Matemática: Una publicación de FCNM – ESPOL
2015, Vol. 13, No.2

CLASIFICACIÓN MEDIANTE MÁQUINAS DE SOPORTE VECTORIAL

Ramírez John[1], Ramos Miriam[2]

Resumen. *Este artículo presenta las principales ideas y conceptos en que se fundamentan las Máquinas de Soporte Vectorial. Uno de estos conceptos es el hiperplano clasificador o de decisión. Otro concepto clave es el de núcleo considerado como una forma de medir la similaridad entre objetos de diferente clase. Sin embargo, el teorema de Mercer es la base matemática para el funcionamiento de las Máquinas de Soporte Vectorial, este posibilita construir clasificadores en espacios cuya dimensión es mayor que la del espacio original. Por lo que, permite separar a los grupos, lo que facilita su clasificación. Además, se presenta una aplicación a datos simulados.*
Palabras claves: Máquinas de Soporte Vectorial, hiperplano, núcleo, espacio de Hilbert.

Abstract. *This article presents the main ideas and concepts behind Support Vector Machines. One of these concepts is the classifier or decisión hyperplane. Another key concept is the kernel considered as a way to measure similarity between objects of different class. However, Mercer's theorem is the mathematical basis for the functioning of Support Vector Machines, it allows to build classifiers in spaces whose dimension is greater than the original space. This allows to separate groups, which facilitates classification. Moreover, an application to simulated data is shown.*
Keywords: Support Vector Machines, hyperplane, kernel, Hilbert space.

Recibido: Mayo 2015.
Aceptado: Junio 2015

1. INTRODUCCIÓN

Una máquina de soporte vectorial es un algoritmo que aprende por medio de ejemplos a etiquetar objetos. Constituye una herramienta para clasificación y regresión desarrollado en la década de 1990; y, no es sino la generalización de un sencillo e intuitivo método de clasificación, el llamado clasificador de margen máximo. Aunque este clasificador es elegante y sencillo, no es posible aplicarlo a la mayoría de los conjuntos de datos debido a que es necesario que las clases puedan separarse por un límite lineal, por esta razón se desarrolló el clasificador de soporte vectorial, mismo que se puede aplicar a una gama más amplia de casos; sin embargo, los últimos estudios incorporan lo que se conoce como máquina de soporte vectorial con el propósito de ajustar los límites de clase no lineales.

Estas máquinas de soporte vectorial están destinadas al ajuste de clasificación binaria, es decir, aquella en la que existen dos clases, pudiendo también extenderse a más de dos; y existiendo además, una conexión muy cercana con métodos estadísticos como la regresión logística.

El concepto de máquina de soporte vectorial descansa en el de clasificador de soporte vectorial y éste a su vez es una generalización de los hiperplanos clasificadores que presentamos a continuación.

[1]*Ramírez John, M.Sc., Profesor, Departamento de Matemáticas, Facultad de Ciencias Naturales y Matemáticas, ESPOL. (e_mail: jramirez@espol.edu.ec)*
[2]*Ramos Miriam, MPC., Profesora, Departamento de Matemáticas, Facultad de Ciencias Naturales y Matemáticas, ESPOL. (e_mail: mvramosb@espol.edu.ec)*

CLASIFICACIÓN MEDIANTE UN HIPERPLANO

Tomando como referencia un espacio p −dimensional, un hiperplano es un subespacio plano afín de dimensión $p-1$; por ejemplo en dos dimensiones, un hiperplano es un subespacio unidimensional, es decir, una línea; mientras que en tres dimensiones, un hiperplano es un plano. Cuando $p>3$, resulta difícil visualizar un hiperplano; sin embargo, la noción de un subespacio plano $(p-1)$ −dimensional se mantiene.

La definición matemática de un hiperplano resulta bastante simple ya que el mismo satisface la ecuación:

$$\beta_0+\beta_1X_1+\beta_2X_2=0 \qquad (0.1)$$

Con parámetros β_0,β_1,β_2; es decir, cualquier $X=(X_1,X_2)^T$ que satisface la ecuación anterior, representa un punto en el hiperplano. Resulta evidente que la ecuación dada representa una recta en el plano (Hastie, Tibshirani, & Friedman, 2013).

La ecuación (0.1) puede extenderse fácilmente a la configuración p-dimensional:

$$\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p=0 \qquad (0.2)$$

Definiendo un hiperplano p-dimensional, con parámetros $\beta_0,\beta_1,\beta_2,...,\beta_p$; para el cual si un punto $X=(X_1,X_2,\ldots,X_p)^T$ satisface la ecuación anterior, entonces este punto pertenece al hiperplano.

En el caso de que X no satisfaga la ecuación (0.2), sino cualquiera de las siguientes desigualdades:

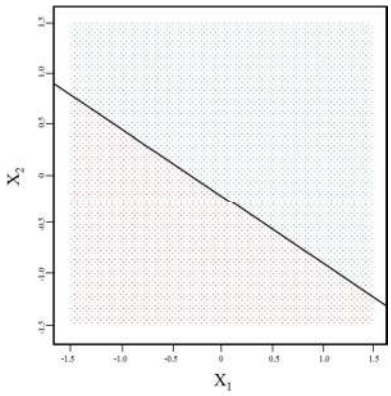
$$\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p>0 \qquad (0.3)$$

ó

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0 \quad (0.4)$$

Se puede concluir que X se localiza a un lado u otro del hiperplano, entendiéndose que el hiperplano divide al espacio *p*-dimensional en dos mitades. Siendo así, se puede fácilmente determinar de qué lado del hiperplano se ubica un punto, simplemente calculando el signo del lado izquierdo de las desigualdades (0.3) y (0.4). Ver Figura 1.

