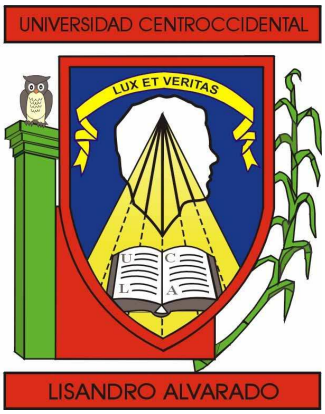
UNIVERSIDAD CENTROCCIDENTAL

"LISANDRO ALVARADO"

Decanato de Ciencias y Tecnología

Licenciatura en Ciencias Matemáticas.



MÁQUINAS DE VECTORES DE SOPORTE CON NORMA-1 VIA MINIMIZACIÓN CONVEXA IRRESTRICTA

Trabajo Especial de Grado presentado por

Br. Ifigenia Yeniree Romero Arenas.

Como requisito final para obtener el título de

Licenciado en Ciencias Matemáticas

Área de Conocimiento: Matemática aplicada.

Tutor: Dr. Javier Hernández Benítez.

Barquisimeto - Venezuela

Febrero 2012

Máquinas de Vectores de Soporte Con Norma-1 Via Minimización Convexa Irrestricta

Br. Ifigenia Y. Romero A.

RESUMEN

Las máquinas vectoriales de soporte (SVMs por sus siglas en inglés: Support vector machines) usando norma 1, modeladas mediante un problema de programación lineal (ver[8]), se puede formular como la minimización sin restricciones de una función convexa diferenciable y cuadrática a trozos en el espacio dual. La función objetivo, la cual tiene gradiente Lipschitz continua y contiene sólo un parámetro finito adicional al problema, se puede minimizar mediante el método de Newton generalizado para obtener una solución exacta del problema SVM. El enfoque del artículo [9], el cual soporta este trabajo, se basa en la formulación de un problema lineal muy general como un problema de minimización sin restricciones y su aplicación a problemas de máquinas vectoriales de soporte.

A Dio ser supremo que le debo lo que soy y seré, es mi guía espiritual y a quien le pido ilumine mi camino siempre. A mi abuela María que desde niña con mucho amor y cariño me guió por el camino correcto y hoy desde el cielo me bendice. A mi Ángel, hermanita Mariangela que desde el cielo sigue llenando mi vida con su dulce risa. A mis padres, Eugenio y Ketty, quienes me dieron el ser y me ayudaron a realizar cada uno de mis sueños, estuvieron cuando los necesite y me dieron la enseñanza para hoy honrarles. A mi segunda madre Maribel, por su incondicional apoyo en todo lo que hago por su comprensión, paciencia y su amor. Eres mi mejor ejemplo de constancia y superación.

Índice general

Índice de figuras

Agradecimientos

A Dios por iluminarme y no permitir que los obstáculos que se me presentaron en el transcurso de esta carrera me vencieran.

A mis padres Eugenio y Ketty que con su cariño comprensión y apoyo guiaron mi camino por el bien. En especial a mi madre, te admiro porque nunca te has dejado vencer por nada, ni nadie, eres mi mejor ejemplo a seguir. Te Amo Madre.

A mi mami Maribel, por ayudarme cuando lo necesite, por sus enseñanzas y por no perder la fe en mi.

A mi hermana Oriana por acompañarme en cada uno de mis sueños, por comprenderme, apoyarme y ayudarme. Te Amo hermanita.

A mis tíos, Marcial, Mirla, Miriam, José Gregorio y Nairodis; ejemplo de superación que estuvieron siempre ayudándome y animándome a no decaer.

A mis Primos; Lorena, Juan Pablo, Miguel, Alba, María Gabriela, Javier, Junior, Rubén, Margareth, Isaac, Paola, Ashleymar, Elismar, Marcial, Camila, An-

drea, Orianthi, Juan Daniel, Mónica y a los que aun no llegan, que este triunfo sea un ejemplo para ellos.

A mi amor, Gleybin quien me ha apoyado y ayudado en los momentos difíciles, por soportarme en lo momentos de estrés, por sus consejos y por ser parte de mi vida. Te adoro chinito.

A mi amiga Yesenia por estar conmigo en este largo camino, dándome sus consejos animándome y ayudándome en momentos difíciles, como también a su familia por permitirme entrar a su hogar y brindarme su apoyo.

A mis amigo, Laura, Yorgeth, Katty, Elismar, Carlos, Yhon. Ander, Edward y Rafael por sus buenos consejos y cooperación durante la carrera y por compartir momentos únicos e inolvidables.

A mi tutor Javier Hernández que con su valiosa enseñanza y paciencia contribuyo con la culminación de este trabajo, sin su ayuda y conocimientos no estaría donde me encuentro ahora. Mil gracias profe.

A la profesora Jurancy Ereú por sus enseñanzas y sus buenos consejos.

A los profesores de la UCLA DCyT que participaron en mi enseñanza durante mi carera.

A la institución UCLA por brindarme su la estadía durante el desarrollo de mi carrera.

A todos los que colocaron su granito de arena para lograr este triunfo.

Introducción

Las máquinas de vectores de soporte son un conjunto de algoritmos de aprendizaje supervisado desarrollados originalmente por Vladimir Vapnik [14]; dado un conjunto de ejemplos de muestras (de entrenamiento) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase. Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensión muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión.

Dado un conjunto de puntos en un espacio, en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo,cuya categoría desconocemos, pertenece a una categoría o a la otra. La SVM busca un hiperplano que separe de forma óptima a los puntos de una clase de la de otra, que eventualmente han podido ser

previamente proyectados a un espacio de dimensión superior.

En ese concepto de "separación óptima" es donde reside la característica fundamental de las SVM: este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. Por eso también a veces se les conoce a las SVM como clasificadores de margen máximo. De esta forma, los puntos del vector que se etiquetados con una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado.

Una de las principales ventajas de las máquinas vectoriales de soporte con norma-1 (SVM) es que a diferencia de SVMs con norma-2, son muy efectivas en la reducción de las características de espacios de entrada para los núcleos lineales y en la reducción del numero de funciones del núcleo (ver [1]) para SVMs no lineales. Con algunas excepciones el método simplex (ver [2]) ha sido el algoritmo exclusivo para solucionar SVMs con norma-1. Lo interesante de este trabajo (ver [16]) es que trata SVM con norma-1 usando paquetes de programación lineal estándar para la solución de su formulación. Para mejoras de nuestros conocimientos no ha habido una formulación completamente exacta de SVMs con norma-1, via minimización diferenciable sin restricción, que es el asunto principal de la presente contribución teórica que planteamos ahora.

La estructura de esta monografía es la siguiente: en el capítulo 1, damos algunos aspectos en la teoría de optimización en la cual se basa lo desarrollado. También damos una pequeña introducción de las máquinas de vectores de soporte.

En el capítulo 2 se muestra como un problema lineal muy general puede ser resuelto como la minimización de una función convexa diferenciable cuadrática a trozos completamente sin restricciones que contiene un sólo parámetro finito. Este resultado se generaliza los resultados que se muestran en [12] donde programas lineales con millones de restricciones fueron resueltos como programas de minimización sin restricción por un método de Newton generalizado (ver [9]). En el capítulo 3 se muestra como configurar una SVMs con norma 1, con núcleo lineal y no lineal como un problema de minimización sin restricción y se establece un método de Newton generalizado para su resolución.

En el capítulo 4 se muestra como resolver el problema de aproximación de una función desconocida basada en un numero determinado de valores de la función usando un numero mínimo de la función núcleo. Esto se logra convirtiendo, de nuevo, problemas de aproximación con norma-1 en un problema de minimización sin restricción. Los resultados computacionales muestran que el enfoque propuesto por [9] es mas rápido de resolver que una programación lineal convencional, CPLEX (ILO, 2003).

Ahora describiremos nuestra notación y daremos a conocer un poco del material a fondo. Todos los vectores son vectores columnas a no ser incorporado a un vector fila por un prima $'$. Para un vector x n-dimensional en el espacio real $\mathbb{R}^n$, x_+ denota el vector en $\mathbb{R}^n$ que todas sus componentes negativas se hacen cero esto corresponde a la proyección de x en el octante no negativo. Para un vector x en $\mathbb{R}^n$, x_* denota el vector en $\mathbb{R}^n$ con componentes $(x_*)_i = 1$ si $x_i > 0$ y cero en otro modo (es decir x_* es el resultado de aplicar la función de paso de componente racional de x). Para $x \in \mathbb{R}^n$, $\|x\|_1$, $\|x\|$, $\|x\|_\infty$, denotara la norma-

1, norma-2, y norma- ∞ de x. Para simplificar eliminamos el 2 para $\|x\|_2$. La notación $A \in \mathbb{R}^{m\times n}$ significara una matriz real $m \times n$ por tal motivo una matriz A^t denota la traspuesta de A, A_i denotara la fila i-esima de A y A_{ij} denotara el elemento ij-esimo de A. Un vector de unos o ceros en un espacio real de dimensiones arbitrarias se denota por e o 0 respectivamente. Para una función cuadrática a trozos tales como, $f(x) = \frac{1}{2}\|(Ax-b)\|^2 + \frac{1}{2}x^T Px$, donde $A \in \mathbb{R}^{m\times n}$, $P \in \mathbb{R}^{n\times n}$, $P^T = P$, P semidefinida positiva y $b \in \mathbb{R}^m$, el Hessiano común no existe porque su gradiente el vector $n \times 1$ $\nabla f(x) = A^T(Ax-b)_+ + Px$ no es diferenciable pero es Lipschitz continua con constante de Lipschitz $\|A\|$ $\|A^T\| + \|P\|$. Sin embargo se puede definir su Hessiano Generalizado (ver [5];[3]; [11]) que es la matriz $n \times n$ simétrica semidefinida positiva:

$$\partial^2 f(x) = A^T diag(Ax-b)_* A + P$$

donde diag $(Ax-b)_*$ denota una matriz diagonal $m \times m$ con elementos en la diagonal $(A_i x - b_i)_*$,con $i = 1...n$. El Hessiano generalizado tiene muchas de las propiedades del Hessiano regular (ver [5]; [3];[11]) en relación con $f(x)$. si el menor valor propio es de $\partial^2 f(x)$ es mayor que alguna constante positiva $\forall x \in \mathbb{R}^n$, entonces $f(x)$ es una función fuertemente cuadrática a trozos convexa en $\mathbb{R}^n$. Un plano separador con respecto a dos los puntos dados de los conjuntos A y B en $\mathbb{R}^n$ es un plano que los intenta separar en dos semiespacios de $\mathbb{R}^n$ tal que cada semiespacio abierto contenga puntos en su mayoría de A o de B. a lo largo de este trabajo la notación $:=$ denotara una definición.

