MÁQUINA DE VECTORES DE SOPORTE

La teoría de las Máquinas de Vectores de Soporte (SVM por su nombre en inglés Support Vector Machine) fue desarrollada por Vapnik basado en la idea de minimización del riesgo estructural (SRM).

Algunas de las aplicaciones de clasificación o reconocimiento de patrones son: reconocimiento de firmas, reconocimiento de imágenes como rostros y categorización de textos y en este trabajo, clasificación del déficit habitacional.

A diferencia de las Redes Neuronales Artificiales (RNA) que utilizan durante la fase de entrenamiento, el principio de Minimización del Riesgo Empírico (ERM), las MVS se basan en el principio de Minimización del Riesgo Estructural (SRM), la cual ha mostrado un mejor desempeño que el ERM, ya que las Máquinas de Vectores de Soporte minimizan un límite superior al riesgo esperado a diferencia del ERM que minimiza el error sobre los datos de entrenamiento (Vapnik, 2000).

La MVS mapean los puntos de entrada a un espacio de características de una dimensión mayor, para luego encontrar el hiperplano que los separe y maximice el margen entre las clases.

Pertenecen a la familia de clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad (introducidos por funciones núcleo o kernel) con un sesgo inductivo muy particular (Carreras, Márquez y Romero, 2004).

La formulación matemática de las Máquinas de Vectores Soporte varía dependiendo de la naturaleza de los datos; es decir, existe una formulación para los casos lineales y, por otro lado, una formulación para casos no lineales.

Es importante tener claro que, de manera general para clasificación, las máquinas de vectores soporte buscan encontrar un hiperplano óptimo que separe las clases.

Las MVS han sido desarrolladas como una técnica robusta para clasificación y regresión aplicado a grandes conjuntos de datos complejos con ruido; es decir, con variables inherentes al modelo que para otras técnicas aumentan la posibilidad de error en los resultados pues resultan difíciles de cuantificar y observar.

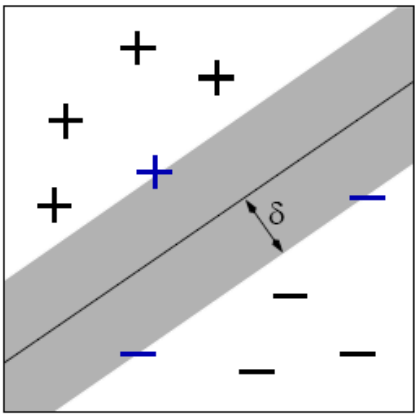
MVS PARA EL CASO NO LINEALMENTE SEPARABLE

Las MVS ofrecen la ventaja de que las mismas pueden ser utilizadas para resolver tanto problemas lineales como no lineales.

SEPARADOR LINEAL

En el caso de ser linealmente separable, las MVS conforman hiperplanos que separan los datos de entrada en dos subgrupos que poseen una etiqueta propia. En medio de todos los posibles planos de separación de las dos clases etiquetadas como {**-1, +1**} , existe sólo un hiperplano de separación óptimo, de forma que la distancia entre el hiperplano óptimo y el valor de entrada más cercano sea máxima (maximización del margen) con la intención de forzar la generalización de la máquina que se esté construyendo.

Aquellos puntos o ejemplos sobre los cuales se apoya el margen máximo son los denominados vectores de soporte. Un ejemplo de este caso se puede observar en la figura 2.1.



MVS Linealmente separable.

SEPARADOR NO LINEAL

Para el caso no lineal existen dos casos que vale la pena mencionar:

a) El primero de estos se presenta cuando los datos pueden ser separables con margen máximo pero en un espacio de características (el cual es de una mayor dimensionalidad y se obtiene a través de una trasformación a las variables del espacio de entrada) mediante el uso de una función kernel.

b) El segundo caso especial de las MVS denominado "Soft Margin" o margen blando, es utilizado cuando no es posible encontrar una trasformación de los datos que permita separarlos linealmente, bien sea en el espacio de entrada o en el espacio de características.