Figura 1.
Separación de observaciones por un hiperplano



2. CLASIFICACIÓN MEDIANTE UN HIPERPLANO DE SEPARACIÓN

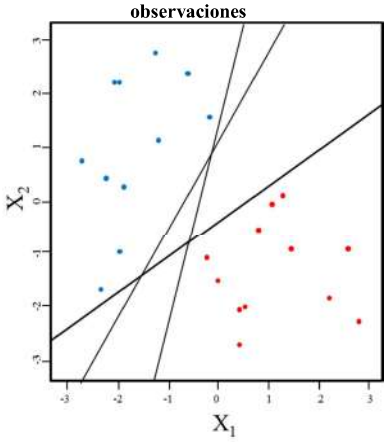
Si ahora se tiene una matriz X de datos de orden $n\,x\,p$ con *n* observaciones en el espacio *p*-dimensional, es decir:

$$x_1 = \begin{pmatrix} x_{11} \\ . \\ . \\ . \\ x_{1p} \end{pmatrix}, \ldots, x_n = \begin{pmatrix} x_{n1} \\ . \\ . \\ . \\ x_{np} \end{pmatrix} \quad (0.5)$$

y si además estas observaciones se dividen en dos clases, es decir, $y_1, \ldots, y_n \in \{-1, 1\}$ donde -1 representa una clase y 1 la otra, contando con un vector *p*-dimensional de características observadas $x^* = \left(x_1^* \ldots x_p^*\right)^T$, el propósito será desarrollar un clasificador basado en los datos que clasifique correctamente la observación de prueba utilizando sus mediciones características (Scholkopf & Smola, Learning with Kernels. Support Vector Machines, Regularization, Optimazation and Beyond, 2001).

Suponiendo que es posible construir un hiperplano que separa las observaciones de acuerdo a sus etiquetas de clase, se pueden observar en la Figura 2 ejemplos de tales hiperplanos de separación.

Figura 2.
Posibles hiperplanos de separación para las mismas observaciones



Se podrían etiquetar las observaciones de la clase azul como $y_i = 1$ y los de la clase roja como $y_i = -1$, con lo cual se establece que un hiperplano de separación tiene la propiedad de que:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} > 0 \; si \; y_i = 1 \quad (0.6)$$

o

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} < 0 \; si \; y_i = -1 \quad (0.7)$$

Con lo cual, un hiperplano de separación cumple con la siguiente propiedad:

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) > 0; \quad i = 1,2,3,\ldots,n \quad (0.8)$$

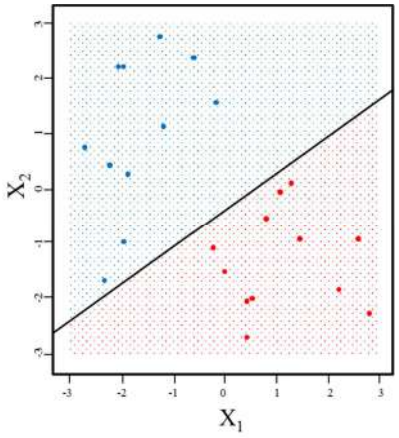
Si un hiperplano de separación existe, es posible utilizarlo para construir un clasificador muy natural, de manera tal que a una observación de prueba se le asigna una clase dependiendo del lado del hiperplano en el que se encuentra.

En la Figura 3 se muestra un ejemplo del clasificador, en el cual se clasifica la observación de prueba x* en base al signo de $f\left(x^*\right) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*$, así si $f(x^*)$ es positivo, entonces se asigna la observación de prueba para la clase 1, y si $f(x^*)$ es negativo, se la asigna a la clase -1. También se puede hacer uso de la magnitud de $f(x^*)$ de manera tal que si $f(x^*)$ está lejos de cero, entonces esto significa que x* se encuentra lejos del hiperplano, y así se tiene seguridad de la asignación de clase para x^*.

Por otro lado, si $f(x^*)$ es cercano a cero, entonces x^* se encuentra cerca del hiperplano, y así tenemos menos certeza acerca de la asignación de clase para x^*. Se puede concluir entonces que un clasificador que se basa en un hiperplano de

separación siempre conduce a un límite de decisión lineal.

Figura 3.
Hiperplano clasificador



3. CLASIFICADOR DE MARGEN MÁXIMO

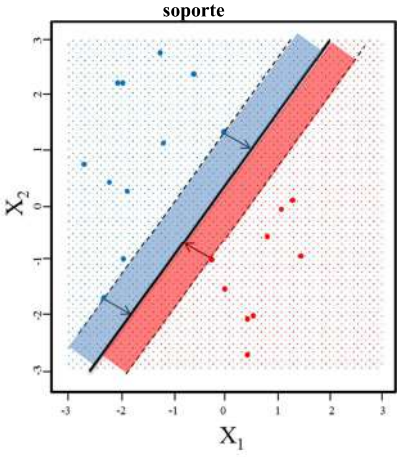
En general, si los datos pueden ser perfectamente separados usando un hiperplano, entonces habrá un número infinito de tales hiperplanos, todo esto debido a que un hiperplano de separación dado, generalmente puede ser desplazado un poco hacia arriba o abajo, o ser girado, sin entrar en contacto con cualquiera de las observaciones. Tres posibles hiperplanos de separación se muestran la Figura 2. Con el fin de construir un clasificador basado en un hiperplano de separación, se debe tener una forma razonable para decidir cuál de los infinitos hiperplanos de separación posibles, utilizar.

