LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs)

RESUMEN
Una Máquina de Soporte Vectorial (SVM) aprende la superficie decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un *kernel* Gaussiano u otro tipo de *kernel* a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento.

PALABRAS CLAVES: Máquina de Soporte Vectorial, *kernel*, aprendizaje, espacio de características.

GUSTAVO A. BETANCOURT
Ingeniero Electricista,
Estudiante Maestría en Ingeniería Eléctrica – FIE
Universidad Tecnológica de Pereira
gustavoa@ohm.utp.edu.co

Grupo de Instrumentación y Control
Facultad de Ingeniería Eléctrica
Universidad Tecnológica

ABSTRACT
A Support Vector Machine (SVM) learns the decision surface from two distinct classes of the input points. As a one class classifier, the support vector data description is able to form a decision boundary around the learned data domain with very little or no knowledge of data points outside the boundary. Data points are mapped by means of a Gaussian kernel or other kind of kernel to a high dimensional feature space, where we search for the maximal separation between classes. This boundary function, when mapped back to data space, can separate into several components, each enclosing a separate cluster.

KEYWORDS: *Support Vector Machines, kernel, learning, feature space.*

1. INTRODUCCIÓN

La teoría de las Máquinas de Soporte Vectorial (SVM por su nombre en inglés *Support Vector Machines*) es una nueva técnica de clasificación y ha tomado mucha atención en años recientes [1]- [2]. La teoría de la SVM esta basada en la idea de minimización de riego estructural (SRM) [3]. En muchas aplicaciones, las SVM han mostrado tener gran desempeño, más que las máquinas de aprendizaje tradicional como las redes neuronales [2] y han sido introducidas como herramientas poderosas para resolver problemas de clasificación.

Una SVM primero mapea los puntos de entrada a un espacio de características de una dimensión mayor (i.e.: si los puntos de entrada están en $\Re^2$ entonces son mapeados por la SVM a $\Re^3$) y encuentra un hyperplano que los separe y maximice el margen *m* entre las clases en este espacio como se aprecia en la Figura 1.

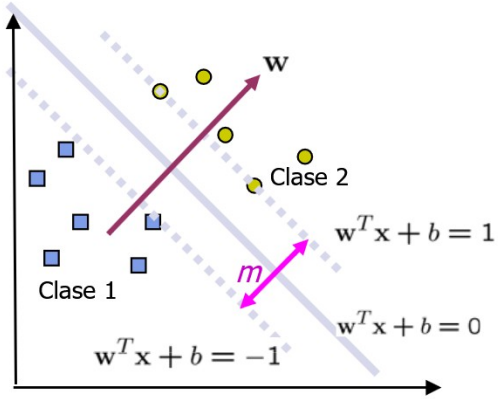

Figura 1. La frontera de desición debe estar tan lejos de los datos de ambas clases como sea.

Maximizar el margen *m* es un problema de programación cuadrática (QP) y puede ser resuelto por su problema dual introduciendo multiplicadores de Lagrange. Sin ningún conocimiento del mapeo, la SVM encuentra el hyperplano óptimo utilizando el producto punto con funciones en el espacio de características que son llamadas *kernels*. La solución del hyperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados vectores de soporte.

Fecha de recepción: 31 Enero de 2005
Fecha de aceptación: 20 Marzo de 2005

Actualmente hay muchas aplicaciones que utilizan las técnicas de las SVM como por ejemplo las de OCR (*Optical Character Recognition*) por la facilidad de las SVMs de trabajar con imágenes como datos de entrada.

2. SVMs

En esta sección se hará una revisión de la teoría básica de las SVM en problemas de clasificación [4]- [5].

2.1. Caso linealmente separable

Supongamos que nos han dado un conjunto *S* de puntos etiquetados para entrenamiento como se aprecia en al Figura 2.

$$(y_1, x_1), \ldots, (y_i, x_i) \qquad (1)$$

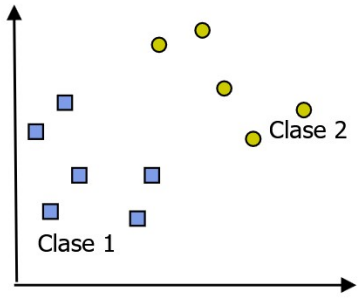

Figura 2. Caso linealmente separable.

