix en calcular dos vectors u_n i v_n per cada paraula w_n, que corresponen a les representacions vectorials de w_n quan aquesta pertany al context i quan és la paraula central, els anomenarem respectivament **vector d'entrada** i **vector de sortida**. Com això és per a cada paraula de les nostres dades, el problema real del model consisteix en calcular dues matrius $\mathcal{V} \in \mathbb{R}^{N' \times N}$ i $\mathcal{U} \in \mathbb{R}^{N \times N'}$, on N' és la mida del nostre espai *embedding*. Anomenarem $\mathcal{V}$ la **matriu d'entrada** de paraules tal que la seva i-èssima columna és el vector d'entrada N'-dimensional v_i, i **matriu de sortida** de paraules a $\mathcal{U}$, la j-èssima fila de la qual correspon al vector de sortida N'-dimensional u_j. El procediment i arquitectura de la xarxa neuronal de dues capes que aplica el model per predir o calcular les matrius és el següent:

Sigui $y \in \mathbb{R}^N$ la representació *one-hot* de la paraula w_n:

1. S'assigna a cada paraula del context $\mathcal{C}$ de mida $2m$ el seu vector *one-hot* de mida N corresponent, obtenint així els vectors: $x^{(n-m)}, ..., x^{(n-1)}, x^{(n+1)}, ..., x^{(n+m)} \in \mathbb{R}^N$.

2. S'aplica als vectors *one-hot* la matriu $\mathcal{V}$, la qual ha estat inicialitzada amb pesos arbitraris:
   $v_{n-m} = \mathcal{V}x^{(n-m)}, ..., v_{n-1} = \mathcal{V}x^{(n-1)}, v_{n+1} = \mathcal{V}x^{(n+1)}, ..., v_{n+m} = \mathcal{V}x^{(n+m)} \in \mathbb{R}^{N'}$.

3. Es calcula el vector mitjà de tots aquests com: $\bar{v} = \frac{v_{n-m}+...+v_{n+m}}{2m} \in \mathbb{R}^{N'}$.

4. Es genera el vector $z = \mathcal{U}\bar{v} \in \mathbb{R}^N$. Intuïtivament, aquest vector és un vector de puntuacions que representa com de probable és que cada paraula del vocabulari sigui la paraula central donades les paraules del context.

5. Es transforma en una distribució de probabilitat aplicant la funció *softmax*: $\tilde{y} = \sigma(z) \in \mathbb{R}^N$.

Cada entrada del vector resultant correspon a la probabilitat de que una paraula w_i tingui un context $\mathcal{C}$, per cada paraula del vocabulari:

$$(\tilde{y})_i = P(w_i|\mathcal{C}) = \sigma_n(u_i^T \bar{v}) = \frac{exp(u_i^T \bar{v})}{\sum_{j \in V} exp(u_j^T \bar{v})}, i \in V. \quad (3.1)$$

Ara, per tal d'actualitzar les matrius d'entrada i de sortida mitjançant el mètode del gradient descendent, cal definir una funció de pèrdua la qual minimitzarem. Com el problema es basa en predir una distribució de probabilitat, es pot calcular l'error entre dues distribucions de probabilitat y i $\bar{y}$ mitjançant l'entropia creuada (*cross-entropy*):

$$H(\tilde{y}, y) = -\sum_{i=1}^{N} y_i log(\tilde{y}_i).$$

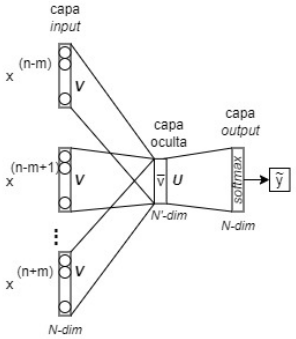

Figura 4 : Arquitectura de la xarxa neuronal que CBOW implementa

Tenint en compte que y és un vector *one-hot*, tots els termes de la suma s'anul·len excepte la posició en què y té un 1 i, per tant, per cada $n \in V$ el nostre problema d'optimització es redueix en minimitzar la següent funció:

$$\begin{aligned} H\big(P(w_n|\mathcal{C}), y\big) &= -log\, P(w_n|\mathcal{C}) \\ &= -log\, \frac{exp(u_n^T \bar{v})}{\sum_{j \in V} exp(u_j^T \bar{v})}, \quad \text{substituint per (3.1)} \\ &= u_n^T \bar{v} - log \sum_{j \in V} exp(u_j^T \bar{v}). \end{aligned}$$

Observem que es poden calcular les derivades parcials respecte qualsevol vector v_k, $\forall k \in \{1, ..., N\}$ de la següent manera:

$$\frac{\partial H\big(P(w_n|\mathcal{C}), y\big)}{\partial v_k} = \frac{1}{2m}\left(u_n - \sum_{i \in V} \frac{exp(u_i^T \bar{v}) u_i}{\sum_{j \in V} exp(u_j^T \bar{v})}\right) = \frac{1}{2m}\left(u_n - \sum_{i \in V} P(w_i|\mathcal{C}) u_i\right).$$

3.2.2 Model *Skip-Gram*

Una altra manera d'enfocar el problema és donada una paraula w_n tractar de predir el seu context $\mathcal{C}$. Es mantenen les definicions de les matrius $\mathcal{V} \in \mathbb{R}^{N' \times N}$ i $\mathcal{U} \in \mathbb{R}^{N \times N'}$ i s'intercanvien les x per les y, és a dir, l'entrada de la xarxa serà el vector *one-hot* $x \in \mathbb{R}^N$ que representa la paraula central, i la sortida de la xarxa seran $y^{(i)} \in \mathbb{R}^N$.

Siguin $y^{(n-m)}, ..., y^{(n-1)}, y^{(n+1)}, ..., y^{(n+m)}$ els vectors *one-hot* dels vectors del context $\mathcal{C}$

1. Primer es construeix el vector *one-hot* de la paraula central w_n, aquest serà $x \in \mathbb{R}^N$.
2. S'aplica la matriu d'entrada: $v_n = \mathcal{V}x \in \mathbb{R}^{N'}$.
3. S'aplica la matriu de sortida per obtenir un vector de puntuacions: $z = \mathcal{U}v_n \in \mathbb{R}$. En aquest cas cada entrada del vector simbolitza quant de probable és que una paraula sigui al context de w_n.
4. Es converteix en una distribució de probabilitat aplicant la funció *softmax*: $\tilde{y} = \sigma(z) \in \mathbb{R}^N$.

A diferència del model CBOW no s'utilitza cap funció d'activació a la capa oculta. Cada entrada del vector resultant correspon a la probabilitat de què una paraula del vocabulari w_i pertanyi al context $\mathcal{C}$ de la paraula central w_n.

$$(\tilde{y})_i = P(w_i \in \mathcal{C}|w_n) = \sigma_n(u_i^T v_n) = \frac{exp(u_i^T v_n)}{\sum_{j \in V} exp(u_j^T v_n)}, i \in V. \tag{3.2}$$

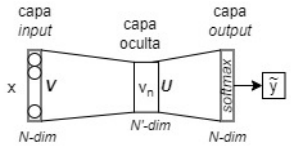

Figura 5 : Arquitectura de la xarxa neuronal que Skip-Gram implementa

Suposant que les probabilitats $P(w_i \in \mathcal{C}|w_n)$ són independents, és a dir, que les paraules del context són generades independentment a la paraula central, podem model·lar la probabilitat de què coneixent la paraula central w_n poguem generar el seu context $\mathcal{C}$ com:

$$P(\mathcal{C}|w_n) = \prod_{i \in \{n-m,...,n+m\}} P(w_i \in \mathcal{C}|w_n) = \prod_{j=0, j \neq m}^{2m} P(w_{n-m+j} \in \mathcal{C}|w_n).$$

Com en el model CBOW hem de definir una funció de pèrdua per tal de generar les matrius d'entrada i de sortida. Es simplificarà la notació $P(w_i \in \mathcal{C}|w_n) := P(w_i|w_n)$. En aquest cas es vol comparar la distribució resultant $\tilde{y}$ amb les distribucions corresponents als vectors *one-hot* $y^{(n-m)}, ..., y^{(n-1)}, y^{(n+1)}, ..., y^{(n+m)}$. El model ho fa mitjançant el mètode de màxima log-verosimilitud que consisteix en trobar els vectors d'entrada, u_i, que maximitzin la log-funció de densitat conjunta o log-verosimilitud, és a dir, maximitzar

la funció:

$$\begin{aligned}
-log\, P(\mathcal{C}|w_n) &= -log \prod_{j=0,j\neq m}^{2m} P(w_{n-m+j}|w_n) \\
&= -log \prod_{j=0,j\neq m}^{2m} P(u_{n-m+j}|v_n) \\
&= -log \prod_{j=0,j\neq m}^{2m} \frac{exp(u_{n-m+j}^T v_n)}{\sum_{k=1}^{N} exp(u_k^T v_n)}, \quad \text{substituint per (3.2)} \\
&= -\sum_{j=0,j\neq m}^{2m} u_{n-m+j}^T v_n + 2m log \sum_{k=1}^{N} exp(u_k^T v_n).
\end{aligned}$$

Notem que

$$\begin{aligned}
-log\, P(\mathcal{C}|w_n) &= -log \prod_{j=0,j\neq m}^{2m} P(u_{n-m+j}|v_n) \\
&= \sum_{j=0,j\neq m}^{2m} -log\, P(u_{n-m+j}|v_n) \\
&= \sum_{j=0,j\neq m}^{2m} H(\tilde{y}, y^{(n-m+j)}),
\end{aligned}$$

on $H(\tilde{y}, y^{(n-m+j)})$ és l'entropia creuada entre el vector de probabilitats $\tilde{y}$ i el vector *one-hot* de la paraula $w^{(n-m+j)}$.