Una elección natural es el hiperplano de margen máximo (también conocido como el hiperplano de separación óptimo), que es el hiperplano de separación que está más alejado de las observaciones de entrenamiento; es decir, es posible calcular la distancia (perpendicular) de cada observación a un hiperplano de separación dado; la más pequeña de tales distancias es la distancia mínima de las observaciones al hiperplano, y se conoce como margen. El hiperplano de margen máximo es el hiperplano de separación para el cual el margen es el más grande, es decir, es el hiperplano que tiene la distancia mínima más lejana a las observaciones de entrenamiento. Entonces podemos clasificar una observación de prueba basada en qué lado del hiperplano de margen máximo se encuentra. Esto se conoce como el clasificador de margen máximo. Se espera que un clasificador que tiene un amplio margen en los datos de entrenamiento también tenga un amplio margen en los datos de prueba, y por lo tanto va a clasificar las observaciones de prueba correctamente. Aunque el clasificador de margen máximo es a menudo exitoso, también puede conducir al sobreajuste cuando p es grande (Hastie, Tibshirani, & Friedman, 2013).

Si $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes del hiperplano de margen máximo, entonces el clasificador de margen máximo clasifica la observación de prueba x^* basado en el signo de $f\left(x^*\right) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*$. La Figura 4 muestra el hiperplano de margen máximo en el conjunto de datos de la Figura 3. El hiperplano de margen máximo que se muestra en la Figura 4 en efecto resulta en una mayor distancia mínima entre las observaciones y el hiperplano de separación, es decir, un margen mayor. En cierto sentido, el hiperplano de margen máximo representa la línea media del "bloque" más amplio que podemos insertar entre las dos clases.

Examinando la Figura 4, vemos que tres observaciones de entrenamiento son equidistantes del hiperplano de margen máximo y yacen a lo largo de las líneas de trazos que indican el ancho del margen. Estas tres observaciones son conocidas como vectores de soporte, ya que son vectores en el espacio p-dimensional y ellos "apoyan" el hiperplano de margen máximo, en el sentido de que si estos puntos fueran movidos ligeramente, entonces el hiperplano de margen máximo se movería también. Curiosamente, el hiperplano de margen máximo depende directamente de los vectores de soporte, pero no en las otras observaciones: un movimiento para cualquiera de las otras observaciones no afectaría el hiperplano de separación, a condición de que el movimiento de la observación no cause que la misma cruce el límite fijado por el margen.

Figura 4.
Hiperplano clasificador de margen máximo y vectores de soporte



4. CONSTRUCCIÓN DEL CLASIFICADOR DE MARGEN MÁXIMO

La tarea de construir el hiperplano de margen máximo estará basada en un conjunto de n observaciones de entrenamiento $x_1, \ldots, x_n \in \mathbb{R}^p$ y las etiquetas asociadas a las clases $y_1, \ldots, y_n \in \{-1,1\}$. El hiperplano de margen máximo es la solución para el siguiente problema de optimización:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{\text{Maximizar } M} \tag{0.9}$$

$$\text{s.a.r} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \tag{0.10}$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) \geq M; \quad \forall i = 1,2,3,\ldots,n \tag{0.11}$$

La restricción en la desigualdad (0.11) respecto a que debe tomar un valor mayor a M $y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) \geq M; \quad \forall i = 1,2,3,\ldots,n$ garantiza que cada observación se ubicará en el lado correcto del hiperplano, siempre que M sea positivo.

La expresión (0.10) no es realmente una limitación en el hiperplano, ya que si $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} = 0$ define un hiperplano, entonces también lo hace $k\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) = 0$ para cualquier $k \neq 0$. Sin embargo, (0.10) añade significado a (0.11); se puede demostrar que con esta restricción la distancia normal (perpendicular) de la i-ésima observación al hiperplano está dada por:

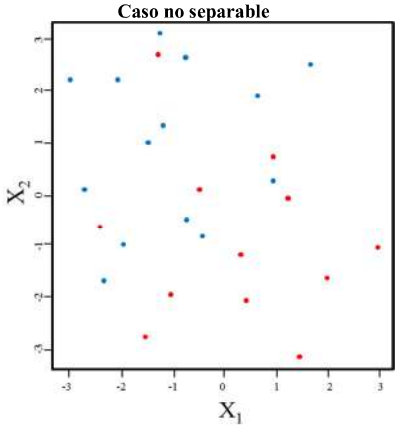
$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) \tag{0.12}$$

Por lo tanto, las restricciones (0.10)-(0.11) asegurarán que cada observación estará en el lado correcto del hiperplano y por lo menos a una distancia M del hiperplano. Por lo tanto, M representa el margen del hiperplano, y el problema de optimización elige $\beta_0, \beta_1, \ldots, \beta_p$ para maximizar M. Esto es exactamente la definición del hiperplano de margen máximo.

5. CASO NO SEPARABLE

El clasificador de margen máximo es una forma muy natural para llevar a cabo la clasificación, si existe un hiperplano de separación; sin embargo, en muchos casos el hiperplano de separación no existe, y entonces no hay un clasificador de margen máximo. En estas circunstancias, el problema de optimización (0.9)-(0.11) no tiene solución con $M > 0$.

Figura 5.
Caso no separable

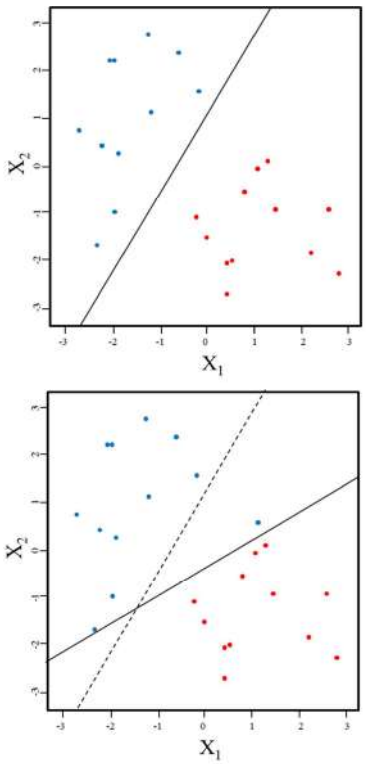


Un ejemplo se muestra en la Figura 5. En este caso, no se pueden separar exactamente las dos clases; sin embargo, se puede ampliar el concepto de un hiperplano de separación con el fin de desarrollar un hiperplano que separa al menos las clases, usando un llamado *margen suave* (James, Witten, Hastie, & Tibshirani, 2013). La generalización del clasificador de margen máximo para los casos no separables se conoce como el clasificador de soporte vectorial.

