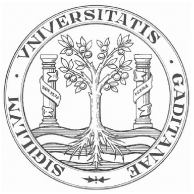


UNIVERSIDAD DE CÁDIZ

FACULTAD DE CIENCIAS

GRADO EN MATEMÁTICAS

PROBLEMAS DE CLASIFICACIÓN MULTICLASES CON SVM

Trabajo de fin de grado presentado por

Marina Aquino del Valle

Tutor: Dr. Nombre apellidos del tutor

Firma de la alumna

Firma de la tutora

Puerto Real, Cádiz, Julio de 2.022

Abstract

A mis padres.

Resumen

Agradecimientos

Pepito Pérez

abril 2023

Índice general

CAPITULO

1

Introduction

Uno de los retos que la humadidad siempre ha querido alcanzar es replicar la inteligencia que caracteriza a nuestra especie, jugando un papel vital para esta hazaña el Machine Learning. Esta disciplina gracias a diversos algoritmos y datos aportados pretende imitar la forma de pensar del ser humano, siendo capaz de modificarse y desarrollarse por sí mismo. Sus aplicaciones son sumamente extensas, y es utilizada en situaciones en las que no nos percatamos de ello, como cuando queremos encontrar la ruta más rápida a nuestro destino y usamos el GPS o al usar traducciones automáticas aplicando PLN (Procesamiento de Lenguaje Natural).

La primera constancia que hay de un sistema de Machine Learning es en 1950 en el conocido Test de Turing, creado por Alan Turing, cuyo objetivo era comprobar si una máquina poseía comportamiento inteligente. Dos años después en 1952 Arthur Samuel escribe el primer algoritmo para jugar a las damas, que a pesar de ser un juego con normas relativamente sencillas, la dificultad reside en adaptarse a los movimientos que realiza el contrincante.

El término Machine Learning nace en la conferencia de Darmouth en 1956, donde se asentarían las bases de lo que se conoce actualmente como Inteligencia Artificial. Seguido de varios años de muchos avances en la materia, a mediados de los 70, el campo sufrió lo que se conoce como el primer “Invierno”, un periodo en el cual hubieron escasos progresos

en este campo, esta etapa duró 6 años desde 1974 hasta 1980. Posteriormente, aparecieron los sistemas expertos basados en reglas, siendo introducido en 1981 el concepto de "Explanation Based Learning" (EBL), este consiste en el análisis de datos de entrenamiento por un ordenador para crear reglas que descarten datos menos importantes. Actualmente, recién salidos del segundo "Invierno" sufrido en los años 1987-1993, el aumento de la potencia de cálculo y la inmensa disponibilidad de datos ha hecho posible que el Machine Learning se encuentre en una etapa explosiva, en la cual los avances son a pasos agigantados.

Existen diferentes aproximaciones al Machine Learning, pero en este trabajo se tratará esta materia bajo un prisma de programación matemática. Dentro del Machine Learning nos encontramos diferentes formas de aprendizaje, Supervisado, No Supervisado y por Refuerzo.

- En el aprendizaje supervisado los datos están etiquetados, esto es, se conoce el valor del atributo objetivo permitiendo así que el algoritmo desarrolle una función que prediga dicho atributo para datos nuevos en los que se desconozca.
- El aprendizaje no supervisado no parte de datos etiquetado, no predicen valores, sino que son usados para agrupar datos y buscar grupos similares entre los datos proporcionados.
- En el aprendizaje por refuerzo la máquina se automodifica a partir de experimentar con los datos, tomará decisiones en función de recompensas y penalizaciones que haya recibido por elecciones tomadas.

En este trabajo se desarrollarán modelos de aprendizaje denominados como SVM (Support Vector Machine) y SVM-Multiclase, explicando los diferentes algoritmos que existen. Por último también se realizarán diversos experimentos computacionales con los algoritmos del modelo SVM-Multiclase que se expondrán en la Sección 3

CAPITULO

2

SUPPORT VECTOR MACHINE

Dentro del aprendizaje supervisado se encuentran un conjunto de modelos que se conocen como SVM (Support Vector Machine) que fueron propuestos por Vladimir Vapnik y sus colaboradores, [6] donde se construye un clasificador de datos binario, es decir, en dos grupos distintos.

El SVM parte de un conjunto de datos, a los que pasaremos a denotar como observaciones, que se expresan en forma de dupla como $(\mathbf{x}_i, y_i)$ para $i = 1, \ldots, n$ con $\mathbf{x}_i \in \Omega \subset \mathbb{R}^m$ e y_i perteneciente a un conjunto formado por dos elementos Y=$\{$-1,1$\}$. Dicho conjunto puede dividirse en dos subgrupos definidos por los valores de y_i, los cuales verifican que $y = 1$, o bien, $y = -1$. A estos dos subgrupos se les pasará a denominar clases, siendo el objetivo del SVM hallar un hiperplano que las separe.

En sus orígenes, el SVM fue desarrollado como un clasificador lineal. Posteriormente mediante el uso de la función kernel, la cual se definirá en la Subsección 2.3 , se pudieron aplicar los casos en donde el separador construido no se correspondía con una función lineal, [11].

Las aplicaciones del SVM se encuentran en diversos campos como medicina ([23],[25]), biología ([8],[19]) y la industria ([24],[20]), en general, este tipo de modelos es usado para resolver problemas de clasificación y regresión. A continuación se expondrá un ejemplo

del empleo del SVM. Suponga que un hospital desea hacer un estudio para diagnosticar el cáncer de mama, en el cual participan 10000 pacientes y esta institución poseía información clínica acerca de estos. Los pacientes se dividían en dos grupos esencialmente, en uno de ellos los integrantes padecían cáncer de mama y en el otro no tenían dicha enfermedad. El objetivo era ver que características podían ayudar a los sanitarios a diagnosticar con mayor eficacia el cáncer de mama. En este caso, las clases corresponderían con adolecer el cáncer de mama o no poseerlo.

El SVM posee diferentes formulaciones en función de la relación que tengan las clases entre sí. Estas relaciones se pueden clasificar esencialmente en dos categorías:

- Separables linealmente: Si se puede definir un hiperplano de separación que permita separar ambas clases entre sí.
- No separables linealmente: Si no existe hiperplano de separación que permita separar una clase de la otra.

A lo largo de esta sección, se expondrán las diferentes formulaciones existentes y el desarrollo de las mismas. Se incluirá un apartado para el caso en el que las clases sean cuasi-separables, esto es, que las clases se puedan separar mediante un hiperplano, pero que existan algunos errores de clasificación. Esencialmente es similar al caso linealmente separables, pero se utilizarán unas variables de holgura que permitirán penalizar a las observaciones que no puedan clasificarse correctamente. Empezaremos por el primero de los casos previamente expuestos.

2.1 CASO LINEALMENTE SEPARABLE

En esta sección se buscará un hiperplano que separe las observaciones en las dos clases existentes, sin encontrarnos errores de clasificación. Este caso es conocido como linealmente separable. Este hiperplano viene dado por la siguiente expresión

$$H(\mathbf{x}) = (v_1 x_1 + ... + v_m x_m) + a =< \mathbf{v}, \mathbf{x} > +a, \tag{2.1}$$

donde $a \in \mathbb{R}$ es una constante, $\mathbf{v} \in \mathbb{R}^m$ es el vector normal que define el hiperplano y $< \cdot, \cdot >$ es el producto escalar entre dos vectores.

El hiperplano nos dividirá el espacio en dos zonas, al sustituir en este los valores de $(\mathbf{x}_i, y_i)$ el resultado será positivo o negativo en función de en que zona se encuentre. En este trabajo para facilitar la notación, en la zona positiva se encontrarán las observaciones con valor $y_i = 1$ y en la zona negativa las observaciones con $y_i = -1$. Por lo tanto se tienen las dos siguientes expresiones

$$< \mathbf{v}, \mathbf{x}_i > + \ a \geq 0, \text{ si } y_i = 1, \quad \text{para} \quad i = 1, \ldots, n,$$

$$< \mathbf{v}, \mathbf{x}_i > + \ a \leq 0, \text{ si } y_i = -1, \quad \text{para} \quad i = 1, \ldots, n.$$

Las ecuaciones anteriores, se pueden comprimir en una sola añadiendo el término y_i, obteniendo la expresión

$$y_i(< \mathbf{v}, \mathbf{x}_i > + a) \geq 0, \qquad \text{para } i = 1, \ldots, n.$$

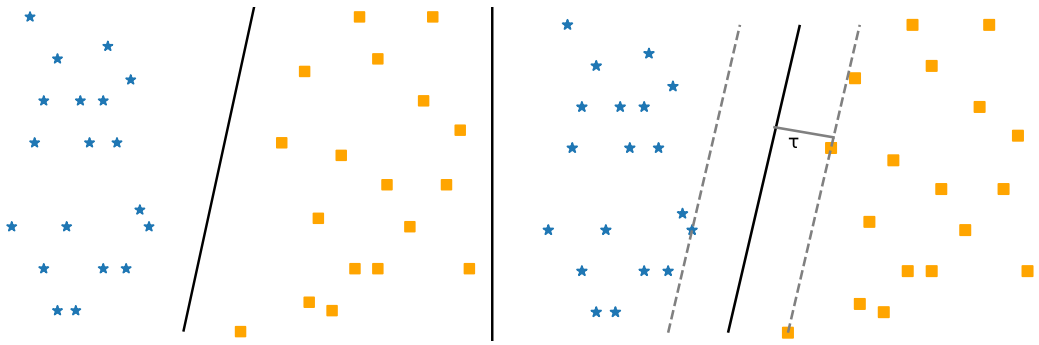


Figura 2.1: En la Figura 2.1a se puede observar un hiperplano que separa dos clases. Por otro lado la Figura 2.1b corresponde con una representación gráfica del concepto de margen dado en la Definición 2.1.1

A continuación se define el concepto de margen, el cual se usará para formular el problema de hallar el hiperplano que separe las clases.

DEFINICIÓN 2.1.1: Margen:

Sea Ω un conjunto de observaciones y H un hiperplano que separe dichas observaciones en dos clases distintas. Llamaremos margen (τ) de un hiperplano H a la distancia mínima que hay entre dicho hiperplano y el elemento más cercano de las clases que separa. Un ejemplo gráfico de esto viene en la Figura 2.1b

Hay infinitos hiperplanos que separen las dos clases. El SVM busca el hiperplano que verifique que su margen sea máximo, por lo tanto la distancia de la observación más cercana al hiperplano será mínima. Antes de definir la distancia de un punto a un hiperplano, explicaremos el concepto de norma dual, ya que será usado posteriormente.

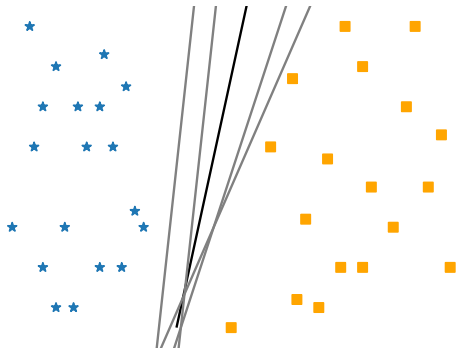


Figura 2.2: En la figura se representan múltiples hiperplanos que separan dos clases

DEFINICIÓN 2.1.2: Norma Dual

Sea $\|\cdot\|$ la norma definida en el espacio donde se encuentran los elementos de la dupla $(\mathbf{x}_i, y_i)$ con $\mathbf{x}_i \in \Omega \subset \mathbb{R}^m$ e $y_i \in \{-1, 1\}$. Definimos la norma dual $\|\cdot\|^\circ$ como

$$\|\mathbf{z}\|^\circ = \max\{< \mathbf{z}, \mathbf{x} >, \|\mathbf{x}\| = 1\}.$$

A continuación daremos la definición de distancia de un punto a un hiperplano.

DEFINICIÓN 2.1.3: Distancia de un punto a un hiperplano

Sea $\|\cdot\|$ una norma definida en $\mathbb{R}^m$. La distancia de un punto $\mathbf{x}$ y un hiperplano $H =< \mathbf{v}, \mathbf{x} > +a$ es la menor distancia entre este punto y los infinitos puntos que constituyen H, lo cual se calcula como

$$D(\mathbf{x}, H) = \frac{|H(\mathbf{x})|}{\|\mathbf{v}\|^\circ},$$

donde $\|\cdot\|^\circ$ denota a la norma dual de $\|\cdot\|$.

En SVM el objetivo es encontrar un hiperplano que separe ambas clases y que maximice la distancia al elemento más cercano, es decir, que maximice el margen, por tanto el problema a resolver sería el siguiente

$$\begin{array}{ll} \max \ \min\limits_{i} & \dfrac{y_i(<\mathbf{v},\mathbf{x}_i>+a)}{\|\mathbf{v}\|^{\circ}} \\ \text{s.a:} & y_i(<\mathbf{v},\mathbf{x}_i>+a)\geq 0, \qquad i=1,\ldots,n. \end{array}$$

La función que calcula la distancia mínima, es decir,

$$D(\mathbf{x},H)=\min_i \frac{y_i H(\mathbf{x}_i)}{\|\mathbf{v}\|^{\circ}}=\min_i \frac{y_i(<\mathbf{v},\mathbf{x}_i>+a)}{\|\mathbf{v}\|^{\circ}},$$

es una función homogénea de grado 0, por tanto $D(\delta\mathbf{x},\delta H)=D(\mathbf{x},H)$ para $\delta>0$, por tanto la solución del problema anterior es equivalente a la solución del siguiente problema

$$\begin{array}{ll} \max \ \min\limits_{i} & \dfrac{\delta y_i(<\mathbf{v},\mathbf{x}_i>+a)}{\|\delta\mathbf{v}\|^{\circ}} \\ \text{s.a:} & \delta y_i(<\mathbf{v},\mathbf{x}_i>+a)\geq 0, \qquad i=1,\ldots,n. \end{array}$$

Busquemos una cota para la restricción anterior. Existen dos posibilidades, que $y_i=1$ o $y_i=-1$, pasamos a desarrollar ambos casos

- Si $y_i=1$, sustituyendo se obtiene que $\delta(<\mathbf{v},\mathbf{x}_i>+a)\geq 0$, por tanto se puede decir que existe un $\epsilon_1\in\mathbb{R}$ con $\epsilon_1\geq 0$ tal que $\delta(<\mathbf{v},\mathbf{x}_i>+a)\geq\epsilon_1$.
- Si $y_i=-1$ entonces se tiene que $-\delta(<\mathbf{v},\mathbf{x}_i>+a)\geq 0$, o lo que es lo mismo, existe un $\epsilon_2\in\mathbb{R}$ con $\epsilon_2\geq 0$ tal que $-\delta(<\mathbf{v},\mathbf{x}_i>+a)\geq-\epsilon_2$.

Podemos concluir que existe un valor $\epsilon\in\mathbb{R}$ con $\epsilon=\min\{\epsilon_1,\epsilon_2\}$, que verifica la expresión $\delta y_i(<\mathbf{v},\mathbf{x}_i>+a)\geq\epsilon$. Podemos escoger el valor de δ como ϵ sin pérdida de generalidad, por tanto, obtendríamos la siguiente ecuación

$$y_i(<\mathbf{v},\mathbf{x}_i>+a)\geq 1, \qquad i=1,\ldots,n. \tag{2.2}$$

Se puede reescribir el problema a resolver como el siguiente

$$\begin{array}{lll} \textbf{(P0)} & \max \ \min\limits_{i} & \dfrac{y_i H(\mathbf{x}_i)}{\|\mathbf{v}\|^{\circ}} \\ & \text{s.a:} & y_i(<\mathbf{v},\mathbf{x}_i>+a)\geq 1, \quad i=1,\ldots,n. \end{array}$$

Veamos una expresión equivalente de la función objetivo del problema **(P0)**, debido a que las restricciones de dicho problema implican que $y_i(<\mathbf{v},\mathbf{x}_i>+a)\geq 1$ para

$i = 1, \ldots, n$, es decir, $y_i H(\mathbf{x}_i) \geq 1$ y como se está minimizando en función de i entonces obtendríamos $\max \frac{1}{\|\mathbf{v}\|^\circ}$.