Composicionalitat de *Skip-Gram*

En aquest apartat es donarà una justificació teòrica sobre com el model Skip-Gram genera els vectors de l'espai *embedding*. Aquest model genera vectors que presenten una propietat anomenada composicionalitat, això vol dir que un vector **v** en l'espai *embedding* serà proper en angle i distància a la composició de vectors que representen conceptes que junts capturen el concepte de **v**. A més a més, es provarà que sota algunes suposicions aquesta composicionalitat és additiva, en concret es provarà que ***v(reina)*** = ***v(rei)*** - ***v(home)*** + ***v(dona)***, on ***v(paraula)*** és el vector que representa a "paraula"en l'espai *embedding*.

Primer definirem el concepte de composicionalitat de paraules. Per això però, necessitem definir abans el concepte de la divergència de Kullback-Leibler o entropia relativa per distribucions de probabilitat discretes.

Definició 3.2 (Divergència de Kullback-Leibler)**.** *Siguin P i Q dues distribucions de probabilitat discretes sobre el mateix espai mostral, $\mathcal{X}$, definim la divergència de Kullback-Leibler com:*

$$D_{KL}(P \parallel Q) = \sum_{x\in\mathcal{X}} P(x)\, log\frac{P(x)}{Q(x)}.$$

Definició 3.3 (Paràfrasi)**.** *Sigui c una paraula del vocabulari i sigui $\mathcal{A} = \{c_1, ..., c_m\}$ un conjunt de paraules, anomenem paràfrasi a la paraula que minimitza la divergència de Kullback-Leibler, això és*

$$\min_{c \in V} D_{KL}\big(P(\cdot|\mathcal{A}) \parallel P(\cdot|c)\big).$$

Intuïtivament, el que vol dir aquesta última definició és que la parafràsi és la paraula que més s'assembla al significat de la composició de les paraules de $\mathcal{A}$, és a dir, $P(w|\mathcal{A})$ és el més semblant a $P(w|c)$ per qualsevol paraula w del vocabulari.

En els resultats que se segueixen se suposaran les dues conjectures següents:

A.1 Per cada paraula c del vocabulari existeix Z_c tal que per cada paraula w:

$$P(w|c) = \frac{1}{Z_c} exp(u_c^T v_w).$$

A.2 Per cada conjunt de paraules $\mathcal{A} = \{c_1, ..., c_m\}$, existeix $Z_{\mathcal{A}}$ tal que per cada paraula w:

$$P(w|\mathcal{A}) = \frac{P(w)^{1-m}}{Z_{\mathcal{A}}} \prod_{i=1}^{m} P(w|c_i).$$

Notació 3.1. *Sigui w una paraula del vocabulari denotarem per $\mathcal{C} = \{w^{(-\Delta)}, ..., w^{(\Delta)}\}$ el seu context i per u_w i v_w els vectors d'entrada i de sortida, tots tres generats pel model Skip-Gram. Es diu que $\mathcal{C}$ és un context de mida de finestra igual a Δ.*

Lemma 3.1. *El model Skip-Gram satisfà **A.1**, i també **A.2** si $m \leqslant \Delta$.*

Demostració. De (3.2) se segueix que el model satisfà **A.1**. Demostrarem que també satisfà **A.2** per $\Delta = m$ ja que pel cas $\Delta < m$ es té prou amb ignorar $\Delta - m$ termes del resultat final. Recordem que el model *Skip-Gram* considera que les paraules d'un conjunt $\{c_1, ..., c_m\}$ es generen independentment a qualsevol paraula:

$$P(c_1, ..., c_m|w) = \prod_{i=1}^{m} P(c_i|w)$$

Aplicant la regla de Bayes:

$$\begin{aligned} P(w|c_1, ..., c_m) &= \frac{P(c_1, ..., c_m|w)P(w)}{P(c_1, ..., c_m)} \\ &= \frac{P(w)}{P(c_1, ..., c_m)} \prod_{i=1}^{m} P(c_i|w) \\ &= \frac{P(w)}{P(c_1, ..., c_m)} \prod_{i=1}^{m} \frac{P(w|c_i)P(c_i)}{P(w)} \\ &= \frac{P(w)^{1-m}}{Z_{\mathcal{A}}} \prod_{i=1}^{m} P(w|c_i), \end{aligned}$$

on $Z_{\mathcal{A}} = 1/(\prod_{i=1}^{m} P(c_i))$. □

Teorema 3.1. *Tot model que satisfà* ***A.1*** *i* ***A.2****, per cada conjunt de paraules* $\mathcal{A} = \{c_1, ..., c_m\}$*, qualsevol paràfrasis* $c \in \mathcal{A}$ *satisfà:*

$$\sum_{w \in V} P(w|c)v_w = \sum_{w \in V} P(w|\mathcal{A})v_w. \tag{3.3}$$

Demostració. Substituïnt (3.2) a la suposició **A.2** obtenim:

$$\begin{aligned} P(w|\mathcal{A}) &= \frac{P(w)^{1-m}}{Z_{\mathcal{A}}} \prod_{i=1}^{m} \frac{exp(u_w^T v_{c_i})}{\sum_{j=1}^{N} exp(u_{w_j}^T v_{c_i})}, \text{ on } V = \{w_j\}_{j=1}^{N} \\ &= \frac{P(w)^{1-m}}{Z_{\mathcal{A}}} exp(\sum_{i=1}^{m} u_{c_i}^T v_w - \sum_{i=1}^{m} log\, Z_{c_i}), \text{ aplicant } e^{log} \text{ al productori} \\ &= \frac{1}{Z} P(w)^{1-m} exp(u_{\mathcal{A}}^T v_w), \end{aligned}$$

on $Z = Z_{\mathcal{A}} \prod_{i=1}^{m} Z_{c_i}$ amb $Z_{c_i} = \sum_{j=1}^{N} exp(u_{w_j}^T v_{c_i})$ i $u_{\mathcal{A}} = \sum_{i=1}^{m} u_{c_i}$.

Ara, minimitzar la divergència de Kullback-Leibler com a funció de $c \in V$ és equivalent a maximitzar l'entropia creuada com a funció de u_c:

$$\begin{aligned} \min_{c \in V} D_{KL}\big(P(\cdot|\mathcal{A}) \parallel P(\cdot|c)\big) &= \max_{u_c,\, c \in V} Z \sum_{i=1}^{N} \frac{exp(u_{\mathcal{A}}^T v_{w_i})}{P(w)^{m-1}} (u_c^T v_{w_i} - logZ_c) \\ &= \max_{u_c,\, c \in V} H\big(P(\cdot|\mathcal{A}), P(\cdot|c)\big). \end{aligned}$$

Com la funció a maximitzar és còncava aleshores els seus punts estacionaris succeeixen quan el gradient s'anul·la, és a dir, $\nabla_{u_c} H\big(P(\cdot|\mathcal{A}), P(\cdot|c)\big) = 0$.

$$\begin{aligned} \nabla_{u_c} H\big(P(\cdot|\mathcal{A}), P(\cdot|c)\big) &= Z \sum_{i=0}^{N} \frac{exp(u_{\mathcal{A}}^T v_{w_i})}{P(w)^{m-1}} \left[v_{w_i} - \frac{\sum_{l=1}^{N} exp(u_c^T v_{w_l}) v_{w_l}}{\sum_{k=1}^{N} exp(u_c^T v_{w_k})} \right] \\ &= \frac{\sum_{l=1}^{N} exp(u_c^T v_{w_l}) v_{w_l}}{\sum_{k=1}^{N} exp(u_c^T v_{w_k})} - Z \sum_{i=1}^{N} \frac{exp(u_{\mathcal{A}}^T v_{w_i}) v_{w_i}}{P(w)^{m-1}} \\ &= \sum_{w \in V} P(w|c) v_w - \sum_{w \in V} P(w|\mathcal{A}) v_w. \end{aligned}$$

Si igualem aquesta última expressió a 0 obtenim el resultat que es desitja. □

Teorema 3.2. *Tot model que satisfà* ***A.1*** *i* ***A.2****, i tal que* $P(w) = 1/|V| = 1/N$*, aleshores per tota* $w \in V$ *la paràfrasi del conjunt* $\mathcal{A} = \{c_1, ..., c_m\}$ *és*

$$u_1 + ... + u_m.$$

Demostració. En el cas concret en què $P(w) = 1/N$, l'equació $\nabla_{u_c} H\big(P(\cdot|\mathcal{A}), P(\cdot|c)\big) = 0$ es simplifica com:

$$Z \sum_{i=1}^{N} exp(u_{\mathcal{A}}^T v_{w_i}) \left[v_{w_i} - \frac{\sum_{l=1}^{N} exp(u_c^T v_{w_l}) v_{w_l}}{\sum_{k=1}^{N} exp(u_c^T v_{w_k})} \right] = 0$$

Observem que $\nabla_{u_c} H\big(P(u_{\mathcal{A}}|\mathcal{A}), P(u_{\mathcal{A}}|c)\big) = 0$, i com H és còncava aleshores $u_{\mathcal{A}}$ és el seu únic màxim. □

El que el Teorema 3.2 vol expressar és que si existeix una paraula c, el vector d'entrada u_c de la qual és igual a la suma dels vectors d'entrada de $c_1, ..., c_m$, aleshores c té el mateix significat (segons la Definició 3.3) que la composició de les paraules $c_1, ..., c_m$. No obstant, és molt poc probable trobar una paraula del vocabulari el vector d'entrada de la qual sigui exactament igual a la suma dels vectors d'entrada de $c_1, ..., c_m$. De forma similar, en el cas general en que no es fa cap suposició sobre la distribució de les paraules del vocabulari, és difícil trobar un vector solució de (3.3) que sigui un vector d'entrada d'una paraula del vocabulari. És així que és necessàri projectar els vectors solució al nostre vocabulari per tal de trobar la paràfrasi. A la pràctica el que es sol fer en concret, és trobar la paraula del vocabulari el vector d'entrada de la qual formi un angle més petit amb el vector solució.