6. CLASIFICADOR DE SOPORTE VECTORIAL

En la Figura 5, se puede notar que las observaciones que pertenecen a dos clases no son necesariamente separables por el hiperplano. De hecho, incluso si un hiperplano de separación existe, entonces habría casos en los cuales un clasificador basado en un hiperplano de separación no podría ser deseable. Un clasificador basado en un hiperplano de separación clasificará necesaria y perfectamente todas las observaciones; esto puede conducir a la sensibilidad en las observaciones individuales.

Figura 6.
Cambio del hiperplano de margen máximo



El incorporar una sola observación en la Figura 6 provocaría un cambio dramático en el hiperplano de margen máximo. El hiperplano de margen máximo resultante no es satisfactorio ya que por un lado, sólo tiene un pequeño margen. Esto es problemático, porque como mencionamos anteriormente, la distancia de una observación del hiperplano puede ser vista como una medida de nuestra confianza en la que la observación fue correctamente clasificada. Por otra parte, el hecho que el hiperplano de margen máximo sea extremadamente sensible a un cambio en una sola observación, sugiere que éste pueda tener un sobreajuste en los datos.

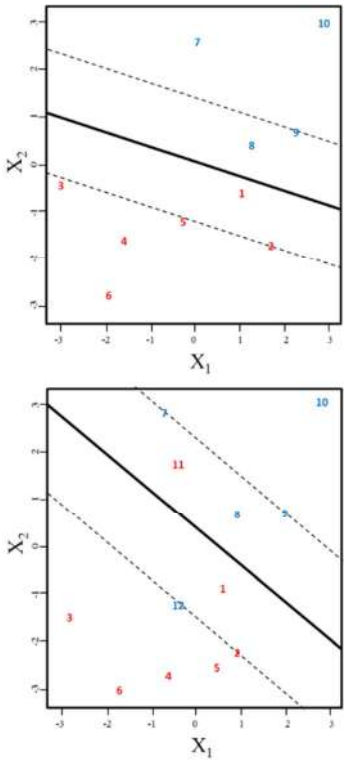
En este caso, podríamos estar dispuestos a considerar un clasificador basado en un hiperplano que no separe perfectamente las dos clases, en beneficio de:

- Mayor robustez a las observaciones individuales, y
- Mejor clasificación de la mayor parte de las observaciones de entrenamiento

Es decir, esto podría ser útil para clasificar incorrectamente algunas observaciones de entrenamiento con el fin de hacer un mejor trabajo en la clasificación de las observaciones restantes.

El *clasificador de soporte vectorial*, a veces llamado *clasificador de margen suave,* hace exactamente esto. En lugar de buscar el margen más largo posible para que cada observación no sólo esté en el lado correcto del hiperplano, sino también en el lado correcto del margen, mejor permitimos que algunas observaciones estén en el lado incorrecto del margen, o incluso en el lado incorrecto del hiperplano.

Figura 7.
Clasificador de soporte vectorial o de margen suave



Un ejemplo es mostrado en la Figura 7. La mayor parte de las observaciones están en el lado correcto del margen; sin embargo, un pequeño subconjunto de las observaciones está en el lado incorrecto del margen.

DETALLES DEL CLASIFICADOR DE SOPORTE VECTORIAL

El clasificador de soporte vectorial clasifica una observación dependiendo de qué lado de un hiperplano ésta se encuentra. El hiperplano se escoge para separar correctamente la mayoría de las observaciones en las dos clases, pero puede clasificar incorrectamente algunas observaciones. La solución al problema de maximización es:

$$\underset{\beta_0,\beta_1,...,\beta_p,\varepsilon_1,\varepsilon_2,...,\varepsilon_n}{\text{Maximizar } M} \tag{0.13}$$

$$\text{s.a.r} \quad \sum_{j=1}^{p}\beta_j^2 = 1 \tag{0.14}$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\right)\geq M\left(1-\varepsilon_i\right);$$
$$\forall i=1,2,3,...,n \tag{0.15}$$

$$\varepsilon_i \geq 0,\ \sum_{i=1}^{n}\varepsilon_i \leq C \tag{0.16}$$

Donde C es un parámetro de ajuste no negativo. Como en (0.11), M es el ancho del margen; buscamos hacer esta cantidad tan grande como sea posible. En (0.15) $\varepsilon_1,...,\varepsilon_n$ son variables de tolerancia que permiten observaciones individuales en el lado incorrecto del margen o del hiperplano. Una vez que se ha resuelto (0.13)-(0.16), se clasifica una observación prueba x^* como antes, simplemente determinando en qué lado del hiperplano se encuentra; es decir, clasificamos la observación prueba basada en el signo de $f(x^*)=\beta_0+\beta_1 x_1^*+\cdots+\beta_p x_p^*$.

