# Tarea 16

Angel Manrique Pozos Flores
Tecnológico Nacional de México,
Blvd. Industrial, Mesa de Otay, 22430 Tijuana, B.C., México.

26 de abril de 2016

Investigar en que consiste el clasificador de soporte vectorial (SVM).

## 1. Maquina de soporte vectorial.

Una maquina de soporte vectorial (SVM) tienen su origen en los años 90 por Vapnik y colaboradores (Boser et al., Cortes &Vapnik) esta se basa en la teoria del aprendizaje estadistico, aunque origanlmente habian sido pensadas para la resolucion de problemas de clasificacion binaria.

Actualmente se utilizan para resolver otros tipos de problemas por ejemplo:

- Problemas de regresion.
- Problemas de agrupamiento.
- Multiclasificacion.

Tambien son diversos los campos en los que podido ser implementada tales como:

- Vision artificial.
- Reconocimiento de caracteres.
- Categorizacion de texto e hipertexto.
- Clasificacion de proteinas.
- Procesamiento de lenguaje natural.

- Analisis de series temporales.

Dentro de la tarea de clasificacion, las SVMs pertenecen a la categoria de los clasificadores lineales, puesto que inducen separadores lineales o hiperplanos, ya sea en el espacio original de los ejemplos de entrada. Si estos son separables o quasi-separables, o en un espacio transformado o espacio de caracteristicas.

Mientras la mayoria de los metodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento, el sesgo inductivo asociado a las SVMs radica en la minimizacion del denominado riesgo estructural.

Donde dado un conjunto de muestras de entrenamiento se puede etiquetar dichas clases y entrenar la SVM para que sea capaz de construir un modelo que predice la clase de una nueva muestra, por ello una SVM puede considerarse como un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo mas amplio posible, donde una buena separacion entre las clases permite que la clasificacion sea mejor, ya que reduce la cantidad de falsos positivos.

El exito de las SVM radica en tres ventajas fundamentales.

- Poseen una fundamentacion matematica solida.
- Estan basadas en el concepto de la minimizacion del riesgo estructural, para la minimizacion de la probabilidad de una clasificacion erronea sobre nuevos ejemplos, lo cual es muy importante cuando solo se dispone de muy pocos datos para entrenamiento.
- Disponen de potentes herramientas y algoritmos para encontrar la solucion de manera rapida y eficiente.

Una SVM primero mapea los puntos de entrada a un espacio de caracteristicas de una dimension mayor, por ejemplo si los puntos de entrada estan en $\mathbb{R}^2$ entonces la SVM los mapea a $\mathbb{R}^3$ y encuentra un hiperplano que los separe, y maximice el margen $m$ entre las clases en este espacio como se muestra en la figura 1.

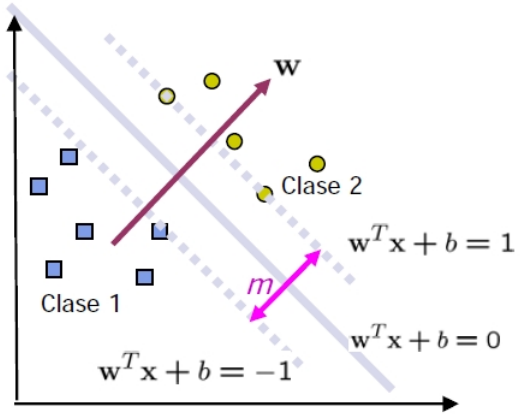


Figura 1: La frontera de decision entre las clases debe estar lo mas lejor posible entre ambas clases con el fin de minimizar el error de clasificacion, donde la linea central representa el hiperplano separador.

Donde maximizar el margen $m$ es un problema de programación cuadrática y puede ser resuelto por su problema dual introduciendo multiplicadores de Lagrange. Donde sin ningun conocimiento del mapeo la SVM es capaz de encontrar el hiperplano optimo utilizando el producto punto con funciones en el espacio de caracteristicas que son llamados *"kernels"*.

La solucion del hiperplano optimo puede ser escrita como la combinacion de unos pocos puntos de entrada que son llamados *vectores de soporte*.

## 2. Caso linealmente separable

Para resolver un problema de clasificacion la SVM debe aprender una superficie de decision adecuada, basandose en el conjunto de datos de entrenamiento, donde la superficie de decision es un hiperplano que separa los patrones de entrenamiento en dos clases.

Si tenemos un conjunto $S$ de puntos etiquetados para entrenamiento como se aprecia en la figura 2.

$$(y_1, x_1), ...(y_i, x_i)) \tag{1}$$

Cada punto de entrenamiento $x_i \in \mathbb{R}^N$ pertenece a alguna de las dos clases donde se le a asignado una etiqueta $y_i \in -1, 1$ para $i = 1 \ldots l$ en la mayoria de los casos, la busqueda de un hiperplano adecuado en un espacio de entrada es dema-

siado restrictivo para ser de uso practico.