Capítulo 1

Preliminares

La teoría de optimización es una rama de las matemáticas que estudia la resolución de problemas que consisten en determinar el valor mínimo o máximo de una función. Su estudio ha tomado una gran importancia en la actualidad evolucionando con el pasar del tiempo, y la ayuda del computador a sido primordial en esto. Gracias a su estudio se han desarrollado diversos algoritmos que ayudan a resolver dichos problemas. Uno de los estudios de gran importancia es ver las características que cumplen las variables que hacen que la función en cuestión alcance su valor óptimo.

Las condiciones de optimalidad nos aportan las características necesarias (o suficientes) para saber si un punto $x \in \mathbb{R}^n$ es o no solución del problema de optimización. Estas condiciones constituyen la base para muchos de los algoritmos, los cuales nos ayudan a resolver diferentes problemas.

1.1. Condiciones de optimalidad

Consideremos el problema de minimizar una función en $\mathbb{R}^n$ sobre un subconjunto de dicho espacio, que está determinado por un conjunto de restricciones. Una formulación general para este problema es:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ s.a. \quad & C_i(x) = 0, \quad i \in \mathcal{E} \\ & C_i(x) \geq 0, \quad i \in \mathcal{I} \end{aligned} \tag{1.1.1}$$

donde f y las funciones C_i son suaves de valores reales en un sub conjunto de $\mathbb{R}^n$, e $\mathcal{I}$ y $\mathcal{E}$ son dos conjuntos finitos de indices. Al igual que antes llamamos f la función objetivo, mientras que C_i, con $i \in \mathcal{E}$ son restricciones de igualdad y C_i, $i \in \mathcal{I}$ son las restricciones de desigualdad. Se define el conjunto factible Ω como el conjunto de puntos $x \in \mathbb{R}^n$ que satisfacen las restricciones, es decir

$$\Omega = \left\{ x \in \mathbb{R}^n \,\middle|\, C_i(x) = 0, \;\; i \in \mathcal{E}; \;\; C_i(x) \geq 0, \;\; i \in \mathcal{I} \right\}, \tag{1.1.2}$$

de modo que podemos volver a escribir (1.1.1) de la forma siguiente

$$\min_{x\in\Omega} \quad f(x). \tag{1.1.3}$$

Además, la función lagrangiana para el problema (1.1.1) se define:

$$\mathcal{L}_(x, \lambda) = f(x) - \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i C_i(x). \tag{1.1.4}$$

El conjunto de restricciones activas $\mathcal{A}(x)$ en $x \in \Omega$ es la union del conjunto $\mathcal{E}$ con los indices de las restricciones de desigualdad activa, es decir, las

restricciones de desigualdad que en x se cumple la igualdad. En otra palabras

$$\mathcal{A}(x) = \mathcal{E} \cup \{i \in \mathcal{I} | C_i(x) = 0\} \tag{1.1.5}$$

Definición 1.1.1 (LICQ (ver [15])) *Dado el punto x^* y el conjunto $\mathcal{A}(x^*)$ definido por* (1.1.5)*, se dice que $\mathcal{A}(x^*)$ verifica las condiciones de calificación de las restricciones de independencia lineales (LICQ, por sus siglas en inglés: linear independence constrain calification) si el conjunto de los gradientes de las restricciones activas $\{\nabla C_i(x^*), i \in \mathcal{A}(x^*)\}$ es linealmente independiente.*

Tenga en cuenta que si se cumple esta condición, ninguno de los gradientes de restricciones activas puede ser cero.

1.1.1. Condición de primer orden

Las condiciones de primer orden están basadas en el gradiente (vector de las primeras derivadas) de la función objetivo y las restricciones.

Teorema 1.1.2 (Condiciones necesarias de primer orden, (ver [15])) *Supongamos que x^* es una solución local de* (1.1.1) *y que (LICQ) se cumple en x^*, entonces existe λ^* un vector multiplicador de lagrange con componentes λ_i^*, $i \in \mathcal{E} \cup \mathcal{I}$,*

de tal manera que las siguiente condiciones se cumplen para (x^*, λ^*)

$$\begin{aligned} \nabla_x \mathcal{L}_(x^*, \lambda^*) &= 0 && (1.1.6)\\ C_i(x^*) &= 0, \quad \forall i \in \mathcal{E} && (1.1.7)\\ C_i(x^*) &\geq 0, \quad \forall i \in \mathcal{I} && (1.1.8)\\ \lambda_i^* &\geq 0, \quad \forall i \in \mathcal{I} && (1.1.9)\\ \lambda_i^* C_i(x^*) &= 0, \quad \forall i \in \mathcal{E} \cup \mathcal{I}. && (1.1.10) \end{aligned}$$

Las expresiones anteriores se les conoce como las condiciones de Karush-Kuhn-Tuker, o condiciones KKT para abreviar. Las condiciones de complementariedad implican que los multiplicadores de lagrange correspondiente a las restricciones con indices $i \notin \mathcal{A}(x^*)$ deben ser ceros. De la primera condición de KKT obtenemos lo siguiente:

$$0 = \nabla_x \mathcal{L}_(x^*, \lambda^*) = \nabla f(x^*) - \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla C_i(x^*) \qquad (1.1.11)$$

Definición 1.1.3 (complementariedad estricta (ver [15])) *Dada una solución local* x^* *de* (1.1.1) *y un vector* λ^* *que satisface las condiciones Karush-Kuhn-Tucker, diremos que se cumple la condición de complementariedad estricta si se satisface (1.1.9) del teorema anterior con* $\lambda_i^* > 0$ *para cada* $i \in \mathcal{I} \bigcap \mathcal{A}(x)$.

Para el problema dado (1.1.1) y el punto x^* como solución, pueden haber muchos vectores λ^* para los que las condiciones KKT se satisfacen. Pero cuando (LICQ) se cumple, el λ^* óptimo es único.

1.2. Dualidad

En esta sección presentamos algunos elementos de la teoría de dualidad para la programación no lineal. Esta teoría es usada para motivar y desarrollar algunos algoritmos importantes, incluyendo el algoritmo del lagrangiano aumentado, de la teoría de métodos de penalidad. En toda su generalidad, la teoría de dualidad va mas allá de la programación no lineal para proporcionar información importante de la optimización no suave convexa e incluso de la optimización discreta. Su especialización para la programación lineal resultó fundamental para el desarrollo del área.

La teoría de dualidad muestra como podemos construir un problema alternativo para las funciones y datos que definen el problema de optimización original. La alternativa del problema dual esta relacionada con el problema original de una manera fascinante. En algunos casos el problema dual es mas fácil para resolver desde el punto de vista computacional que el problema original. En otros casos el dual puede ser usado para obtener fácilmente una cota inferior del valor óptimo de la función objetivo del problema primal. Como se ha señalado anteriormente el dual también se ha utilizado para diseñar algoritmos que resuelvan el problema primal. Suponemos que existen m-restricciones de desigualdad etiquetadas $1, \ldots, m$ y escritas en (1.1.1) como sigue:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ s.a \quad & C_i(x) \geq 0 \quad i = 1, \ldots, m \end{aligned} \tag{1.2.1}$$

Si unimos las restricciones en una función vectorial

$$C(x) := (C_1(x), C_2(x), \ldots, C_m(x))^T \tag{1.2.2}$$

en este caso $C\mathbb{R}^n \mapsto \mathbb{R}^m$ y en consecuencia el problema (1.2.1) se puede formular de la manera siguiente:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ s.a \quad & C(x) \geq 0_{m\times 1} \end{aligned} \tag{1.2.3}$$

la función lagrangiana del problema anterior queda:

$$\mathcal{L}_(x, \lambda) = f(x) - \lambda^T C(x) \tag{1.2.4}$$

Definamos la función objetivo dual $q : \mathbb{R}^n \mapsto \mathbb{R}$ como sigue:

$$q(\lambda) := \text{ínf}\, \mathcal{L}_(x, \lambda) \tag{1.2.5}$$

En muchos problemas este ínfimo es $-\infty$ para algunos valores de λ. Definamos el dominio de q como el conjunto de valores λ para los que q es finito, tal es:

$$D := \{x | q(\lambda) > -\infty\} \tag{1.2.6}$$

Notemos que el cálculo del ínfimo en (1.2.5) requiere encontrar un mínimo global de la función $\mathcal{L}_(x, \lambda)$ para el λ dado, labor que puede ser extremadamente difícil en la práctica. Sin embargo cuando la función f y $-C_i$ son funciones convexas y $\lambda \geq 0$, la función $\mathcal{L}_(\cdot, \lambda)$ es también convexa en este caso todo mínimo local es un mínimo global. Así los cálculos de $q(\lambda)$ se convierte en una opción mas práctica. El problema dual de (1.2.3) es definido como sigue:

$$\begin{aligned} \max_{x\in\mathbb{R}^n} \quad & q(\lambda) \\ s.a \quad & \lambda \geq 0 \end{aligned} \tag{1.2.7}$$

El siguiente teorema muestra que cualquier valor del problema dual es una cota inferior del valor óptimo del problema primal

Teorema 1.2.1 (Dualidad Débil (ver [15])) *Para cualquier* $\overline{x}$ *factible para* (1.2.3) *y cualquier* $\overline{\lambda}$, *tenemos que* $q(\overline{\lambda}) \leq f(\overline{x})$.

el siguiente resultado muestra como el multiplicador de lagrange óptimo (1.2.3) son soluciones del problema dual (1.2.7) bajo ciertas condiciones

Teorema 1.2.2 (ver [15]) *Supongamos que* $\overline{x}$ *es una solución de* (1.2.3) *y que* f *y* $-C_i$, $i = 1, \ldots, m$ *son funciones convexas en* $\mathbb{R}^n$ *que son diferenciables en* $\overline{x}$. *Entonces cualquier* $\overline{\lambda}$ *para el cual* $(\overline{x}, \overline{\lambda})$ *satisfacen las condiciones de KKT es solución de* (1.2.7)

Teorema 1.2.3 (ver [15]) *Supongamos que* f *y* $-C_i$, $i = 1, \ldots, m$ *son convexas y continuamente diferenciable en* $\mathbb{R}^n$. *Supongamos que* $\overline{x}$ *es una solución de* (1.2.3) *en el que LICQ se cumple. Suponga que* $\widehat{\lambda}$ *resuelve* (1.2.7) *y que el ínfimo en* $\inf_x \mathcal{L}(x, \widehat{\lambda})$ *es alcanzado en* $\widehat{x}$. *Supongamos además que* $\mathcal{L}(\cdot, \widehat{\lambda})$ *es una función estrictamente convexa , entonces* $\overline{x} = \widehat{x}$, *que es,* $\widehat{x}$ *es la única solución de* (1.2.3) *y* $f(\overline{x}) = \mathcal{L}(\widehat{x}, \widehat{\lambda})$.