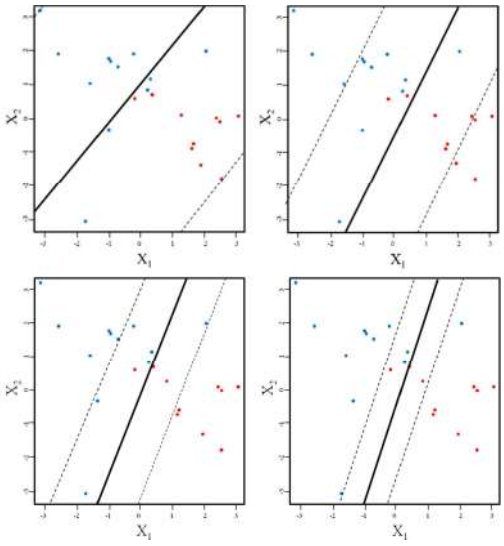
El problema (0.13)-(0.16) parece complejo, pero la agudeza de su comportamiento puede hacerse a través de una serie de simples observaciones que se presentan a continuación. En primer lugar la variable de tolerancia ε_i nos dice donde se encuentra la i-ésima observación, en relación al hiperplano y al margen. Si $\varepsilon_i=0$ entonces la i-ésima observación está en el lado correcto del margen. Si $\varepsilon_i>0$ entonces la i-ésima observación está en el lado incorrecto del margen, y decimos que la i-ésima observación ha transgredido el margen. Si $\varepsilon_i>1$ entonces está en el lado incorrecto del hiperplano.

Al considerar el papel del parámetro de ajuste C, se tiene que en (0.16), C limita la suma de las ε_i 's, por lo que determina el número y la gravedad de las violaciones al margen (y para el hiperplano) que vamos a tolerar. Podemos pensar en C como un presupuesto para la cantidad que el margen puede ser violado por n observaciones. Si $C=0$, entonces no hay presupuesto para violaciones al margen, y tiene que ser el caso que $\varepsilon_1=\varepsilon_2=\cdots=\varepsilon_n=0$, en cuyo caso (0.13)-(0.16) simplemente alcanza al problema de optimización del hiperplano de margen máximo (0.9)-(0.12). (Por supuesto, existe un hiperplano de margen máximo sólo si las dos clases son separables.) Para $C>0$ no más de C observaciones pueden estar en el lado equivocado del hiperplano, porque si una observación está en el lado equivocado del hiperplano entonces $\epsilon_i>1$, y (0.16) requiere que $\sum_{i=1}^{n}\varepsilon_i\leq C$. Como el presupuesto C aumenta, se vuelve más tolerante a las violaciones del margen, por lo que el margen se ampliará. Por el contrario, como C disminuye, nos volvemos menos tolerantes a violaciones al margen y así el margen se estrecha. Un ejemplo se muestra en la Figura 8.

En la práctica, C es tratada como un parámetro de ajuste que generalmente se elige a través de la validación cruzada. Al igual que con los parámetros de ajuste, C controla la compensación sesgo-varianza de la técnica de aprendizaje estadístico. Cuando C es pequeño, se buscan estrechos márgenes que rara vez son transgredidos; esto equivale a un clasificador que es altamente ajustable a los datos, que pueden tener poco sesgo pero de gran varianza (James, Witten, Hastie, & Tibshirani, 2013). Por otra parte, cuando C es más grande, el margen es más ancho y se permiten más violaciones a la misma; esto equivale a ajustar los datos en forma menos estricta y obtener un clasificador que es potencialmente más sesgado pero puede tener menor varianza.

El problema de optimización (0.13)-(0.16) tiene una propiedad muy interesante: resulta que sólo observaciones que, o bien se encuentran en el margen o que violan el margen afectarán al hiperplano, y por lo tanto al clasificador obtenido. En otras palabras, una observación que se encuentra estrictamente en el lado correcto del margen no afecta al clasificador de soporte vectorial. Cambiar la posición de dicha observación no cambiaría el clasificador en absoluto, a condición de que su posición se mantenga en el lado correcto del margen. Las observaciones que se encuentran directamente en el margen, o en el lado equivocado del margen de su clase, se conocen como vectores de soporte. Estas observaciones afectan al clasificador de soporte vectorial.

Figura 8.
Clasificadores de soporte vectorial y su relación sesgo-varianza



El hecho de que sólo los vectores de soporte afectan el clasificador está alineado con la afirmación anterior de que C controla la compensación del sesgo-varianza del clasificador de soporte vectorial. Cuando el parámetro de ajuste C es grande, entonces el margen es ancho, muchas observaciones violan el margen, y así hay muchos vectores de soporte. En este caso, muchas observaciones están involucradas en la determinación del hiperplano. El panel superior izquierdo en la Figura 8 ilustra esta configuración: este clasificador tiene una baja varianza (ya que muchas observaciones son vectores de soporte) pero potencialmente alto sesgo. En contraste, si C es pequeño, entonces habrá un menor número de vectores de soporte y por lo tanto el clasificador resultante tendrá sesgo bajo pero alta varianza. El panel inferior derecho de la Figura 8 ilustra esta configuración, con sólo ocho vectores de soporte.

El hecho de que la regla de decisión de clasificador de soporte vectorial se basa sólo en un potencialmente pequeño subconjunto de las observaciones de entrenamiento (los vectores de apoyo) significa que es bastante robusto para el comportamiento de las observaciones que están lejos del hiperplano. Esta propiedad es distinta de algunos de los otros métodos de clasificación existentes, tales como el análisis discriminante lineal (LDA). Es importante recordar que la regla de clasificación LDA depende de la media de todas las observaciones dentro de cada clase, así como de la matriz de covarianzas entre clases calculada utilizando todas las observaciones. Por el contrario, la regresión logística, a diferencia de LDA, tiene muy baja sensibilidad a las observaciones lejos de la frontera de decisión. De hecho, se puede afirmar que el clasificador de soporte vectorial y la regresión logística están estrechamente relacionados.