Tornant a l'exemple inicial, es vol trobar la paraula relacionada amb *rei* la qual comparteixi la mateixa relació que *home* i *dona*. Aquesta idea l'escrivim com $h::d:rei::?$, on $?$ simbolitza la paraula que busquem. Segons el que hem vist, el fet que h i d estiguin relacionades equival a què h és la paràfrasi d'un conjunt de paraules $\{d, R\}$, on R és un conjunt de paraules que captura la relació entre h i d. De la mateixa forma r és la paràfrasi de $\{?, R\}$. Pel Teorema 3.1 tenim que R i $?$ han de satisfer les equacions:

$$\sum_{l \in V} P(l|m) v_l = \sum_{l \in V} P(l|\{w, R\}) v_l$$
$$\sum_{l \in V} P(l|k) v_l = \sum_{l \in V} P(l|\{?, R\}) v_l$$

Trobar $?$ equival a resoldre dos sistemes d'equacions no lineals. En cas que suposem que les paraules del vocabulari segueixen una distribució uniforme, pel Teorema 3.2

$$u_m = u_w + u_R$$
$$u_k = u_? + u_R$$

que ens dona la relació que buscàvem

$$u_? = u_k + (u_w - u_m).$$

4 Implementació i anàlisi de models SVM per multiclassificació

Tota la implementació que dona lloc als resultats que es tractaran en aquest capítol ha estat realitzada en llenguatge *Python* i es pot trobar a l'apèndix C d'aquest treball.

4.1 Conjunt de dades: preprocessament

El conjunt de dades d'entrada que s'utilitzarà consisteix en 6317 frases de text extretes d'una base de dades clíniques d'anotacions provinents de diferents entorns assistencials. Això implica la presència de moltes abreviacions i faltes ortogràfiques a causa de la rapidesa amb la que aquests informes s'escriuen. El procediment d'extracció ha consistit en la cerca del termes "M1" i "M0", que són codis que s'utilitzen per indicar metàstasi o no metàstasi respectivament, i l'extracció de cinc paraules per davant i cinc paraules per darrere d'aquest. Les frases del conjunt d'entrada son el resultat d'aplicar aquest procediment sobre tots els textos informatitzats de la base de dades. Cadascuna ha estat processada eliminant signes de puntuació, accents i caràcters especials. A més s'han exclós les paraules formades només per lletra ja que s'ha considerat que no aportaven informació rellevant. Per últim, s'han convertit totes les lletres en minúscules.

Les etiquetes o conjunt de sortida és $Y = \{0, 1, 2\}$, cada número corresponent a una classe. Les frases que neguen la presència de metàstasi són assignades a la classe 0, a les que afirmen que el pacient té metàstasi se les assigna la classe 2, i per últim, la classe 1 és una classe neutra que conté totes aquelles frases que no aporten informació suficent com per diagnosticar un "M0" o un "M1". El nostre conjunt de dades consta de 999 frases de la classe 1, 1250 frases de la classe 0 i 4068 frases de la classe 2.

A continuació es mostren tres exemples de frases, prèviament preprocessades, per cada classe*:

- *operat de neo pancrees amb dubtoses m1 hepatiques vs hepatocarcinoma mes probable doncs*
- *diagnostica encefalopatia hepatica en pacient amb m1 hepatiques en progressio servei de*
- *sense objectivar cap lesio sospitosa de m1 ni fractures es continua estudi amb*

Per tal d'avaluar l'efectivitat del model, el conjunt de dades s'ha de dividir en un conjunt d'entrenament i un conjunt de prova. El primer s'utilitza perquè el model aprengui i el segon per avaluar la seva capacitat de generalització sobre dades que encara no ha vist. Les proporcions solen ser de 80%-20% o 70%-30%, respectivament. Aquest pas és indispensable en qualsevol tècnica d'aprenentatge automàtic i és també molt important fer aquesta separació de forma representativa i prevenint fugues d'informació. En el nostre cas hem de repartir les dades en dos conjunts que continguin la mateixa proporció de casos de cada classe i, a més, no continguin frases d'un mateix pacient, és a dir, totes les frases d'un pacient han d'estar al mateix conjunt ja que sino en el moment d'avaluar el

*Cal recalcar que les frases no contenen informació real sinó que han estat inventades intentant capturar l'estructura sintàctica de les frases del conjunt de dades.

model amb el conjunt de prova aquest simplement hauria après a identificar el pacient no a etiquetar la frase.

4.2 Mètriques de multiclassificació

En la classificació binària tenim dues classes, una d'elles se l'anomena la classe positiva (1) i l'altra la negativa (0). Seguint aquesta nomenclatura el classificador pot cometre dos tipus d'encerts i dos tipus d'errors: els veritables positius (VP) o veritables negatius (VN) i falsos postius (FP) o falsos negatius (FN). Tal com els seus noms indiquen els dos primers tipus corresponen als casos en què el model fa una predicció encertada i els dos últims a quan el model s'equivoca. A continuació definim tres mètriques bàsiques en l'avaluació del rendiment de qualsevol model d'aprenentatge automàtic.

$$\text{Exactitud} := \frac{\text{VP} + \text{VN}}{\text{VP+VN+FP+FN}}, \qquad \text{Precisió} := \frac{\text{VP}}{\text{VP+FP}}, \qquad \text{Sensibilitat} := \frac{\text{VP}}{\text{VP+FN}}$$

L'exactitud és simplement la proporció d'encerts que el model ha realitzat. D'altra banda, notem que la precisió fa referència a la proporció de classes predites com a positives que són realment positives, mentre que la sensibilitat representa la proporció de classes positives que són correctament classificades. Per tant, el model s'ha d'optimitzar segons els resultats de la precisió si es volen minimitzar els falsos positius i segons la sensibilitat si es volen minimitzar els falsos negatius. En el nostre cas, si els resultats de les prediccions estan lligats al subministrament d'un tractament o medicació, és molt greu assignar a un pacient un tractament més suau que no ataqui la malaltia eficientment, però també és prou greu administrar-li un més agressiu mentre que no el necessita. Per tant, ens interessa que el nostre classificador sigui capaç de minimitzar tant els falsos positius com els falsos negatius. És per això que ens fixarem en la mitjana harmònica de la precisió i de la sensibilitat, l'anomenada puntuació F1.

$$\text{F1} := \frac{2 \cdot (\text{precisió} \cdot \text{sensibilitat})}{\text{precisió} + \text{sensibiltat}}$$

Un problema de multiclassificació s'aborda escalonant-lo en subproblemes de classificació binària, en el cas de les màquines de suport vectorial ja hem vist que el model troba un hiperplà separador per cada classe, considerant-la com a positiva mentre que agrupa totes les altres classes com a negatives. Per tant, en el nostre cas concret es generarà un hiperplà que separarà els punts que pertanyen a la classe 0 de la unió de punts de les classes 1 i 2, i així mateix per les altres classes, obtenint així tres hiperplans separadors o classificadors a més de tres mètriques per cada classe. Segons la distribució de classes al nostre conjunt de dades i quina importància li donem a cadascuna, es pot calcular la mètrica d'èxit total de tres formes diferents:

- Macro: calcula la mètrica d'èxit de cada classificador per separat per posteriorment calcular la seva mitjana aritmètica. És útil quan es vol donar igual importància a totes les classes ja que no té en compte el desequilibri de classes.
- Ponderada: calcula la mètrica d'èxit de cada classificador per separat i després pren la seva mitjana ponderada segons el nombre d'exemples de cada classe, és a dir, multiplica cada mètrica per el nombre d'ocurrències de cada classe dividit entre el

nombre total de mostres. És útil quan es vol que el model tingui més en compte els errors comesos en les classes majoritàries, és a dir, que tingui en compte el desequilibri de classes.

- Micro: acumula els recomptes de veritables positius, falsos positius i falsos negatius de tots els classificadors per després calcular la mètrica corresponent amb el nombre total de recomptes.

En el nostre cas utilitzarem la F1 ponderada ja que el nostre conjunt de dades no és del tot balancejat i volem donar més importància a les classes 0 i 2, que són les quals contenen més exemples.

4.3 Optimització d'hiperparàmetres i comparació de models

4.3.1 Resultats Word2Vec

Abans d'experimentar amb el classificador de vectors de suport, es donarà una justificació del model de vectorització de *Word2Vec* que s'ha triat (CBOW o Skip-Gram), i s'exploraran algunes relacions semàntiques reflectides per aquest entre els *word embeddings* que representen les paraules del conjunt d'entrenament. Utilitzarem la simulitud cosinus, freqüentment emprada en anàlisi de text i recuperació de informació, per mesurar com de semblants són dos vectors de l'espai *embedding*, és a dir, quanta relació sostenen dues paraules. Donats dos vectors u, v la similitud cosinus es calcula com

$$S_C(u, v) := |cos(\theta)| = \left| \frac{\langle u, v \rangle}{\|u\|_2 \cdot \|v\|_2} \right|,$$

on el numerador és el producte escalar euclidià entre els dos vectors i el denominador és el producte entre les normes euclidianes d'aquests. Clarament valors més propers a 1 indiquen més similitud, mentre que valors més propers a 0 indiquen dissimilitud.

S'ha pogut demostrar que el model CBOW produeix millors resultats en conjunts de dades grans on hi ha paraules que tenen significats similars en contexts similars, per tant, representa amb més exactitud les paraules més freqüents en el corpus. D'altra banda el model Skip-Gram captura amb més exactitud les relacions semàntiques entre paraules menys freqüents ja que cada paraula s'utilitza per predir les seves paraules veïnes. A més a més, s'ha estudiat que funciona millor en conjunts de dades més petits. Tot i així, l'efectivitat i capacitat de representació vectorial d'aquests dos models pot tenir diferents resultats segons la naturalesa de les dades i la tasca que es vol dur a terme. És així que abans de triar-ne un basant-me només en altres estudis i resultats teòrics, vaig voler experimentar amb ambdós per determinar quin representava amb més realitat les paraules del conjunt de dades.