Una forma diferente para la dualidad que es conveniente para los cálculos, es conocida como el dual Wolfe, se puede decir de la siguiente manera

$$\begin{aligned} \max_{x,\lambda} \quad & \mathcal{L}(x, \lambda) \\ s.a \quad & \nabla \mathcal{L}(x, \lambda) = 0 \quad \lambda \geq 0. \end{aligned} \tag{1.2.8}$$

El siguiente resultado explica la relación de el dual Wolfe para (1.2.3)

Teorema 1.2.4 (ver [15]) *Supongamos que* f *y* $-C_i$, $i = 1, \ldots, m$ *son convexas y continuamente diferenciables en* $\mathbb{R}^n$, *suponga que* $(\overline{x}, \overline{\lambda})$ *es un par solución de* (1.2.3) *en el que LICQ se cumple. Entonces* $(\overline{x}, \overline{\lambda})$ *resuelve el problema* (1.2.8).

Las funciones cuadráticas son de gran importancia en el estudio de optimización; puesto que son curvas suaves, diferenciables y convexas, donde su comportamiento es de gran utilidad al momento de hallar la solución optima (mínimos y máximos). Asimismo en los diversos algoritmo estas funcionan muy bien y es por ello que estudiaremos este tipo de funciones.

Teorema 1.2.5 (ver [7]) *Consideremos un problema de programación con una función objetivo cuadrática y restricciones lineales. Sea* $\tilde{b}$ *un vector en* $\mathbb{R}^n$, $\tilde{c}$ *un vector en* $\mathbb{R}^m$, $\tilde{u}$ *un vector en* $\mathbb{R}^m$, $\tilde{C}$ *una matriz simétrica de orden* $n \times n$ *y* $\tilde{A}$ *una matriz de orden* $m \times n$. *Así el primal para el problema de programación cuadrática esta dado por:*

$$\begin{aligned} \min_{x \in X} \quad & \frac{1}{2} x^T \tilde{C} x - \tilde{b}^T x \\ s.a. \quad & \tilde{A} x \leq \tilde{c}, \end{aligned} \tag{1.2.9}$$

y el dual de Wolfe que viene dado por:

$$\begin{aligned} \max \quad & -\frac{1}{2} x^T \tilde{C} x - \tilde{c}^T \tilde{u} \\ s.a \quad & \tilde{C} x + \tilde{A}^T \tilde{u} = \tilde{b} \\ & \tilde{u} > 0. \end{aligned} \tag{1.2.10}$$

1.3. El método de penalidad cuadrática

Uno de los enfoques fundamentales de la optimización con restricciones es reemplazar el problema original por una función de penalidad que consiste en agregar un término para cada restricción, que es positivo cuando el punto x actual no cumple con esta restricción y cero en caso contrario.

La mayoría de las orientaciones de estos métodos define una sucesión de funciones de penalidad, en la que los términos de penalidad para restricciones son multiplicados por un coeficiente positivo. Al hacer este coeficiente más grande las restricciones de penalidad son más suaves, lo que obliga al minimizador de la función de penalidad a estar cada ves más cerca de la region factible para el problema original.

Estos enfoques se conocen como los métodos de penalidad exterior ya que el término de penalidad para cada restricción es cero cuando x no es factible con respecto a esas restricciones. A menudo los minimizadores de la función de penalidad no son factible con respecto a el problema original y el enfoque de factibilidad se logra cuando el límite del parámetro de penalidad es cada ves mas grande.

La función de penalidad más simple de este tipo es la función de penalidad cuadrática, en el que los términos de penalidad son los cuadrados para las restricciones, se describe la mayor parte de nuestra discusión para el problema con restricciones de igualdad

$$\begin{aligned} \min_{x} \quad & f(x) \\ s.a \quad & C_i(x) = 0, \quad i \in \mathcal{E} \end{aligned} \tag{1.3.1}$$

que es un caso especial de (1.1.1). La función de penalidad cuadrática $Q(x;\mu)$ para esta formulación es:

$$Q(x;\mu) = f(x) + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} C_i^2(x), \tag{1.3.2}$$

donde $\mu > 0$ es el parámetro de penalidad.

El hecho de que $\mu \to 0$ hace que las restricciones de penalidad sean mas suave y obliga a el minimizador de la función de penalidad a estar mas cerca de la region factible. Tiene sentido considerar una sucesión de valores $\{\mu_k\}$ con $\mu_k \downarrow 0$ cuando $k \to \infty$ y buscar el minimizador aproximado x_k de $Q(x_k;\mu_k)$ para cada k. Debido a que los términos de penalidad en (1.3.2) son suaves, podemos utilizar técnicas de optimización irrestricta para la búsqueda de x_k. Los minimizadores aproximados x_k, x_{k-1}, entre otros, pueden se utilizados para la minimización de $Q(\cdot;\mu_{k+1})$ en la iteración $k+1$. Al elegir la sucesión $\{\mu_k\}$ y los puntos de partida con prudencia, puede ser posible llevar a cabo en solo un paso la minimización irrestricta para cada valor de μ_k.

Para el problema general de optimización:

$$\begin{aligned} \min_{x} \quad & f(x) \\ s.a \quad & C_i(x) = 0, \quad i \in \mathcal{E} \\ & C_i(x) \geq 0, \quad i \in I \end{aligned} \tag{1.3.3}$$

que contiene restricciones de desigualdad, así como restricciones de igualdad, se puede definir la función cuadrática de penalidad como:

$$Q(x;\mu) = f(x) + \frac{1}{2\mu}\sum_{i\in\mathcal{E}} C_i^2(x) + \frac{1}{2\mu}\sum_{i\in\mathcal{I}}([C_i(x)]^-)^2, \tag{1.3.4}$$

donde $\mu > 0$, es el parámetro de penalidad y $[C(x)]^- = \text{máx}\{-C_i(x), 0\}$

Nota 1.3.1 *la función* $[C(x)]_+ = \text{máx}\{C_i(x), 0\}$ *así tenemos*

$$[C(x)]^- = -[C(x)]_+$$

1.4. Maquinas vectoriales de soportes (SVM)

Las maquinas de vectores de soportes (SVM por sus siglas en inglés) es una técnica de clasificación y han sido introducida como una poderosa herramienta para resolver problemas de clasificación con una gran cantidad de puntos. Una Maquina de vectores de soporte mapea los puntos de entrada a un espacio de características de una dimensión mayor y encuentra un hiperplano que separa los datos; es decir los puntos que son etiquetados con una categoría estarán a un lado del hiperplano y los que se encuentren en la otra categoría estarán al otro lado, asimismo las SVM maximizan el margen entre dos clases de puntos.

Maximizar el margen entre dos clases de puntos se puede moderar como un problema de programación cuadrática y puede ser resuelto por su problema dual,

introduciendo multiplicadores de lagrange, además consiste en maximizar la distancia entre el hiperplano separador y el valor de entrada mas cercano es decir el vector soporte; que es un punto donde se apoya el margen máximo.(Observe figura 1.1)

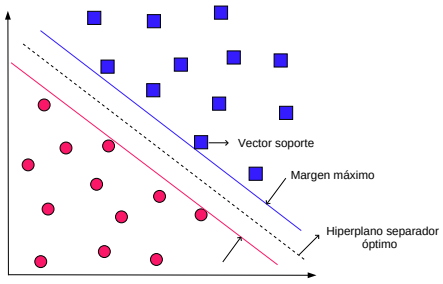


Figura 1.1: Margen máximo

Ahora bien, Consideremos el problema de optimización modelado como un problema de clasificación de maquinas de vectores de soporte SVM estándar como sigue:

$$\begin{array}{lll} \underset{(\omega,\gamma,y)\in\mathbb{R}^{n+1+m}}{\text{mín}} & \nu e^T y + f(\omega) & \\ s.a. & D(A\omega - \gamma e) + y & \geq e \\ & y & \geq 0 \end{array} \tag{1.4.1}$$

donde $f(\omega)$ es una función convexa, diferenciable que por lo general son norma o seminormas, la matriz A de orden $m \times n$ representa los m-puntos en $\mathbb{R}^n$, D representa la matriz diagonal de orden $m \times m$ de las clases de puntos $\{+1\}$ ó $\{-1\}$, e es un vector columna de puros (1) de dimension arbitraria, aquí ω es el vector normal a los planos separadores:

$$\begin{aligned} x^T\omega - \gamma \quad +y = \quad +1 \\ x^T\omega - \gamma \quad -y = \quad -1 \end{aligned} \tag{1.4.2}$$

donde γ es quien determina su posición respecto al origen.(Observe figura (1.2))

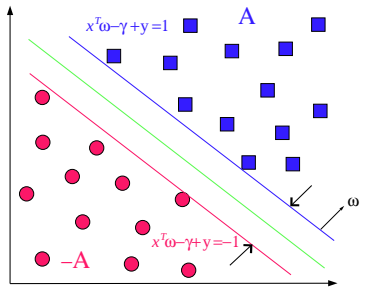


Figura 1.2: SVM Estandar

1.4.1. Caso linealmente separable

En el caso linealmente separable las maquinas vectoriales de soporte SVM están formadas por hiperplano que separa por completo los datos de entrada en dos subgrupos de puntos que poseen una etiqueta como {1} o {-1}. Además en la (1.4.1) cuando $y = 0$ estamos es presencia del caso separable, por lo que existe solo un hiperplano óptimo:(Observe figura 1.3)

$$x^T\omega = \gamma \tag{1.4.3}$$

y este se halla al maximizar el margen entre los planos separadores siguientes:

$$\begin{aligned} x^T\omega - \gamma &= +1 \\ x^T\omega - \gamma &= -1 \end{aligned} \tag{1.4.4}$$