En este trabajo se usará la 2-norma, debido a que la norma dual de la 2-norma es ella misma, en lo que sigue no utilizaremos la notación de norma dual. Si se hubiese establecido otra norma el desarrollo que sigue tendría que hacerse con la norma dual de la norma elegida. De lo dicho previamente en este párrafo se deduce que el margen sea máximo es equivalente a encontrar el menor valor de $\|\mathbf{v}\|^\circ$. Por tanto, el objetivo será minimizar $\|\mathbf{v}\|_2$, para simplificar cálculos se usará como función objetivo la expresión $\frac{1}{2}\|\mathbf{v}\|_2^2$ en vez de $\|\mathbf{v}\|_2$, dado que es equivalente minimizar una expresión que otra, así pues tenemos el siguiente problema.

$$\begin{aligned} \textbf{(P1)} \quad & \min \quad \frac{1}{2}\|\mathbf{v}\|_2^2 \\ & \text{s.a:} \quad y_i(< \mathbf{v}, \mathbf{x}_i > + a) \geq 1, \quad i = 1, \ldots, n. \end{aligned}$$

Al obtener la expresión del hiperplano óptimo veremos que éste se puede deducir a partir de los vectores soporte, los cuales definiremos a continuación.

DEFINICIÓN 2.1.4: Vectores Soporte

Sea $(\boldsymbol{x}_i, y_i)$ con $\boldsymbol{x}_i \in \mathbb{R}^m$ e $y_i \in \{1, -1\}$ tal que $i = 1, \ldots, n$. Existe un hiperplano H que separa la clase $y = 1$ de $y = -1$. Llamaremos vectores soporte, a aquellos valores de la dupla $(\boldsymbol{x}_i, y_i)$ que verifiquen la igualdad en la ecuación (2.2).

Nos encontramos ante un problema con $n + 1$ variables y n restricciones, usaremos la relajación Lagrangiana para facilitar la resolución del problema, para ello se definirá lo que es la relajación de un problema.

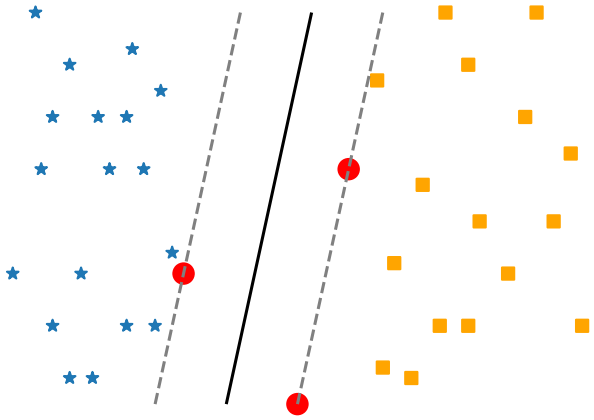


Figura 2.3: En la figura, los tres puntos rojos representan a los vectores soporte del hiperplano que separa las dos clases.

DEFINICIÓN 2.1.5: Relajación de un problema

Sea un problema de programación entera

$$\textbf{(PP)} \quad \begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.a:} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n, \\ & \mathbf{x} \in X \in \mathbb{Z}^m, \end{aligned}$$

se dice que

$$\textbf{(RP)} \quad \begin{aligned} \min \quad & q(\mathbf{x}) \\ \text{s.a:} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n, \\ & \mathbf{x} \in T \in \mathbb{Z}^m, \end{aligned}$$

es una relajación de **(PP)** *si se verifica:*

- $X \subseteq T$.
- $q(\mathbf{x}) \leq f(\mathbf{x}), \forall \mathbf{x} \in X$.

Al querer aplicar la relajación lagrangiana, la relajación del problema vendrá dada por la función de Lagrange que se define a continuación. Así se pasa a un problema sin restricciones y con $2n+1$ variables, correspondiendo n de ellas a multiplicadores de Lagrange.

DEFINICIÓN 2.1.6: Función de Lagrange

Sea un problema de programación entera

$$\begin{aligned} \textbf{(PP)} \quad \min \quad & f(\mathbf{x}) \\ \text{s.a:} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n, \\ & \mathbf{x} \in \mathbb{Z}^m, \end{aligned}$$

su función de Lagrange viene dada por

$$L(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}),$$

donde $\alpha_i \geq 0$ para $i = 1, \ldots, n$ son los multiplicadores de Lagrange.

En este sentido la función de Lagrange del problema **(P1)** es la siguiente

$$L(\mathbf{v}, a, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{v}\|_2^2 - \sum_{i=1}^{n} \alpha_i (y_i(< \mathbf{v}, \mathbf{x}_i > + a) - 1), \tag{2.3}$$

con $\alpha_i \geq 0$. El símbolo negativo del sumatorio en la expersión (2.3) se debe a que las restricciones en el problema **(P1)**, a diferencia del problema **(PP)** dado en la Definición de función de Lagrange (2.3) , son mayores o iguales que 0.

Veamos que el siguiente problema **(RLP)** es una relajación de **(PP)** que llamaremos relajación lagrangiana

$$\begin{aligned} \textbf{(RLP)} \quad \min \quad & f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}) \\ & \mathbf{x} \in \mathbb{Z}^m, \quad i = 1, \ldots, n. \end{aligned}$$

COROLARIO 2.1.1:

El problema **(RLP)** *es una relajación de* **(PP)** *, para $\alpha_i \geq 0$ con $i = 1, \ldots, n$.*

DEMOSTRACIÓN:

Para la demostración, veremos si se verifica la definición de relajación dada en 2.1.5. Si **(RLP)** fuera relajación de **(PP)** se debería verificar que la región factible del primer problema debe ser, al menos, tan grande como la del problema **(PP)**, esto se verifica, ya que el

conjunto de puntos factibles del problema **(PP)** son puntos factibles en el problema **(RLP)**, puesto que no hay ninguna restricción y el conjunto de valores son los mismos.

Por otro lado, se debe cumplir que la función objetivo del problema **(RLP)** sea menor o igual que la de **(PP)**, esto es cierto ya que la función $g_i(\mathbf{x}) \leq 0$ entonces $\sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}) \leq 0$.

$$f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}) \leq f(\mathbf{x}).$$

□

Obtenida la relajación del problema **(PP)**, pasamosa a desarollar el problema dual de **(PP)** ya que su resolución es más sencilla y como demostraremos más adelante, su resolución implica también obtener la solución del problema primal **(PP)**. Para conseguir el problema dual de **(PP)** definimos antes una función a la que llamaremos función dual y cuya expresión será la siguiente

$$\Theta(\boldsymbol{\alpha}) = \inf_{\mathbf{x}} \; L(\mathbf{x}, \boldsymbol{\alpha}) = \inf_{\mathbf{x}} \; f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}).$$

Por lo tanto la expresión del problema dual es la siguiente

$$\begin{aligned} \textbf{(PD)} \quad \max_{\boldsymbol{\alpha}} \quad \inf_{\mathbf{x}} \quad & f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}) \\ & \mathbf{x} \in \mathbb{Z}^m, \quad \alpha_i \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

Sustituyendo los datos del caso que estamos tratando obtenemos

$$\begin{aligned} \textbf{(PD1)} \quad \max_{\boldsymbol{\alpha}} \quad \inf_{\mathbf{v}} \quad & \frac{1}{2}\|\mathbf{v}\|_2^2 - \sum_{i=1}^{n} \alpha_i (y_i(< \mathbf{v}, \mathbf{x}_i > + a) - 1) \\ & \alpha_i \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

A continuación se expondrá el Teorema de Karush-Kuhn-Tacker, el cual aplicaremos para buscar la solución el problema primal, pero antes de pasar a ello se verán conceptos básicos necesarios para la demostración de dicho teorema. Por otro lado, se mostrará la relación entre la solución del problema dual **(PD1)** y el problema primal **(P1)**. Para esto último, se necesita demostrar el siguiente teorema.

TEOREMA 2.1.1:

Sean $\boldsymbol{\alpha}$ y $\boldsymbol{x}$ vectores que satisfacen las restricciones del problema dual **(PD)** *y del primal* **(PP)** *respectivamente, es decir, $g_i(\boldsymbol{x}) \leq 0$ y $\alpha_i \geq 0$, con $i = 1, \ldots, n$, entonces $\varphi(\boldsymbol{\alpha}) \leq f(\boldsymbol{x})$, siendo*

$$\varphi(\boldsymbol{\alpha}) = \inf_{\mathbf{x}} \mathbf{L(x,\boldsymbol{\alpha})} \ .$$

DEMOSTRACIÓN:

Para la demostración bastará con desarrollar la expresión.

$$\varphi(\boldsymbol{\alpha}) = \inf_{\mathbf{x}\in\mathbb{R}^m} L(\mathbf{x}, \boldsymbol{\alpha}) = \inf_{\mathbf{x}\in\mathbb{R}^m} f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}).$$

Por hipótesis, sabemos que $\alpha_i \geq 0$ y $g_i(\mathbf{x}) \leq 0$ para $i = 1, \ldots, n$ y por tanto $\alpha_i g_i(\mathbf{x}) \leq 0$, luego $\sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}) \leq 0$. Aplicando esta desigualdad obtenemos

$$\varphi(\boldsymbol{\alpha}) = \inf_{\mathbf{x}\in\mathbb{R}^m} f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}) \leq \inf_{\mathbf{x}\in\mathbb{R}^m} f(\mathbf{x}).$$

Por último, se deduce que

$$\inf_{\mathbf{x}\in\mathbb{R}^m} f(\mathbf{x}) \leq f(\mathbf{x}).$$

□

Del Teorema 2.1.1 se pueden deducir 2 corolarios.

COROLARIO 2.1.2:

El problema dual **(PD)** *está acotado superiormente por el problema primal* **(PP)**.

DEMOSTRACIÓN:

Puesto que la función $\varphi(\boldsymbol{\alpha})$ corresponde con la función objetivo del problema dual y $f(\mathbf{x})$ con la del problema primal, aplicando el Teorema 2.1.1 tenemos que

$$\varphi(\boldsymbol{\alpha}) \leq f(\mathbf{x}).$$

□

COROLARIO 2.1.3:

Sea **(PP)** *el problema primal dado en la Definición 2.1.5 con* $\boldsymbol{x}^*$ *un punto factible y* **(PD)** *su correspondiente problema dual con* $\boldsymbol{\alpha}^*$ *un punto factible de este. Se verifica que si* $\varphi(\boldsymbol{\alpha}^*) = f(\boldsymbol{x}^*)$*, entonces* $\boldsymbol{\alpha}^*$ *y* $\boldsymbol{x}^*$ *son soluciones óptimas de* **(PD)** *y* **(PP)** *respectivamente*.

DEMOSTRACIÓN:

Sabemos que $\varphi(\boldsymbol{\alpha}) \leq f(\mathbf{x})$ por el Teorema 2.1.1 para todo $\mathbf{x}$ factible primal y $\boldsymbol{\alpha}$ factible dual. Entonces

$$\sup_{\boldsymbol{\alpha}} \ \varphi(\boldsymbol{\alpha}) \leq \inf_{\mathbf{x}} \ f(\mathbf{x}),$$

con lo cual sí existe $\mathbf{x}^*$ y $\boldsymbol{\alpha}^*$ soluciones primal y dual factibles respectivamente tales que

$$\varphi(\boldsymbol{\alpha}) = f(\mathbf{x}).$$

COMENTARIO: MODIFICAR EL PÁRRAFO SIGUIENTE COMO SE DICE EN LA CORRECCIÓN

es decir, que $\varphi(\boldsymbol{\alpha})$ es el menor valor entre los supremos, entonces es el máximo. Análogamente vemos que el valor de $f(\mathbf{x})$ es un mínimo.
Al verificarse las funciones objetivos, $\boldsymbol{\alpha}$ y $\mathbf{x}$ son soluciones factibles de sus respectivos problemas. □

A partir del Corolario 2.1.3, se puede deducir que si se cumplen las hipótesis del corolario, es posible saber la solución de un problema y de su problema dual partiendo solo de una ellas. En el caso que se trata en esta subsección vamos a resolver el problema **(PD1)**, y a partir de este, obtendremos la solución de **(P1)**.

Pasemos a ver algunos conceptos que se necesitan conocer para la demostración del Teorema de Karush-Kuhn-Tucker.

DEFINICIÓN 2.1.7: Función Convexa

Sea f: $X \subseteq \mathbb{R}^m \to \mathbb{R}$ una función derivable con X un conjunto convexo no vacío. Decimos que f es una función convexa en χ si y solo si $\forall \boldsymbol{x}, \boldsymbol{y} \in X,\ \lambda \in [0,1]$ se tiene que $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$.

Decimos que la función f es estrictamente convexa cuando $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) < \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$.

En este caso, es fácil comprobar que las restricciones y la función objetivo del problema **(PD1)** son convexas, y al tratarse de funciones polinómicas son diferenciables.

A continuación, se definen lo que es un punto crítico y un vector gradiente.

DEFINICIÓN 2.1.8: Punto Crítico

Sea $f : X \subset \mathbb{R}^m \to \mathbb{R}$ una función derivable y sea $\boldsymbol{x} = (x_1, \ldots, x_m) \in X$ un punto. Se dice que $\boldsymbol{x}$ es un punto crítico, si sus derivadas parciales son cero, es decir

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} = 0, \ \ \textit{para } i = 1, \ldots, m.$$

DEFINICIÓN 2.1.9: Vector Gradiente

Sea $f : X \subset \mathbb{R}^m \to \mathbb{R}$ una función derivable y sea $\boldsymbol{x} = (x_1, \ldots, x_m) \in X$ un punto. Se define el vector gradiente de la función f en el punto $(x_1, \ldots, x_m)$ como

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_m} \right).$$

Por último veremos el siguiente teorema, para comprobar que en una función convexa, el mínimo local es global.

TEOREMA 2.1.2:

Sea $f : X \subset \mathbb{R}^m \to \mathbb{R}$ una función derivable y convexa, entonces se verifica que el mínimo local de f es también un mínimo global.

DEMOSTRACIÓN:

Supongamos que **x** es un mínimo local, entonces existe $r \in \mathbb{R}$ con $r > 0$ tal que $f(\mathbf{x}) \leq f(\mathbf{y})$ para todo $\mathbf{y} \in X \cap B(\mathbf{x}, r)$ siendo $B(\mathbf{x}, r)$ una bola centrada en el punto **x** con radio r.

Supongamos también que existe un mínimo global $\mathbf{u} \in X$, por tanto, $f(\mathbf{u}) \leq f(\mathbf{x})$. Consideremos un vector $\mathbf{z} \in X$ construido de la siguiente forma

$$\mathbf{z} = (1 - \lambda)\mathbf{x} + \lambda\mathbf{u},$$

con $\lambda > 0$ suficientemente pequeño para pertenecer a la bola, entonces $\mathbf{z} \in X \cap B(\mathbf{x}, r)$, pero esto se contradice con que **x** sea mínimo local en $B(\mathbf{x}, r)$, puesto que

$$f(\mathbf{z}) = (1 - \lambda)f(\mathbf{x}) + \lambda(\mathbf{u}) < (1 - \lambda)f(\mathbf{x}) + \lambda f(\mathbf{x}) = f(\mathbf{x}).$$

Concluimos así que si **x** es mínimo local en X, entonces también un mínimo global.