Una solucion es mapear el espacio de entrada en un espacio de caracteristicas de una dimension mayor y buscar un hiperplano optimo ahi, donde, sea $z = \varphi(x)$ la notacion del correspondiente vector en el espacio de caracteristicas con un mapeo $\varphi$ de $\mathbb{R}^N$ a un espacio de caracteristicas $Z$ se busca encontrar un hiperplano de la forma.

$$w \cdot z + b = 0 \tag{2}$$

Definido por el par (w, b), tal que, se pueda separar el punto $x_i$ de acuerdo a la funcion.

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

Donde $w \in Z$ y $b \in \mathbb{R}$, mas precisamente, el conjunto $S$ se dice que es linealmente separable si existe $(w,b)$ tal que, las inecuaciones:

$$\begin{array}{ll} (w \cdot z_i + b) \geq 1, & y_i = 1 \\ (w \cdot z_i + b) \leq -1, & y_i = -1 \\ con & i = 1 \ldots l \end{array} \tag{4}$$

Sean validas para todos los elementos del conjunto S, para el caso linealmente separable de S, podemos encontrar un unico hiperplano optimo, para el cual el margen entre las proyecciones de los puntos de entrenamiento de dos clases diferentes es maximizado.

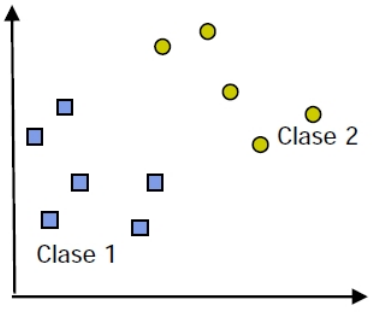


Figura 2: Caso linealmente separable.

# 3. Caso no linealmente separable

Si el conjunto S no es linealmente separable, es permitido hacer una reformulacion de la SVM, para tratar con datos que no son linealmente separables, el analisis previo puede ser generalizado introduciendo algunas variables no-negativas $\xi \geq 0$ donde la ecuacion (4) se modifica quedando:

$$y_i(w \cdot z_i + b) \leq 1 - \xi_i, i = 1 \ldots l \tag{5}$$

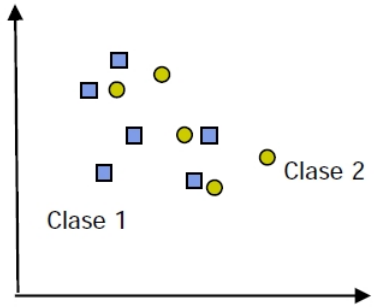


Figura 3: Caso no linealmente separable.

Los $\xi_i \neq 0$ en la nueva ecuacion (5) son aquellos para los cuales el punto $x_i$ no satisface (4) el termino $\sum_{i=1}^{l} \xi_i$ se puede tomar como una medida del error de la clasificacion.

El problema del hiperplano optimo es entonces redefinido como la solucion al problema.

$$\begin{array}{ll} Min\left(\frac{1}{2} w \cdot w + C \sum_{i=1}^{l} \xi_i\right) & \\ y_i(w \cdot z_i + b) \geq 1 - \xi_i, & i = 1 \ldots l \\ \xi_i \geq 0, & i = 1 \ldots l. \end{array} \tag{6}$$

Donde C es una constante, este parametro puede ser definido como un parametro de regularizacion, este es el unico parametro libre de ser ajustado en la formulacion de la SVM, el ajuste de este parametro puede hacer un balance entre la maximizacion del margen y la violacion a la clasificacion. Buscando el hiperplano optimo en (6) es un problema que puede ser resuelto contruyendo un Lagrangiano y transformandolo en el dual.

$$\begin{array}{ll} Max & W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j z_i \cdot z_j \\ \sum_{i=1}^{l} y_i \alpha_i = 0, & 0 \leq \alpha_i \leq C \\ i = 1 \ldots l & \end{array} \tag{7}$$

Donde $\alpha = (\alpha_1 \dots \alpha_l)$ es un vector de multiplicadores de Lagrande positivos asociados a las constantes en la ecuacion (5).

El teorema de Khun - Tucker juega un papel importante en la teoria de las SVM de acuerdo a dicho teorema la solucion de $\alpha_i$ del problema en (7) satisface.

$$\overline{\alpha}(y_i(\overline{w} \cdot z_i + \overline{b}) - 1 + \overline{\xi_i}) = 0, i = 1 \dots l \tag{8}$$

$$(C - \overline{\alpha_i})\xi_i = 0, i = 1 \dots l \tag{9}$$

De esta igualdad se deduce que los unicos valores $\alpha_i \neq 0$ son aquellos para los cuales las constantes en (5) son satisfechas con el signo de igualdad.