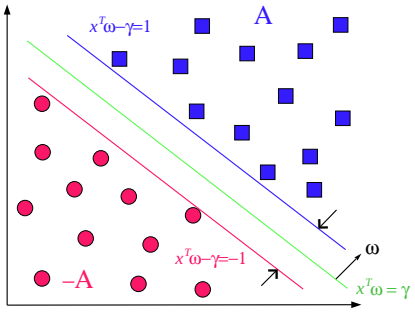


Figura 1.3: Conjunto Linealmente Separable

Observe en (1.4.2) que cuando $y = 0$ ocurre:

$$\begin{aligned} x^T\omega - \gamma &< +1 \\ x^T\omega - \gamma &> -1. \end{aligned} \tag{1.4.5}$$

1.4.2. Caso no linealmente separable

Para el caso donde los datos no pueden ser separados linealmente a través de un hiperplano óptimo, (Observe figura (1.4)) tendríamos en la ecuación (1.4.1) que si la variable de holgura $y > 0$ entonces el caso es no separable y es por ello que las ecuaciones (1.4.2) nos quedan:

$$\begin{aligned} x^T\omega - \gamma \quad +y \geq \quad +1 \\ x^T\omega - \gamma \quad -y \leq \quad -1 \end{aligned} \tag{1.4.6}$$

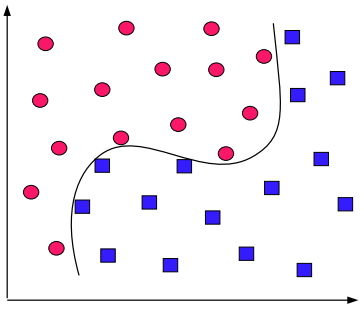


Figura 1.4: Conjunto no Linealmente Separable

Sin embargo, existe la posibilidad de separar los puntos pero en un espacio de mayor dimension, esto ocurre al transformar los datos de entrada para así aplicar los mismos razonamientos que las maquinas vectoriales de soporte lineales con margen máximo.

Por otra parte, la transformación de los datos de un espacio entrada a otro de mayor dimensión se logra mediante el uso de la función núcleo. En este sentido una función núcleo es un producto interno en el espacio de características que tiene su equivalente en el espacio de entrada dado por:

$$K(x, x^T) = \langle \phi(x), \phi(x^t) \rangle \tag{1.4.7}$$

donde k es una función simétrica definida positiva que cumple las condiciones de Mercer.

A continuación mostramos, un programa matemático para un núcleo general $K(A, A^T)D = A$, así el problema (1.4.1) se transforma:

$$\begin{array}{lll} \underset{(\omega,\gamma,y)\in\mathbb{R}^{n+1+m}}{\text{mín}} & \nu e^T y + f(\omega) & \\ s.a. & D(K(A,A^T)D\omega - \gamma e) + y \geq e & \\ & y \geq 0 & \end{array} \tag{1.4.8}$$

En la figura (1.5) se muestra en efecto de la función núcleo para una SVM no linealmente separable

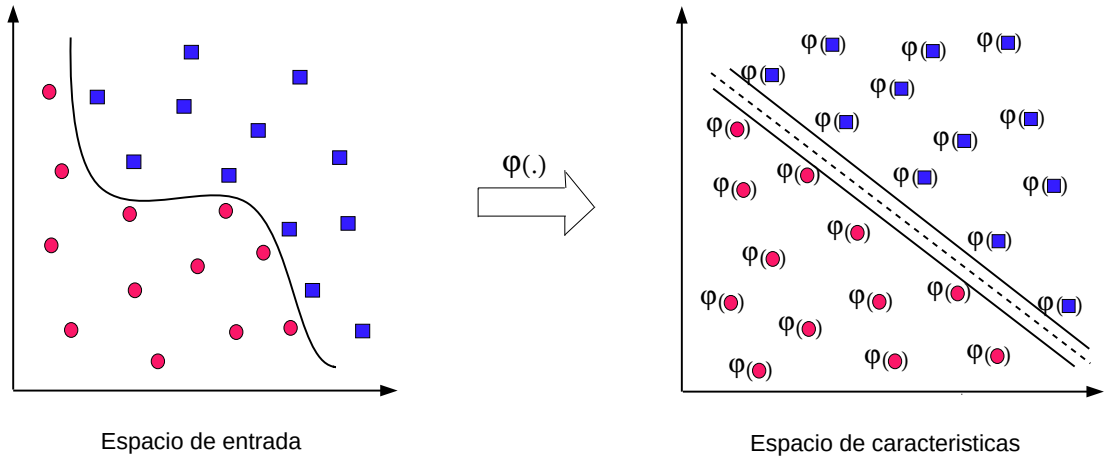


Figura 1.5: Efecto del núcleo.

Tipos de funciones Kernel (Núcleo)

Polinomial-homogénea:

$$K(x_i, x_j) = (x_i \Delta x_j)^n \tag{1.4.9}$$

Perceptron:

$$K(x_i, x_j) = \|x_i - x_j\| \tag{1.4.10}$$

Función de base radial Gaussiana:

$$K(x_i, x_j) = \exp\left(\frac{-(x_i - x_j)^2}{2(sigma)^2}\right) \tag{1.4.11}$$

Sigmoid:

$$K(x_i, x_j) = \tanh(x_i x_j - \theta) \tag{1.4.12}$$

Capítulo 2

Programación lineal como problemas de minimización irrestricta

La programación lineal constituye un importante campo en la optimización, pues muchos problemas prácticos pueden plantearse como problemas de programación lineal. Esta es una técnica matemática que pretende optimizar (maximizar o minimizar) una función objetivo, sujeta a una serie de restricciones donde tanto la función objetivo como las restricciones son lineales o afines. Además es una herramienta significativa en la toma de decisiones.

Consideremos en esta sección un programa lineal muy general (PL) que contiene variables sin restricción no negativas así como restricciones de igualdad y

desigualdades. Se va a mostrar como obtener una solución exacta de este (PL) por una minimización de una función cuadrática, diferenciable, a trozos sin restricción que contiene un solo parámetro finito. Comenzamos con el siguiente programa lineal primal:

$$\begin{array}{lrcl} \underset{(x,y)\in\mathbb{R}^{n+l}}{\text{mín}} & c^T x + d^T y & & \\ s.a. & Ax + By & \geq & b \\ & Ex + Gy & = & h \\ & x & \geq & 0 \end{array} \tag{2.0.1}$$

donde $c \in \mathbb{R}^n$, $d \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{mxl}$, $E \in \mathbb{R}^{k\times n}$, $G \in \mathbb{R}^{kxl}$, $b \in \mathbb{R}^m$ y $h \in \mathbb{R}^k$ y su dual es

$$\begin{array}{lrcl} \underset{(u,v)\in\mathbb{R}^{m+k}}{\text{máx}} & b^T u + h^T v & & \\ s.a. & A^T u + E^T v & \leq & c \\ & B^T u + G^T v & = & d \\ & u & \geq & 0 \end{array} \tag{2.0.2}$$

El problema de penalidad exterior para el programa lineal dual es:

$$\begin{array}{ll} \text{mín}_{(u,v)\in\mathbb{R}^{m+k}} & \epsilon(-b^T u - h^T v) + \frac{1}{2}(||(A^T u + E^T v - c)_+||^2 + \\ & +||B^T u + G^T v - d||^2 + ||(-u)_+||^2) \end{array} \tag{2.0.3}$$

Esto se debe a el problema general de optimización con restricción (1.3.3) y su función de penalidad (1.3.4) quedando de la siguiente manera:

$$\begin{aligned} f(x) + \frac{1}{2\mu}\sum_{i\in\mathcal{E}} C_i^2(x) + \frac{1}{2\mu}\sum_{i\in\mathcal{I}}([C_i(x)]^-)^2 &= \frac{1}{\mu}\left(\mu f(x) + \frac{1}{2}\left[\sum_{i\in\mathcal{E}} C_i^2(x) + \sum_{i\in\mathcal{I}}([C_i(x)]^-)^2\right]\right) \\ &= \tfrac{1}{\mu}\left(\mu f(x) + \tfrac{1}{2}\left[\|C_{\mathcal{E}}(x)\|^2 + \|(C_{\mathcal{I}}(x))^-\|^2\right]\right) \\ &= \tfrac{1}{\mu}\left(\mu f(x) + \tfrac{1}{2}\left[\|C_{\mathcal{E}}(x)\|^2 + \|(-C_{\mathcal{I}}(x))_+\|^2\right]\right) \end{aligned}$$

Así obtenemos

$$\text{mín} \qquad \frac{1}{\mu}\left(\mu f(x) + \frac{1}{2}\left[\|C_{\mathcal{E}}(x)\|^2 + \|(-C_{\mathcal{I}}(x))_+\|^2\right]\right)$$

Para el problema de penalidad exterior (2.0.3)

$\epsilon = \mu$,

$C_i(x) = B^T u + G^T v - d = 0$ si $i \in \mathcal{E}$,

$C_i(x) = -A^T u - E^T v + c$ y $C_i(x) = u$ si $i \in \mathcal{I}$,

$f(x) = b^T u + h^T v$,

$\|(C_i)(x)\|^2 = \|B^T u + G^T v - d\|^2$,

$\|(-C_i^+)\|^2 = \|(A^T u + E^T v - c)_+\|^2$, $\|(-C_i^+)\|^2 = \|(-u)_+\|^2$.

Resolviendo el problema de penalidad (2.0.3) para una sucesión del parámetro de penalidad ϵ_i convergiendo a cero producirá una solución para el problema

lineal dual (2.0.2). Sin embargo no vamos ha hacer esto debido a las inexactitudes inherentes a los métodos asintóticos de penalización exterior, y el hecho de que esto solo daría una solución dual aproximada pero no una solución primal. En su lugar vamos a resolver el problema de penalización exterior para un valor finito del parámetro de penalidad ϵ y de esta solución dual se extrae una solución primal exacta utilizando la siguiente proposición.