Emprant el model CBOW ens trobem amb què la majoria de similituds cosinus entre dues paraules del conjunt d'entrenament són més altes que 0.8, per tant el model no està capturant correctament les similituds semàntiques entre paraules, ja que les representa totes semblants. Aquest resultat només varia si s'utilitza el paràmetre *sample* en la definició del model, el qual controla la freqüència amb què s'aplicarà el submostrig de paraules freqüents durant l'entrenament. El submostreig de paraules freqüents és una tècnica utilizada per reduir l'impacte de paraules molt freqüents que no proporcionen tanta informació semàntica com paraules que ho són menys. Un valor més alt de *sample*

significa que s'aplicarà menys submostreig, el que implica que es conservaran més paraules freqüents en el conjunt d'entrenament. L'interval de valors útils d'aquest parametre és $(0{,}10^{-5})$, però, tot i utilitzant el valor més alt possible i intentant experimentar amb els altres paràmetres, s'obté que pràcticament totes les similituds cosinus tenen un valor més petit que 0.5. Aquest fet ens fa descartar l'ús del model CBOW.

Ara, emprant el model Skip-Gram havent optimitzat alguns paràmetres, obtenim que les similituds cosinus de 1358 paraules úniques del vocabulari del conjunt d'entrenament segueixen aproximadament una distribució normal.

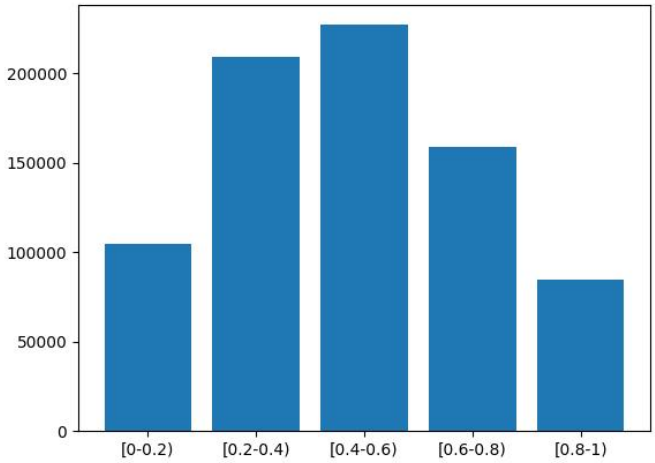

Figura 6 : Diagrama de barres on l'eix horitzontal representa els intervals de similituds

Per entendre les proporcions d'aquest diagrama notem que si el vocabulari té 1358 paraules, aleshores tenim $C^2_{1358} = \frac{1358!}{2 \cdot 1356!} = \frac{1358 \cdot 1357}{2} = 921403$ parelles de paraules i, per tant, aquesta és la quantitat total de similituds cosinus. Per acabar observem la següent taula amb algunes similituds que demostren que el model ha capturat correctament les relacions semàntiques i gramaticals que ens interessen.

(u, v)	$S_C(u, v)$
$(adenocarcinoma\,,\,tumor)$	0.88794
$(adenocarcinoma\,,\,adk)$	0.96838
$(suggestiva\,,\,suggestives)$	0.82815
$(suggestives\,,\,sospitoses)$	0.93820
$(suggestives\,,\,compatibles)$	0.77313
$(sospitoses\,,\,compatibles)$	0.89070
$(evidencia\,,\,signes)$	0.94362
$(hepatica\,,\,hep)$	0.98086
$(hepatica\,,\,hepatiques)$	0.97895
$(ecografia\,,\,eco)$	0.92999
$(folfox\,,\,folfiri)$	0.92513
$(inici\,,\,inicia)$	0.87153
$(recaiguda\,,\,recidiva)$	0.82722

Notem que el model ha capturat relacions de nombre, *hepatica* i *hepatiques*, de derivació, *inici* i *inicia*, a més de relacions entre paraules i les seves abreviacions, *adenocarcinoma* i *adk* o *hepatica* i *hep*. Era important que el model captés la semblança entre les paraules *suggestives*, *sospitoses* i *compatibles*, ja que moltes frases positives per metàstasi comencen per "s'han trobat imatges suggestives/compatibles/sospitoses de M1..." i és important que el classificador les identifiqui correctament. Altres relacions semàntiques interessants són les trobades entre *folfox* i *folfiri*, dos tractaments de quimioteràpia, *recaiguda* i *recidiva* o *evidencia* i *signes*.

4.3.2 Resultats classificadors de suport vectorial

Per donar amb el classificador de suport vectorial més adient pel nostre problema s'ha experimentat amb els quatre kernels introduïts a l'apartat 1.3: polinòmic, lineal, gaussià i sigmoide. Per cada nucli s'ha entrenat un classificador per diferents paràmetres especificats i s'ha calculat una mètrica d'èxit per cada entrenament per finalment quedar-se amb aquells paràmetres que obtenien una millor mètrica.

No obstant, ens interessa poder visualitzar d'alguna manera com el model està classificant els punts sense recaure només en les mètriques d'èxit definides anteriorment. El problema de SVM és que la dimensió de l'espai característic és massa alta com per poder representar gràficament els seus elements i l'hiperplà separador. Existeixen tècniques de reducció de dimensionalitat basades en detectar les característiques més rellevants, però d'aquesta manera perdem dimensions que podrien aportar informació imprescindible per visualitzar correctament la classificació. Una bona tècnica introduïda a [15] que ens permet visualitzar el límit de decisió i totes les classificacions sense pèrdua de dimensionalitat i sense importar com de alta sigui la dimensió del nostre espai característic, és l'anomenat histograma univariat de projeccions. Aquest consisteix en fer un histograma de la distribució dels valors de sortida de la funció solució (2.3) sobre les dades d'entrenament o les dades de prova. Aquest tipus de tècnica proporciona una visió discreta basada en la freqüència de les dades en intervals específics. Per tal d'obtenir una representació més suau i contínua, emprarem la tècnica d'**estimació de la densitat de nucli (KDE)**. Aquesta utilitza un mètode no paramètric per estimar la funció de densitat de probabilitat d'una variable aleatoria basada en funcions nucli com a pesos.

Primer de tot es mostraran els resultats de les mètriques d'exactitud i puntuació F1 per cada nucli, tant sobre el conjunt d'entrenament com sobre el de prova.

Lineal		
Conjunt	Exactitud	F1
Entrenament	0.780	0.789
Prova	0.753	0.763

Polinòmic		
Conjunt	Exactitud	F1
Entrenament	0.827	0.835
Prova	0.767	0.775

Gaussià		
Conjunt	Exactitud	F1
Entrenament	0.817	0.824
Prova	0.767	0.777

Sigmoide		
Conjunt	Exactitud	F1
Entrenament	0.734	0.747
Prova	0.749	0.759

El fet que les mètriques del conjunt de prova no s'allunyin massa de les del conjunt d'entrenament mostra que el model es bo generalitzant sobre dades noves. A més, notem que en el cas dels nuclis polinòmic i gaussià s'aconsegueix assolir més d'un 0.8 tant en l'exactitud com en la puntuació F1, en el conjunt d'entrenament. Tot i així, com el nostre conjunt de dades és bastant desbalancejat, per veure quin tipus d'errors està cometent el model és útil visualitzar l'anomenada **matriu de confusió**, la qual representa en cada fila el nombre de prediccions de cada classe i en cada columna el nombre d'instancies en la classe real. Un dels avantatges que presenta és que permet veure quines classes en concret el model està confonent, en el nostre cas per exemple, com s'ha explicat a l'apartat 4.2, no és el mateix confondre una classe 2 amb una 0 o que una 2 amb una 1. A la següent imatge podem observar les corresponents matrius de confusió del conjunt de prova per cada nucli.

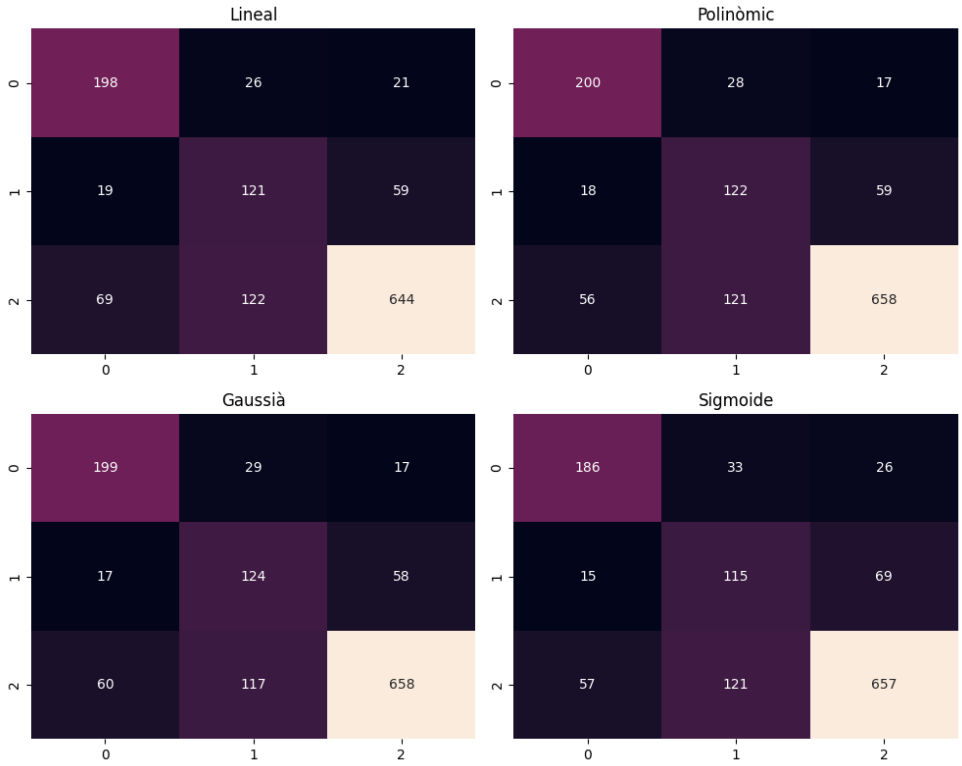

Figura 7 : Matrius de confusió dels quatre models entrenats

Observem que els errors més freqüents són les confusions entre les classes 1 i 2. És obvi que és més greu confondre la classe 1 amb la 2 ja que en el context del problema significaria fer un diagnòstic positiu de metàstasi a algú qui podria no tenir-ne, en canvi a l'inversa simplement es queda com una incògnita.