□

Ahora con los conceptos anteriores explicados, podemos pasar a demostrar el Teorema de Karush-Kuhn-Tucker.

TEOREMA 2.1.3: KARUSH-KUHN-TUCKER

Sea $\boldsymbol{x}^*$ *un punto factible para el problema* **(PP)** *, las funciones* $f : \mathbb{R}^m \to \mathbb{R}$*,* $g_i : \mathbb{R}^m \to \mathbb{R}$ *con* $i = 1, \ldots, n$ *funciones convexas y diferenciables. Si existen constantes* $\alpha_i \geq 0$*, tales que*

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_j} + \sum_{i=1}^{n} \alpha_i \frac{\partial g_i(\mathbf{x}^*)}{\partial x_j} = 0, \text{ para } j = 1, \ldots, m, \tag{2.4}$$

$$\alpha_i g_i(\mathbf{x}^*) = 0, \text{ para } i = 1, \ldots, n, \tag{2.5}$$

entonces el punto $\boldsymbol{x}^*$ *es un mínimo global del problema primal.*

DEMOSTRACIÓN:

Sea

$$L(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}),$$

la función de Lagrange correspondiente al problema **(PP)**, al ser f y g_i funciones convexas por hipótesis, $L(\mathbf{x}, \boldsymbol{\alpha})$ también lo es por ser suma de funciones convexas.

Se puede observar que la expresión (2.4) es equivalente a que el vector gradiente de $L(\mathbf{x}^*, \boldsymbol{\alpha})$ sea 0, es decir $\nabla L(\mathbf{x}^*, \boldsymbol{\alpha}) = 0$, luego $\mathbf{x}^*$ es un punto crítico. Al ser L una función convexa el punto crítico se trata de un mínimo y por tanto $L(\mathbf{x}^*, \boldsymbol{\alpha}) \leq L(\mathbf{x}, \boldsymbol{\alpha})$ para $\forall \mathbf{x} \in \mathbb{R}^m$.

Como $\alpha_i g_i(\mathbf{x}^*) = 0$ para $i = 1, \ldots, n$, entonces $\sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}^*) = 0$. Por otro lado, sabemos que para el problema **(PP)** se verifica la restricción $g_i(x) \leq 0$ para $i = 1, \ldots, n$ con $\mathbf{x} \in \Omega$ y $\alpha_i \geq 0$, luego $\alpha_i g_i(x) \leq 0$.

Aplicando (2.4) y (2.5) tenemos que

$$f(\mathbf{x}^*) = f(\mathbf{x}^*) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}^*) = L(\mathbf{x}^*, \alpha) \leq L(\mathbf{x}, \alpha) = f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x}) \leq f(\mathbf{x}).$$

□

La primera de las condiciones del Teorema de Karush-Kuhn-Tucker (2.4) corresponde con el ínfimo de la función de Lagrange, que en el caso que se está estudiando es en función de las variables $\mathbf{v}$ y a, por otro lado, la segunda condición es necesaria, ya que se busca que la función objetivo del problema primal y la del dual sean iguales para aplicar el Corolario 2.1.3.

Aplicamos la primera de las condiciones KKT, como tenemos solamente la función de Lagrange hacemos las derivadas parciales con respecto $\mathbf{v}$ y a:

$$\frac{\partial L(\mathbf{v}^*, a^*, \boldsymbol{\alpha}^*)}{\partial \mathbf{v}} = \mathbf{v}^* - \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i = 0, \qquad i = 1, \ldots, n, \tag{2.6}$$

$$\frac{\partial L(\mathbf{v}^*, a^*, \boldsymbol{\alpha}^*)}{\partial a} = -\sum_{i=1}^{n} \alpha_i^* y_i = 0, \qquad i = 1, \ldots, n. \tag{2.7}$$

De la ecuación (2.6) se puede deducir que:

$$\mathbf{v}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i. \tag{2.8}$$

A continuación aplicamos la segunda de las condiciones KKT obteniendo lo siguiente:

$$\alpha_i^*(1 - y_i(< \mathbf{v}^*, \mathbf{x}_i > + a^*)) = 0. \tag{2.9}$$

Sustituimos las ecuaciones (2.7) y (2.8) en (2.3):

$$\begin{aligned}
L(\mathbf{v}, a, \boldsymbol{\alpha}) &= \frac{1}{2} < \mathbf{v}^*, \mathbf{v}^* > - \sum_{i=1}^{n} \alpha_i^*(y_i(< \mathbf{v}^*, \mathbf{x}_i > + a^*) - 1) = \\
&= \frac{1}{2} < \mathbf{v}^*, \mathbf{v}^* > - \sum_{i=1}^{n} \alpha_i^* y_i < \mathbf{v}^*, \mathbf{x}_i > - \sum_{i=1}^{n} \alpha_i^* y_i a^* + \sum_{i=1}^{n} \alpha_i^* = \\
&= \frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i \right) \left(\sum_{j=1}^{n} \alpha_j^* \mathbf{x}_j y_j \right) - \left(\sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i \right) \left(\sum_{j=1}^{n} \alpha_j^* \mathbf{x}_j y_j \right) + \sum_{i=1}^{n} \alpha_i^* = \\
&= -\frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i \right) \left(\sum_{j=1}^{n} \alpha_j^* \mathbf{x}_j y_j \right) = \\
&= \sum_{i=1}^{n} \alpha_i^* - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i^* \alpha_j^* y_i y_j < \mathbf{x}_i \mathbf{x}_j > .
\end{aligned}$$

Como resultado de la primera condición podemos construir el problema dual que se muestra a continuación, en función únicamente de los multiplicadores de Lagrange

$$\begin{aligned}
\textbf{(PD2)} \quad \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j < \mathbf{x}_i \mathbf{x}_j > \\
\text{s.a:} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \\
& \alpha_i \geq 0, \quad i, j = 1, \ldots, n.
\end{aligned}$$

Al resolver el problema **(PD2)** se obtiene el valor de $\boldsymbol{\alpha}$, que sustituyendo en la expresión (2.8) nos daría el valor de $\mathbf{v}^*$, por tanto solo faltaría conocer el valor de a para conseguir la expresión del hiperplano definido como (2.2).

Volviendo a la expresión deducida de la segunda condición del Teorema KKT (2.9), podemos concluir que cuando $\alpha_i > 0$, entonces $1 - y_i(< \mathbf{v}^*, \mathbf{x}_i > + a^*) = 0$, por lo que

$$y_i(< \mathbf{v}^*, \mathbf{x}_i > + a^*) = 1. \tag{2.10}$$

Las observaciones $(\mathbf{x}_i, y_i)$ que verifiquen dicha expresión son lo que previamente definimos como vectores soporte. De la expresión (2.10) se puede despejar el valor de a

$$a^* = y_i - < \mathbf{v}^*, \mathbf{x}_i >,$$

es decir, la traslación del hiperplano depende de los vectores soporte y por tanto la expresión del hiperplano óptimo viene dado en función de estos. El hiperplano óptimo sería el siguiente

$$H(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^* y_i < \mathbf{x}_i, \mathbf{x} > + y_i - < \mathbf{v}^*, \mathbf{x}_i >, \qquad i = 1, \ldots, n. \tag{2.11}$$

Se concluye por la Definición 2.1.4 que las duplas $(\mathbf{x}_i, y_i)$ para $i = 1, \ldots, n$ que les corresponda $\alpha_i > 0$ son vectores soporte, por tanto el hiperplano de separación está definido a partir de vectores soporte.

Hallado el hiperplano de separación gracias a la aplicación de la relajación Lagrangiana y al Teroema de KKT, se podría separar el conjunto de observaciones $(\mathbf{x}_i, y_i)$ para $i = 1, \ldots, n$ en las dos clases existentes, siempre que las observaciones sean linealmente separables. A continuacion veremos el caso en el que el conjunto de datos es cuasi-separable.

2.2 CASO CUASI-SEPARABLE

Como se explicó a principio de sección, el tipo de formulación del SVM depende de la relación que tengan las observaciones $(\mathbf{x}_i, y_i)$, y por tanto de si se pueden separar las clases con hiperplanos o no. En esta subsección se desarrollará el caso en el que no es posible separar las clases mediante un hiperplano, y como resultado se encontrarán algunas observaciones mal clasificadas.

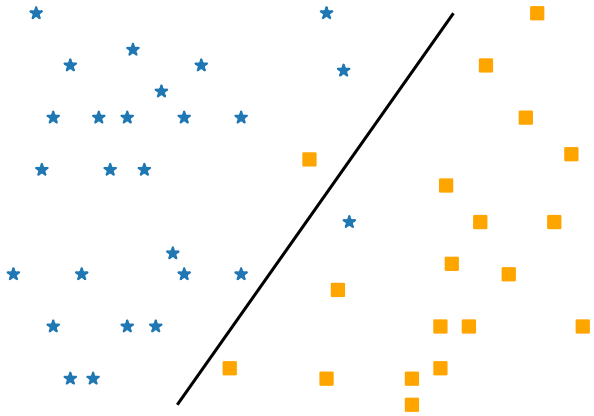


Figura 2.4: Ejemplo de observaciones mal clasificadas después de tratar de separar las clases mediante un hiperplano.

Dichos errores de clasificaciones serán expresados mediante una nueva variable, $\xi_i \geq 0$ para $i = 1, \ldots, n$ la cual estará asociada a la observación i de la dupla $(\mathbf{x}, \mathbf{y})$ y su valor corresponde con la desviación al hiplerplano separador. En función del valor de ξ_i, se puede extraer información de la posición de la observación $(\mathbf{x}_i, y_i)$. En concreto distinguimos los casos $\xi_i = 0$, $\xi_i \in (0, 1)$, $\xi_i \geq 1$. Más concretamente:

- Si $\xi_i = 0$ entonces $(\mathbf{x}_i, y_i)$ está bien clasificado y nos encontraríamos en el caso anteriormente explicado.
- Si $\xi_i \in (0, 1)$, $(\mathbf{x}_i, y_i)$ se encontraría en el lado correcto del hiperplano separador, pero la distancia a este sería inferior al margen, es decir, la observación se encontraría entre el hiperplano separador y el hiperplano soporte.
- Si $\xi_i \geq 1$ entonces $(\mathbf{x}_i, y_i)$ se encontraría en el lado opuesto del hiperplano que separa a su clase, por lo que estaría mal clasificado.

Si tuviéramos una serie de observaciones que no fueran linealmente separables, no podríamos aplicar el modelo dado por el problema **(P1)**, puesto que dicho problema sería infactible, por lo que el hiperplano separador no tendría la expresión (2.1). El uso de la variable $\boldsymbol{\xi}$, previamente definida, implica que algunas observaciones pueden hallarse en un semiespacio distinto al de su clase, por ello existe la posibilidad de que la distancia de la observación al hiperplano separador sea inferior a 1, a diferencia del caso en el que las observaciones fueran linealmente separables. Por tanto la expresión del hiperplano separador sería la siguiente

$$y_i(< \mathbf{v}, \mathbf{x}_i > +a) \geq 1 - \xi_i.$$

Se puede deducir que cuantas más observaciones se encuentren mal clasificadas, mayor será $\sum_{i=1}^{n} \xi_i$, de modo que la función objetivo no puede ser análoga a la del caso anterior, ya que ahora no solo hay que maximizar el margen, sino que también hay que minimizar la suma de $\boldsymbol{\xi}_i$. La función objetivo sería la siguiente

$$\frac{1}{2}\|\mathbf{v}\|_2^2 + C\sum_{i=1}^{n} \xi_i, \tag{2.12}$$

con C una constante que permite regular en que grado se prima minimizar $\|\mathbf{v}\|_2$ sobre minimizar la suma de ξ_i o viceversa. Así si C es un valor muy elevado, usualmente conducirán a valores de ξ_i pequeños. En el caso contrario, valores de C pequeños admitirían valores de ξ_i elevados. El problema a resolver sería

$$\begin{aligned} \textbf{(PPC)} \quad \min \quad & \frac{1}{2}\|\mathbf{v}\|_2^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s.a:} \quad & y_i(< \mathbf{v}, \mathbf{x}_i > +a) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

Al igual que se razonó en el caso linealmente separable, aplicamos la relajación lagrangiana, siendo la función de Lagrange en este caso

$$L(\mathbf{v}, a, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{v}\|_2^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i[y_i(< \mathbf{v}, \mathbf{x}_i > +a) - 1 + \xi_i] - \sum_{i=1}^{n} \beta_i \xi_i, \tag{2.13}$$

ahora se tienen dos tipos diferentes de multiplicadores de lagrange, $\alpha_i, \beta_i \geq 0$ para $i = 1, \ldots, n$, debido a que el problema **(PPC)** tiene dos grupos de restricciones.

De nuevo como en el caso anterior, aplicamos el Teorema de Karush-Kuhn-Tucker 2.1.3 . Primero se desarrollará la primera de las condiciones, (2.4), para ello hacemos la derivada de (2.13) con respecto a las variables $\mathbf{v}$, a y ξ_i:

$$\frac{\partial L(\mathbf{v}^*, a^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta})}{\partial \mathbf{v}} = \mathbf{v}^* - \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i = 0, \qquad i = 1, \ldots, n, \tag{2.14}$$

$$\frac{\partial L(\mathbf{v}^*, a^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta})}{\partial a} = -\sum_{i=1}^{n} \alpha_i^* y_i = 0, \qquad i = 1, \ldots, n, \tag{2.15}$$

$$\frac{\partial L(\mathbf{v}^*, a^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\xi}_i} = C - \alpha_i^* - \beta_i = 0, \qquad i = 1, \ldots, n. \tag{2.16}$$

De las expresiones anteriores se pueden sacar algunas conclusiones. De (2.14) concluimos que:

$$\mathbf{v}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i, \quad i = 1, \ldots, n. \tag{2.17}$$

Por otro lado de (2.15) y (2.16) se deduce respectivamente que:

$$\sum_{i=1}^{n} \alpha_i^* y_i = 0, \qquad i = 1, \ldots, n, \tag{2.18}$$

$$C = \alpha_i^* + \beta_i, \qquad i = 1, \ldots, n. \tag{2.19}$$

A continuación construimos, realizando los cálculos pertinentes, la segunda condición del Teorema KKT, (2.5). En el caso que se está tratando se obtienen dos ecuaciones, una por cada restricción de nuestro problema **(PPC)**:

$$\alpha_i^* [y_i(< \mathbf{v}^*, \mathbf{x}_i > + a^*) - 1 + \xi_i^*] = 0, \qquad i = 1, \ldots, n, \tag{2.20}$$

$$\beta_i \xi_i^* = 0, \qquad i = 1, \ldots, n. \tag{2.21}$$

Sustituimos (2.17) , (2.18) y (2.19) en la expresión (2.13):

$$\begin{aligned}
L(\mathbf{v}, a, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\|\mathbf{v}\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(<\mathbf{v}, \mathbf{x}_i> +a) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i\xi_i = \\
&= \frac{1}{2} <\mathbf{v}, \mathbf{v}> - \sum_{i=1}^{n}\alpha_i y_i <\mathbf{v}, \mathbf{x}_i> + \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\xi_i(C - \alpha_i - \beta_i) = \\
&= \frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i\right)\left(\sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j\right) - \left(\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i\right)\left(\sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j\right) + \sum_{i=1}^{n}\alpha_i = \\
&= -\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j <\mathbf{x}_i, \mathbf{x}_j> + \sum_{i=1}^{n}\alpha_i.
\end{aligned}$$

Se consigue una expresión en función de $\alpha_i \geq 0$ para $i = 1, \ldots, n$, siendo esta variable desconocida. A partir de (2.19) se deduce que $\alpha_i \leq C$, debido a $\beta_i \geq 0$ para $i = 1, \ldots, n$. Luego el problema a resolver es el siguiente

$$\begin{aligned}
\textbf{(PDC)} \quad \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j <\mathbf{x}_i, \mathbf{x}_j> + \sum_{i=1}^{n}\alpha_i \\
\text{s.a:} \quad & \sum_{i=1}^{n}\alpha_i^* y_i = 0, \\
& 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n.
\end{aligned}$$

Análogo al caso de clases linealmente separables para hallar el hiperplano separador hay que sustituir la expresión (2.17) en (2.1) obteniendo

$$H(x) = \sum_{i=1}^{n}\alpha_i^* y_i <\mathbf{x}_i, \mathbf{x}> + a. \tag{2.22}$$

El valor de a depende de α_i para $i = 1, \ldots, n$ como se puede observar en la expresión anterior, esto se debe a que $\alpha_i \in [0, C]$ para $i = 1, \ldots, n$. Estudiaremos para que valores de α_i está definida a.