Proposición 2.0.1 (ver [9]) *Supongamos que el PL primal* (2.0.1) *es factible entonces el problema de penalidad exterior para el dual* (2.0.3) *es factible para todo* $\epsilon > 0$. *Para cualquier* $\epsilon \in (0, \overline{\epsilon}]$ *para algún* $\overline{\epsilon} > 0$ *, cualquier solución (u,v) de* (2.0.3) *genera una solución exacta para el PL primal* (2.0.1) *como sigue*

$$\begin{aligned} x = &\quad \epsilon^{-1}(A^T u + E^T v - c)_+ \\ y = &\quad \epsilon^{-1}(B^T u + G^T v - d) \end{aligned} \tag{2.0.4}$$

además (x,y) minimiza

$$||x||^2 + ||y||^2 + ||Ax + By - b||^2 \tag{2.0.5}$$

sobre el conjunto solución del PL primal (2.0.1).

Demostración: El problema de minimización de penalidad exterior dual (2.0.3) puede escribirse en forma equivalente:

$$\begin{array}{lll} \min\limits_{(u,v,z_1,z_2)\in\mathbb{R}^{m+k+n+m}} & \epsilon(-b^T u - h^T v) + \frac{1}{2}(\|z_1\|^2 + \|B^T u + G^T v - d\|^2 + \|z_2\|^2) & \\ s.a. & -A^T u - E^T v + c + z_1 & \geq 0 \\ & u + z_2 & \geq 0. \end{array} \tag{2.0.6}$$

La justificación de esto es que en un mínimo de (2.0.6) las variables z_1 y z_2 son no negativas, de lo contrario si cualquiera de las componentes de estas variables es negativas la función objetivo puede ser estrictamente reducida mediante el ajuste de esa componente a cero mientras que se mantiene la restricción factible. Por lo tanto una solución de (2.0.6) es $z_1 = (A^T u + E^T v - c)_+$ y $z_2 = (-u)_+$.

En la ecuación (1.2.9) del teorema (1.2.6) se tiene que la función objetivo viene dada por:

$$\frac{1}{2}x^T \tilde{C} x - \tilde{b}^T x$$

Además la función objetivo de (2.0.6) que es un programa cuadrático convexo es la siguiente:

$$\begin{aligned} &\epsilon(-b^T u - h^T v) + \tfrac{1}{2}(\|z_1\|^2 + \|B^T u + G^T v - d\|^2 + \|z_2\|^2) = \\ &\quad = \epsilon(-b^T u - h^T v) + \tfrac{1}{2}(z_1^T I_n z_1 + (B^T u + G^T v - d)^t (B^T u + G^T v - d) + z_2^T I_m z_2) \\ &\quad = -\epsilon b^T u - \epsilon h^T v + \tfrac{1}{2}(z_1^T I_n z_1 + u^T B B^T u + v^T G G^T v + u^T B G^T v + v^T G B^T u - u^T B d \\ &\qquad - v^T G d - d^T B^T u - d^T G^T v + d^T d + z_2^T I_m z_2) \end{aligned}$$

En formato matricial, si tomando

$$x = \begin{bmatrix} u \\ v \\ z_1 \\ z_2 \end{bmatrix},$$

obtenemos

$$\frac{1}{2}x^T\tilde{C}x = \frac{1}{2}\begin{bmatrix} u^T & v^T & z_1^T & z_2^T \end{bmatrix}\begin{bmatrix} BB^T_{m\times m} & BG^T_{m\times k} & 0_{m\times n} & 0_{m\times m} \\ GB^T_{k\times m} & GG^T_{k\times k} & 0_{k\times n} & 0_{k\times m} \\ 0_{n\times m} & 0_{n\times k} & I_n & 0_{n\times m} \\ 0_{m\times m} & 0_{m\times k} & 0_{m\times n} & I_m \end{bmatrix}\begin{bmatrix} u \\ v \\ z_1 \\ z_2 \end{bmatrix},$$

y también

$$-\tilde{b}^T x = -\begin{bmatrix} (d^TB^T + \epsilon b^T)_{1\times m} & (d^TG^T + \epsilon h^T)_{1\times k} & 0_{1\times n} & 0_{1\times m} \end{bmatrix}\begin{bmatrix} u \\ v \\ z_1 \\ z_2 \end{bmatrix}$$

así

$$\epsilon(-b^Tu - h^Tv) + \tfrac{1}{2}(\|z_1\|^2 + \|B^Tu + G^Tv - d\|^2 + \|z_2\|^2) = \tfrac{1}{2}x^T\tilde{C}x - \tilde{b}^Tx$$

Usando las restricciones de (2.0.6) se tiene que:

$$\begin{aligned} -A^Tu - E^Tv + c + z_1 &\geq 0 \\ u + z_2 &\geq 0. \end{aligned}$$

Esto implica

$$\begin{aligned} A^Tu + E^Tv - I_nz_1 + 0_{n\times m}z_2 &\leq c \\ -I_mu + 0_{m\times k} + 0_{m\times n}z_1 - I_mz_2 &\leq 0. \end{aligned}$$

Finalmente, en forma matricial

$$\begin{bmatrix} A^T & E^T & -I_n & \mathbb{0}_{n\times m} \\ -I_m & \mathbb{0}_{m\times k} & \mathbb{0}_{m\times m} & -I_m \end{bmatrix} \begin{bmatrix} u \\ v \\ z_1 \\ z_2 \end{bmatrix} \leq \begin{bmatrix} c \\ \mathbb{0}_{m\times 1} \end{bmatrix}.$$

Ahora en la ecuación (1.2.10) del teorema (1.2.6), la función objetivo viene dada por

$$-\frac{1}{2}x^T\tilde{C}x - \tilde{c}^T\tilde{u}$$

ahora

$$\begin{aligned} -\tfrac{1}{2}x^T\tilde{C}x &= -\tfrac{1}{2}\begin{bmatrix} u^T & v^T & z_1^T & z_2^T \end{bmatrix} \begin{bmatrix} BB^T & BG^T & \mathbb{0}_{m\times m} & \mathbb{0}_{m\times n} \\ GB^T & GG^T & \mathbb{0}_{k\times n} & \mathbb{0}_{k\times m} \\ \mathbb{0}_{n\times m} & \mathbb{0}_{n\times k} & I_n & \mathbb{0}_{n\times m} \\ \mathbb{0}_{m\times m} & \mathbb{0}_{m\times k} & \mathbb{0}_{m\times n} & I_m \end{bmatrix} \begin{bmatrix} u \\ v \\ z_1 \\ z_2 \end{bmatrix} \\ &= -\tfrac{1}{2}\begin{bmatrix} u^TBB^T + v^TBG & u^TGB^T + v^TG^TG & z_1^TI_n & z_2^TI_m \end{bmatrix} \begin{bmatrix} u \\ v \\ z_1 \\ z_2 \end{bmatrix} \\ &= -\tfrac{1}{2}[u^TBB^Tu + v^TBGu + u^TGB^Tv + v^TG^TGv + z_1^TI_nz_1 + z_2^TI_mz_2] \\ &= -\tfrac{1}{2}[\|z_1\|^2 + \|B^Tu\|^2 + \|G^Tv\|^2 + 2v^TGB^Tu + \|z_2\|^2] \end{aligned}$$

tomando

$$\tilde{u} = \begin{bmatrix} r \\ s \end{bmatrix}$$

obtenemos también

$$\begin{aligned} -\tilde{c}^T\tilde{u} &= -\begin{bmatrix} \tilde{c}^T & 0_{m\times 1} \end{bmatrix}\begin{bmatrix} r \\ s \end{bmatrix} \\ &= -\tilde{c}^T r. \end{aligned}$$

Para las matrices de restricciones $\tilde{C}x + \tilde{A}^T\tilde{u} = \tilde{b}$ tenemos que:

$$\begin{aligned} \tilde{C}x &= \begin{bmatrix} B^TB & BG^T & 0_{m\times m} & 0_{m\times n} \\ GB^T & G^TG & 0_{k\times n} & 0_{k\times m} \\ 0_{n\times m} & 0_{n\times k} & I_n & 0_{n\times m} \\ 0_{m\times m} & 0_{m\times k} & 0_{m\times n} & I_m \end{bmatrix}\begin{bmatrix} u \\ v \\ z_1 \\ z_2 \end{bmatrix} \\ &= \begin{bmatrix} B^TBu + BG^Tv \\ GB^Tu + G^TGv \\ z_1 \\ z_2 \end{bmatrix}. \end{aligned}$$

También

$$\tilde{A}^T\tilde{u} = \begin{bmatrix} A & -I_n \\ E & 0_{k\times m} \\ -I_n & 0_{n\times m} \\ 0_{m\times n} & -I_m \end{bmatrix}\begin{bmatrix} r \\ s \end{bmatrix} = \begin{bmatrix} Ar - s \\ Er \\ -r \\ -s \end{bmatrix}.$$

Además

$$\tilde{b}^T = \begin{bmatrix} d^TB^T + \epsilon b^T & d^TG^T + \epsilon h^T & 0_{1\times n} & 0_{1\times m} \end{bmatrix},$$

y como

$$\tilde{C}x + \tilde{A}^T\tilde{u} = \tilde{b}^T$$

tenemos

$$\begin{aligned} B^TBu + BG^Tv + Ar - I_n s &= d^TB^T + \epsilon b^T \\ GB^Tu + G^TGv + Er &= d^TG^Tv + \epsilon h^T \\ z_1 - r &= 0_{n\times 1} \\ z_2 - s &= 0_{m\times 1}. \end{aligned}$$

Así

$$\begin{aligned} -\epsilon b^T + B(B^Tu + G^Tv - d) + Ar - s &= 0 \\ -\epsilon h^T + G(B^Tu + G^Tv - d) + Er &= 0 \\ z_1 &= r \\ z_2 &= s \end{aligned}$$

Luego el programa cuadrático convexo queda como sigue:

$$\begin{aligned} \max_{(u,v,z_1,z_2,r,s)\in\mathbb{R}^{m+k+n+m+n+m}} \quad & -\tfrac{1}{2}(||z_1||^2 + ||B^Tu||^2 + ||G^Tv||^2 + 2v^TGb`tu - ||d||^2 + ||z_2||^2) - c^Tr \\ s.a \quad & -\epsilon b^T + B(B^Tu + G^Tv - d) + Ar - s = 0 \\ & -\epsilon h^T + G(B^Tu + G^Tv - d) + Er = 0 \\ & z_1 = r \geq 0 \\ & z_2 = s \geq 0. \end{aligned} \tag{2.0.7}$$