Ara veiem com és l'histograma univariant de les projeccions del conjunt d'entrenament del classificador de nucli polinòmic, pel qual hem obtingut millors resultats.

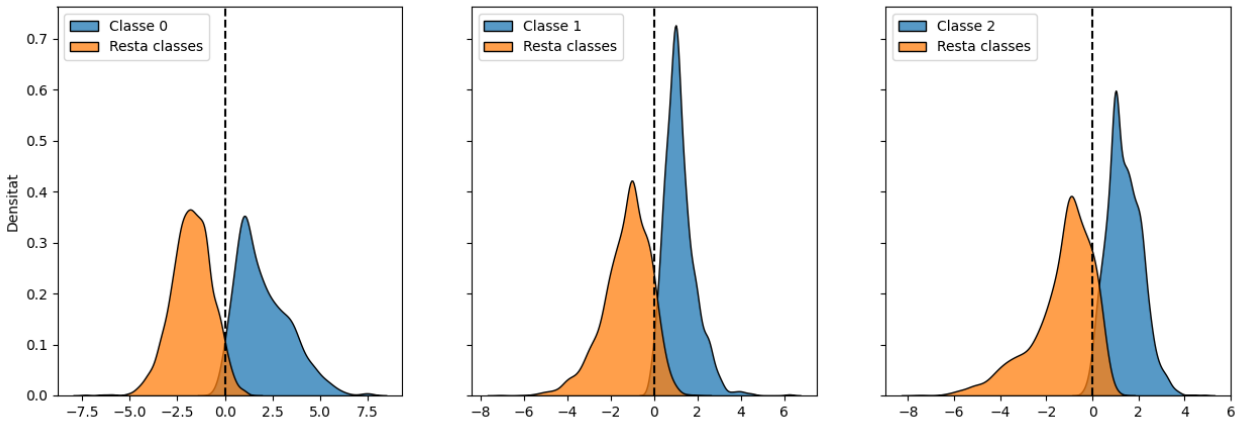

Figura 8 : KDE de la distribució de la funció solució del conjunt d'entrenament amb el classificador de nucli polinòmic

El gràfic és fàcil d'interpretar: l'eix horitzontal correspon a la distància d'un punt del conjunt de dades fins al límit de decisió del classificador de suport vectorial (indicat amb una línea discontínua vertical en x=0). Un model de màquines de suport vectorial té un marge funcional igual a 1 (recordem que aquesta és una restricció del problema d'optimització) i els vectors de suport són aquells punts que recauen dins d'aquest marge. Com podem observar en els gràfics de la **Figura 8** existeix certa fuita de punts d'una classe cap a l'altre costat del límit de decisió, això és degut a l'ús del paràmetre de regularització C que permet certes classificacions errònies per tal de què el model sigui bo generalitzant i no se sobreajusti. Observem també que es poden visualitzar els veritables positius o negatius i els falsos positius o negatius; els punts en el color blau que traspassen el límit de decisió són falsos negatius, i els de color vermell que el traspassen són falsos positius, la resta de punts que es mantenen en el costat adient són els veritables positius i veritables negatius. Tal com hem pogut comprovar a la matriu de confusió de la **Figura 7**, les classes 1 i 2 són les que al model li costa més diferenciar de les altres ja que la distribució vermella envaeix molt més el límit de decisió que en el gràfic de la classe 0.

Si el KDE pel conjunt de prova és similar a les del conjunt d'entrenament aleshores significa que el model és bo generalitzant sobre dades que encara no ha vist. Per últim, observem a la **Figura 9** que efectivament les gràfiques són bastant semblants a les del conjunt d'entrenament, per tant, reafirmem que el model és bo generalitzat a dades noves.

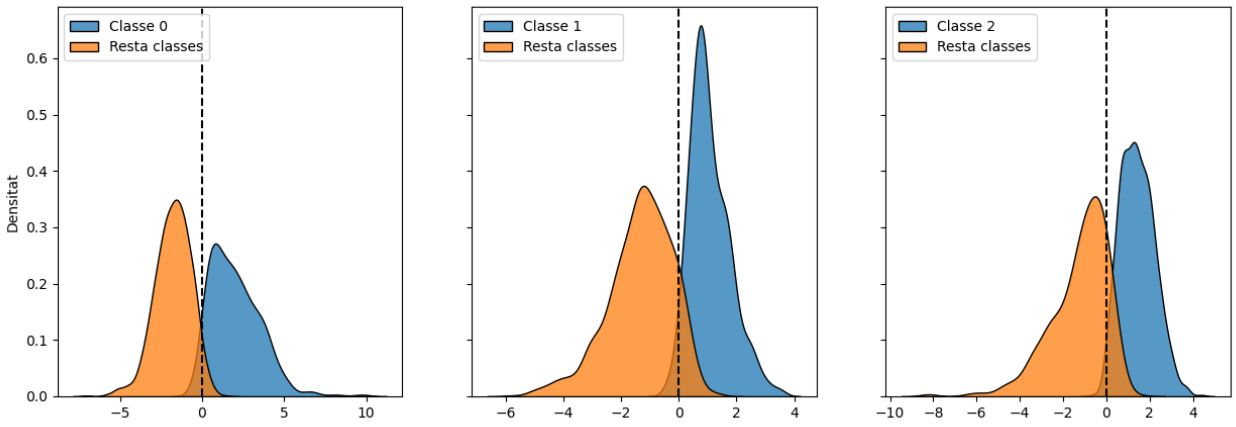

Figura 9 : KDE de la distribució dels marges funcionals del conjunt de prova amb el classificador de nucli polinòmic

4.4 Conclusions i futures investigacions

Tot i no haver assolit un 80% d'exactitud en el conjunt de prova, hem pogut posar en pràctica alguns conceptes teòrics i fer observacions interessants sobre els resultats dels models per, conjuntament, acabar de formar una idea clara sobre la seva metodologia i procés d'aprenentatge. A més cal dir que els resultats són bastant positius partint d'una base de dades relativament petita. En el context clínic i assistencial en el que es desenvolupa l'estudi mencionat a la introducció, l'exactitud assolida no és suficient com per posar en producció l'algoritme. Tot i així, per valorar l'alternativa de l'ús d'un mètode més potent i efectiu com poden ser les xarxes neuronals abans seria imprescindible comptar amb un conjunt de dades més gran. Les frases de les classes 0 i 2 encara segueixen

patrons fàcilment detectables per qualsevol model d'aprenentetge automàtic, en canvi, la classe 1 representa totes aquelles frases que no presenten cap afirmació, per tant, el seu vocabulari és molt més ampli. La falta de mostres de la classe 1 crec que provoca que al model li sigui més difícil diferenciar-la de les altres.

D'altra banda, tant important és optimitzar els paràmetres dels classificadors SVM com ho és generar uns *word embeddings* representatius. Per tant, seria interessant ser capaços d'avaluar d'una manera més precisa la qualitat dels vectors numèrics generats pel mètode *Skip-Gram* per tal de poder ajustar més els paràmetres del model. En [12] s'exposen dos tipus d'avaluacions: les intrínseques i les extrínseques. El primer tipus consisteix en avaluar els vectors sobre petites tasques específiques com per exemple completar analogies entre vectors tal com s'ha explicat al final de l'apartat de "Composicionalitat de Skip-Gram" al Capítol 3. Una avaluació intrínteca ha de retornar un valor que indiqui el rendiment dels vectors sobre la tasca d'avaluació que s'està emprant. Altres tècniques d'avaluació intrínseca de *word embeddings* es poden trobar en [14]. El segon tipus consisteix en avaluar els vectors sobre la tasca real que es vol dur a terme, en el nostre cas la classificació dels vectors en tres classes.

No obstant, el tamany del nostre conjunt de dades podria continuar essent un obstacle determinant a l'hora de millorar el model. Una solució: augmentar el nostre conjunt de dades. Hi ha diverses tècniques que es poden emprar, com la substitució de paraules per sinònims o paraules relacionades, la inversió de l'ordre de les paraules de les frases o la eliminació aleatòria de paraules. Caldria analitzar quins efectes tindrien ja que, per exemple, algunes no tenen en compte l'ordre de les paraules en les frases i el seu ús podria empitjorar el model. També hi ha l'alternativa, molt més costosa, de trobar alguna base de dades clíniques sobre metàstasis de colon i fer etiquetar les frases per professionals en la matèria, o directament generar a mà noves frases.

En resum, es considera que els resultats són satisfactoris tenint en compte el tamany del conjunt de dades amb el que s'ha treballat. Per futurs treballs es tindran en compte els resultats obtinguts i s'intentarà combinar l'algoritme amb altres indicadors o processos per tal de donar-li una utilitat ja que es creu potencialment que la pot tenir.

A Operadors lineals

En aquest apartat denotarem un espai de Hilbert com $\mathcal{H}$ i suposarem que és una generalització al cas de dimensió infinita de l'espai $\mathbb{R}^n$. Siguin $\mathcal{H}_1, \mathcal{H}_2$ dos espais de Hilbert, un operador lineal no és res més que una aplicació lineal $T : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$. Escriurem $T(x) := Tx$ per simplificar la notació.