El punto $x_i$ correspondiente con $\overline{\alpha_i} > 0$ es llamado *vector de soporte*. Pero para el caso no separable existen dos tipos de vectores de soporte.

- En el caso $0 < \overline{\alpha_i} < C$, el correspondiente vector de soporte $x_i$ satisface las igualdades $y_i(\overline{w} \cdot z_i + \overline{b}) = 1$ y $\xi_i = 0$.
- En el caso $\overline{\alpha_i} = C$, el correspondiente $\xi_i$ es diferente de 0 y el correspondiente vector de soporte $x_i$ no satisface (4), por ello nos referimos a estos vectores de soporte como errores, el punto $x_i$ correspondiente con $\overline{\alpha_i} = 0$ es clasificado correctamente y esta claramente alejado del margen de decisión como se observa en la figura 14.

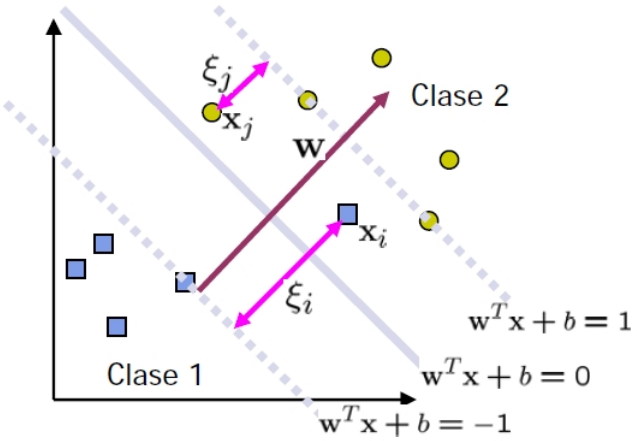


Figura 4: Aparicion del parametro de error $\xi_i$ en el error de clasificacion.

Donde para construir el hiperplano optimo $\overline{w} \cdot z + \overline{b}$ utilizamos:

$$\overline{w} = \sum_{i=1}^{l} \overline{\alpha_i} y_i z_i \tag{10}$$

Y el escalar $b$ puede ser determinado de las condiciones de Kunh-Tucker(9), donde la funcion de decision generalizada de las ecuaciones (3, 10) es tal que:

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

# 4. Función del Kernel

La forma mas simple de llevar a cabo la separacion es mediante una linea recta, un plano recto o un hiperplano N-Dimensional, sin embargo, no se suelen presentar casos idealizados como estos, por ello la representacion utilizando las funciones kernel nos da una solucion, ya que proyecta la informacion a un espacio de caracteristicas de mayor dimension el cual aumenta la capacidad computacional de las maquinas de aprendizaje lineal, otra forma de visualizar este proceso es mediante la figura 5.

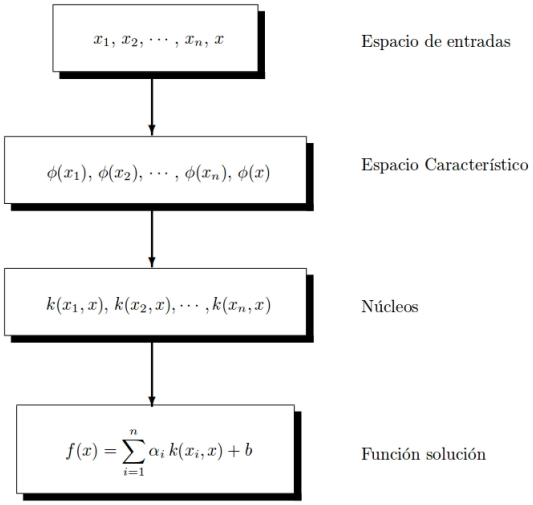


Figura 5: Las maquinas de vectores de soporte transforman inicialmente, el espacio de entradas en un espacio característico de dimension superior, dado esto, es capaz de construir la funcion de clasificacion lineal optima dentro de este nuevo espacio.

Al usar $\Phi : X \longrightarrow H$ se trabaja en un nuevo espacio $H$ por lo cual el vector solucion $w$ se encuentra en este espacio. Por tanto, puede ocurrir que sobre el

conjunto $X$ inicial no se tenga definida ningun tipo de estructura y la funcion $\Phi$ sirve para dar una estructura a los datos y poder aplicar una adecuada clasificación.

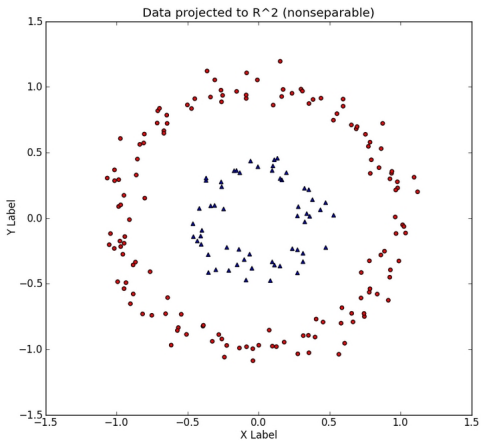


Figura 6: Aqui mostramos un espacio de datos en $\mathbb{R}^2$ que no es linealmente separable.