Que se puede escribir en forma equivalente

$$\begin{aligned} -\min_{(u,v,z_1,z_2,r,s)\in\mathbb{R}^{m+k+n+m+n+m}} \quad & \tfrac{1}{2}(||z_1||^2 + ||B^Tu||^2 + ||G^Tv||^2 + 2v^TGb`tu - ||d||^2 + ||z_2||^2) - c^Tr \\ s.a \quad & -b + B(\tfrac{B^Tu+G^Tv-d}{\epsilon}) + A\tfrac{r}{\epsilon} = \tfrac{s}{\epsilon} \geq 0 \\ & -h + G(\tfrac{B^Tu+G^Tv-d}{\epsilon}) + E\tfrac{r}{\epsilon} = 0 \\ & r \geq 0. \end{aligned} \tag{2.0.8}$$

Note que una solución del problema de penalidad exterior (2.0.6) y el Dual de Wolfe correspondiente (2.0.7) esta dada por:

$$\begin{aligned} r &= z_1 = (A^T u + E^T v - c)_+ \\ s &= z_2 = (u)_+. \end{aligned} \tag{2.0.9}$$

Definamos ahora:

$$\begin{aligned} x &:= \frac{r}{\epsilon} = \frac{1}{2}(A^T u + E^T v - c)_+ \\ y &:= \frac{1}{\epsilon}(B^T u + G^T v - d), \end{aligned} \tag{2.0.10}$$

donde la igualdad en (2.0.10) sigue de (2.0.9). Sustituyendo (2.0.10) en (2.0.8)

Obtenemos lo siguiente:

$$\|r\|^2 = \|\epsilon x\|^2 = \epsilon^2 \|x\|^2$$

también

$$\begin{aligned} \epsilon^2\|y\|^2 &= \|B^t u + G^T v - d\|^2 \\ &= (B^T u + G^T v - d)^T (B^T u + G^T v - d) \\ &= (u^T B + v^T G - d^T)(B^T u + G^T v - d) \\ &= u^T B B^T u + u^T B G^T v - u^T B d + v^T G B^T u + v^T G G^T v - \\ &\quad - v^T G d - d^T B u - d^T G^T v + d^T d \end{aligned}$$

además

$$\begin{aligned} ||B^T u||^2 + ||G^T v||^2 + 2v^T G B^T u &= u^T B B^T u + v^T G G^T v + 2v^T G B^T u \\ &= \epsilon^2 ||y||^2 + u^T B d + v^T G d + d^T B u + d^T G^T v - d^T d \\ &= \epsilon^2 ||y||^2 + 2u^T B d + 2v^T G d - 2d^T d + d^T d \\ &= \epsilon^2 ||y||^2 + 2(u^T B + v^T G - d^T) d + ||d||^2 \\ &= \epsilon^2 ||y||^2 + 2d^T (B^T u + G^T v - d) + ||d||^2 \\ &= \epsilon^2 ||y||^2 + 2\epsilon d^T y + ||d||^2 \end{aligned}$$

y

$$\begin{aligned} ||s||^2 &= s^T s \\ &= \epsilon^2 (Ax + By - b)^T (Ax + By - b) \\ &= \epsilon^2 ||Ax + By - b||^2 \end{aligned}$$

Luego la función objetivo queda

$$\frac{1}{2}\left(\epsilon^2 ||x||^2| + \epsilon^2 ||y||^2 + 2\epsilon d^T y + ||d||^2 - ||d||^2 + \epsilon^2 ||Ax + B - b||^2\right) + c^T \epsilon x,$$

esto implica

$$\epsilon\left(c^T x + d^T y + \frac{\epsilon}{2}\left(||x||^2 + ||y||^2 + ||Ax + By - b||^2\right)\right)$$

y como $\epsilon \geq 0$ se tiene que

$$c^T x + d^T y + \frac{\epsilon}{2}\left(||x||^2 + ||y||^2 + ||Ax + By - b||^2\right).$$

Finalmente el problema queda de la siguiente manera

$$
\begin{array}{lll}
-\min\limits_{(x,y)\in\mathbb{R}^{n+l}} & c^T x + d^T y + \frac{\epsilon}{2}\left(||x||^2 + ||y||^2 + ||Ax + By - b||^2\right) & \\
s.a & Ax + By \geq b & \\
 & Ex + Gy = h & \\
 & x \geq 0 &
\end{array}
\tag{2.0.11}
$$

Tenga en cuenta que $0 \leq r = \epsilon x$ y que $0 \leq s = \epsilon(Ax + By - b)$ que sigue de las limitaciones de (2.0.8) y de las definición (2.0.10) de x y y

El programa cuadrático (2.0.11) es factible, porque el programa lineal (2.0.1) es factible pues existen valores x e y que satisfacen todas y cada una de las restricciones. Además tiene solución para cualquier $\epsilon \geq 0$ porque su función objetivo esta acotada inferiormente ya que es una función cuadrática fuertemente convexa en (x,y). Donde la función objetivo dual del problema de minimización de penalidad exterior (2.0.3) o equivalente (2.0.6) es acotada inferiormente por el negativo de la función objetivo (2.0.11) por teorema (1.2.1), por tanto (2.0.3) tiene solución para cualquier $\epsilon > 0$. Por la teoría de perturbación de la programación lineal (ver [10]), se deduce que para $\epsilon \in (0, \overline{\epsilon}]$ y para algún $\overline{\epsilon} > 0$, (x,y) como se define en (2.0.10) o equivalente (2.0.4), resuelve el programa lineal (2.0.1) además minimiza la expresión (2.0.5) sobre el conjunto solución del programa original (2.0.1).

■

Una demostración más directa de la proposición 2.0.1, pero igual de laboriosa,

se puede dar al mostrar que las condiciones de optimalidad necesarias y suficientes de KKT para (2.0.11) se derivan de las condiciones de optimalidad necesarias y suficientes para que el ajuste de el gradiente del problema de penalidad exterior (2.0.3) sea igual a cero. No damos prueba aquí porque no justificamos como surgieron los resultados

Capítulo 3

SVMs con norma-1 como problemas de minimización irrestricta

Consideremos primero el problema lineal de clasificación binaria SVM donde en la ecuación (1.4.1) tomamos $f(\omega) = \|\omega\|_1$ una función cuadrática, convexa, diferenciable la cual maximiza el margen

$$\begin{aligned} \min_{(\omega,\gamma,y)\in\mathbb{R}^{n+1+m}} \quad & \nu e^T y + \|\omega\|_1 \\ s.a \quad & D(A\omega - \gamma e) - y \geq e \\ & y \geq 0 \end{aligned} \tag{3.0.1}$$

Donde con cierto abuso de notación de representación multiple, dejamos que la matriz A de orden $m \times n$ que en esta sección representa m puntos en $\mathbb{R}^n$, estos puntos se separan en la mayor medida posible gracias a un plano de separación

$$x^T\omega = \gamma \tag{3.0.2}$$

Donde D representa una matriz diagonal de orden $m \times m$ de las clases de puntos $+1$ ó -1, e es un vector columna de puros unos (1) de dimension adecuada, aquí γ es quien determina la posición respecto a el origen, ω es un vector normal a los planos separadores, y representa la ubicación de las rectas en el espacio de entrada e indica si el problema es separable o no, esto es si y es cero estaríamos en el caso separable, si $y > 0$ estaríamos en el caso no separable. El termino $e^T y$ en la función objetivo minimiza el error de clasificación ponderado con el parámetro positivo ν mientras que el termino $\|\omega\|_1$ maximiza el margen con norma-∞ entre los planos separadores $x^T\omega = \gamma + 1$ y $x^T\omega = \gamma - 1$ que obliga a que las dos clases de puntos de la matriz A sean separados. Se usa $\|\omega\|_1$ en la función objetivo de (3.0.1) en lugar de norma-2 estándar con término cuadrático $\|\omega\|^2$ puesto que esta reduce de manera mas rápida los vectores soportes en el espacio de entrada, mientras que las SVM estándar con norma-2 en general no elimina ningún vector soporte.(Observe figura (3.1))

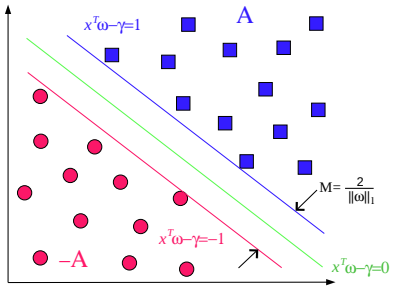


Figura 3.1: SVM con norma-1

Convertimos (3.0.1) en un programa lineal explícito como en [4] mediante el establecimiento de

$$\omega = p - q, \quad p \geq 0, \quad q \geq 0 \tag{3.0.3}$$

Se traduce en el programa lineal

$$\begin{array}{lll} \underset{(p,q,\gamma,y)\in\mathbb{R}^{n+n+1+m}}{\text{mín}} & \nu e^T y + e^T(p-q) & \\ s.a & D(A(p-q)-\gamma e)-y & \geq e \\ & y & \geq 0 \end{array} \tag{3.0.4}$$

Notemos de inmediato que este programa lineal tiene solución porque es factible y su función objetivo esta acotada inferiormente por cero. Por tanto (2.0.1) se puede utilizar para producir el siguiente problema de reformulación sin restricción

Proposición 3.0.2 (ver [9]) *El problema dual del programa lineal* (3.0.4) *esta dado por*

$$\begin{array}{lrl} \underset{u\in\mathbb{R}^n}{\text{máx}} & e^T u & \\ s.a & DA^T u & \leq e \\ & -DA^T u & \leq e \\ & -Deu & = 0 \\ & u & \leq \nu e \\ & u & \geq 0, \end{array}$$

y el problema de penalidad exterior esta dado por:

$$\begin{aligned} \min_{u\in\mathbb{R}^n} \quad & -\epsilon e^T u + \tfrac{1}{2}(\|(A^T DU - e)_+\|^2 + \|(-A^T Du - e)_+\|^2 + \\ & +(-e^T Du)^2 + \|(u - \nu e)_+\| + \|(-u)_+\|^2) \end{aligned} \tag{3.0.5}$$

Tiene solución para todo $\epsilon \geq 0$ *para cualquier* $\overline{\epsilon} \in (0, \overline{\epsilon}]$ *para algún* $\overline{\epsilon} \geq 0$, *cualquier solución u de* (3.0.5) *genera una solución exacta del problema* (3.0.1) *como sigue*

$$\begin{aligned} \omega = p - q &== \epsilon^{-1}\left((A^T Du - e)_+ - (-A^T Du - e)_+\right) \\ \gamma &== \epsilon^{-1} e^T Du \\ y &== \epsilon^{-1}(u - \nu e)_+ \end{aligned} \tag{3.0.6}$$

Además estos (ω, γ, y) *minimizan la expresión*

$$\|\omega\|^2 + \gamma^2 + \|D(A\omega - \gamma e) + y - e\|^2 \tag{3.0.7}$$

sobre el conjunto de solución del problema de clasificación SVM con norma 1 (3.0.1).