Existen tres posibles casos, $\alpha_i = 0$, $\alpha_i = C$ o $0 < \alpha_i < C$. Analicemos estas tres situaciones.

- Si $\alpha = 0$, por la expresión (2.19) se obtiene que $\beta_i = C$, y puesto que $\beta_i\xi_i^* = 0$ (consecuencia de la segunda condición del Teorema KKT 2.1.3) se deduce que $\xi_i^* = 0$. Por tanto, se puede concluir que todas las observaciones $(\mathbf{x}_i, y_i)$ con $\alpha_i = 0$ se encuentran bien clasificadas.

- Si $\alpha = C$ entonces obtenemos que $\beta_i = 0$ de la expresión (2.19), por tanto ξ_i puede ser mayor que cero debido a (2.21). Si $xi_i > 0$, entonces la observación $(\mathbf{x}_i, y_i)$ está mal clasificada puede ocurrir dos situaciones, que se encuentre mal clasificado porque se encuentre en el lado opuesto del hiperplano separador como en la Figura 2.5b, o que se encuentre en el lado correcto del hiperplano separador, pero no se encuentre en el semiespacio definido por el hiperplano soporte de su clase como se muestra en la Figura 2.5a .

- Si $\alpha \in (0, C)$, se puede deducir de la expresión (2.19) que $\beta_i \neq 0$, por lo que se concluye que $\xi_i = 0$ de (2.21). De la expresión (2.20) se deduce que

$$y_i(< \mathbf{v}^*, \mathbf{x}_i > + a^*) - 1 = 0,$$

  por lo que las observaciones $(\mathbf{x}_i, y_i)$ con un valor asociado de α perteneciente al intervalo $(0, C)$ son vectores soporte.

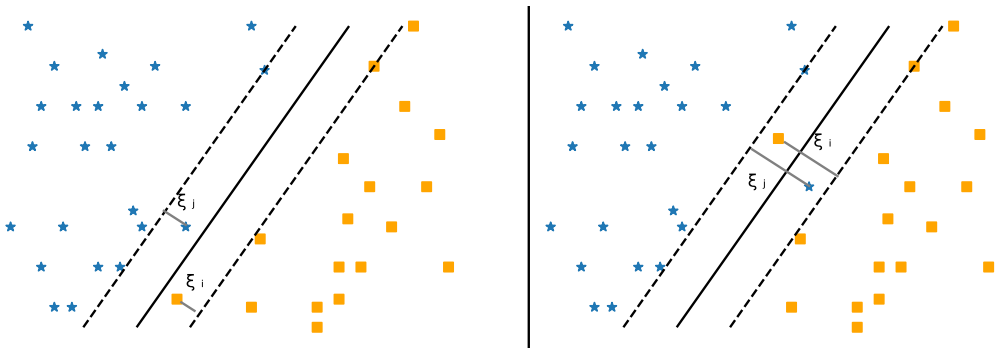


Figura 2.5: Las imágenes corresponden a los casos en los que x_i se encuentre bien clasificado 2.5a o que esté mal clasificado 2.5b

Después de formular el problema en el caso cuasi-separable, de manera muy similar al caso anterior, y desarrollar las distintas soluciones en función de los posibles valores de α_i y de ξ_i, solo nos faltaría por estudiar el caso en el que las observaciones sean linealmente no separables pero no se use un modelo lineal que posibilite la utilización de desviaciones, el cual se expondrá en la siguiente sección.

2.3 CASO LINEALMENTE NO SEPARABLE

En las secciones anteriores se ha explorado la posibilidad de que exista un hiperplano que separe ambas clases, y en caso de que no existiera se permitían errores de clasificación, es decir, había observaciones mal clasificadas. No siempre obtendremos buenos resultados permitiendo errores de clasificación, por ello, en esta sección se buscará transformar el espacio en el que se encuentren las observaciones $(\mathbf{x}_i, y_i)$ con $\mathbf{x}_i \in \Omega \subset \mathbb{R}^m$ e $y_i \in \{0, 1\}$ para $i = 1, \ldots, n$, en otro espacio usualmente de mayor dimensión en el que sí se puedan separar las clases mediante un hiperplano.

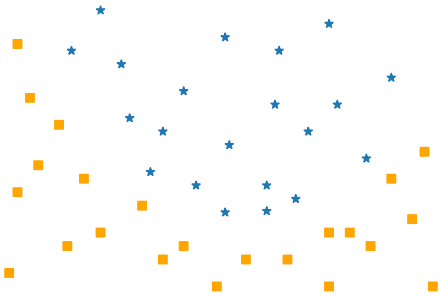
Para ello definiremos la función no lineal Φ que será la que transformará los elementos $\mathbf{x} \in \Omega \subset \mathbb{R}^n$ a otra dimensión superior

$$\begin{aligned} \mathbf{\Phi} : \Omega \subset \mathbb{R}^n \rightarrow & \quad W \subset \mathbb{R}^k \\ \mathbf{x} \rightarrow & \quad (\Phi_1(\mathbf{x}), \ldots, \Phi_k(\mathbf{x})), \end{aligned}$$

donde Φ_i con $i = 1, \ldots, k$ son funciones no lineales por tanto ahora el hiperplano no lineal que se busca es

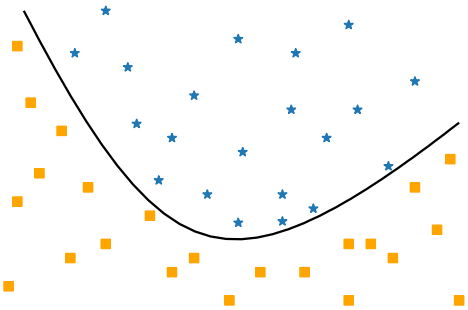
$$H(\mathbf{x}) = (v_1\Phi_1(\mathbf{x}) + \cdots + v_k\Phi_k(\mathbf{x})) =< \mathbf{v}, \mathbf{\Phi}(\mathbf{x}) > .$$

Antes de indagar en la metodología para hallar el hiperplano, veremos un ejemplo que permitirá comprender mejor lo anteriormente explicado. Se considera el siguiente conjunto de observaciones.



Como se puede observar, no hay forma de separar los conjuntos formados por las estrellas y los cuadrados mediante un hiperplano, sin embargo, una parábola si podría separar estos conjuntos.

Para llevar las observaciones a un espacio superior, definimos una función Φ, en este caso, su expresión es $\Phi(x_1, x_2) = (x_1, x_2, x_3)$ con $x_3 = 3$ si el resto de la división $\frac{x_2}{2}$ es 0 y $x_3 = -3$ en caso contrario. Tras aplicar esta función la representación de los datos se correspondería con la figura 2.6a, mientras que la imagen 2.6b representa a los conjuntos separados mediante un hiperplano en una dimensión superior.



COMENTARIO: No me salen las imágenes como antes, tengo que mirar porque.

(a) **(b)**

Figura 2.6

Después de este ejemplo ilustrativo, pasaremos a definir lo que denotaremos a partir de ahora como función Kernel y será utilizada en lugar del producto escalar $< \cdot, \cdot >$.

DEFINICIÓN 2.3.1: Función Kernel

Sea $\mathbf{\Phi} : \Omega \subset \mathbb{R}^n \to W$ *una función que transforma los valores de* Ω *a un espacio* W. *Se define la función Kernel como la función* $K : \Omega \times \Omega \to \mathbb{R}$ *con* $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \Omega$ *tal que verifique*

$$K(\mathbf{x}, \mathbf{y}) =< \mathbf{\Phi}(\mathbf{x}), \mathbf{\Phi}(\mathbf{y}) >= \Phi_1(\mathbf{x})\Phi_1(\mathbf{y}) + \cdots + \Phi_k(\mathbf{x})\Phi_k(\mathbf{y}).$$

En el siguiente teorema se verá que es posible transforma un conjunto de entrada de

dimensión finita a otro de dimensión infinita.

TEOREMA 2.3.1: Aronszajn

Para cualquier función $K : \Omega \times \Omega \to \mathbb{R}$ que sea simétrica y semidefinida positiva, existe un espacio de Hilbert y una función $\Phi : \Omega \to W$ tal que

$$K(\mathbf{x}, \mathbf{y}) =< \Phi(\mathbf{x}), \Phi(\mathbf{y}) >, \quad \forall \mathbf{x}, \mathbf{y} \in \Omega.$$

La demostración del anterior teorema se puede encontrar en [3]. De este teorema se deduce que no es necesario saber la expresión de la función Φ para conocer K, basta comprobar que K es simétrica y definida positiva para poder decir que existe un producto escalar de funciones Φ que lo defina.

Teniendo en cuenta que en el espacio transformado W sí existe un hiperplano que separe las clases, se puede deducir que el problema a resolver sería análogo al caso en el que los datos sean linealmente separables pero teniendo en cuenta la función de transformación Φ, con lo cual se tendría

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j < \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_j) > \\ \text{s.a:} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \quad i, j = 1, \ldots, n. \end{aligned}$$

Aplicando la función Kernel $K(\cdot, \cdot)$ definida previamente, el problema a resolver sería el siguiente

$$\begin{aligned} \textbf{(PD-NS)} \quad \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.a:} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

COMENTARIO: CAMBIAR EL PÁRRAFO DE ABAJO COMO LO QUIERE MARTA.

El hiperplano óptimo para separar las clases en el espacio W vendría dado por la expresión (2.2), su deducción es similar al caso anterior donde el hiperplano venía dado por la expresión (2.22), con la sustitución de la función Kernel, en donde se encontraba el producto escalar

$$\mathbf{v}^* = \sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}).$$

Equivalentemente al caso linealmente separable, se deduce que la variable a^* viene dada por

$$a^* = y_i - K(\mathbf{v}^*, \mathbf{x}_i).$$

Luego el hiperplano separador le corresponde la siguiente expresión

$$H(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + y_i - K(\mathbf{v}^*, \mathbf{x}_i).$$

En conclusión, hemos hallado un hiperplano que separa las dos clases con ayuda de la función Φ, que transporta las observaciones a otro espacio, y aplicando la función Kernel la cual sustituirá al producto escalar usado en los casos anteriores. Salvo a estos dos cambios, el problema a resolver es análogo al caso donde las observaciones sean linealmente separables.

CAPITULO

3

SVM MULTICLASE

Desarrollado el algoritmo SVM, se ha observado que una de las condiciones necesarias para su uso es que haya únicamente dos clases en el conjunto de datos, pero en la realidad también nos encontramos situaciones en la que se buscará dividir los datos en un número mayor de clases; un ejemplo de esto sería la siguiente situación. Supongamos que tenemos una cuenta de correo electrónico y queremos que los correos entrantes se clasifique en función de quien ha escrito el correo, el asunto que tenga... y se archiven en las carpetas denominadas como Amigos, Familia, Trabajo, Hobbies y Otros, como se puede deducir, en este caso se busca clasificar los datos en cinco clases y por tanto no podríamos aplicar el algoritmo SVM.

En esta sección nos centraremos en desarrollar modelos basados en el SVM que clasifiquen utilizando un número mayor de clases, para ello introducimos lo que denominaremos como SVM-Multiclase, un conjunto de modelos similares al SVM con la diferencia de que la variable y_n perteneciente a la dupla $(\mathbf{x}_n, y_n)$ no pertenece a un conjunto binario, sino a un conjunto de k elementos, $\{0, 1, \ldots, k\}$, así el conjunto de observaciones se puede dividir en k clases distintas. En la literatura existen varios enfoques distintos para abordar el SVM-Multiclase, estos pueden clasificarse en dos grandes categorías:

- Indirecto: Busca utilizar los clasificadores binarios que se han visto en la Sección 2 , aplicándolos para la creación de diversas clases artificiales independientes (por

ejemplo pertenecer o no una determinada clase) que clasificaran los datos en varias variables.

- Directo: El objetivo es desarrollar nuevos modelos basados en el SVM en los que se considere simultáneamente la clasificación en las distintas clases, pudiendo así introducir las observaciones sin necesidad de alterarlas mediante la creación de clases artificiales.

A continuación se desarrollarán ambos métodos junto con ejemplos para facilitar su comprensión, se empezará con el método Indirecto

3.1 MÉTODO INDIRECTO

Como se ha explicado anteriormente, en el método indirecto el objetivo es clasificar las observaciones en varias clases utilizando múltiples clasificadores binarios, para ello se construirán nuevas clases artificiales independientes que permitirán construir el clasificador que aplicaremos a las observaciones. Se usan principalmente tres métodos para obtener las dos clases artificiales necesarias para aplicar el clasificador binario.

- One vs One (OVO)
- One vs All (OVA)
- Directed Acyclic Graph SVM (DAGSVM)

En las siguientes subsecciones se analizarán en que consiste cada método empezando por OVO y terminando por DAGSVM.

3.1.1 One vs One (OVO)

Para aplicar este método, se utiliza el SVM para separar las clases dos a dos, para ello se escogen dos clases de las k disponibles a las que se le aplicará el modelo SVM, obteniendo así un hiperplano de separación entre estas dos clases. Este proceso se realizará con cada pareja de clases, es decir $\binom{k}{2}$, lo que corresponde con $\frac{k(k-1)}{2}$ veces. Para localizar la clase a la que pertenece una nueva observación existen diferentes métodos, nosotros usaremos el primero de los que exponemos a continuación

- Regla Max-Wins [13]: Este método funciona como un sistema de votos en el cual cada clasificador binario da un voto a la clase a la que pertenecería según el hiperplano separador. Los votos son contados y la clase con mayor número de votos será la elegida. En caso de empate se suman las confianzas de cada clase para cada pareja y la que tenga el mayor resultado será la elegida.

- Weighted voting strategy (WV): Este método es similar a la regla Max-Wins, pero en este caso cada clasificador binario vota a ambas clases y el peso de cada voto vendrá dado por la función de decisión

$$\text{clase de } \mathbf{x} = arg \max_{1,\ldots,k} < \mathbf{v}_r, \mathbf{x} > + a_r \tag{3.1}$$

  donde $< \mathbf{v}_r, \mathbf{x} > + a_r$ es el hiperplano que separa la clase r del resto de clases. La clase correspondiente con el voto de mayor peso será a la que pertenezca la observación, en caso de empate la clase es escogida al azar.