Definició A.1. Definim la **norma** d'un operador lineal $T : \mathcal{H}_1 \longmapsto \mathcal{H}_2$ com:

$$\|T\| = \mathit{mín}\,\{M \in \mathbb{R}^+ \,:\, \|Tx\|_{\mathcal{H}_2} \leqslant M\|x\|_{\mathcal{H}_1}\, \forall x \in X\} \tag{A.1}$$

Diem que un operador lineal $T : \mathcal{H}_1 \longmapsto \mathcal{H}_2$ és **acotat** si $\|T\| < \infty$. Denotarem el conjunt d'operadors lineals i acotats entre $\mathcal{H}_1$ i $\mathcal{H}_2$ com $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$, i en el cas en què $\mathcal{H}_1 = \mathcal{H}_2 := \mathcal{H}$ escriurem simplement $\mathcal{L}(\mathcal{H})$.

Observació A.1. Notem que la desigualtat $\|Tx\|_{\mathcal{H}_2} \leqslant M\|x\|_{\mathcal{H}_1}$ és trivial per $x = 0$ i en cas contrari equival a què $\|Tx\|_{\mathcal{H}_2}/\|x\|_{\mathcal{H}_1} \leqslant M$, per tant podem escriure la expressió de norma d'un operador com

$$\|T\| = sup\left\{\frac{\|Tx\|_{\mathcal{H}_2}}{\|x\|_{\mathcal{H}_1}} \,:\, x \in X \setminus \{0\}\right\}.$$

Per $x \in X \setminus \{0\}$ tenim que $\|Tx\|_{\mathcal{H}_2}/\|x\|_{\mathcal{H}_1} = \|T(x/\|x\|_{\mathcal{H}_1})\|_{\mathcal{H}_2}$ i el conjunt $\{x/\|x\|_{\mathcal{H}_1} : x \in X \setminus \{0\}\}$ és la esfera unitat $S = \{z \in X \,:\, \|z\|_{\mathcal{H}_1} = 1\}$ i podem reescriure

$$\|T\| = sup\,\{\|Tz\|_{\mathcal{H}_2} \,:\, z \in S\} = sup\,\{\|Tz\|_{\mathcal{H}_2} \,:\, z \in X,\, \|z\|_{\mathcal{H}_1} = 1\}. \tag{A.2}$$

El conjunt que apareix en (A.2) augmenta si substituïm l'esfera unitat per la bola unitat $B = \{x \in X \,:\, |, \|x\|_{\mathcal{H}_1} \leqslant 1\}$, no obstant, el seu suprem no varia degut a què $x = \|x\|_{\mathcal{H}_1} z$ amb $z \in S$, i per tant, $\|Tx\|_{\mathcal{H}_2} = \|x\|_{\mathcal{H}_1}\|Tz\|_{\mathcal{H}_2} \leqslant \|Tz\|_{\mathcal{H}_2}$ i, per tant,

$$\|T\| = sup\,\{\|Tx\|_{\mathcal{H}_2} \,:\, x \in B\} = sup\,\{\|Tx\|_{\mathcal{H}_2} \,:\, x \in X,\, \|x\|_{\mathcal{H}_1} \leqslant 1\}. \tag{A.3}$$

Observació A.2. Per últim, demostrarem aquesta última expressió de la norma d'un operador T, que ens servirà per resultats posteriors

$$\|T\| = sup\,\{|\langle Tx, y\rangle_{\mathcal{H}_2}| \,:\, \|x\|_{\mathcal{H}_1} = \|y\|_{\mathcal{H}_2} = 1\}. \tag{A.4}$$

Siguin $x \in \mathcal{H}_1$ i $y \in \mathcal{H}_2$, si $\|x\|_{\mathcal{H}_1} = 1$ i $\|y\|_{\mathcal{H}_2} = 1$, aleshores per la desigualtat de *Cauchy-Schwarz*, tenim que $|\langle Tx, y\rangle_{\mathcal{H}_2}| \leqslant \|Tx\|_{\mathcal{H}_2}\|y\|_{\mathcal{H}_2} = \|Tx\|_{\mathcal{H}_2} \leqslant \|T\|$. Això implica que $\|T\| \geqslant sup\,\{|\langle Tx, y\rangle_{\mathcal{H}_2}| \,:\, \|x\|_{\mathcal{H}_1} = \|y\|_{\mathcal{H}_2} = 1\}$.
Per veure l'altra desigualtat prenem $x \in \mathcal{H}_1$ tal que $\|x\|_{\mathcal{H}_1} = 1$. Llavors si $Tx = 0$ clarament $\|Tx\|_{\mathcal{H}_2} \leqslant |\langle Tx, y\rangle_{\mathcal{H}_2}|$, suposem ara que $Tx \neq 0$, aleshores

$$\|Tx\|_{\mathcal{H}_2} = \left\langle Tx, \frac{Tx}{\|Tx\|_{\mathcal{H}_2}}\right\rangle_{\mathcal{H}_2} \leqslant sup\,\{|\langle Tx, y\rangle_{\mathcal{H}_2}| \,:\, \|x\|_{\mathcal{H}_1} = \|y\|_{\mathcal{H}_2} = 1\},$$

d'on deduïm que $\|T\| \leqslant sup\,\{|\langle Tx, y\rangle_{\mathcal{H}_2}| \,:\, \|x\|_{\mathcal{H}_1} = \|y\|_{\mathcal{H}_2} = 1\}$. En resum, comptem amb quatre expressions diferents per la norma d'un operador lineal T donades per (A.1), (A.2), (A.3) i (A.4).

Definició A.2. Sigui $T \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$. Un operador $T^* : \mathcal{H}_2 \longmapsto \mathcal{H}_1$ es diu que és un **adjunt** de T si:

$$\langle Tf, g\rangle = \langle f, T^*g\rangle, \forall f, g \in \mathcal{H}.$$

Definició A.3. Un operador $T \in \mathcal{L}(\mathcal{H})$ és **autoadjunt** si $T^* = T$. Un operador autoadjunt en $\mathcal{H}$ és **positiu** si

$$\langle Tf, f\rangle \geqslant 0, \forall f \in \mathcal{H}.$$

Notació A.1. A partir d'ara sempre que es parli d'un operador autoadjunt en $\mathcal{L}(\mathcal{H})$, no s'especificarà $\|\cdot\|_{\mathcal{H}}$ ja que s'entén que la norma o producte escalar està definit sempre sobre l'espai de Hilbert $\mathcal{H}$ al ser T un endomorfisme.

Proposició A.1. Si $T \in \mathcal{L}(\mathcal{H})$ és autoadjunt aleshores

$$\|T\| = sup\,\{|\langle Tx, x\rangle| \,:\, \|x\| = 1\}.$$

Demostració. Anomenem $\alpha = sup\,\{|\langle Tx, x\rangle| \,:\, \|x\| = 1\}$. Per (A.4) tenim que $\alpha \leqslant \|T\|$. Provem llavors l'altra desigualtat. Per cada $x \in \mathcal{H}$ amb $\|x\| = 1$ veurem que $\|Tx\| \leqslant \alpha$. Suposem que $Tx \neq 0$ (el cas $Tx = 0$ és trivial). Definim també $u = ax + \frac{1}{a}Tx$ i $v = ax - \frac{1}{a}Tx$ amb $a \in \mathbb{R} \setminus \{0\}$.

$$\begin{aligned}\langle Tu, u\rangle &= \langle aTx + \frac{1}{a}T^2x, ax + \frac{1}{a}Tx\rangle \\ &= a^2\langle Tx, x\langle + \frac{1}{a^2}\langle T^2x, Tx\rangle + \langle Tx, Tx\rangle + \langle T^2x, x\rangle \\ &= a^2\langle Tx, x\rangle + \frac{1}{a^2}\langle T^2x, Tx\rangle + 2\langle Tx, Tx\rangle,\end{aligned}$$

on a la última igualtat estem aplicant que $\langle T^2x, x\rangle = \langle Tx, T^*x\rangle = \langle Tx, Tx\rangle$ per ser T autoadjunt.
Anàlogament, $\langle Tv, v\rangle = a^2\langle Tx, x\rangle + \frac{1}{a^2}\langle T^2x, Tx\rangle - 2\langle Tx, Tx\rangle$. Després, $\|Tx\|^2 = \frac{1}{4}(\langle Tu, u\rangle - \langle Tv, v\rangle) \leqslant \alpha\frac{1}{4}(\|u\|^2 + \|v\|^2)$. Calculem $\|u\|^2$ i $\|v\|^2$:

$$\begin{aligned}\|u\|^2 = \langle u, v\rangle &= \langle ax + \frac{1}{a}Tx, ax + \frac{1}{a}Tx\rangle \\ &= a^2\|x\|^2 + \frac{1}{a^2}\|Tx\|^2 + 2\langle x, Tx\rangle.\end{aligned}$$

De la mateixa manera obtenim que $\|v\|^2 = a^2\|x\|^2 + \frac{1}{a^2}\|Tx\|^2 - 2\langle x, Tx\rangle$. Així, $\|Tx\|^2 \leqslant \alpha\frac{1}{4}(\|u\|^2 + \|v\|^2) \leqslant \alpha\frac{1}{2}(a^2\|x\|^2 + \frac{1}{a^2}\|Tx\|^2)$. Si prenem $a^2 = \|Tx\|$ aleshores deduïm que $\|Tx\| \leqslant \alpha$ com volíem. □

Definició A.4. Un operador lineal $T \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ és **compacte** si tota successió acotada $(x_n)_n$ conté una subsuccessió $(x_{n_k})_k$ tal que $(Tx_{n_k})_k$ convergeix.