7. MÁQUINAS DE SOPORTE VECTORIAL

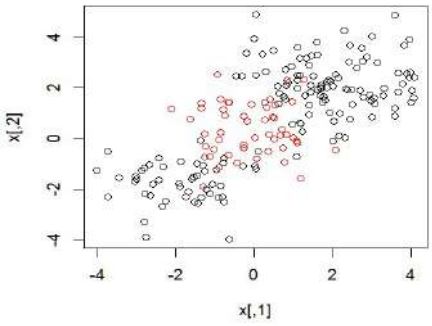
NÚCLEOS

La clasificación requiere identificar observaciones que tengan bastante similitud. Es decir, requiere la noción de semejanza o similitud. Al clasificar se utiliza la información obtenida de datos que a priori se conoce su clasificación (clasificación supervisada), para catalogar datos nuevos o datos que no intervinieron en la obtención del modelo, en función de su similitud con los datos iniciales.

Una función útil para medir la semejanza entre datos es la función producto interno, pues a partir de ella se deriva la noción de distancia o similitud.

El clasificador de soporte vectorial descrito anteriormente se utiliza usualmente en problemas de clasificación cuando la frontera de separación es lineal. En la práctica, la mayoría de veces la frontera de separación de las clases no es lineal (Figura 9), en este caso los resultados obtenidos no se ajustarán a la realidad.

Figura 9.
Frontera de separación no lineal



Una forma de enfrentar este problema es el de tomar funciones de las variables originales. Así por ejemplo se podrían tomar en cuenta los cuadrados (y/o cubos) de los predictores, para capturar la no linealidad del problema. Así, en vez de considerar p características $X_1,\dots X_p$, podríamos considerar $2p$ características $X_1,\dots X_p, X_1^2,\dots,X_p^2$. Las ecuaciones (0.13)-(0.16) se convertirán en:

$$\max_{\beta_0,\beta_{11},\dots,\beta_{p1},\dots\beta_{p2},\varepsilon_1,\dots,\varepsilon_n} M$$

s.a.r

$$y_i\left(\beta_0+\sum_{j=1}^{p}\beta_{j1}x_{ij}+\sum_{j=1}^{p}\beta_{j2}x_{ij}^2\right)\geq M\left(1-\varepsilon_i\right),$$

$$\sum_{i=1}^{n}\varepsilon_i\leq C,\quad \varepsilon_i\geq 0,\quad \sum_{j=1}^{p}\sum_{k=1}^{2}\beta_{jk}^2=1 \qquad (0.17)$$

La frontera de decisión en el espacio de las $2p$ características sigue siendo lineal, pero en el espacio original es no lineal.

Se puede pensar en ampliar el espacio de las característica no sólo utilizando cuadrados y cubos, sino también polinomios de mayor grado, incluso considerar interacciones de los predictores (X_iX_j). Esta es la idea base de las máquinas de soporte vectorial (MSV).

El clasificador de soporte vectorial se puede expresar como:

$$f(x)=\beta_0+\sum_{i=1}^{n}\alpha_i\left\langle x,x_i\right\rangle \qquad (0.18)$$

Donde hay n parámetros α_i, uno por cada observación de entrenamiento. Para estimar estos

parámetros y β_0 se necesitan los $\binom{n}{2}$ productos internos $\langle x_i, x_{i'} \rangle$ entre todos los pares de observaciones de entrenamiento.

Para evaluar la función clasificadora f en un nuevo punto x, es necesario calcular todos los productos $\langle x, x_i \rangle$ entre la nueva observación y los datos de entrenamiento. Sin embargo, los α_i son diferentes de cero sólo para los vectores de soporte. Así, si notamos por Γ al conjunto de observaciones que son a su vez vectores de soporte, tenemos:

$$f(x) = \beta_0 + \sum_{i \in \Gamma} \alpha_i \langle x, x_i \rangle \qquad (0.19)$$

Esto significa que para evaluar f en x sólo es necesario el cálculo de un pequeño número de productos internos.

Las MSV son una extensión de los clasificadores de soporte vectorial que resultan al ampliar el espacio de características mediante el uso de núcleos, que son una generalización del producto interno usual $\langle x_i, x_{i'} \rangle$ presente en (0.19).

Si en (0.19) reemplazamos los productos internos por una generalización de la forma

$$K(x_i, x_{i'}) \qquad (0.20)$$

Donde K es una función denominada núcleo que mide el grado de similaridad de las observaciones $x_i, x_{i'}$. Un caso particular de K es el producto interno usual en $\mathbb{R}^p$:

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \qquad (0.21)$$

Conocido como núcleo lineal, debido a que el clasificador de soporte vectorial es lineal en las características. El núcleo lineal cuantifica el grado de semejanza entre dos variables o características mediante la correlación entre sus componentes. Así, variables ortogonales según el producto interno usual equivalen a variables no correlacionadas. Además, la raíz cuadrada del producto interno proporciona la distancia entre las variables.

La idea de utilizar el producto interno usual se puede generalizar a espacios vectoriales más generales que $\mathbb{R}^p$ en los que sean completos y donde sea posible definir un producto interno, los denominados espacios de Hilbert. Estos espacios se asemejan al espacio euclídeo en el sentido de que poseen una geometría inherente al producto interno. Una geometría que permite medir distancias, tamaños y ángulos (Berlinet & Thomas-Agnan, 2004).