Supongamos que disponemos de un conjunto de observaciones $(\mathbf{x}_i, y_i)$ con $\mathbf{x}_i \in \Omega$ e $y_i \in \{0, 1, 2\}$, por lo tanto en este caso no se puede aplicar el algoritmo SVM, ya que poseemos tres clases en lugar de dos. La representación de los datos es la siguiente

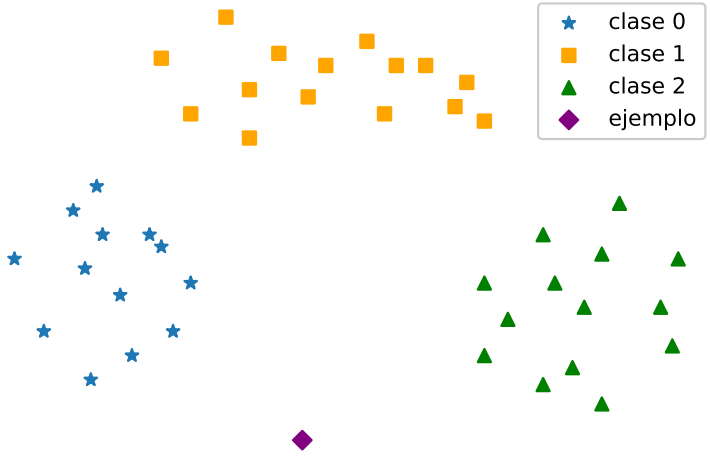


Figura 3.1: En la imagen se puede observar las tres clases en las que se dividen los datos, clase estrella, clase triángulo y clase cuadrado y un dato representada con un diamante morado que queremos comprobar la clase pertenece.

Para poder aplicar el algoritmo SVM necesitamos tener únicamente dos clases, para esto veremos todas las posibles parejas de clases existentes, es decir $\binom{3}{2} = 3$. Los casos posibles son la clase estrella-clase cuadrado, clase estrella-clase triángulo y clase cuadrado-clase triángulo. Al aplicar el algoritmo SVM en cada uno de los casos anteriores obtendríamos un hiperplano de separación para cada pareja, obteniendo el siguiente escenario

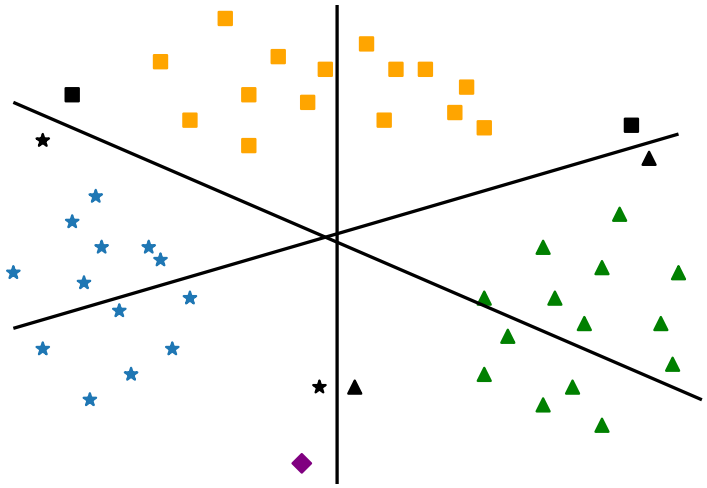


Figura 3.2: En la representación mostrada arriba se aprecian las distintas clases y los hiperplanos que separan cada pareja de clases existente.

La clase cuadrado no tiene ningún voto, la clase estrella tiene un voto, dado por el hiperplano estrella-cuadrado y por último la clase triángulo tiene dos votos, uno del hiperplano triángulo-estrella y otro del hiperplano triangulo-cuadrado, por tanto la clase que tiene más votos es la clase cuadrado, por lo que concluimos que es la clase a la que pertenece la observación que estábamos estudiando.

El ejemplo anterior corresponde con el caso en el que las observaciones $(\mathbf{x}_i, y_i)$ son linealmente separables, pero como hemos discutido en la sección anterior no es el único caso posible. Si los datos fueran cuasi-separables o linealmente no separables el desarrollo del ejemplo sería equivalente, pero aplicando el algoritmo SVM correspondiente.

3.1.2 One vs All (OVA)

A diferencia del método OVO, en este caso se escoge una clase r en concreto y el objetivo es separarlo de las $(k-1)$ clases restantes, es decir, se crean dos clases artificiales, perte-

necer a la clase r o no pertenecer a la clase r, así se cumplen los requisitos necesarios para aplicar el SVM. Si en el conjunto de observaciones hay k clases diferentes, este proceso se repetirá k veces, una por cada clase existente $\binom{k}{1}$.

Para hallar la clase a la que pertenece una observación existen diferentes métodos, pero en este trabajo se realizará con una estrategia de votos, su funcionamiento es el siguiente; cada clasificador binario vota a ambas clases, el peso de cada voto vendrá dado por la función de decisión 3.1 . La clase correspondiente con el voto de mayor peso será a la que pertenezca la observación, en caso de empate la clase escogida es al azar.

Como en el caso anterior se desarrollará un ejemplo para comprender los pasos a seguir para resolver el problema. Supongamos que tenemos un conjunto de observaciones linealmente separables $(\mathbf{x}_i, y_i)$ con $\mathbf{x}_i \in \Omega$ e $y_i \in \{0, 1, 2\}$, al no poder aplicar el algoritmo SVM, se procederá a construir dos clases artificiales, una de las cuales estará formada por dos de las tres clases previamente definidas, en total se construirán seis clases aritficiales, dos por cada clase existente, aplicando así el algoritmo SVM tres veces.

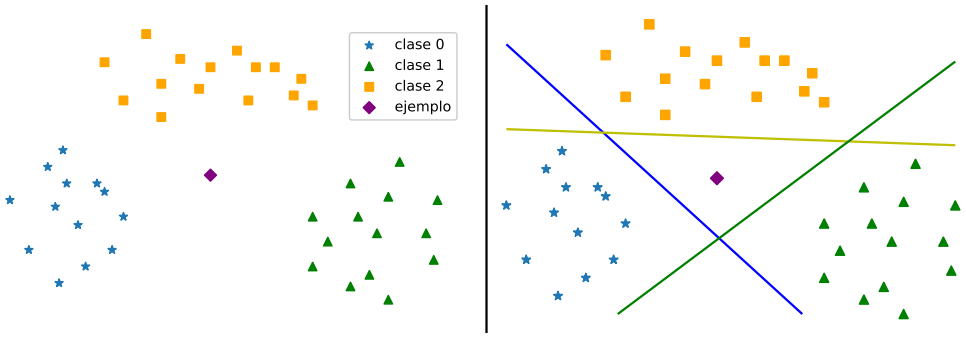


Figura 3.3: La Figura 3.3a muestra el conjunto de observaciones de las que disponemos junto con el ejemplo a estudiar representado por un diamante morado. Por otro lado en 3.3b se puede observar las clases con los hiperplanos que las separan del resto.

En la Figura 3.3b el color de cada hiperplano representa la clase que separa de las otras dos, teniendo por tanto tres hiperplanos resultantes del algoritmo SVM, uno por cada clase existente. Ya obtenidos los tres hiperplanos, comprobemos a que clase pertenece la observación representada por un diamante morado.

- El hiperplano azul sería el siguiente $-2{,}00057x - 1{,}10168 = y$, al sustituir el valor del ejemplo se obtendría el valor $-3{,}2013$.
- El hiperplano naranja tendría la expresión $-0{,}07861x + 5{,}74053 = y$, tras sustituir el ejemplo a estudiar obtenemos el valor $-2{,}29556$.
- El hiperplano verde tiene la siguiente formulación $1{,}64151x + 1{,}24594 = y$ y al sustituir el ejemplo que estamos estudiando se tiene el valor $-3{,}40311$.

Para saber la clase a la que pertenece la observación bajo estudio aplicamos la expresión 3.1 y obtenemos que pertenece a la clase cuadrado.

El caso en el que las clases sean cuasi-separables o no separables linealmente es análogo al ejemplo que acabamos de mostrar, con la diferencia de que los hiperplanos que se buscan serán los resultantes de aplicar la formulación del SVM adecuada.

3.1.3 Directed Acyclic Graph SVM (DAGSVM)

En este método desarrollado por J.Platt, N.Cristianini y J Shawe-Taylor [22] se clasifica cada observación $(\mathbf{x}_i, y_i)$ con $i = 1, \ldots, n$ para ver en que clase se encuentra, esto se realizará creando un árbol cuyos nodos representarán hiperplanos separadores entre dos clases distintas y los nodos finales corresponderán con cada una de las k clases que poseen las observaciones. Para la elección de los nodos primero es necesario tener una estructura jerárquica de las clases, así estas se podrán ordenar en una lista, y a partir de ella escoger las parejas de clases que corresponderán a cada nodo. Encontrar la forma de estructurar las clases ha sido un campo estudiado ampliamente [14], [21],[10],[26], en este trabajo daremos una estructura fija en las clases para su mejor compresión. A continuación se explicará como se eligen las parejas de clases para cada nodo

- Nodo Raíz: Para el nodo raíz se escogen las clases que se encuentran en los extremos de la lista, es decir, el primer y último elemento.
- Nodos Intermedios: Para facilitar la explicación supongamos que los clases que formaban el hiperplano en el nodo anterior se encontraban en la posición l y b con $l < b$. En esta ocasión existen dos posibilidades

- En el nodo anterior se halla eliminado la clase posicionada más alta en la lista (b): En este caso, en el nodo actual tendríamos el hiperplano que separa la clase que no se ha eliminado (l) y la enfrentaríamos a la clase que se encuentra en una posición más abajo de la clase eliminada previamente, es decir, $b-1$.
- En el nodo anterior se halla eliminado la clase posicionada más abajo en la lista (l):En este caso, en el nodo actual se encontraría el hiperplano que separa la clase que no se ha eliminado(b) y la clase que se encuentra en una posición más arriba de la clase eliminada previamente, es decir, $l+1$.

Antes de pasar a un ejemplo veremos como el orden de la lista puede afectar a la elección de la clase para una observación. Supongamos que tenemos el siguiente conjunto de datos

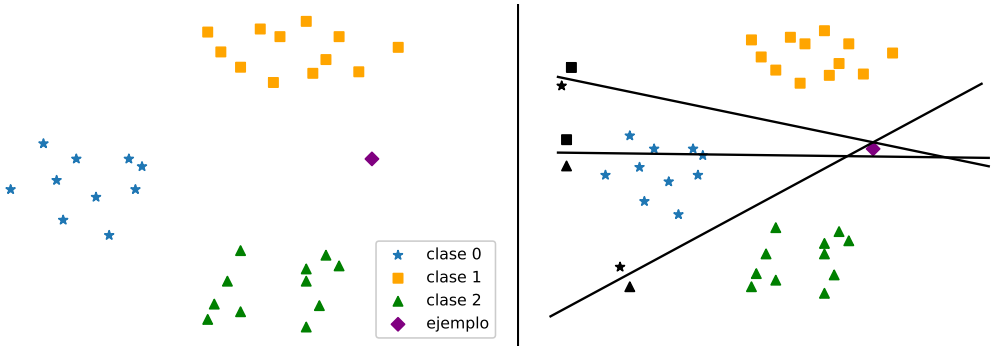


Figura 3.4: La Figura 3.4a muestra el conjunto de observaciones de las que disponemos junto con el ejemplo a estudiar representado por un diamante morado. Por otro lado en 3.4b se puede observar las clases con los hiperplanos que separa cada pareja de clases.

Primero veamos el caso en el cual el orden jerárquico de las clases es $\{0,1,2\}$, el árbol tendría la forma de la Figura 3.5a. Con esta lista obtenemos que el ejemplo se encuentra en la clase 2. Sin embargo si tuviéramos la lista $\{2,0,1\}$ el árbol sería el representado en 3.5b y el ejemplo se encontraría en la clase 1. En el siguiente ejemplo se hará un desarrollo más exhaustivo de la elección de la clase a la que pertenecerá una nueva observación.

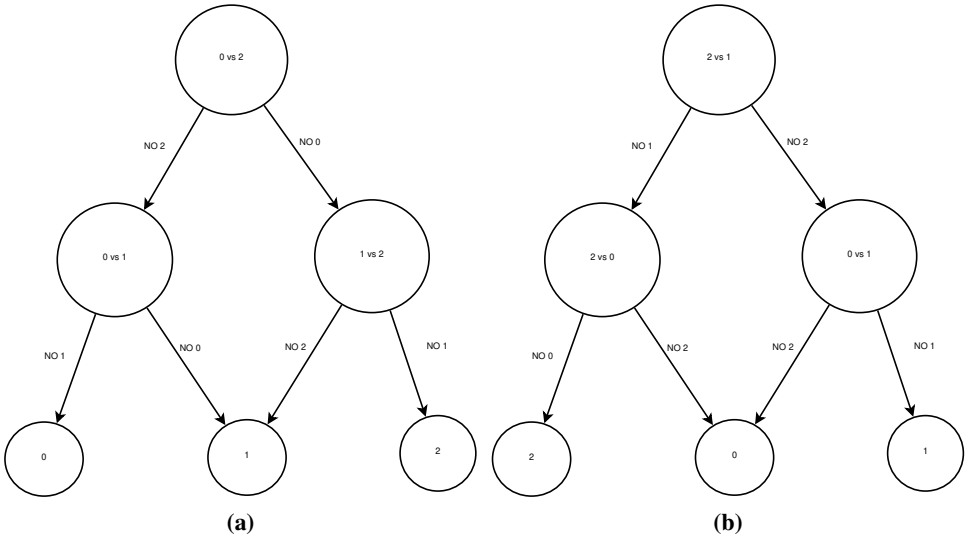


Figura 3.5: La Figura 3.5a se muestra el árbol formado a partir de la lista $\{0,1,2\}$. Por otro lado en 3.5b se puede observar el árbol desarrollado para la lista $\{2,0,1\}$.

Supongamos que disponemos de unas observaciones $(\mathbf{x}_i, y_i)$ con $\mathbf{x}_i \in \Omega$ e $y_i \in \{0, 1, 2, 3\}$, cuya representación vendría dada por la siguiente imagen.

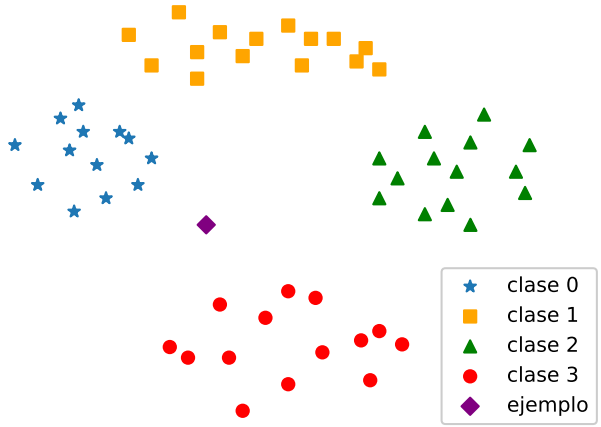


Figura 3.6

Queremos hallar la clase en la que se encuentra nuestra nueva observación denotada en la Figura 3.6 como un diamante morado, para ello se irán escogiendo todas las posibles parejas de clases y descartando clases en las que no se encuentra la nueva observación.

Como se ha explicado anteriormente la elección de la pareja se realizará en un proceso arborescente, partiendo de un nodo raíz, una representación de esto sería la dada en la Figura 3.7.