Observemos aquí la similitud entre nuestro problema de penalización sin restricción (3.0.5) y el problema correspondiente:

$$\begin{aligned} \min_{u\in\mathbb{R}^n} \quad & f(u) = -\epsilon e^T u + \tfrac{1}{2}(\|(A^T DU - e)_+\|^2 + \|(-A^T Du - e)_+\|^2 + \\ & +(-e^T Du)^2 + \|(u - \nu e)_+\| + \alpha\|(-u)_+\|^2) \end{aligned} \tag{3.0.8}$$

Pero también observamos una mayor diferencia en esta última, un parámetro de penalidad α que multiplica el termino $||(-u)_+||^2$ que en la ecuación (3.0.5) no aparece y este parámetro esta obligado a tender a ∞ para obtener una solución exacta del problema original (3.0.1). Por lo tanto la solución obtenida por (3.0.8) para cualquier α finito es solo aproximada como se ha señalado, sin embargo la solución de (3.0.5) minimiza la expresión (3.0.7) en lugar de ser simplemente una aproximación de la solución por lo menos como con norma-2.

Como es en el caso en:

$$\begin{aligned} \min_{(p,q,\gamma,y)\in\mathbb{R}^{n+n+1+m}} \quad & \nu e^T y + e^T(p+q)\tfrac{\epsilon}{2}\left(||p||^2 + ||q||^2 + \gamma^2 + ||y||^2\right) \\ s.a \quad & D(A(p-q) - \gamma e) + y \geq e \\ & p, q, y \geq 0 \end{aligned}$$

Sin embargo el método de Newton generalizado prescrito en [4] para una sucesión $\alpha \uparrow \infty$ es aplicable con $\alpha = 1$, así podemos aplicar este resultado en (3.0.8)

Para ello vamos a denotar $f(u)$ como la función de penalidad exterior (3.0.5), además definamos el gradiente y el Hessiano generalizado como en la introducción así tenemos que:

$$f(u) = -\epsilon e^T u + \frac{1}{2}\left(||(A^T DU - e)_+||^2 + ||(-A^T Du - e)_+||^2 + (-e^T Du)^2 + ||(u - \nu e)_+|| + ||(-u)_+||^2\right)$$

Entonces el gradiente queda de la siguiente manera

$$\begin{aligned}\nabla f(u) &= -\epsilon e + (A^T D)^T (A^T Du - e)_+ - (A^T D)^T (-A^T Du - e)_+ + \\ &\quad + Dee^T Du + (u - \nu e)_+ - (-u)_+ \\ &= -\epsilon e + DA(A^T Du - e)_+ - DA(-A^T Du - e)_+ + Dee^T Du + \\ &\quad + (u - \nu e)_+ - (-u)_+.\end{aligned} \tag{3.0.9}$$

Así el Hessiano generalizado nos queda:

$$\begin{aligned}\partial^2 f(u) &= DAdiag(A^T Du - e)_* A^T D + DAdiag(-A^T Du - e)_* A^T D + Dee^T D + \\ &\quad + diag(u - \nu e)_* + diag(-u)_* \\ &= DAdiag[(A^T Du - e)_* + (-A^T Du - e)_*]AD + Dee^T D + diag[(u - \nu e)_* + \\ &\quad + (-u)_*] \\ &= DAdiag[(|A^T Du| - e)_*]A^T D + Dee^T D + diag[(u - \nu e)_* + (-u)_*]\end{aligned} \tag{3.0.10}$$

Donde la ultima igualdad proviene de la igualdad:

$$(a - 1)_* + (-a - 1)_* = (|a| - 1)_* \tag{3.0.11}$$

En el caso que el conjunto sea no linealmente separable al manejar un núcleo simétrico no lineal $K(A, B)$ que mapea $\mathbb{R}^{m+n} \times \mathbb{R}^{n+l}$ en $\mathbb{R}^{m+l}$ y que genera en lugar de un plano separador, una superficie separadora no lineal

$$K(x^T, A^T)D\upsilon = \gamma. \tag{3.0.12}$$

Todo lo que necesita hacer es la sustituir:

$$A \rightarrow K(A, A^T)D, \tag{3.0.13}$$

Además para un núcleo lineal $K(A, A^T) = AA^T$, tenemos que $\omega = A^T D\upsilon$, donde $\upsilon \in \mathbb{R}^m$ es una variable dual (ver [8]) y de la programación lineal primal de la SVM (3.0.4) se convierte en $\omega = p - q = A^T D\upsilon$ y minimizando la norma 1 de υ en lugar de la norma-1 de ω en la función objetivo:

$$\begin{array}{ll} \underset{(\upsilon,\gamma,y)\in\mathbb{R}^{m+1+m}}{\text{mín}} & \nu e^T y + \|\upsilon\|_1 \\ s.a & D(AA^T D\upsilon - \gamma e) + y \geq e \\ & y \geq 0 \end{array} \tag{3.0.14}$$

Estableciendo

$$\upsilon = r - s, \quad r \geq 0, s \geq 0 \tag{3.0.15}$$

el problema lineal (3.0.14) se convierte en

$$\begin{array}{ll} \underset{(r,s,\gamma,y)\in\mathbb{R}^{m+m+1+m}}{\text{mín}} & \nu e^T y + e^T(r + s) \\ s.a & D(AA^T D(r - s) - \gamma e) + y \geq e \\ & r, s, y \geq 0 \end{array} \tag{3.0.16}$$

que es el núcleo lineal de la SVM en términos de la variable dual $\upsilon = r - s$. Si reemplazamos el núcleo AA^T en (3.0.16) por el núcleo no lineal $K(A, A^T)$ obtenemos el programa lineal con núcleo no lineal

$$\begin{array}{ll} \underset{(r,s,\gamma,y)\in\mathbb{R}^{m+m+1+m}}{\text{mín}} & \nu e^T y + e^T(r+s) \\ s.a & D(K(A,A^T)D(r-s)-\gamma e)+y \geq e \\ & r,s,y \geq 0 \end{array} \tag{3.0.17}$$

Inmediatamente notamos q el programa lineal (3.0.4) es idéntico a el programa lineal (3.0.17) si hacemos la sustitución en (3.0.13).

Finalmente unas palabras sobre la elección del ϵ en la (2.0.1) y (3.0.2) computacionalmente en ([4]) esto no parece ser crítico y es efectivamente dirigido de la siguiente manera por ([6]), si para dos valores sucesivos de ϵ, $\epsilon^1 \geq \epsilon^2$, la solución correspondiente al ϵ-perturbado del programa cuadrático (2.0.11) son iguales, entonces bajo ciertas condiciones estas soluciones sucesivas e iguales constituyen una solución de los programas lineales (2.0.1) y (3.0.1) que también minimiza las perturbaciones cuadráticas (2.0.5) y (3.0.7) estos resultados pueden ser implementados computacionalmente utilizando un ϵ que disminuye por algún factor de rendimiento da la misma solución para (2.0.1) o (2.0.2). En nuestros resultados computacionales estos en cualquiera de los dos fueron $4x10^{-4}$ o 10^{-6}

Se da ahora el algoritmo de Newton generalizado para resolver el problema de minimización irrestricta (3.0.5) de la síguete manera:

Algoritmo 3: Algoritmo de Newton generalizado para (3.0.5)

Sea $f(u)$, $\nabla f(u)$ y $\partial^2 f(u)$ definidos por (3.0.5), (3.0.9) y (3.0.10). Establecer los valores de los parámetros ν, ϵ, γ, tolerancia *tol*, y *imax* (comúnmente:

$\epsilon \in [10^{-6}\ ,\ 4x10^{-4}]$ para SVMs lineal y $\epsilon\ \in\ [\ 10^{-9}\ , 1]$ para SVMs no lineales, $tol = 10^{-3}$, $imax = 50$ mientras que ν y γ son establecidos por un procedimiento de ajuste) Comenzaremos con cualquier $u^0 \in \mathbb{R}^m$, para $i = 0, 1, ...$

(I)

$$\begin{aligned} u^{i+1} &= u^i - \lambda_i(\partial^2 f(u^i) + \delta I)^{-1}\nabla f(u^i) \\ &= u^i + \lambda_i d^i, \end{aligned}$$

donde el tamaño de paso Armijo

$$\lambda_i = \text{máx}\left\{1, \frac{1}{2}, \frac{1}{4}, \dots\right\}$$

es tal que

$$f(u^i) - f(u^i + \lambda_i d^i) \geq -\frac{\lambda_i}{4}\nabla f(u^i)^T d^i \tag{3.0.18}$$

y d^i es la dirección de Newton modificada

$$d^i = -(\partial^2 f(u^i) + \delta I)^{-1}\nabla f(u^i). \tag{3.0.19}$$

En otras palabras comenzaremos con $\lambda_i = 1$ y multiplicándose λ_i por $\frac{1}{2}$ hasta (3.0.18) se cumple.

(II) Detenerse si $\|u^i - u^{i+1}\| < tol$ o $i \geq max$ sino establecer $i = i + 1$ e ir a I

(III) Definimos la solución del problema SVM con norma 1 (3.0.1) por lo menos con perturbación cuadrática (3.0.7) por (3.0.6) con $u = u^i$.