Puede suceder que la similitud entre dos variables sea más notoria en algún espacio diferente de características original (por lo general $\mathbb{R}^p$), notado por H, en un espacio de Hilbert más general que se notará por V. Se define la función de mapeo:

$$\begin{aligned} \varphi : H &\to V \\ x &\to v = \varphi(x) \end{aligned} \qquad (0.22)$$

La función φ transforma a las características originales en elementos del espacio V. Así:

$$K(x_1, x_2) = v_1 \cdot v_2 = \varphi(x_1) \cdot \varphi(x_2)$$

A partir de (0.23) se puede decir que un núcleo es una función que se puede representar por medio de un producto interno en algún espacio de Hilbert V. Lo que permite la utilización de conceptos geométricos por abstracto que sea V. Puesto que la aplicación φ se puede elegir libremente, siempre que cumpla ciertas condiciones, se dispone de una amplia gama de medidas de similitud. Las condiciones sobre φ las proporciona el teorema de Mercer[1]:

Teorema de Mercer: Sea $x \in \mathbb{R}^p$ un espacio de Hilbert V y una aplicación

$$\begin{aligned} \varphi : \mathbb{R}^p &\to V \\ x &\to v = \varphi(x) \end{aligned} \qquad (0.24)$$

Entonces, el producto interno en V tiene la representación equivalente:

$$\sum_i \varphi_i(x_1)\varphi_i(x_2) = K(x_1, x_2) \qquad (0.25)$$

Donde φ_i es la i-ésima componente del mapeo φ y K es una función simétrica definida positiva.

La inversa del teorema también es cierta. Es decir, para cualquier función definida positiva K, existe un espacio en el cual define un producto interno (Berlinet & Thomas-Agnan, 2004).

La aplicación φ constituye un isomorfismo entre dos espacios de Hilbert y por ende establece una relación biunívoca entre elementos de ambos espacios que preserva el producto interno.

[1] Enunciado por el matemático inglés James Mercer en 1909

El teorema de Mercer nos asegura que se puede definir un mapeo a un nuevo espacio de Hilbert, que en general tendrá dimensiones mayores que el del espacio de características original, en el cual se puede definir una similitud, mediante la cual las clases se separen más fácilmente.

Cuando el clasificador de soporte vectorial se combina con un núcleo se obtiene un clasificador denominado *máquina de soporte vectorial*, que tiene la forma:

$$f(x) = \beta_0 + \sum_{i \in \Gamma} \alpha_i K\left(x, x_i\right) \qquad (0.26)$$

Al utilizar el clasificador con este tipo de núcleos se obtienen fronteras de decisión más flexibles que las lineales, puesto que el espacio en el que se trabaja es de dimensión más alta que el espacio de los datos originales. El teorema de Mercer no indica la como construir el espacio V. Es decir, no indica cómo construir la aplicación φ y por ende el núcleo *K*. Sin embargo, en la práctica se han hallado varios núcleos que permiten resolver problemas de clasificación bastante complejos. Los núcleos más utilizados son:

Núcleo polinomial de grado $d \in \mathbb{R}$:

$$K\left(x_1, x_2\right) = \left(1 + \sum_{j=1}^{p} x_{1j} x_{2j}\right)^d \qquad (0.27)$$

Núcleo radial

$$K\left(x_1, x_2\right) = \exp\left(-\gamma \sum_{j=1}^{p} \left(x_{1j} - x_{2j}\right)^2\right) \qquad (0.28)$$

Tangente hiperbólica

$$K\left(x_1, x_2\right) = \tanh\left(-\gamma\left(x_1 \cdot x_2\right) + c\right) \qquad (0.29)$$

MEDIDAS DE RENDIMIENTO DE UNA MSV

Las medidas de precisión en la clasificación de una MSV son el error de clasificación y la exactitud predictiva, y se las obtiene de la matriz o tabla de clasificación:

Tabla 1. Clase real vs Clase Predicha

	Clase Real	
Clase Predicha	Positiva	Negativa
Positiva	Verdaderos positivo (VP)	Falsos positivos (FP)
Negativa	Falsos negativos (FN)	Verdaderos negativos (VN)

Sea $n = VP + FP + VN + FN$

Exactitud predictiva:

$$\text{Exactitud} = \frac{VP + VN}{n} \qquad (0.30)$$

Error de clasificación:

$$\text{Error} = \frac{FP + FN}{n} \qquad (0.31)$$

Estas medidas dependen de cómo están distribuidos los datos, se definen otras medidas que permiten evaluar la calidad del clasificador sobre cada una de las clases de forma independiente. Estas medidas son la precisión y la sensibilidad:

$$\text{Especificidad} = \frac{VN}{VN + FP} \qquad (0.32)$$

$$\text{Sensibilidad} = \frac{VP}{VP + FN} \qquad (0.33)$$

La precisión mide la exactitud al clasificar los casos como positivos, cuando éstos son clasificados correctamente. La sensibilidad mide la completitud o exactitud positiva, que indica cuantos casos de esta clase fueron clasificados correctamente.

8. MSV CUANDO HAY MÁS DE DOS CLASES

Hasta el momento se ha descrito un clasificador para identificar las clases cuando los individuos pertenecen a dos categorías. Sin embargo, se puede utilizar este tipo de clasificadores cuando los individuos pertenecen a más de una clase.

Hay dos enfoques a saber: uno vs uno y uno vs todos:

CLASIFICACIÓN UNO VS UNO

Supongamos que existen $N > 2$ clases. En la clasificación *uno vs uno* se construyen $\binom{N}{2}$ máquinas de soporte vectorial, cada una de las cuales compara un par de clases. Por ejemplo, una MSV podría comparar la *i*-ésima clase codificada como *i*, la *j*-ésima clase codificada como *j*. Clasificamos a una observación utilizando cada uno de los $\binom{N}{2}$ clasificadores. Cada uno de ellos va a asignarla a una clase. La clase que se le asigna a la observación es la que más se repite.

CLASIFICACIÓN UNO VS TODOS

El enfoque *uno vs todos* es un procedimiento alternativo para cuando el número de categorías *N* es mayor a *2*. Se trata de ajustar *N* MSV, cada vez comparando una de las *N* clases comparándola con las restantes *N*-1 clases.

Sean $\beta_{0k}, \beta_{1k}, ..., \beta_{pk}$ los parámetros que resultan de ajustar un MSV comparando la *k*-ésima clase (codificada como +1), comparándola con las clases restantes (codificadas como -1).