Cada punto de entrenamiento $x_i \in \Re^N$ pertenece a alguna de dos clases y se le ha dado una etiqueta $y_i \in \{-1,1\}$ para *i = 1, . . . , l*. En la mayoría de los casos, la búsqueda de un hyperplano adecuado en un espacio de entrada es demasiado restrictivo para ser de uso práctico. Una solución a esta situación es mapear el espacio de entrada en un espacio de características de una dimensión mayor y buscar el hyperplano óptimo allí. Sea $z = \varphi(x)$ la notación del correspondiente vector en el espacio de características con un mapeo φ de $\Re^N$ a un espacio de características *Z*. Deseamos encontrar el hyperplano

$$w \cdot z + b = 0 \qquad (2)$$

Definido por el par (*w, b*), tal que podamos separar el punto x_i de acuerdo a la función

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1 & y_i = 1 \\ -1 & y_i = -1 \end{cases} \qquad (3)$$

Donde $w \in Z$ y $b \in \Re$. Más precisamente, el conjunto *S* se dice que es linealmente separable si existe (*w, b*) tal que las inecuaciones

$$\begin{cases} (w \cdot z_i + b) \geq 1, & y_i = 1 \\ (w \cdot z_i + b) \leq -1, & y_i = -1 \end{cases} \quad i = 1, \ldots, l \qquad (4)$$

Sean válidas para todos los elementos del conjunto *S*. Para el caso linealmente separable de *S*, podemos encontrar un único hyperplano óptimo, para el cual, el margen entre las proyecciones de los puntos de entrenamiento de dos diferentes clases es maximizado.

2.2. Caso no linealmente separable

Si el conjunto *S* no es linealmente separable, violaciones a la clasificación deben ser permitidas en la formulación de la SVM.

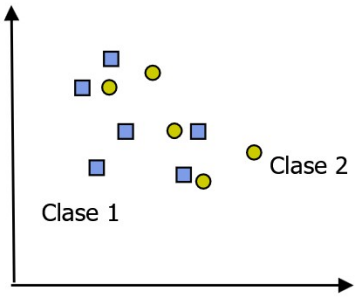

Figura 3. Caso no linealmente separable.

Para tratar con datos que no son linealmente separables, el análisis previo puede ser generalizado introduciendo algunas variables no-negativas $\xi_i \geq 0$ de tal modo que (4) es modificado a

$$y_i(w \cdot z_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l. \qquad (5)$$

Los $\xi_i \neq 0$ en (5) son aquellos para los cuales el punto x_i no satisface (4). Entonces el término $\sum_{i=1}^{l} \xi_i$ puede ser tomado como algún tipo de medida del error en la clasificación.

El problema del hyperplano óptimo es entonces redefinido como la solución al problema

$$\min \left\{ \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} \xi_i \right\}$$
$$s.a \quad y_i \left(w \cdot z_i + b \right) \geq 1 - \xi_{i,} \quad i = 1, \ldots, l \qquad (6)$$
$$\xi_i \geq 0, \quad i = 1, \ldots, l$$

Donde C es una constante. El parámetro C puede ser definido como un parámetro de regularización. Este es el único parámetro libre de ser ajustado en la formulación de la SVM. El ajuste de éste parámetro puede hacer un balance entre la maximización del margen y la violación a la clasificación. Más detalles se pueden encontrar en [5], [6].

Buscando el hyperplano óptimo en (6) es un problema QP, que puede ser resuelto construyendo un Lagrangiano y transformándolo en el dual

$$\begin{aligned} Max \quad & W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j z_i \cdot z_j \\ s.a \quad & \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1,\ldots,l \end{aligned} \tag{7}$$

Donde $\alpha = (\alpha_1, \ldots \alpha_l)$ es una vector de multiplicadores de Lagrange positivos asociados con las constantes en (5).

El teorema de Khun-Tucker juega un papel importante en la teoría de las SVM. De acuerdo a este teorema, la solución $\bar{\alpha}_i$ del problema (7) satisface:

$$\bar{\alpha}_i \left(y_i \left(\bar{w} \cdot z_i + \bar{b} \right) - 1 + \bar{\xi}_i \right) = 0, \quad i = 1,\ldots,l \tag{8}$$

$$\left(C - \bar{\alpha}_i \right) \xi_i = 0, \quad i = 1,\ldots l \tag{9}$$

De esta igualdad se deduce que los únicos valores $\bar{\alpha}_i \neq 0$ (9) son aquellos que para las constantes en (5) son satisfechas con el signo de igualdad. El punto x_i correspondiente con $\bar{\alpha}_i > 0$ es llamado *vector de soporte*. Pero hay dos tipos de vectores de soporte en un caso no separable. En el caso $0 < \bar{\alpha}_i < C$, el correspondiente vector de soporte x_i satisface las igualdades $y_i \left(\bar{w} \cdot z_i + \bar{b} \right) = 1$ y $\xi_i = 0$. En el caso $\bar{\alpha}_i = C$, el correspondiente ξ_i es diferente de cero y el correspondiente vector de soporte x_i no satisface (4). Nos referimos a estos vectores de soporte como errores. El punto x_i correspondiente con $\bar{\alpha}_i = 0$ es clasificado correctamente y esta claramente alejado del margen de desición. Figura 4.