Proposició A.2. Sigui $(T_n)_n$ una successió d'operadors compactes. Si $T \in \mathcal{L}(\mathcal{H})$ i $lim_{n\to\infty} \| T_n - T \| = 0$, aleshores T és compacte.

Demostració. Sigui $(x_n)_n$ una successió amb $\|x_n\| \leqslant 1$, per $n \in \mathbb{N}$. Hem de trobar una subsuccessió $(x_{n_k})_k$ tal que $(T_{n_k})_k$ sigui convergent. Seguirem un procediment diagonal.

Sigui $(x_{1,n})_n$ una subsuccessió de $(x_n)_n$ tal que $(T_1 x_{1,n})_n$ convergeix, digue'm a u_1.

Sigui $(x_{2,n})_n$ una subsuccessió de $(x_{1,n})_n$ tal que $(T_2 x_{2,n})_n$ convergeix, digue'm a u_2.

Seguint aquest procediment, suposem que escollim $(x_{k,n})_n$ tal que $(T_k x_{k,n})_n$ convergeix a u_k, sigui $(x_{k+1,n})_n$ una subsuccessió de $(x_{k,n})_n$ tal que $(T_{k+1} x_{k+1,n})_n$ convergeix a u_{k+1} $(k \in \mathbb{N})$.

Considerem la successió diagonal $(x_{k,k})_k$. Està clar que $lim_{n\to\infty} T_k x_{n,n} = u_k$. El nostre objectiu ara és provar que $(Tx_{k,k})_k$ és de Cauchy. Donat $\epsilon > 0$, fixem un índex $k \in \mathbb{N}$ tal que $\|T_k - T\| < \epsilon/3$, tenim que per qualsevols $m, n \in \mathbb{N}$,

$$\begin{aligned}\|Tx_{m,m} - Tx_{n,n}\| &\leqslant \|(T - T_k)x_{m,m}\| + \|T_k x_{m,m} - T_k x_{n,n}\| + \|(T - T_k)x_{n,n}\| \\ &\leqslant \frac{2\epsilon}{3} + \|T_k x_{m,m} - T_k x_{n,n}\|.\end{aligned}$$

Com $lim_{n\to\infty} T_k x_{n,n} = u_k$, aleshores existeix $N \in \mathbb{N}$ tal que $\|T_k x_{m,m} - T_k x_{n,n}\| < \epsilon/3$ $(m, n \in \mathbb{N},\ m, n \geqslant N)$. □

Definició A.5. Sigui $T \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ un operador lineal. Anomenem espai imatge o **rang** del operador T al subespai de $\mathcal{H}_2$ definit com $Im\,T = \{Tx : x \in \mathcal{H}\}$. Diem que T és de **rang finit** si la dimensió del seu espai imatge és finita.

Proposició A.3. Sigui $T \in \mathcal{L}(\mathcal{H})$ un operador de rang finit, aleshores T és compacte.

Demostració. Sigui $n = dim(Im\,T)$ i $\{v_1, ..., v_n\}$ una base ortonormal de $Im\,T$, per cada $x \in \mathcal{H}$ tenim que

$$Tx = \sum_{k=1}^{n} \alpha_k(x) v_k, \tag{A.5}$$

on $\alpha_1(x), ..., \alpha_n(x)$ són escalars unívocament determimats per x que satisfan

$$\alpha_k(x) = \langle Tx, v_k\rangle = \langle x, T^* v_k\rangle,\ \forall k \in \{1, ..., n\}$$

Per tant podem reescriure (A.5) com:

$$Tx = \sum_{k=1}^{n} \langle x, T^* v_k\rangle v_k$$

Veiem que l'operador T és compacte veient que ho és l'operador $\tilde{T}x = \langle x, y\rangle z$, per $x, y, z \in \mathcal{H}$. Suposem que $(x_n)_n$ satisfà $\|x_n\| \leqslant 1$ per tot $n \in \mathbb{N}$. Com $|\langle x_n, y\rangle| \leqslant \|y\|$, la successió d'escalars $(x_n, y)_n$ conté una parcial $(x_{n_k}, y)_k$ convergent digue'm a λ, aleshores $lim_{k\to\infty} \tilde{T} x_{n_k} = lim_{k\to\infty} \langle x_{n_k}, y\rangle z = \lambda z$. És fàcil deduïr de la Definició A.4 que la suma d'operadors compactes és també un operador compacte. □

Definició A.6. Sigui $T \in \mathcal{L}(\mathcal{H})$ un operador positiu i sigui $\{f_i\}_{i\in A}$ una base ortonormal

de $\mathcal{H}$. Definim la **traça** de T, $tr(T)$, com la sèrie de termes no negatius:

$$tr(T) = \sum_{i \in A} \langle T f_i, f_i \rangle.$$

Definició A.7. Diem que $T \in \mathcal{L}(\mathcal{H})$ és un **operador de Hilbert-Schmidt** si $tr(\|T\|^2) = tr(T^*T) < \infty$.

Proposició A.4. Tot operador de Hilbert-Schmidt és compacte.

Demostració. Sigui T un operador de Hilbert-Schmidt i sigui $\mathcal{B} = \{f_i\}_{i \in A}$ una base ortonormal de $\mathcal{H}$.

$$\begin{aligned}
\|T\|^2 &= sup\{\, \|Tx\|^2 \,:\, \|x\| \leqslant 1\} \\
&= sup\{\, |\langle Tx, Tx\rangle| \,:\, \|x\| \leqslant 1\} \\
&= sup\{\, |\langle TT^*x, x\rangle| \,:\, \|x\| \leqslant 1\} \qquad (T^* \text{ és un adjunt de } T) \\
&\leqslant \sum_{i \in A} \langle TT^* f_i, f_i \rangle \\
&= \sum_{i \in A} \|T f_i\|^2
\end{aligned}$$

Sigui ara $\epsilon > 0$. Triem $f_1, ..., f_n \in \mathcal{B}$ tals que:

$$\sum_{f_i \notin \{f_1, ..., f_n\}} \|T f_i\|^2 < \epsilon^2$$

Considerem la projecció P_n de $\mathcal{H}$ en el subespai generat per $\{f_1, ..., f_n\}$ i definim l'operador de rang finit (l'espai imatge del cual té dimensió finita) $B_n = TP_n$. Aleshores:

$$\|T - B_n\|^2 = \| T(I - P_n)\|^2 \leqslant \sum_{f_i \notin \{f_1, ..., f_n\}} \|T f_i\|^2 < \epsilon^2$$

Com $(B_n)_{n \in \mathbb{N}}$ són operadors de rang finit aleshores per la Proposició A.3 són compactes, i per tant es dedueix de la Proposició A.2 que T és compacte. □

B Teoria espectral d'operadors autoadjunts i compactes

Definició B.1. Sigui $T \in \mathcal{L}(\mathcal{H})$. Diem que un escalar $\lambda \in \mathbb{R}$ és un valor propi si existeix $x \in \mathcal{H} \setminus \{0\}$ tal que $Tx = \lambda x$. Al vector x se l'anomena vector propi de valor propi λ.

Proposició B.1. Els vectors propis associats a diferents valors propis d'un operador autoadjunt són ortogonals.

Demostració. Siguin $\lambda, \mu \in \mathbb{R}$ dos valors propis diferents de T i siguin $x, y \neq 0$ els seus vectors propis associats. Aleshores:

$$\lambda\langle x, y\rangle = \langle Tx, y\rangle = \langle x, Ty\rangle = \mu\langle x, y\rangle.$$

Com $\lambda \neq \mu$ tenim que $\langle x, y\rangle = 0$. □

Teorema B.1. Sigui $T \in \mathcal{L}(\mathcal{H})$ un operador autoadjunt i compacte, aleshores almenys un dels escalars $\|T\|$ o $-\|T\|$ és un valor propi de l'operador T.

Demostració. Suposarem que $T \neq 0$ (el cas $T = 0$ és trivial). Per la Proposició A.1 sabem que $\|T\| = sup\,|\langle Tx, x\rangle| : x \in X, \|x\| = 1\}$, per tant, existeix una successió $(x_n)_n$ en $\mathcal{H}$ amb $\|x_n\| = 1$ per tot $n \in \mathbb{N}$ tal que

$$lim_{n\to\infty}|\langle Tx_n, x_n\rangle| = \|T\|.$$

Sigui $\alpha \in \mathbb{R}$ amb $|\alpha| = \|T\|$ i $(x_{n_k})_k$ una subsuccessió tal que $lim_{k\to\infty}\langle Tx_{n_k}, x_{n_k}\rangle = \alpha$. Seguidament veiem que $(Tx_{n_k} - \alpha x_{n_k}) \xrightarrow[k\to\infty]{} 0$.

$$\begin{aligned}\|Tx_{n_k} - \alpha x_{n_k}\|^2 &= \|Tx_{n_k}\|^2 + \alpha^2\|x_{n_k}\|^2 - \alpha\langle Tx_{n_k}, x_{n_k}\rangle - \alpha\langle x_{n_k} Tx_{n_k}\rangle \\ &= \|Tx_{n_k}\|^2 + \alpha^2 - 2\alpha\langle Tx_{n_k}, x_{n_k}\rangle \\ &\leqslant 2\alpha^2 - 2\alpha\langle Tx_{n_k}, x_{n_k}\rangle \xrightarrow[k\to\infty]{} 0.\end{aligned}$$

Ara, com T és compacte i $(x_{n_k})_k$ és una successió de vectors unitaris, aleshores existeix una subsuccessió de $(Tx_{n_k})_k$, la denotem per $(Tx_{n_{k_m}})_m$ tal que $lim_{m\to\infty}Tx_{n_{k_m}} = y \in \mathcal{H}$. Veiem que $(x_{n_{k_m}})_m$ convergeix a $\frac{1}{\alpha}y$:

$$\|\alpha x_{n_{k_m}} - y\| \leqslant \|\alpha x_{n_{k_m}} - Tx_{n_{k_m}}\| + \|Tx_{n_{k_m}} - y\| \xrightarrow[m\to\infty]{} 0.$$