MVS CON MARGEN MÁXIMO EN EL ESPACIO DE CARACTERÍSTICAS

Hay casos donde los datos no pueden ser separados linealmente a través de un hiperplano óptimo en el espacio de entrada. En muchas situaciones, los datos, a través de una transformación no lineal del espacio de entradas, pueden ser separados linealmente pero en un espacio de características y se pueden aplicar los mismos razonamientos que para las MVS lineal con margen máximo.

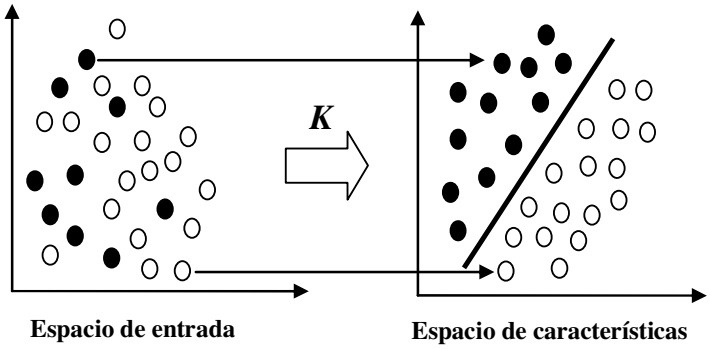
La trasformación de los datos de un espacio inicial a otro de mayor dimensión se logra mediante el uso de la función kernel.

Una función núcleo o kernel es un producto interno en el espacio de características, que tiene su equivalente en el espacio de entrada (Gunn, 1997).

$$K(x, x') = < \Phi(x), \Phi(x') >$$

donde *K,* es una función simétrica positiva definida que cumple las condiciones de Mercer (mayor detalle se puede ver en Gunn, 1997.

De manera gráfica se puede observar en la figura como la función kernel permite realizar la separación y el traslado de los datos al espacio de características.



MVS no linealmente separable inducida por una función kernel.

Entre los kernels más comunes, se encuentran: la función lineal, polinomial, RBF (Radial Basis Function), ERBF (Exponential Radial Basis Function), entre otros.

El problema de optimización a resolver para las MVS con margen blando está definido por un modelo de programación cuadrática con restricciones; es decir:

Maximizar $\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i x_j)$

Sujeto a: $\sum_{i=1}^{N} \alpha_i y_i = 0$

$\alpha_i \geq 0 \, , 1 \leq i \leq 0$

donde $K(x,y)$ es la función núcleo.

De acuerdo con Benavidez, *et al* (2006), este problema de optimización se resuelve introduciendo los multiplicadores de Lagrange, así los datos de entrenamiento sólo aparecerán en forma de una combinación de vectores y la resolución del

problema, se puede hallar resolviendo el problema dual dado por las ecuaciones que preceden.

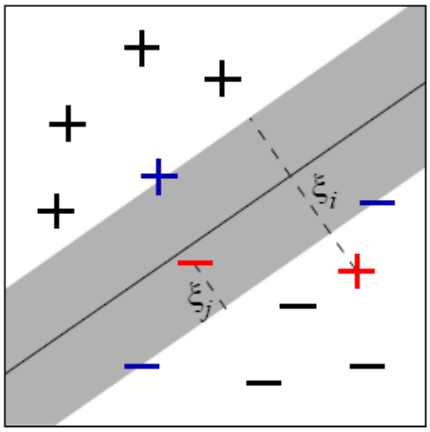
Según Betancourt (2005), la principal característica de la formulación dual de las MVS es que, mediante el uso de multiplicadores de Lagrange es posible representar el hiperplano deseado como combinación lineal de los propios datos y no en términos de la base del espacio vectorial en que aparecen los datos.

El análisis denominado Karush Kuhn Tucker (KKT) demuestra que la gran mayoría de los coeficientes de Lagrange son cero, y que sólo pueden ser distintos de cero para los vectores de soporte, puntos que se encuentran exactamente a la distancia marcada por el margen. Al dualizar el modelo de maximización del margen se transforma en un problema de minimización de una función cuadrática convexa sujeta a restricciones lineales.