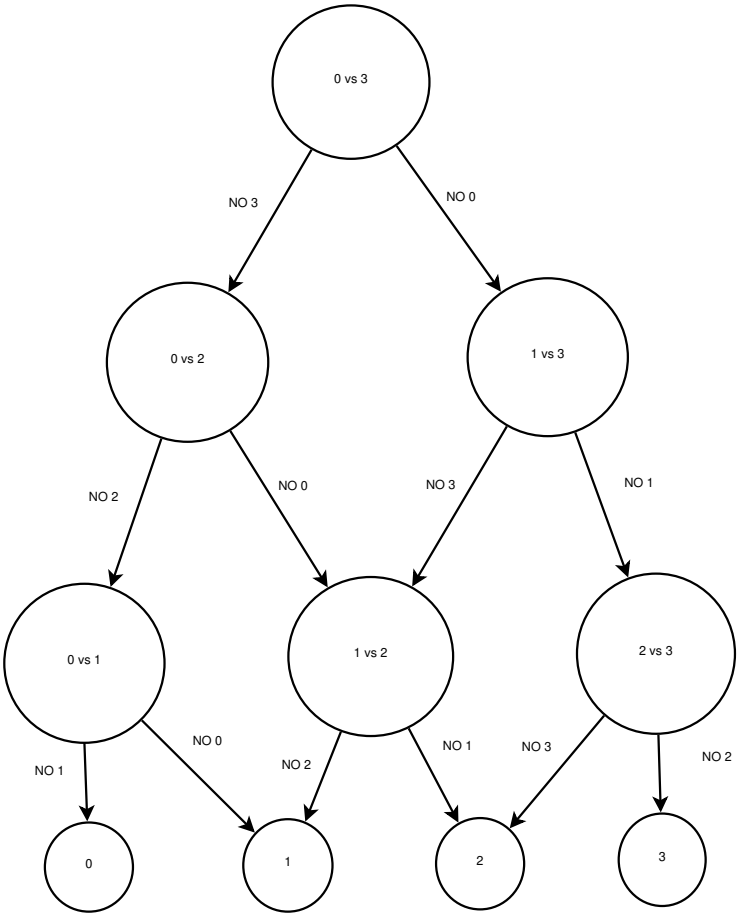


Figura 3.7: En la representación mostrada arriba se aprecian las distintas clases y los hiperplanos que separan cada pareja de clases existente.

En el nodo raíz, es decir, el que se encuentra en lo más alto del árbol, se busca un hiperplano que separe la clase 0 de la clase 3, este hiperplano podría ser el representado en la Figura 3.8a, descartamos la clase 3, puesto que el ejemplo no se encuentra en ese lado del hiperplano. Siguiendo el árbol de la figura 3.7 ahora veríamos si el elemento bajo estudio se encuentra en la clase 0 o en la 2, lo que correspondería con la Figura 3.8b

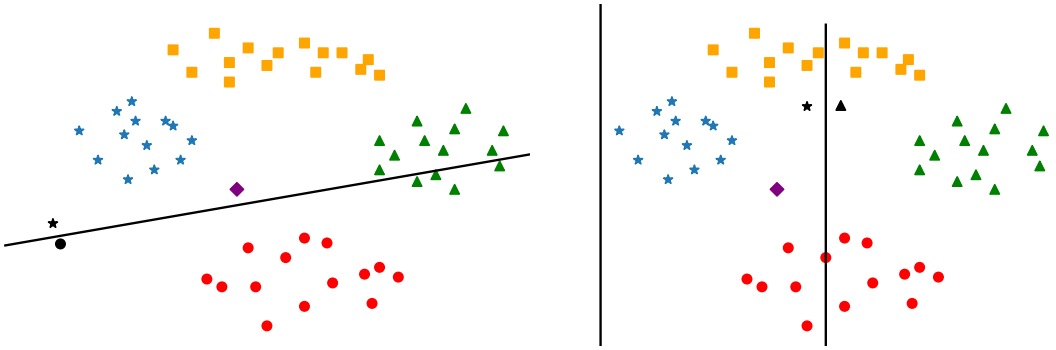


Figura 3.8: En la Figura 3.8b se distingue un hiperplano que separa la clase 0 (estrellas), de la clase 3 (círculos). En cambio en 3.8a se puede observar un hiperplano que separa la clase 0 (estrellas) de la clase 2 (triángulos).

Como resultado del hiperplano de la Figura 3.8b descartamos la clase 2 y solo nos quedaría comprobar si pertenece a la clase 0 o a la clase 1, el hiperplano de separación de estas dos clases es el mostrado en 3.9, y como se puede observar en la imagen, la observación a estudiar pertenece a la clase 0. Con esto hemos llegado al final del árbol representado en 3.7 y podemos concluir que el ejemplo referido como diamante morado pertenece a la clase 0 (estrellas).

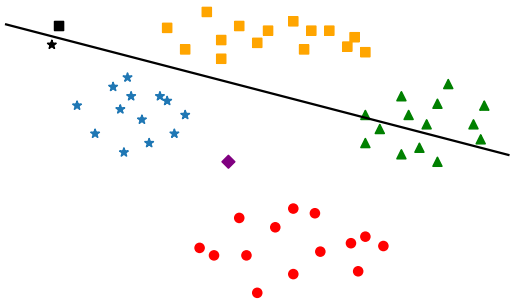


Figura 3.9: Hiperplano de separación entre la clase 0 (estrellas) y la clase 1 (cuadrados).

Hemos visto los tres métodos más usados para agrupar las observaciones que conocemos en dos grupos diferentes y así aplicar el algoritmo SVM, en la siguiente sección se estudiará el método directo, en el que se desarrollarán nuevos modelos basado en el SVM para que diferencia de manera simultánea entre todas las clases, en lugar de aplicarlo en subgrupos de dos clases.

3.2 MÉTODO DIRECTO

El objetivo de estos métodos es desarrollar un modelo que permita separar las k clases de manera simultánea. En este trabajo estudiaremos principalmente dos modelos, el de Weston-Watkins y Cramer-Singer.

3.2.1 Modelo Weston-Watkins

En esta sección se desarrollará el modelo que fue propuesto simultáneamente por J.Weston y C.Watkins en [28] y en otro artículo por E.Bredensteiner y K.Bennett [7], el análisis se realizará para el caso en que las observaciones sean cuasi-separables, debido a que engloba el caso de linealmente separable y linealmente no separable. En este modelo el clasificador que se utilizará funciona de manera parecida al definido en la Sección 2 , pero en lugar de ser un solo hiperplano, será una intersección entre los hiperplanos resultantes de separar una clase en concreto del resto de clases que tendrán nuestras observaciones.

COMENTARIO: Quiero poner una imagen, pero no he terminado de programar el algoritmo en Python.

Pasemos a analizar el algoritmo propuesto por Weston-Watkins, primero supongamos que disponemos de un conjunto de observaciones $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ siendo $\mathbf{x}_i \in \Omega \subset \mathbb{R}^m$ e $y_i \in Y = \{1, \ldots, k\}$, con $i = 1, \ldots, n$. A diferencia del caso cuasi-separable expuesto en la Sección 2.2 , la variable de holgura que definiremos tendrá dos índices, $\xi_i^r \geq 0$ para $i \in \{1, \ldots, n\}$ y $r \in \{1, \ldots, k\}$, el contador i esta asociado a la clase de la observación $(\mathbf{x}_i, y_i)$, por otro lado el contador r corresponde con una de las otras clases que no son y_i. Para facilitar notación diremos que $\xi_i^{y_i} = 2$, es decir el caso en el que $r = y_i$.

El objetivo del problema continua siendo maximizar el margen, y por tanto minimizar $\|\mathbf{v}\|_2$ o equivalentemente $\frac{1}{2}\|\mathbf{v}\|_2^2$. A la función objetivo debemos añadir la suma de las variables de holgura que hemos definido, por lo que tendríamos que la función objetivo vendría dada por

$$\min \quad \frac{1}{2}\|\mathbf{v}\|_2^2 + C\sum_{i=1}^{n}\sum_{r=1}^{k}\xi_i^r. \tag{3.2}$$

Veamos cual sería la restricción para este problema. Para mejor comprensión se empezará estudiando el caso en el que solo existan dos clases y lo generalizamos posteriormente a k clases. Supongamos que nuestras observaciones poseen únicamente dos clases que denotaremos como i, j, por ende la observación solo podrá encontrarse en un lado del hiperplano, es decir,

$$< \mathbf{v}, \mathbf{x}_i > + a_i \geq 1 - \xi \quad \text{si} \quad y_i = i,$$

$$< \mathbf{v}, \mathbf{x}_i > + a_j \leq \xi - 1 \quad \text{si} \quad y_i = j.$$

Las expresiones anteriores se pueden deducir a partir de la ecuación (2.12) . En el caso que nos atañe, queremos ver si la observación está en su clase asociada y_i, o en otra de las clases existentes. Tras generalizar obtenemos dos expresiones

$$< \mathbf{v}_{y_i}, \mathbf{x}_i > + a_{y_i} \geq 1 - \xi_i \quad \text{si} \quad y_i = i,$$

$$< \mathbf{v}_r, \mathbf{x}_i > + a_r \leq \xi_r - 1 \quad \text{si} \quad y_i = r,$$

la variable $\mathbf{v}_{y_i}$ representa el hiperplano que separa la clase y_i del resto de clases y la variable $\mathbf{v}_r$ el hiperplano que separa la clase r del resto de las clases existentes.

Para agrupar las dos expresiones anteriores restamos ambas desigualdades, obteniendo

$$< \mathbf{v}_{y_i}, \mathbf{x}_i > + a_{y_i} - < \mathbf{v}_r, \mathbf{x}_i > - a_r \geq 2 - \xi_i^r, \tag{3.3}$$

donde ξ_i^r corresponde con la suma de las expresiones ξ_i, ξ_r que son las variables de holgura los hiperplanos que separan la clase y_i, r respectivamente del resto. Con la expresión (3.3) sí se tienen en cuenta todas las clases existentes en el grupo de observaciones y se realizaría un total de nk veces, siendo n el número de observaciones y k el número de clases totales. El problema a resolver sería el siguiente

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{r=1}^{k}\|\mathbf{v}_r\|_2^2 + C\sum_{i=1}^{n}\sum_{r=1}^{k}\xi_i^r \\ \text{s.a:} \quad & < \mathbf{v}_{y_i}, \mathbf{x}_i > + a_{y_i} - < \mathbf{v}_r, \mathbf{x}_i > - a_r \geq 2 - \xi_i^r, \quad i = 1, \ldots, n, \quad r \in \{1, \ldots, k\} \\ & \xi_i^r \geq 0. \end{aligned}$$

Como ya hemos explicamos anteriormente, resolveremos el problema dual usando la relajación Lagrangiana, para ello se necesitará conocer cual es su función de Lagrange,

teniendo esta la siguiente expresión

$$\begin{aligned} L(\mathbf{v}_r, a_r, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\sum_{r=1}^{k}\|\mathbf{v}_r\|_2^2 + C\sum_{i=1}^{n}\sum_{r=1}^{k}\xi_i^r - \sum_{i=1}^{n}\sum_{r=1}^{k}\beta_i^r\xi_i^r - \\ - \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r[< \mathbf{v}_{y_i}, \mathbf{x}_i > + a_{y_i} - < \mathbf{v}_r, \mathbf{x}_i > - a_r - 2 + \xi_i^r], \end{aligned} \tag{3.4}$$

en este caso, los multiplicadores de Lagrange serán $\alpha_i^r \geq 0$, $\beta_i^r \geq 0$ para $i = 1, \ldots, n$ y $r \in \{1, \ldots k\}\backslash y_i$. En los casos en que $y_i = r$, es decir los casos en que queremos separar la clase y_i de la clase y_i, se impondrán los siguientes valores

$$\alpha_i^{y_i} = 0, \quad \beta_i^{y_i} = 0, \quad i = 1, \ldots, n, \tag{3.5}$$

estos valores son consecuencia de que es debido a que es fútil separar una clase de sí misma.

A continuación aplicamos el Teorema KKT 2.1.3 , en la primera de las condiciones de este teorema se usan las derivadas parciales, en este caso respecto a las variables $\mathbf{v}_r, a_r$ y ξ_i^r. Como en la expresión de la función de Lagrange (3.4) tenemos las variables $\mathbf{v}_{y_i}, a_{y_i}$ introducimos unas variables con las que consideraremos si $y_i = r$, y por tanto fuera necesario derivar también $\mathbf{v}_{y_i}$ y a_{y_i}

$$A_i = \sum_{r=1}^{k}\alpha_i^r, \quad c_i^r = \begin{cases} 1 \text{ si } y_i = r \\ 0 \text{ si } y_i \neq r \end{cases}$$

Procedemos a realizar los cálculos de la primera de las condiciones del Teorema KKT (2.4)

$$\frac{\partial L(\mathbf{v}_r, a_r, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \mathbf{v}_r} = \mathbf{v}_r + \sum_{i=1}^{n}\alpha_i^r\mathbf{x}_i - \sum_{i=1}^{n}A_i c_i^r\mathbf{x}_i = 0, \qquad i = 1, \ldots, n, \tag{3.6}$$

$$\frac{\partial L(\mathbf{v}_r, a_r, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial a_r} = -\sum_{i=1}^{n}A_i c_i^r + \sum_{i=1}^{n}\alpha_i^r = 0, \qquad i = 1, \ldots, n, \tag{3.7}$$

$$\frac{\partial L(\mathbf{v}_r, a_r, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \xi_i^r} = -\alpha_i^r + C - \beta_i^r = 0, \qquad i = 1, \ldots, n. \tag{3.8}$$

De las derivadas anteriores se pueden deducir las siguientes expresiones

$$\mathbf{v}_r = \sum_{i=1}^{n}A_i c_i^r\mathbf{x}_i - \sum_{i=1}^{n}\alpha_i^r, \qquad i = 1, \ldots, n, \tag{3.9}$$

$$\sum_{i=1}^{n} A_i c_i^r = \sum_{i=1}^{n} \alpha_i^r, \qquad i = 1, \ldots, n, \tag{3.10}$$

$$C = \alpha_i^r + \beta i^r \qquad i = 1, \ldots, n. \tag{3.11}$$

De la expresión (3.11) se puede obtener la siguiente acotación para α_i^r

$$0 \leq \alpha_i^r \leq C,$$

debido a que $\beta_i^r \geq 0, \alpha_i^r \geq 0$ para $i = 1, \ldots, n$ y $r \in \{1, \ldots, k\}$. Sustituimos la expresión (3.9) y (3.11) en (3.4)

$$\begin{aligned} L(\mathbf{v}_r, a_r, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \frac{1}{2} \sum_{r=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{n} (c_i^r A_i - \alpha_i^r)(c_j^r A_j - \alpha_j^r) < \mathbf{x}_i, \mathbf{x}_j > + C \sum_{i=1}^{n} \sum_{r=1}^{k} \xi_i^r - \\ & - \sum_{i=1}^{n} \sum_{r=1}^{k} \beta_i^r \xi_i^r - \sum_{i=1}^{n} \sum_{r=1}^{k} \alpha_i^r \xi_i^r - \sum_{i=1}^{n} \sum_{r=1}^{k} \alpha_i^r \Bigg[\sum_{j=1}^{n} (c_j^{y_i} A_j - \alpha_j^{y_i}) < \mathbf{x}_i, \mathbf{x}_j > - \\ & - a_r - \sum_{j=1}^{n} (c_j^r A_j - \alpha_j^r) < \mathbf{x}_i, \mathbf{x}_j > + a_{y_i} - 2 \Bigg] = \\ = & - \sum_{i=1}^{n} \sum_{r=1}^{k} \alpha_i^k \Bigg[\sum_{j=1}^{n} (c_j^{y_i} A_j - \alpha_j^{y_i}) < \mathbf{x}_i, \mathbf{x}_j > - \sum_{j=1}^{n} (c_j^r A_j - \alpha_j^r) < \mathbf{x}_i, \mathbf{x}_j > + \\ & + a_{y_i} - a_r - 2 \Bigg] + \frac{1}{2} \sum_{r=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{n} (c_i^r A_i - \alpha_i^r)(c_j^r A_j - \alpha_j^r) < \mathbf{x}_i, \mathbf{x}_j > . \end{aligned} \tag{3.12}$$