Per tant, $x_{n_{k_m}} \xrightarrow[m\to\infty]{} \frac{y}{\alpha}$. I per la continuïtat de T:

$$y = lim_{m\to\infty}Tx_{n_{k_m}} = \frac{1}{\alpha}Ty \iff Ty = \alpha y.$$

A més, $y \neq 0$ ja que $\|y\| = lim_{m\to\infty}\|\alpha x_{n_{k_m}}\| = \|T\| \neq 0$. Per tant α és un valor propi de T. □

Teorema B.2. (Teorema espectral per operadors autoadjunts i compactes) Sigui $T \in \mathcal{L}(\mathcal{H})$ un operador compacte i autoadjunt aleshores existeix un sistema ortonormal de vectors propis $\{x_n\}_{n\in\mathbb{N}}$ i una corresponent successió de valors propis $\{\lambda_n\}_{n\in\mathbb{N}}$, tal que per cada $x \in \mathcal{H}$:

$$Tx = \sum_{n=1}^{\infty} \lambda_n \langle x, x_n\rangle x_n.$$

Demostració. Sigui $\mathcal{H}_1 = \mathcal{H}$ i $T_1 = T$, aplicant el Teorema B.1 existeix un vector propi x_1 de valor propi λ_1 tals que $\|x_1\| = 1$ i $|\lambda_1| = \|T_1\|$. Busquem el següent vector propi en $\{x_1\}^\perp$ ja que per la Proposició B.2 sabem que dos vectors propis de valors propis diferents són ortogonals.

Definim $\mathcal{H}_2 = \{x_1\}^\perp$ i $T_2 = T|_{\mathcal{H}_2}$. $\mathcal{H}_2$ és un espai tancat per ser l'ortogonal d'un conjunt i com $\{x_1\}$ és T_2-invariant aleshores $\{x_1\}^\perp$ és també T_2-invariant, per tant, $T_2 \in \mathcal{L}(\mathcal{H}_2)$. La propietat emprada sobre espais invariants és fàcil de veure ja que si $v \in \{x_1\}^\perp$, aleshores $\langle Tu, v\rangle = 0 = \langle u, Tv\rangle$ per tot $u \in \{x_1\}$ ja que també $Tu \in \{x_1\}$. Ara, com T_1 és compacte aleshores T_2 ho serà també. A més, T_2 és autoadjunt doncs $\langle T_2x, x\rangle = \langle Tx, x\rangle \in \mathbb{R}$ per tot $x \in \mathcal{H}_2$. Si $T_2 \neq 0$ tornem a aplicar el Teorema B.1 i obtenim que existeix $\lambda_2 \in \mathbb{R}$ i $x_2 \in \mathcal{H}_2$ amb $\|x_2\| = 1$ tal que $T_2x_2 = \lambda_2x_2$ i $|\lambda_2| = \|T_2\|$, i així, $|\lambda_2| \leqslant |\lambda_1|$.

Prenem $\mathcal{H}_3 = \{x_1, x_2\}^\perp$ i $T_3 = T|_{\mathcal{H}_2}$ i reiterem el procès fins arribar a $T_n = 0$ o bé una quantitat infinit numerable de vegades. Així mateix obtindrem una successió finita o numerable de valors propis $\{\lambda_n\}_{n\in\mathbb{N}}$ tals que $|\lambda_n| \leqslant |\lambda_{n+1}|$ $\forall n \in \mathbb{N}$, i un conjunt $\{x_n\}_{n\in\mathbb{N}}$ de vectors propis ortogonals.

Per últim, provem la representació de l'operador. Sigui $x \in \mathcal{H}$ definim $y_n = x - \sum_{k=1}^n \langle x, x_k\rangle x_k$. Distingim dos casos segons si la successió $\{\lambda_n\}_{n\in\mathbb{N}}$ és finita o infinita numerable:

- Cas 1: Si existeix $n \in \mathbb{N}$ tal que $T_n = 0$, aleshores $T_ny_n = 0 = Tx - \sum_{k=1}^n \langle x, x_k\rangle Tx_k$ i, com a conseqüència, aplicant que els $\{x_n\}$ són vectors propis tenim que $Tx = \sum_{k=1}^n \lambda_k\langle x, x_k\rangle x_k$.
- Cas 2: $T_n \neq 0$ per tot $n \in \mathbb{N}$, aleshores

$$\begin{aligned}\|Tx - \sum_{k=1}^{n} \lambda_k\langle x, x_k\rangle x_k\| = \|T_ny_n\| \geqslant \|T_n\|\|y_n\| = |\lambda_n|\|y_n\| \\ = |\lambda_n|(\|x\|^2 - \sum_{k=1}^{n} |\langle x, x_k\rangle|^2)^{1/2} \geqslant |\lambda_n|\|x\| \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Per tant, $Tx = \sum_{k=1}^{\infty} \lambda_k\langle x, x_k\rangle x_k$. □

C Implementació

Tot el codi corresponent als resultats del Capítol 4 es pot trobar escanejant el següent codi QR, que està enllaçat amb un repositori de *GitHub*:

o també enganxant el següent link al navegador:

`https://github.com/ElenaBlanco5/PLN-Word2Vec-SVC`.

Referències

[1] Cristianini N.; Shawe-Taylor J.: *An introduction to support vector machines: and other kernel-based methods*. Cambridge University Press, 2000. ISBN 9780511801389.

[2] Steinwart I.; Christmann A.: *Support Vector Machines*. Editors: Jordan M.; Kleinberg J.; Schölkopf B. Springer, 2008. ISBN 9780387772417.

[3] Chollet F.: *Deep learning with Python*. Segona edició. Manning Publications, 2021. ISBN 9781617296864.

[4] Aizerman M.; Braverman E.; Rozonoer L.: *Theoretical foundations of the potential function method in pattern recognition learning*. 1964, pp. 821-837.

[5] Thickstun J.: *Mercer's Theorem*. Universitat de Washington. Disponible a: `<homes.cs.washington. edu/ thickstn/docs/mercer.pdf>`, 2019. pp. 3-4.

[6] Steinwart, I.: *On the influence of the kernel on the consistency of support vector machines*. Volum: 2. Disponible a: `<jmlr.org/papers/volume2/steinwart01a/steinwart01a.pdf>`, novembre de 2001. pp. 74-75.

[7] Ghojogh B., Ghodsi A., Karray F., Crowley M.: *Reproducing Kernel Hilbert Space, Mercer's Theorem, Eigenfunctions, Nystrom Method, and Use of Kernels in Machine Learning: Tutorial and Survey*, `arXiv:2106.08443v1`, juny de 2021.

[8] Kuhn H.; Tucker A.: *Nonlinear programming*. A: Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics, pp.481-492. University of California Press, 1951.

[9] Rudin, C.: *Kernels*. Notes del curs "Prediction: Machine Learning And Statistics" de l'Institut de Tecnologia de Massachusetts. `MIT 15.097`, 2012. pp. 10-14.

[10] Zhang B.; Srihari S. N.: *Properties of Binary Vector Dissimilarity Measures*. Departament d'Enginyeria i Informàtica, Universitat Estatal de Nova York, Buffalo. Disponible a: `<cedar.buffalo.edu/papers/articles/CVPRIP03_propbina.pdf>`, 2003. pp. 1-2.

[11] Chaubard F.; Fang M.; Genthial G.; Rohit M.; Socher R.: *CS224n: Natural Language Processing with Deep Learning Lecture Notes: Part I*. Notes del curs "CS224N: Natural Language Processing with Deep Learning" de l'Universitat de Stanford. Disponible a: `<cs224d.stanford.edu/lecture_notes/LectureNotes1.pdf>`, març 2016.

[12] Chaubard F.; Fang M.; Genthial G.; Rohit M.; Socher R.: *CS224D: Deep Learning for NLP Lecture Notes: Part II*. Notes del curs "CS224d: Deep Learning for Natural Language Processing" de l'Universitat de Stanford. Disponible a: `<cs224d.stanford.edu/lecture_notes/LectureNotes2.pdf>`, març 2016.

[13] Gittens A.; Achlioptas D.; Mahoney M. W.: *Skip- Gram- Zipf+ Uniform= Vector Additivity*. A: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canadà. Association for Computational Linguistics. `aclanthology:P17-1007`, juliol 2017.

[14] Gladkova A.; Drozd A.: *Intrinsic Evaluations of Word Embeddings: What Can We Do Better?*. A: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. Association for Computational Linguistics, Berlín, Alemania. aclanthology:W16-2507, agost 2016.

[15] Cherkassky V.; Dhar S.: *Simple Method for Interpretation of High-Dimensional Nonlinear SVM Classification Models*. A: Proceedings of The 2010 International Conference on Data Mining, Las Vegas, Nevada. Disponible a: <researchgate.net/publication/220705021_Simple_Method_for_Interpretation_of_High-Dimensional_Nonlinear_SVM_Classification_Models>, gener 2010.

[16] Marrero I.: *Clases de operadores II*. Notes del curs: Teoría de operadores. Departament d'Anàlisi Matemàtica de la Universitat de La Laguna. Disponible a: <campusvirtual.ull.es/ocw/pluginfile.php/18433/mod_resource/content/3/tema3/3-clasoperadores-II.pdf>. 2011-2012. pp. 9-18.

[17] Fernández B.: *Teoría espectral de operadores compactos y autoadjuntos en espacios de Hilbert*. Treball dirigit per: Fernando Cobos Díaz. Departament d'Anàlisi Matemàtica de l'Universitat Complutense de Madrid. Disponible a: <ucm.es/data/cont/docs/193-2016-10-17-MemorioBecaColaboraci%C3%B3nBlancaFernandezBesoy.pdf>. 2015-2016. pp. 23-27.