Sea x^* una observación a clasificarse. Sea asigna esta observación a la clase en la cual la cantidad $\beta_{0k} + \beta_{1k}x_1^* + ... + \beta_{pk}x_p^*$ sea mayor, puesto que esto es un indicativo de que la observación pertenece a esa clase antes que a las otras clases.

Para el caso en que existan más de un clase, una medida de la exactitud de la clasificación está dada por:

$$\text{Exactitud} = \frac{\sum_{i=1}^{N} a_{ii}}{n} \qquad (0.34)$$

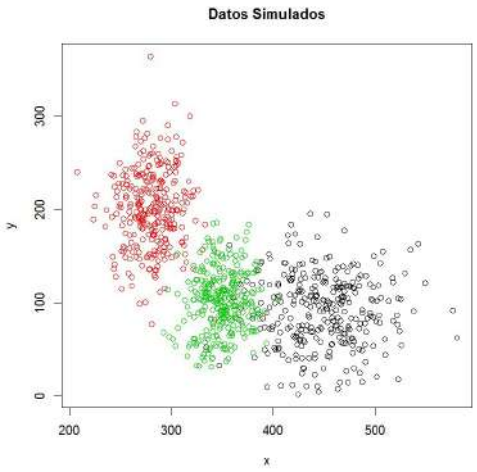
Dónde:
N es el número de clases
n es el número de individuos
a_{ii} es el número de casos que pertenecen a la clase *i* y que el clasificador le asignó (correctamente) como perteneciente a la clase *i*.

9. APLICACIÓN

Como un ejemplo de aplicación, tenemos un conjunto de datos bivariantes (Figura 10), divididos en tres categorías o grupos:

Mediante simulación generamos un data frame denominado *datos* con 900 observaciones divididas en tres grupos de 300 observaciones.

Figura 10.
Distribución en categorías de un conjunto de datos bivariantes



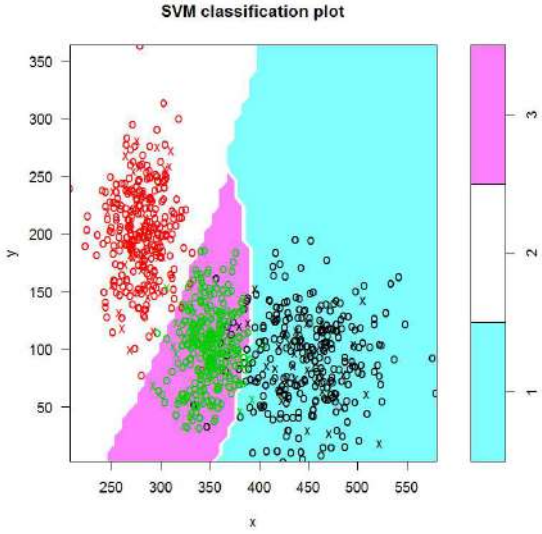
Creamos una muestra de 200 observaciones para entrenar al clasificador:

```
elec=sample(1:nrow(datos),200)
muestra=datos[elec,]
train=datos[-elec,]
```

Utilizamos la librería e1071 y la función MSV o SVM(inglés) con el núcleo radial, para estimar los parámetros del modelo

```
library(e1071);
modelo=svm(grupo~y+x,data=train,method="C-classification",
kernel="radial",cost=11,gamma=.12)
```

Figura 11.
Clasificación de datos por códigos de color



Examinamos el comportamiento del clasificador, para ello obtenemos la tabla de clasificación para de los datos contenidos en la muestra de entrenamiento

```
predic=data.frame(predict(modelo,muestra))
muestra=cbind(muestra,predic=predic)
table(muestra$grupo,muestra$predict.modelo..
muestra.)
```

	1	**2**	**3**
1	66	0	1
2	0	54	0
3	0	1	78

De donde:

$$Exactitud = \frac{66+54+78}{200} = \frac{198}{200} = 0.99$$

Lo que indica que es un excelente clasificador.

10. CONCLUSIONES

El componente principal de las MSV es el núcleo, y su existencia está garantizada, bajo ciertas condiciones, por el teorema de Mercer. La utilización de núcleos para mapear en espacios de mayor dimensión, es conocida como el truco del núcleo (kernel trick) y ha facilitado de la implementación de MSV en diferentes procedimientos y mecanismos gracias a su rapidez de cálculo. Este truco, permite la formulación de variantes no-lineales de cualquier algoritmo que pueda ser expresado en forma de productos internos en un espacio de Hilbert.

Como podemos apreciar en el ejemplo del presente trabajo, el clasificador obtenido al utilizar un núcleo radial es bastante bueno, pues su exactitud en la muestra de entrenamiento es del 99%, algo difícil de lograr con clasificadores lineales. Además, la muestra utilizad para estimar los coeficientes del clasificador es pequeña, 200, individuos. Estos resultados validan la preferencia de las MSV frente a los métodos clásicos del análisis discriminante.

REFERENCIAS BIBLIOGRÁFICAS

[1] Berlinet, A., & Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability Spaces*. New York: Springer US.

[2] Dalgaard, P. (2008). I*ntroductory Statistics with R*. Second Eddition. New York: Springer.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Second Edition. New York: Springer.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with applications in R*. New York: Springer.

[5] Lantz, B. (2013). *Machine Laerning with R*. London: Packt Publishing.

[6] Scholkopf, B., & Muandet, K. (2013). *One-Class Support Measure Machines for Group Anomaly Detection*. arXiv:1303, 1-10.

[7] Scholkopf, B., & Smola, A. (2001). *Learning with Kernels. Support Vector Machines, Regularization, Optimazation and Beyond*. Cambridge MA: MIT Press.