MVS CON MARGEN BLANDO

Este tipo particular de las MVS trata aquellos casos donde existe datos de entrada erróneos, ruido o alto solapamiento de las clases en los datos de entrenamiento, donde se puede ver afectado el hiperplano clasificador, por esta razón se cambia un poco la perspectiva y se busca el mejor hiperplano clasificador que pueda tolerar el ruido en los datos de entrenamiento.

Según Benavidez, *et al*., (2006) esto se logra relajando las restricciones presentadas en el caso lineal, introduciendo variables de holgura no-negativas $\xi_i \geq 0$.



MVS con margen blando

De manera que, matemáticamente el problema queda definido como:

Minimizar $\frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i$

Sujeto a: $yi(< \omega, x_i > +b) \geq 1 - \xi_i$

$\xi_i \geq 0$, $i = 1, \ldots, l$

En la función de optimización que se debe solucionar para el modelo de MVS con margen blando se incluye un término de regularización que depende de las variables de holgura, el cual, a su vez depende de la magnitud de las mismas y del margen. Además, este término incluye una constante *C,* que determina la holgura del margen blando. El valor de este parámetro *C*, el cual debe ser estimado a priori, depende del evaluador. La elección de un valor para este parámetro y el tipo de función kernel influyen en el desempeño de las MVS (Benavidez, *et al* ,2006).

Siguiendo el mismo procedimiento utilizado en el caso linealmente separable, la resolución de este problema en 2.3, viene dada por la búsqueda de los multiplicadores de Lagrange, para esto se construye un Lagrangiano y se resuelve en el problema dual:

$$\textit{Maximizar} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

$$\textit{sujeto a}: \quad \sum_{i=1}^{N} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C \qquad 1 \leq i \leq N$$

La función a maximizar es la misma que para el caso de margen máximo, a diferencia de la restricción $0 \leq \alpha_i \leq C$. Esto implica que los datos o patrones que cumplen la condición de tener valores

$\alpha_i \geq C$ tienen el mismo comportamiento en la MVS con margen máximo. Es decir, que las MVS con margen máximo se pueden obtener con $C = \infty$. Esto significa que mientras más grande es el valor de *C*, más estricta es la MVS al momento de permitir errores, penalizándolos con mayor rigurosidad (Betancourt, 2005).

VENTAJAS DE LAS MVS

Las Máquinas de Vectores Soporte tienen ciertas características que las han puesto en ventaja respecto a otras técnicas populares de clasificación y/o regresión.

Una de dichas características que vale la pena mencionar es que las mismas pertenecen a las disciplinas de aprendizaje automático o aprendizaje estadístico. La idea que hay detrás de este tipo de aprendizaje es la de hacer que las máquinas puedan ir aprendiendo, a través de ejemplos; las salidas correctas para ciertas entradas.

La diferencia más notable de las máquinas de vectores de soporte con respecto a otros algoritmos de aprendizaje, es la aplicación de un nuevo principio inductivo, que busca la minimización del riesgo estructural, además del uso de una función núcleo o kernel, atribuyéndole una gran capacidad de generalización, incluso cuando el conjunto de entrenamiento es pequeño.

Se dice que tanto la capacidad de generalización cómo el proceso de entrenamiento de la máquina no dependen necesariamente del número de atributos, lo que permite un excelente comportamiento en problemas de alta dimensionalidad (Aranguren, 2008).

POSIBLES PROBLEMAS CON LAS MVS

Uno de los más comunes es lo que se conoce como "*Overtraining*" o sobre entrenamiento, el cual ocurre cuando se han aprendido muy bien los datos de entrenamiento pero no se pueden clasificar bien ejemplos nunca antes vistos (datos de verificación), es decir, una mala generalización del modelo.

Otro problema que se puede presentar cuando no se ha aprendido muy bien la característica de los datos de entrenamiento, por lo que se hace una mala clasificación. El experimentador debe tener en cuenta estas consideraciones a la hora de ajustar el modelo ya que de ello depende la exactitud y éxito de la predicción.