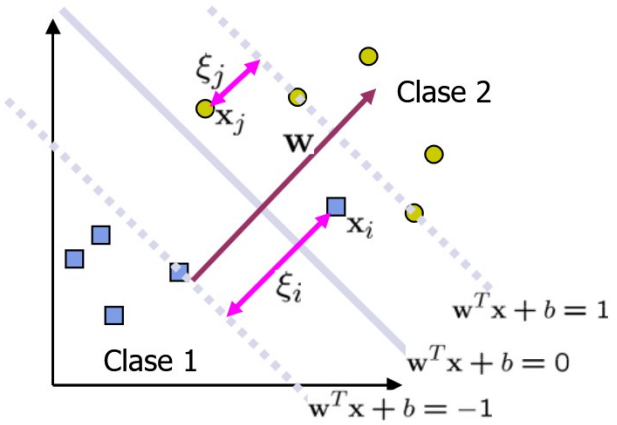

Figura 4. Aparición del parámetro de error ξ_i en el error de clasificación.

Para construir el hyperplano óptimo $\bar{w} \cdot z + \bar{b}$, se utiliza

$$\bar{w} = \sum_{i=1}^{l} \bar{\alpha}_i y_i z_i \tag{10}$$

Y el escalar *b* puede ser determinado de las condiciones de Kuhn-Tucker (9).

La función de desición generalizada de (3) y (10) es tal que

$$f(x) = sign\left(w \cdot z + b \right) = sign\left(\sum_{i=1}^{l} \alpha_i y_i z_i \cdot z + b \right) \tag{11}$$

2.3. Truco del Kernel para el caso no linealmente separable

Como no tenemos ningún conocimiento de φ, el cálculo del problema (7) y (11) es imposible. Hay una buena propiedad de la SVM la cual es que no es necesario tener ningún conocimiento acerca de φ. Nosotros sólo necesitamos una función *K*(·, ·) llamada *kernel* que calcule el producto punto de los puntos de entrada en el espacio de características *Z*, esto es

$$z_i \cdot z_j = \varphi(x_i) \cdot \varphi(x_j) = K(x_i, x_j) \tag{12}$$

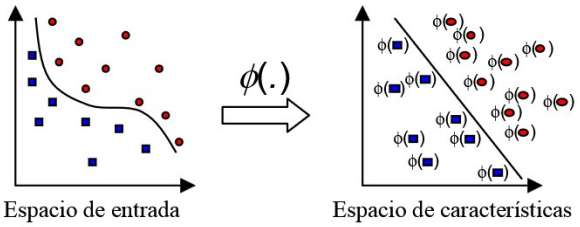

Figura 5. Idea del uso de un *kernel* para transformación del espacio de los datos.

Las Funciones que satisfacen el teorema de Mercer pueden ser usadas como productos punto y por ende pueden ser usadas como kernels. Podemos usar el kernel polinomial de grado *d*

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^d \qquad (13)$$

Para construir un clasificador SVM.

Entonces el hyperplano no lineal de separación puede ser encontrado como la solución de

$$Max \quad W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$
$$s.a \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1, \ldots, l \qquad (14)$$

y la función de decisión es

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x_j) + b \right) \qquad (15)$$

4. EJEMPLO DE APLICACIÓN

Para el siguiente ejemplo se considerará el conjunto de datos conocido como *twospirals* (dos espirales) conformado por 200 puntos en dos clases que son no linealmente separables como se muestra en la Figura 6.

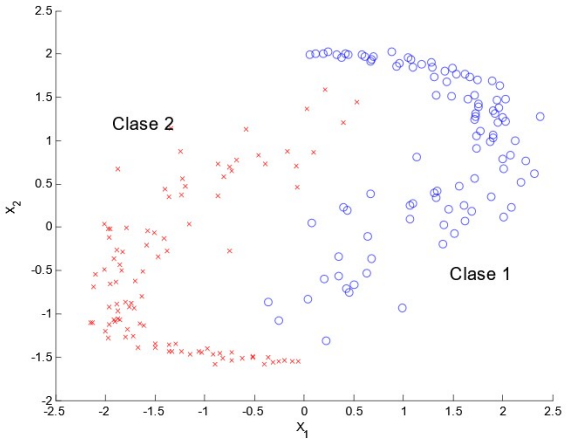

Figura 6. Conjunto de datos twospirals para el entrenamiento del clasificador SVM.