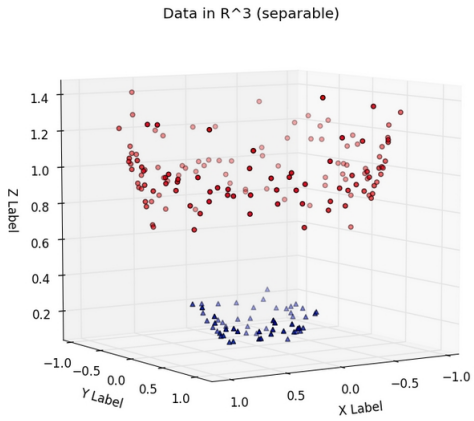


Figura 7: La misma base de datos antes mostrada pero cambiando el espacio a uno en $\mathbb{R}^3$.

Donde observamos que al cambiar el espacio a un espacio mayor de $\mathbb{R}^N$ somos capaces de poder observar los datos de tal manera que podemos encontrar la forma de hacer la clasificacion, donde los datos sean linealmente separables.

## 4.1. Tipos de funciones

Existen para ello kernels especificos para cambiar dicha dimension algunos de los mas utilizados son:

- Polinomial - Homogenea:

$$K(x_i, x_j) = (x_i \cdot x_j)^2 \tag{12}$$

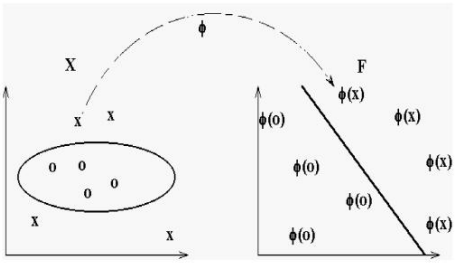


Figura 8: Polinomial-Homogenea

- Perceptron:

$$K(x_i, x_j) = \| x_i - x_j \| \tag{13}$$

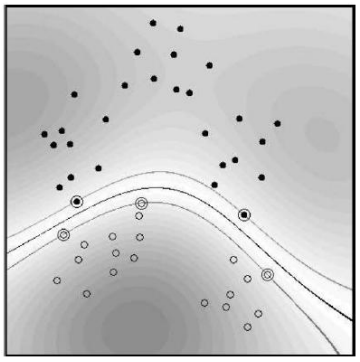

Figura 9: Perceptron.

- Funcion de base radial gaussiana, separada por un hiperplano en el espacio transformado:

$$K(x_i, x_j) = \exp\left( - \frac{(x_i - x_j)^2}{2\sigma^2} \right) \tag{14}$$

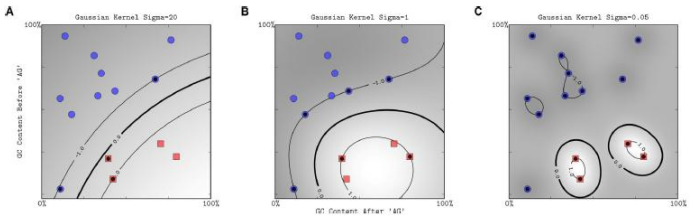


Figura 10: Radial Gaussiana.

- Funcion Sigmoid:

$$\tanh((x_i, x_j) + r) \tag{15}$$

Desafortunadamente escoger el kernel correcto no es una tarea trivial, ya que dependera especificamente de la tarea que se este realizando, no importando que kernel se utilice es necesario pulir los parametros para poder obtener un buen desempeño en el clasificador, una tecnica de parametrizaciones muy popular es la que incluye la validacion de K-Fold Cross.

# 5. Bibliografía


1 Jordan, Michael I., and Romain Thibaux. "The Kernel Trick."Lecture Notes. 2004. Web. 5 Jan. 2013. *http://tinyurl.com/homswfj*

2 Hofmann, Martin. "Support Vector Machines – Kernels and the Kernel Trick". Notes. 26 June 2006. Web. 7 Jan. 2013. *http://tinyurl.com/h3hc7gl*

3 Balcan, Maria Florina. "8803 Machine Learning Theory: Kernels". Lecture Notes. 9 March. 2010. Web. 6 Jan. 2013. *http://tinyurl.com/zjnobl9*

4 J. Enrique, Tutorial sobre maquinas de vector de soporte (SVM), Departamento de inteligencia artificial, ETS de Ingenieria Informatica, UNED, 11 Julio 2014. *http://tinyurl.com/jl7zrna*