Antes de continuar con el desarrollo de la función de Lagrange veamos que se verifica la siguiente igualdad

$$\sum_{i=1}^{n} \sum_{r=1}^{k} \alpha_i^r a_{y_i} = \sum_{i=1}^{n} \sum_{r=1}^{k} \alpha_i^r a_r, \tag{3.13}$$

para ello, desarrollaremos primero el lado izquierdo de la igualdad. Lo que se busca es transformar la variable a_{y_i} en a_m, es decir, en lugar de considerar el término independiente del hiperplano asociado a cada observación, solo consideraremos el hiperplano asociado a cada clase. Se puede observar que por la definición dada previamente de la variable c_i^r se verifica la siguiente igualdad

$$\sum_{i=1}^{n} \sum_{r=1}^{k} c_i^r a_{y_i} = \sum_{r=1}^{k} a_r,$$

por tanto mediante la expresión anterior y con el uso de la definición de A_i se puede ver que se verifican las siguientes igualdades

$$\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r a_{y_i} = \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r a_r c_i^r = \sum_{r=1}^{k} a_r \sum_{i=1}^{n} c_i^r \alpha_i^r = \sum_{r=1}^{k} a_r \sum_{i=1}^{n} c_i^r A_i. \tag{3.14}$$

Por otro lado desarrollamos el lado derecho de la igualdad (3.13)

$$\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r a_r = \sum_{r=1}^{k} a_r \sum_{i=1}^{n}\alpha_i^r. \tag{3.15}$$

Aplicando la expresión (3.10) confirmamos que las dos expresiones anteriores, (3.14) y (3.15) son iguales, y por tanto al estar restándose en la función de Lagrange se anulan. Continuamos desarrollando la función de Lagrange 3.12

$$\begin{aligned}
L(\mathbf{v}_r, a_r, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = &-\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r\Bigg[\sum_{j=1}^{n}(c_j^{y_i}A_j - \alpha_j^{y_i}) < \mathbf{x}_i, \mathbf{x}_j > - \sum_{j=1}^{n}(c_j^r A_j - \alpha_j^r) < \mathbf{x}_i, \mathbf{x}_j > + \\
&+ a_{y_i} - a_r - 2\Bigg] + \frac{1}{2}\sum_{r=1}^{k}\left(\sum_{i=1}^{n}(c_i^r A_i - \alpha_i^r)\mathbf{x}_i\right)\left(\sum_{j=1}^{n}(c_j^r A_j - \alpha_j^r)\mathbf{x}_j\right) = \\
= &-\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r\Bigg[\sum_{j=1}^{n}(c_j^{y_i}A_j - \alpha_j^{y_i}) < \mathbf{x}_i, \mathbf{x}_j > - \sum_{j=1}^{n}(c_j^r A_j - \alpha_j^r) < \mathbf{x}_i, \mathbf{x}_j > - \\
&-2\Bigg] + \frac{1}{2}\sum_{r=1}^{k}\left(\sum_{i=1}^{n}(c_i^r A_i - \alpha_i^r)\mathbf{x}_i\right)\left(\sum_{j=1}^{n}(c_j^r A_j - \alpha_j^r)\mathbf{x}_j\right) = \\
= &-\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}\left[\alpha_i^r c_i^{y_i} A_j - \alpha_i^r \alpha_j^{y_i} - \alpha_i^r c_j^r A_j + \alpha_i^r \alpha_j^r\right] < \mathbf{x}_i, \mathbf{x}_j > + 2\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r + \\
&+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}\left[c_i^r c_j^r A_i A_j - \alpha_j^r c_i^r A_i - \alpha_i^r c_j^r A_j + \alpha_i^r \alpha_j^r\right] < \mathbf{x}_i, \mathbf{x}_j > = \\
= &\, 2\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}\Bigg[\frac{1}{2}c_i^r c_j^r A_i A_j - \frac{1}{2}\alpha_j^r c_i^r A_i - \frac{1}{2}\alpha_i^r c_j^r A_j + \frac{1}{2}\alpha_i^r \alpha_j^r - \\
&-\alpha_i^r c_i^{y_i} A_j + \alpha_i^r \alpha_j^{y_i} + \alpha_i^r c_j^r A_j - \alpha_i^r \alpha_j^r\Bigg] < \mathbf{x}_i, \mathbf{x}_j > .
\end{aligned}$$

A continuación empleamos en la ecuación anterior la expresión $\sum_r c_i^r A_i \alpha_j^r = \sum_r \alpha_j^r A_j \alpha_i^r$, esto se verifica debido a que

$$\sum_{r=1}^{k} c_i^r A_i \alpha_j^r = \sum_{r=1}^{k} c_i^r \alpha_i^r \alpha_j^r = \sum_{r=1}^{k} c_j^r \alpha_i^r \alpha_j^r = \sum_{r=1}^{k} c_j^r A_j \alpha_i^r,$$

con lo que obtendríamos

$$\begin{aligned}
L(\mathbf{v}_r, a_r, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= 2\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}\left[\frac{1}{2}c_i^r c_j^r A_i A_j - \frac{1}{2}\alpha_j^r c_i^r A_i - \frac{1}{2}\alpha_i^r c_j^r A_j + \right.\\
&\left. + \frac{1}{2}\alpha_i^r\alpha_j^r - \alpha_i^r\alpha_j^r - \alpha_i^r c_i^{y_i} A_j + \alpha_i^r\alpha_j^{y_i} + \alpha_i^r c_j^r A_j\right] < \mathbf{x}_i, \mathbf{x}_j >= \\
&= 2\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}\left[\frac{1}{2}c_i^r c_j^r A_i A_j - \frac{1}{2}c_i^r A_i\alpha_j^r - \frac{1}{2}c_j^r A_j\alpha_i^r + \right.\\
&\left. + c_j^r A_j\alpha_i^r - c_j^{y_i} A_j\alpha_i^r + \alpha_i^r\alpha_j^{y_i} - \frac{1}{2}\alpha_i^r\alpha_j^r\right] < \mathbf{x}_i, \mathbf{x}_j >= \\
&= 2\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}\left[\frac{1}{2}c_i^r c_j^r A_i A_j - \frac{1}{2}c_i^r A_i\alpha_j^r + \frac{1}{2}c_j^r A_j\alpha_i^r - \right.\\
&\left. - c_j^{y_i} A_j\alpha_i^r + \alpha_i^r\alpha_j^{y_i} - \frac{1}{2}\alpha_i^r\alpha_j^r\right] < \mathbf{x}_i, \mathbf{x}_j >= \\
&= +\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}\left[\frac{1}{2}c_i^r c_j^r A_i A_j - c_j^{y_i} A_j\alpha_i^r + \alpha_i^r\alpha_j^{y_i} - \frac{1}{2}\alpha_i^r\alpha_j^r\right] < \mathbf{x}_i, \mathbf{x}_j > + \\
&+ 2\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r.
\end{aligned}$$

Por tanto el problema a resolver sería el siguiente

$$\begin{aligned}
\textbf{(PWW)} \quad \min \quad & 2\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}\left[\frac{1}{2}c_i^r c_j^r A_i A_j - c_j^{y_i} A_j\alpha_i^r + \alpha_i^r\alpha_j^{y_i} - \frac{1}{2}\alpha_i^r\alpha_j^r\right] < \mathbf{x}_i, \mathbf{x}_j > \\
\text{s.a:} \quad & \sum_{i=1}^{n}\alpha_i^r = \sum_{i=1}^{n} A_i c_i^r, \\
& \xi_i^r \geq 0, \quad 0 \leq \alpha_i^r \leq C, \quad i = 1, \ldots, n, \quad r \in \{1, \ldots, k\}\backslash y_i.
\end{aligned}$$

Este modelo no es el único existente, como explicamos a principio de la Sección 3.2 en este trabajo se desarrollará el modelo propuesto por Cramer-Singer que estudiaremos a continuación.

3.2.2 Modelo Crammer-Singer

Otro modelo para clasificar las k clases simultáneamente es el realizado por K.Crammer y Y.Singer [12], en este partimos de un conjunto de observaciones $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ siendo $\mathbf{x}_i \in \Omega \subset \mathbb{R}^m$ e $y_i \in Y = \{1, \ldots, k\}$, con $i = 1, \ldots, n$, para explicar esta formulación pasaremos a desarrollar el clasificador que vamos a utilizar en esta sección.

Defininmos una matriz M de dimensión $k \times m$ cuyas filas equivaldrían al valor de $y \cdot \mathbf{v}$, con $\mathbf{v}$ el vector normal del hiperplano H dado en la Definición 2.1 y cada columna correspondería con una de las k clases existentes. El clasificador que aplicaremos es una función $C_{\mathbf{M}} : \Omega \subset \mathbb{R}^n \rightarrow Y$ que asocia un valor y a la variable $\mathbf{x}_i$, la expresión del clasificador vendrá dada por

$$C_M(\mathbf{x}_i) = arg \max_{r=1}^{k} \{M_r \cdot \mathbf{x}_i\}, \quad \text{siendo } M_r \text{ la fila r-ésima de } \mathbf{M}.$$

Decimos que $(\mathbf{x}_i, y_i)$ está mal clasificado cuando $C_M(\mathbf{x}_i) \neq y_i$. Definimos $[\pi]$ como 1 cuando π es cierto y 0 en caso contrario. Por tanto el error empírico de un problema mutliclase es la cantidad de veces que una observación está mal clasificada dividido por el número total de observaciones, es decir,

$$\epsilon_S(M) = \frac{1}{n} \sum_{i=1}^{n} [C_M(\mathbf{x}_i) \neq y_i]. \tag{3.16}$$

La definición que se ha dado de clasificador es buena cuando el número total de clases es igual a dos, pero cuando este es mayor o iguala a tres es menos eficiente, por ello para aplicar el clasificador a un problema multiclase reemplazamos $arg \max_{r=1}^{k} \{M_r \cdot \mathbf{x}_i\}$ por

$$\max_r \{M_r \mathbf{x}_i + 1 - \delta_{y_i,r}\} - M_{y_i} \mathbf{x}_i, \tag{3.17}$$

donde $\delta_{p,d}$ es un parámetro cuyo valor es 1 si $p = d$ y 0 en caso contrario. La expresión (3.17) es 0 cuando $M_{y_i} \mathbf{x}_i \geq M_r \mathbf{x}_i + 1$, para todo $r \in \{1, \ldots, k\} \backslash y_i$, esto ocurre cuando $(\mathbf{x}_i, y_i)$ se encuentra bien clasificado. Por otro lado, cuando el máximo que se busca no coincide con la clase y_i, nos encontramos dos situaciones

COMENTARIO: Las desigualdades de abajo no son realmente así ¿no?, sería con $max\{M_r x_i\}$. Me baso en el documento de Cramer-Singer pag 4, ultimo párrafo. Por otro lado quiero poner unas imágenes en lugar de las usar las figuras 2.5a, 2.5b de la

sección anterior, pero antes de añadirles necesito saber si es correcto lo que os estoy preguntando.

- $M_r\mathbf{x}_i - M_{y_i}\mathbf{x}_i \leq 0$ en cuyo caso $(\mathbf{x}_i, y_i)$ está bien clasificado, pero se encuentra fuera del hiperespacio adecuado, como se muestra en la Figura (PONER IMAGEN)
- $M_r\mathbf{x}_i - M_{y_i}\mathbf{x}_i \geq 0$ en esta situación $(\mathbf{x}_i, y_i)$ se encuentra mal clasificado, como se muestra en la Figura (PONER IMAGEN)

Con la expresión (3.17) se puede obtener una cota superior del error empírico, siendo esta

$$\epsilon_S(M) \leq \frac{1}{n}\sum_{i=1}^{n}[\max_r\{M_r\mathbf{x}_i + 1 - \delta_{y_i,r}\} - M_{y_i}\mathbf{x}_i].$$

Mediante esta cota superior del error empírico se puede ver si las observaciones son linealmente separables, cuasi-separables o linealmente no separables. En este trabajo estudiaremos el caso cuasi-separable, puesto que es una generalización de los casos linealmente separables y linealmente no separables.

La cota superior (3.17) verifica que si las observaciones son cuasi-separables entonces $\max_r\{M_r\mathbf{x}_i + 1 - \delta_{y_i,r}\} - M_{y_i}\mathbf{x}_i = \xi_i$ para $i \in \{1, \ldots, n\}$, con $\xi_i \geq 0$ una variable de holgura, puesto que existirá un error de clasificación asociado a cada observación. De esta expresión se deduce que $M_{y_i}\mathbf{x}_i - M_r\mathbf{x}_i + \delta_{y_i,r} \geq 1 - \xi_i$.

Por tanto la expresión del problema sería la siguiente

$$\begin{aligned} \textbf{(PCS1)} \quad \min \quad & \frac{1}{2}\beta\sum_{r=1}^{k}\|M_r\|_2^2 + \sum_{i=1}^{n}\xi_i \\ \text{s.a:} \quad & M_{y_i}\mathbf{x}_i - M_r\mathbf{x}_i + \delta_{y_i,r} \geq 1 - \xi_i, \\ & i = 1, \ldots, n, \quad r \in \{1, \ldots, k\}, \end{aligned}$$

siendo $\beta \in \mathbb{R}$ una constante reguladora entre la matriz $\mathbf{M}$ y la variable de holgura $\boldsymbol{\xi}$.

En el problema anterior no encontramos a simple vista la restricción $\xi_i \geq 0$, pero debido a que este problema se define para cada $i = 1, \ldots, n$ y $r = \in \{1, \ldots, k\}$, el caso en el que $i = r$ la restricción tiene la siguiente expresión

$$M_{y_i}\mathbf{x}_i - M_{y_i}\mathbf{x}_i + 1 \geq 1 - \xi_i \qquad \text{o lo que es lo mismo} \qquad \xi_i \geq 0.$$

Tras esta puntualización aplicamos la relajación Lagrangiana, por lo que necesitamos conocer la expresión de su función de Lagrange, que es la siguiente

$$L(M_r, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\beta \sum_{r=1}^{k} \|M_r\|_2^2 + \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_i^r (M_{y_i}\mathbf{x}_i - M_r\mathbf{x}_i + \delta_{y_i,r} - 1 + \xi_i), \quad (3.18)$$

con $\alpha_i^r \geq 0$ para $i = 1, \ldots, n$ y $r = 1, \ldots, k$ los multiplicadores de lagrange.