Para el primer intento se utilizará un kernel RBF (*Radial Basis function*) o Gaussiano con $\sigma = 1$ y C=100 entrenando la SVM con el algoritmo SMO (*Sequential Minimal Optimizer*) para SVM binaria.

El entrenamiento de la SVM se hizo con el SPRToolbox [7] de Matlab®.

Para el primer intento se tienen los siguientes datos:

α = [4.3475 0.1072 19.8084 75.8903 86.2008 -94.5622 -0.5096 -0.1663 -1.2177 -84.3138 -5.5846],

b=0.1484

Vectores de Soporte=

0.0625	2.0007
2.5263	1.2379
0.3844	0.3465
-0.3321	-0.3822
0.1254	-1.4875
-0.0573	-1.5489
-2.1932	-0.4391
-1.6895	-0.4672
-1.9075	-0.4692
-0.2136	-0.1409
0.2116	1.2649

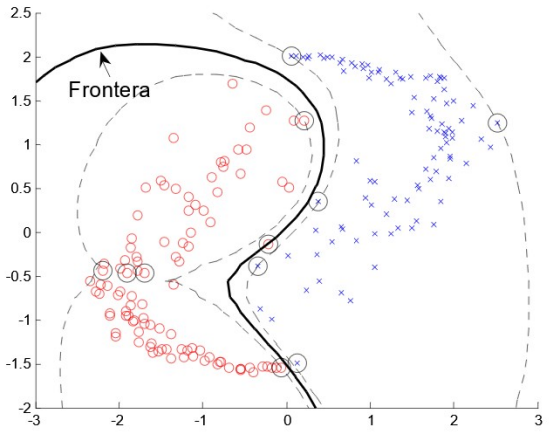

Figura 7. Se aprecian los vectores de soporte marcados con circulo, las márgenes en punteado y la frontera de separación entre clases para $\sigma = 1$ y C=100.

Ahora se refina más la frontera de desición con $\sigma = 1$ y C=1000

α = [5.0305 0.1825 22.7773 85.4252 60.4037 -69.9884 -4.1113 -93.0078 -6.7117]

b= -0.0223

Vectores de soporte=

0.0625	2.0007
2.5263	1.2379
0.3844	0.3465
-0.3321	-0.3822
0.1254	-1.4875
-0.0573	-1.5489
-1.3457	-0.6016
-0.2136	-0.1409
0.2116	1.2649

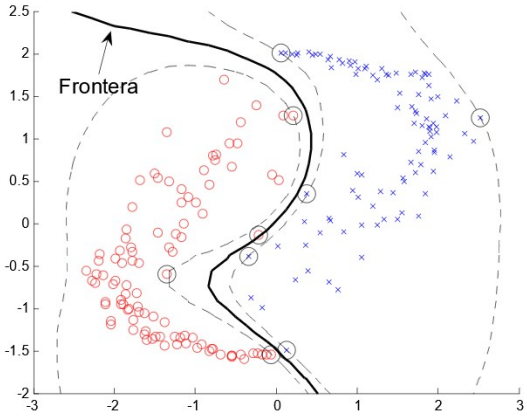

Figura 8. Se aprecian los vectores de soporte marcados con circulo, las márgenes en punteado y la frontera de separación entre clases para $\sigma = 1$ y C=1000.

Ahora se refina más la frontera de desición con $\sigma = 1$ y C=10000

α =[5.4094 0.2671 22.1051 80.4316 49.6585 -58.8928 -4.3958 -87.2509 -7.3322]

b= -0.0552

Vectores de Soporte=

0.0625	2.0007
2.5263	1.2379
0.3844	0.3465
-0.3321	-0.3822
0.1254	-1.4875
-0.0573	-1.5489
-1.3457	-0.6016
-0.2136	-0.1409
0.2116	1.2649

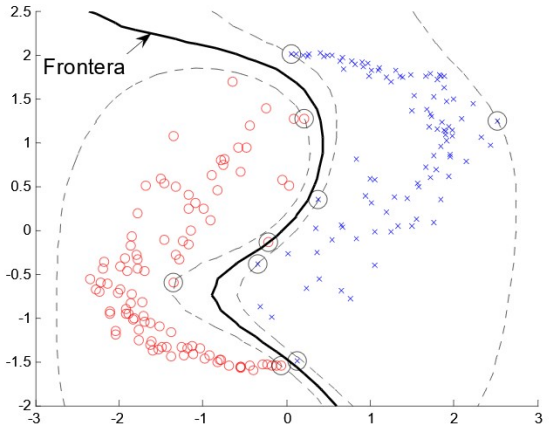

Figura 8. Se aprecian los vectores de soporte marcados con circulo, las márgenes en punteado y la frontera de separación entre clases para $\sigma = 1$ y C=10000.