Ahora daremos un resultado de la convergencia de este algoritmo

Proposición 3.0.3 (ver [9]) *Sea* $tol = 0$, $imax = \infty$ *y sea* $\epsilon > 0$ *suficientemente pequeño cada punto de acumulación* $\overline{u}$ *de la sucesión* $\{u^i\}$ *generadas por el algoritmo 3 resuelve el problema de penalization exterior* (3.0.5)*. El correspondiente* $(\overline{\omega}, \overline{\gamma}, \overline{y})$ *que obtenemos mediante el ajuste u a* $\overline{u}$ *en* (3.0.6) *es una solución exacta para la SVM primal con norma 1* (3.0.1)

Capítulo 4

Aproximación de la función núcleo como problema de minimización irrestricta

Ahora mostraremos otra aplicación de las máquinas de vectores de soporte que en lugar de usar estrategias de clasificación se usara una estrategia de regresión la cual consiste en aproximar una función desconocida y se va ajustar mediante una función lineal:

Sea $f : \mathbb{R}^n \mapsto \mathbb{R}$ una función desconocida. Queremos aproximar $f(x^i) = b_i$ con $i = 1 \ldots m$.

Para ello consideremos la matriz A de orden $m \times n$ que representa m puntos

en $\mathbb{R}^n$ es decir:

$$A = \begin{bmatrix} x^1 \\ \vdots \\ x^m \end{bmatrix},$$

además

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Supóngase

$$f(x) \approx \omega^T x + \gamma$$

con $\gamma \in \mathbb{R}$.

Pero como queremos aproximar f a b entonces ocurre:

$$\omega^T x + \gamma e \approx b$$

así

$$A\omega + \gamma e - b \approx 0 \tag{4.0.1}$$

como $\omega \in CL\{x^1, \ldots, x^m\}$ así tenemos

$$\omega = \upsilon_1 x^1 + \upsilon_2 x^2 + \ldots + \upsilon_m x^m$$

entonces

$$\omega = A^T \upsilon$$

con $\upsilon \in \mathbb{R}^m$

sustituyendo esto en (4.0.1) obtenernos

$$AA^T\upsilon + \gamma e - b \approx 0$$

Ajustando los puntos de los datos mediante una combinación lineal de la función núcleo simétrica tenemos:

$$K(A, A^T)\upsilon + \gamma e - b \approx 0$$

donde el parámetro desconocido υ y γ son determinados por minimización con norma-1 de la aproximación ponderada de error con $\nu > 0$ y la norma-1 de υ de la siguiente manera:

$$\min_{(\upsilon,\gamma)\in\mathbb{R}^{m+1}} \nu\|K(A, A)^T\nu + \gamma e - b\|_1 + \|\upsilon\|_1. \tag{4.0.2}$$

Estableciendo

$$\begin{aligned} &\upsilon = r - s, && r \geq 0, s \geq 0, \\ &K(A, A^T)\upsilon + \gamma e - b = y - z, && y \geq 0, z \geq 0. \end{aligned} \tag{4.0.3}$$

Obtenemos el siguiente programa lineal

$$\begin{aligned} \min_{(s,r,\gamma,y,z)\in\mathbb{R}^{n+n+1+m+m}} \quad & \nu e^T(y + z) + e^T(r + s) & & \\ s.a \quad & K(A, A^T)(r - s) + \gamma e - y + z &= & b \\ & r, s, y, z &\geq & 0, \end{aligned} \tag{4.0.4}$$

que es similar a la formulación de programación lineal del clasificador SVM con núcleo no lineal (3.0.17) con restricciones de igualdad en lugar de restricciones de desigualdad. Observemos también que este programa lineal tiene solución porque es factible y su función objetivo es acotada inferiormente por cero. Por tanto la proposición 3.0.2 se puede utilizar para la siguiente reformulación del problema.

Proposición 4.0.4 (ver [9]) *De el problema lineal* (4.0.4) *su dual viene dado por:*

$$\begin{array}{lrcl} \underset{u\in\mathbb{R}^m}{\text{mín}} & b^T u & & \\ s.a & K(A,A^T)^T u & \leq & e \\ & -K(A,A^T)^T u & \leq & e \\ & -u & \leq & \nu e \\ & u & \leq & \nu e \\ & e^T u & = & 0 \end{array} \tag{4.0.5}$$

que es lo mismo:

$$\begin{array}{lrcl} \underset{u\in\mathbb{R}^m}{\text{mín}} & b^T u & & \\ s.a & -e \leq K(A,A^T)^T u & \leq & e \\ & -\nu e \leq u & \leq & \nu e \\ & e^T u & = & 0 \end{array} \tag{4.0.6}$$

Además el problema de penalización exterior sin restricción para la aproximación de SVM con norma 1 de (4.0.4) *es:*

$$\begin{aligned} \text{mín } f(u) \;=\; & -\epsilon b^T u + \tfrac{1}{2}(\|K(A,A^T)^T u - e)_+\|^2 + \|(-K(A,A^T)u - e)_+\|^2 + \\ & + \|(-u - \nu e)_+\|^2 + \|(u - \nu\nu e)_+\|^2 + (e^T u)^2, \end{aligned} \tag{4.0.7}$$

Tiene solución para todo $\epsilon \geq 0$ *para cualquier* $\epsilon \in (0, \overline{\epsilon}]$ *para algún* $\overline{\epsilon} \geq 0$, *cualquier solución u de* (4.0.7) *genera una aproximación exacta del problema de*

clasificación SVM con norma 1 de la siguiente manera

$$\begin{aligned} \upsilon = r - s &= \epsilon^{-1}\left((K(A, A^T)u - e)_+ - (-K(A, A^T)u - e)_+\right) \\ \gamma &= \epsilon^{-1} e^T u \\ y &= \epsilon^{-1}(-u - \nu e)_+ \\ z &= \epsilon^{-1}(u - \nu e)_+ \end{aligned} \tag{4.0.8}$$

Además estos (r, s, γ, y, z) *minimizan:*

$$\|r\|^2 + \|s\|^2 + \gamma^2 + \|y\|^2 + \|z\|^2, \tag{4.0.9}$$

sobre el conjunto solución de el problema de clasificación SVM con norma 1 (4.0.4)

Los resultados computacionales utilizados por la formulación de programación lineal (4.0.2) con conocimientos previos en ([13]) pero utilizando el método simplex de la solución efectiva para resolver el problema de aproximación. La formulación de minimización sin restricción (4.0.7) es otro método de solución que también puede manejar este tipo de problemas sin conocimientos previos correspondientes pero con modificaciones sencillas.

Date set	Algorithm	Iters	Time	Train %	Test %	Feat	Eps
Ionosphere	NLPSVM	69	0.1769	92.6254	83.8016	20.6	4e-6
Ionosphere	CPLEX		0.179	92.6255	85.4841	25.1	
Ionosphere	LPNewton	30.7	0.0767	89.6169	87.1825	9.6	1e-1
351x34							
BUPA Liver	NLPSVM	100	0.1069	70.1791	67.916	5.9	4e-4
BUPA Liver	CPLEX		0.2278	70.4994	67.2941	6	
BUPA Liver	LPNewton	63.3	0.0623	69.1814	67.563	5.2	1e-6
345x6							
Prima Indians	NLPSVM	93.2	0.2169	73.5809	72.6692	6.8	4e-4
Prima Indians	CPLEX		1.1707	76.8086	75.2683	5.8	
Prima Indians	LPNewton	40.6	0.0904	76.0563	75.0051	4.6	1e-6
768x8							
Cleveland	NLPSVM	42.2	0.0515	85.6742	84.1609	7.5	4e-4
Cleveland	CPLEX		0.1409	85.9348	84.1609	8.4	
Cleveland	LPNewton	25.3	0.028	85.7478	84.5287	7.1	1e-6
297x13							
Housing	NLPSVM	66.6	0.0891	83.9049	83.8078	9.1	4e-4
Housing	CPLEX		0.363	86.8035	84.3882	10.5	
Housing	LPNewton	57.4	0.0781	85.6626	83.2078	7.7	1e-6
506x13							
Galaxy Dim	NLPSVM	97.5	1.097	94.4392	94.4415	5.9	4e-4
Galaxy Dim	CPLEX		12.5357	95.5153	95.5153	11.5	
Galaxy Dim	LPNewton	39.2	0.4297	94.4948	94.5131	4.8	1e-6
4192x14							

Bibliografía

[1] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Procceding of the Fifteenth International Conference (ICML '98)*, pages 82–90, San Francisco, California, 1998.

[2] G. B. Dantzig. *Linear Programming and extensions*. Princeton University, Princeton, New Jersey, 1963.

[3] F. Facchinei. Minimization of sc^1 functions and the maratos effect. *Operations Research Letters*, 1995.

[4] G. Fung and O. L. Mangasarian. A feature selection newton method for support vector machine classification. *computational optimization and applicatiions*, 28(2):185–202, 2004.

[5] J. B. Hiriart-Urruty, J. J. Strodiot, and V. H. Nguyen. Generalized hessian matrix and second-order optimality for problems with c^{L_1} data. *Applied Mathematics and Optimization*, 11:43–56, 1984.

[6] S. Lucidi. A new result in the theory and computation of the least-norm

solution of a linear program. *Journal of optimization and Aplications*, 5:103–117, 1987.

[7] O. L. Mangasarian. *Nonlinear Programming*. SIAM, Philadelphia, PA, 1994.

[8] O. L. Mangasarian. Generalized support vector machines. In B. Schölkopf A. Smola, P. Bartlett and Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135–146. MIT Press, Cambridge, MA, 2000.

[9] O. L. Mangasarian. Exact 1-norm support vector machines via unscontrained convex differenciable minimization. *J. of Machine Learning Research*, 7:1517–1530, 2006.

[10] O. L. Mangasarian and R.R. Meyer. Nolinear perturbation of linear programs. *Journal on Control and Optimization*, 1979.

[11] O.L. Mangasarian. A finite newton method for classification problems. *Optimization Methods and sofware*, 2001.

[12] O.L. Mangasarian. A newton method for linear programming. *Journal of Optimization Theory and Aplications*, 121:1–18, 2004.

[13] Shavlik J.W. Wild E. W. Mangazarian, O.L. Knowledge-based kernel aproximation. *Journal of Machine Learning Research*, 5:03–05, 2004.

[14] V. N. Vapnik. *The nature of Stadistical Learning Theory*. Springer, New York, 1995.

[15] S. Wright and J. Nocedal. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, 2^{nd} edition, 2006.

[16] J. Zhu, S. Rosset, T. Hastie, and Tibshirani. 1-norm support vector machines. In Sebastian Thrun Lawrence K. Saul and Bernhard Schölkopf, editors, *Advances in neural information processing systems*. MIT press, 2004.