Empleando un razonamiento similar al usado en la Sección 2.1 aplicamos el Teorema de Karush-Kuhn-Tucker 2.1.3 , para ello realizamos las derivadas parciales de la función de Lagrange con respecto a las variables $\boldsymbol{\xi}, \mathbf{M}$.

$$\frac{\partial L(M_r, \boldsymbol{\xi}, \boldsymbol{\alpha})}{\partial \xi_i} = 1 - \sum_{r=1}^{k} \alpha_i^r = 0, \qquad i = 1, \ldots, n, \quad (3.19)$$

$$\frac{\partial L(M_r, \boldsymbol{\xi}, \boldsymbol{\alpha})}{\partial M_r} = \beta M_r + \sum_{i=1}^{n} \alpha_i^r \mathbf{x}_i - \sum_{i=1, y_i=r}^{n}\sum_{r=1}^{k} \alpha_i^r \mathbf{x}_i = 0, \qquad r = 1, \ldots, k. \quad (3.20)$$

Desarrollemos las expresiones anteriores, de la primera de ellas se puede deducir que

$$\sum_{r=1}^{k} \alpha_i^r = 1, \quad i = 1 \ldots, n. \quad (3.21)$$

De la expresión (3.20) deducimos

$$\beta M_r + \sum_{i=1}^{n} \alpha_i^r \mathbf{x}_i - \sum_{i=1, y_i=r}^{n}\sum_{r=1}^{k} \alpha_i^r \mathbf{x}_i = \beta M_r + \sum_{i=1}^{n} \alpha_i^r \mathbf{x}_i - \sum_{i=1, y_i=r}^{n} \left(\sum_{r=1}^{k} \alpha_i^r \right) \mathbf{x}_i =$$
$$= \beta M_r + \sum_{i=1}^{n} \alpha_i^r \mathbf{x}_i - \sum_{i=1, y_i=r}^{n} \mathbf{x}_i = \beta M_r + \sum_{i=1}^{n} \alpha_i^r \mathbf{x}_i - \sum_{i=1}^{n} \delta_{y_i,r} \mathbf{x}_i = 0, \quad i = 1 \ldots, n,$$

donde despejando la variable M_r obtenemos

$$M_r = \frac{1}{\beta} \left(\sum_{i=1}^{n} (\delta_{i,r} - \alpha_i^r) \mathbf{x}_i \right), \quad r = 1 \ldots, k. \quad (3.22)$$

A continuación sustituimos las expresiones (3.21) y (3.22) en (3.18) obteniendo

$$L(M_r, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\beta\sum_{r=1}^{k}\|M_r\|_2^2 + \sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r M_{y_i}\mathbf{x}_i + \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r M_r\mathbf{x}_i -$$
$$- \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r\delta_{y_i,r} + \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r - \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r\xi_i =$$
$$= \frac{1}{2}\beta\sum_{i=1}^{k}\|M_r\|_2^2 + \sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r M_{y_i}\mathbf{x}_i + \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r M_r\mathbf{x}_i +$$
$$+ \sum_{i=1}^{n}1 - \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r\delta_{y_i,r} =$$
$$= \underbrace{\frac{1}{2}\beta\sum_{r=1}^{k}\|M_r\|_2^2}_{(S_1)} \underbrace{- \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r M_{y_i}\mathbf{x}_i}_{(S_2)} + \underbrace{\sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r M_r\mathbf{x}_i}_{(S_3)} - \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r\delta_{y_i,r} + n.$$

Tras el desarrollo anterior, continuaremos con la explicación de cada uno de los grupos que se han definido como $(S_1), (S_2), (S_3)$. Comenzamos con (S_1), sustituimos la expresión (3.22)

$$(S_1) = \frac{1}{2}\beta\sum_{r=1}^{k} < M_r, M_r > = \frac{1}{2}\beta\sum_{r=1}^{k}\left[\frac{1}{\beta}\sum_{i=1}^{n}(\delta_{y_i,r} - \alpha_i^r)\mathbf{x}_i\right]\left[\frac{1}{\beta}\sum_{j=1}^{n}(\delta_{y_j,r} - \alpha_j^r)\mathbf{x}_j\right] =$$
$$= \frac{1}{2\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k}(\delta_{y_i,r} - \alpha_i^r)(\delta_{y_j,r} - \alpha_j^r).$$

Tras obtener la expresión (S_1), pasaremos a hacer lo mismo con (S_2) aplicando en su desarrollo la definición de la variable $\delta_{i,r}$, esto es, $\delta_{i,r} = 1$ si $i = r$ y $\delta_{i,r} = 0$ en caso contrario.

$$(S_2) = \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r M_{y_i}\mathbf{x}_i = \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_i^r\mathbf{x}_i\frac{1}{\beta}\sum_{j=1}^{n}(\delta_{y_j,y_i} - \alpha_{y_j}^{y_i})\mathbf{x}_j =$$
$$= \frac{1}{\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > (\delta_{y_j,y_i} - \alpha_{y_j}^{y_i})\sum_{r=1}^{k}\alpha_i^r = \frac{1}{\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > (\delta_{y_j,y_i} - \alpha_{y_j}^{y_i}) =$$
$$= \frac{1}{\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k}\delta_{y_i,r}(\delta_{y_j,r} - \alpha_{y_j}^r).$$

Por último desarrollamos (S_3)

$$(S_3) = \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_i^r M_r \mathbf{x}_i = \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_i^r \mathbf{x}_i \frac{1}{\beta}\sum_{j=1}^{n}(\delta_{j,r} - \alpha_j^r)\mathbf{x}_j =$$
$$= \frac{1}{\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k} \alpha_i^r(\delta_{i,r} - \alpha_i^r).$$

Ya obtenidas las expresiones de $(S_1), (S_2), (S_3)$ las sustituimos en (3.23)

$$\begin{aligned} L(M_r, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k}(\delta_{y_i,r} - \alpha_i^r)(\delta_{y_j,r} - \alpha_j^r) - \\ - \frac{1}{\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k} \delta_{y_i,r}(\delta_{y_j,r} - \alpha_{y_j}^r) - \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_i^r \delta_{y_i,r} + n + \\ + \frac{1}{\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k} \alpha_i^r(\delta_{i,r} - \alpha_i^r) \\ = \frac{1}{2\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k}(\delta_{y_i,r} - \alpha_i^r)(\delta_{y_j,r} - \alpha_j^r) - \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_i^r \delta_{y_i,r} + n - \\ - \frac{1}{\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k}(\delta_{y_i,r} - \alpha_i^r)(\delta_{y_j,r} - \alpha_j^r) = \\ = -\frac{1}{2\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k}(\delta_{y_i,r} - \alpha_i^r)(\delta_{y_j,r} - \alpha_j^r) - \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_i^r \delta_{y_i,r} + n. \end{aligned}$$

En el problema a resolver hemos excluido el término n de la ecuación anterior por ser una constante y no influir en la solución del problema, luego el problema pasaría a ser el siguiente

$$\begin{aligned} \textbf{(PCS)} \quad \text{mín} \quad & -\frac{1}{2\beta}\sum_{i=1}^{n}\sum_{j=1}^{n} < \mathbf{x}_i, \mathbf{x}_j > \sum_{r=1}^{k}(\delta_{y_i,r} - \alpha_i^r)(\delta_{y_j,r} - \alpha_j^r) - \sum_{i=1}^{n}\sum_{r=1}^{k} \alpha_i^r \delta_{y_i,r} \\ \text{s.a:} \quad & \sum_{r=1}^{k} \alpha_i^r = 1, \\ & \alpha_i^r \geq 0, \quad i = 1, \ldots, n, \quad r \in \{1, \ldots, k\}. \end{aligned}$$

Hemos analizado 5 métodos diferentes para el modelo de SVM-Multiclase, a continuación se hará una recopilación de las propiedades de cada método, con lo que se podrán ver con más facilidad similitudes y diferencias entre estos

Métodos	Número de Clasficadores a entrenar	Forma de abordar aprendizaje	Forma de abordar test	Término independiente hiperplano
OVO (One vs One)	$\frac{k(k-1)}{2}$	Usando múltiples clasificadores binarios	Max-Wins	Si
OVA (One vs All)	k	Usando múltiples clasificadores binarios	Se selecciona la clase con el voto más alto	Si
DAGSVM (Directed Acyclil Graph)	$\frac{k(k-1)}{2}$	Usando múltiples clasificadores binarios	Similar al método OVO	Si
W-W (Weston-Watkins)	1	Resolviendo un solo problema de optimización	Usa el clasificador	Si
C-S (Crammer-Singer)	1	Resolviendo un solo problema de optimización	Usa el clasificador	No

CAPITULO

4

EXPERIMENTOS COMPUTACIONALES

En esta sección se analizarán los métodos explicados en la Sección 3 estos son, OVO, OVA, DAGSVM, Modelo Weston-Watkins y Modelo Crammer-Singer, el objetivo es comparar los resultados obtenidos al aplicar dichos métodos. Los experimentos que se llevarán a cabo en esta sección han sido realizado en el entorno Jupyter usando el lenguaje de programación Python.

Para los modelos OVO, OVA y Crammer-Singer se han utilizado librerías del paquete sklearn que los resuelven mientras que por otro lado para los modelos DAGSVM y Weston-Watkins se ha tenido que desarrollar el código para resolverlos, puesto que no existen librerías como en los casos anteriores que los resuelvan.

4.1 CONJUNTOS DE ENTRENAMIENTO

Para realizar los experimentos computacionales se han usado 3 bases de datos las cuales se pueden encontrar en el repositorio UCI, estas bases de datos contienen una pequeña cantidad de atributos debido al incremento en tiempo computacional que supondría hacer los cálculos con más atributos. A continuación se explicará brevemente las bases de datos que se utilizarán.

- IRIS: Es una de las bases más usadas en reconocimiento de patrones. Esta base de datos posee 150 plantas de género Iris, en concreto a lo largo de las 150 observaciones nos encontramos con 3 clases distintas, Iris Setosa, Iris Versicolour e Iris Virginica, cada una de ellas con 50 de las 150 observaciones de la base de datos. Posee 4 atributos que son longitud y anchura del sépalo y del pétalo.

- Contraceptive Method Choice (CMC): En este base de datos se posee información de 1473 mujeres casadas, en concreto se han estudiado 10 características de estas mujeres como la edad, nº de hijos que ha tenido, su educación, la educación de su pareja... La variable que determina la clase es que método anticonceptivo usan, existen tres posibilidades, que no usen, un método de larga duración o un método de corta duración.

- Car Evolution Database (CED): Esta base de datos dispone de información sobre 1728 coches, en concreto el estudio provee 6 atributos para cada coche como número de puertas, precio de compra, capacidad del vehículo..., y así clasifica los coches en 4 clases diferentes.

La tabla de abajo muestra un pequeño resumen de las propiedades que conocemos de las bases de datos, la variable **m** corresponde con el número de datos, **d** con la cantidad de atributos y **n** el número de clases.

Bases de datos	**m**	**d**	**n**
IRIS	150	4	3
CMC	1473	10	3
CED	1728	6	4

4.2 RESULTADOS PREVIOS

COMENTARIO: Quiero escribir la tarjeta que estoy usando para hacer las operaciones en python..

Para valorar la eficiencia de cada método, se utilizarán los valores de ACC, el parámetro C que da mayor precisión a cada método y el tiempo que tarda el programa en realizar el

entrenamiento del modelo y la obtención de estos valores. A continuación explicaremos con más detalle los dos primeros valores que se han expuesto.

- ACC: Este valor nos indica la calidad del modelo, esto es, el porcentaje de elementos clasificados correctamente.

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}$$

siendo VP,VN,FP y FNv denonimados como verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos respectivamente y corresponden a los valores que se encuentran en la matriz de confusión.

		Valor Predicho	
		1	0
Valor Real	1	Verdadero Positvo	Falso Negativo
	0	Falso Positivo	Verdadero Negativo

- Parámetro C: Con este parámetros se penaliza el uso de las variable de holgura ξ. Para la obtención del parámetro C utilizaremos la técnica de la validación cruzada (Cross Validation), para este proceso se busca dividir los datos en T pliegues o conjuntos, de estos T pliegues se utilizan (T-1) para el entrenamiento del modelo y el restante para la fase de testeo, el algoritmo en cuestión se entrenará y probará T veces, cada vez que se utiliza un nuevo conjunto como conjunto de prueba. Los parámetros que se probarán de C son los siguientes

$$C = \{2^{-4}, 2^{-3}, 2^{2}, 2^{-1}, 1, 2, 2^{2}, 2^{3}, 2^{4}\}$$

COMENTARIO: Tengo que añadir las tablas con los resultados, pero estoy programando los métodos DAGSVM y WW, que no vienen en librerías de python, y la división en grupos de 5 con una proporción de clases parecida en cada grupo.

CAPITULO

5

POR AHORA NADA

5.1 Lo siguiente

5.2 Otra cosa más

5.2.1 Una subcosa

CAPITULO

6

Conclusiones

Bibliografía

[1] S. Abe. Analysis of multiclass support vector machines. *International Conference on Computational Intelligence for Modelling Control and Automation*, 01 2003.

[2] N. Agarwal, V. Balasubramanian, and C. Jawahar. Improving multiclass classification by deep networks using dagsvm and triplet loss. *Pattern Recognition Letters*, 112, 07 2018.

[3] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. 26

[4] G. Betancourt. Las máquinas de soporte vectorial (svms). *Scientia Et Technica*, 01 2005.

[5] J. A. Blanco. Víctor and P. Justo. Optimal arrangements of hyperplanes for svm-based multiclass classification. *Advances in Data Analysis and Classification*, 14, 07 2019.

[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92)*, Pittsburgh, PA, USA, July 1992. ACM Press. 3

[7] E. Bredensteiner and K. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12, 09 1999. 39

[8] Y.-D. Cai, P.-W. Ricardo, C.-H. Jen, and K.-C. Chou. Application of svm to predict membrane protein types. *Journal of Theoretical Biology*, 226(4):373–376, 2004. 3

[9] E. Carmona. *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. 11 2016.

[10] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46, 05 2001. 34

[11] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20, 1995. 3

[12] K. Crammer, Y. Singer, N. Cristianini, J. Shawe-Taylor, and B. Williamson. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res*, 2, 01 2002. 45

[13] J. H. Friedman. Technical report, Department of Statistics, Stanford University. 31

[14] T. Gao and D. Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *2011 International Conference on Computer Vision*, pages 2072–2079, 2011. 34

[15] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[16] B. Kijsirikul and N. Ussivakul. Multiclass support vector machines using adaptive directed acyclic graph. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, volume 1, pages 980–985 vol.1, 2002.

[17] Y. Ma and G. Guo. *Support Vector Machines Applications*. 9 2014.

[18] J. Martínez, C. Iglesias, J. Matías, J. Taboada, and M. Araújo. Solving the slate tile classification problem using a dagsvm multiclassification algorithm based on svm binary classifiers with a one-versus-all approach. *Applied Mathematics and Computation*, 230:464–472, 2014.

[19] W. S. Noble. 1 support vector machine applications in computational biology. 3

[20] D. A. Otchere, T. O. Arbi Ganat, R. Gholami, and S. Ridha. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ann and svm models. *Journal of Petroleum Science and Engineering*, 200:108182, 2021. 3

[21] P. Panda and K. Roy. Attention tree: Learning hierarchies of visual features for large-scale image recognition. *CoRR*, abs/1608.00611, 2016. 34

[22] J. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin dags for multiclass classification. *Advances in neural information processing systems*, 12, 03 2000. 34

[23] S. Rakhmetulayeva, K. Duisebekova, A. Mamyrbekov, D. Kozhamzharova, G. Astaubayeva, and K. Stamkulova. Application of classification algorithm based on svm for determining the effectiveness of treatment of tuberculosis. *Procedia Computer Science*, 130:231–238, 2018. The 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018) / The 8th International Conference on Sustainable Energy Information Technology (SEIT-2018) / Affiliated Workshops. 3

[24] Rostami, Hamidey, Dantan, Jean-Yves, and Homri, Lazhar. Review of data mining applications for quality assessment in manufacturing industry: support vector machines. *Int. J. Metrol. Qual. Eng.*, 6(4):401, 2015. 3

[25] Y. Shen, C. Wu, C. Liu, Y. Wu, and N. Xiong. Oriented feature selection svm applied to cancer prediction in precision medicine. *IEEE Access*, 6:48510–48521, 2018. 3

[26] F. Takahashi and S. Abe. Optimizing directed acyclic graph support vector machines. *Artificial Neural Networks in Pattern Recognition (ANNPR)*, 01 2003. 34

[27] Y.-C. F. Wang and D. Casasent. A support vector hierarchical method for multiclass classification and rejection. In *2009 International Joint Conference on Neural Networks*, pages 3281–3288, 2009.

[28] J. Weston and C. Watkins. Multi-class support vector machine. 03 1999. 39