Si ahora se superpondrán los 3 gráficos se podrá apreciar la evolución de la frontera de decisión cuando $C \to \infty$

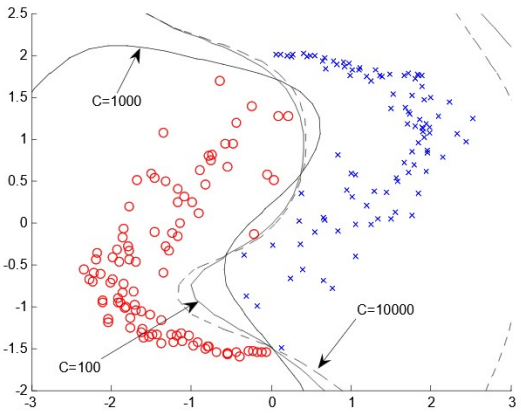

Figura 9. Evolución de la frontera de decisión para diferentes valores de C.

Idealmente la frontera óptima se logra cuando $C = \infty$.

Es de importancia notar dos cosas: El modelamiento de la SVM es tal que no necesita de toda la totalidad de puntos disponibles para hallar una solución al problema de maximización de la separación entre clases, lo que lleva a la segunda anotación, y es que el tiempo de entrenamiento para todos estos casos no supera 1 segundo en el peor de los casos, lo cual presenta una ventaja frente de otros métodos de clasificación que utilizan un porcentaje alto del conjunto de entrada (o entrenamiento) para poder hallar una separación entre clases.

5. CONSIDERACIONES ALREDEDOR DE LAS SVM

Las SVM son básicamente clasificadores para 2 clases.

Se puede cambiar la formulación del algoritmo QP para permitir clasificación multiclase. Más comúnmente, los datos son divididos "inteligentemente" en dos partes de diferentes formas y una SVM es entrenada para cada forma de división. La clasificación multiclase es hecha combinando la salida de todos los clasificadores

Fortalezas

El entrenamiento es relativamente fácil.

No hay óptimo local, como en las redes neuronales.

Se escalan relativamente bien para datos en espacios dimensionales altos.

El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente.

Datos no tradicionales como cadenas de caracteres y árboles pueden ser usados como entrada a la SVM, en vez de vectores de características.

Debilidades

Se necesita una "buena" función *kernel,* es decir, se necesitan metodologías eficientes para sintonizar los parámetros de inicialización de la SVM.

6. CONCLUSIONES

Una lección aprendida en las SVM: un algoritmo lineal en el espacio de características es equivalente a un algoritmo no lineal en el espacio de entrada.

Algoritmos clásicos pueden ser generalizados a su versión no lineal yendo al espacio de características.

El kernel de análisis de componente principal, el kernel de análisis de componente independiente, el kernel de análisis de correlación canónico, el kernel k-means, las SVM en 1D son algunos ejemplos.

Las SVM son una buena alternativa a las redes neuronales.

Dos conceptos claves de SVM: maximizar el margen y el truco del kernel.

Muchas investigaciones se encuentran activas alrededor de áreas relacionadas con SVM.

Muchas implementaciones de SVM se encuentran disponibles en la red para hacer pruebas en conjuntos de datos propios.

7. BIBLIOGRAFIA

[1] C. Burges B. Schölkopf and A. Smola. Advances in kernel methods: Support vector machines. Cambridge, MA: MIT Press, 1999.

[2] C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, vol. 2, no. 2, 1998.

[3] V.Ñ. Vapnik. The nature of statistical learning theory. New York: Springer-Verlag, 1995.

[4] C. Cortes and V.Ñ. Vapnik. Support vector networks. Machine Learning, vol. 20, pp 273-297, 1995.

[5] V.Ñ. Vapnik. Statistical learning theory. New York: Wiley, 1998.

[6] N. Cristianini and J. Shawe-Taylor. Support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridbe MA, ISBN 0-521-78019-5, 2000.

[7] Vojtěch Franc and Václav Hlaváč. Statistical pattern recognition toolbox for matlab userís guide. Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University Technická 2, 166 27 Prague 6, Czech Republic, [http://cmp.felk.cvut.cz], 2004.

[8] Vojislav Kecman. Learning and soft computing: Support vector machines, neural networks, and fuzzy logic models. The MIT Press Computer Science, ISBN 0-262-11255-8, 2001.