MÁQUINAS DE SOPORTE VECTORIAL: METODOLOGÍA Y APLICACIÓN EN R

VARGAS, JOSÉ[1], CONDE, BELÉN[2], PACCAPELO, VALERIA[3], ZINGARETTI, LAURA[4]

*1 UNVM-UNC (*donjmvargas@gmail.com*)*

*2 EEA Marcos Juárez del INTA (*bmabconde12@gmail.com*)*

*3 Monsanto Argentina (*mpacc@monsanto.com*)*

*4 UNVM (*m.lau.zingaretti@gmail.com*)*

RESUMEN

Máquinas de Soporte Vectorial (SVM de la expresión en inglés *Support Vector Machines*) es un método de *clasificación supervisada* cuyo objetivo es determinar la frontera óptima entre dos grupos, pudiendo extenderse a un número mayor. En el caso más simple, caso linealmente separable, existe una distancia positiva entre ambos grupos. Así, es posible determinar el hiperplano que maximiza la distancia de cada grupo a éste. En el caso más complejo, caso no separable linealmente, los grupos se superponen espacialmente siendo imposible la separación por medio de un hiperplano. SVM apela a una transformación no lineal del espacio en otro de mucha mayor dimensión en el cual, por efecto del aumento de dimensión y la no linealidad de la transformación, los dos grupos así transformados pueden ser separados como en el primer caso por un hiperplano. La transformación inversa deforma dicho hiperplano en una frontera no lineal que separa a los dos grupos en el espacio original. La transformación que separa perfectamente ambos grupos no sólo contiene la información que distingue a los grupos, sino que también contiene el ruido de los datos. Una manera de lograr que la frontera separe sólo por información relevante, es tolerar que algunos puntos queden mal asignados a un grupo. Para ello se introducen *variables de holgura* que cuantifican la distancia de un punto mal clasificado a la frontera de separación. Desde una perspectiva matemática, ambos problemas se formulan como problemas de programación cuadrática y sólo se necesita conocer el producto interno de los puntos transformados que, como función de los datos originales, se denomina *función kernel*. En este trabajo se presentan el método de SVM para separar dos grupos y se aplica la metodología en el entorno del software R para los dos casos mencionados.

Palabras claves: clasificación supervisada, máquinas de soporte vectorial (SVM), variables de holgura, función kernel

1. INTRODUCCIÓN

Máquinas de Soporte Vectorial (SVM de la expresión en inglés Support Vector Machines) es un método de clasificación supervisada. SVM emplea un algoritmo de optimización para determinar la frontera óptima entre dos grupos; se puede generalizar para múltiples grupos. El caso más simple es el linealmente separable, donde existe una distancia positiva entre ambos grupos y es posible elegir un hiperplano de separación maximizando la distancia de éste a cada grupo. Para decidir a qué grupo pertenece una nueva observación, se considera el signo de la función que define al hiperplano de separación.

En el caso más complejo (no separable linealmente), los grupos se superponen espacialmente haciendo imposible la separación por medio de un hiperplano. Una alternativa es cambiar la frontera lineal del hiperplano por una no lineal. SVM lo hace de una manera indirecta que resulta computacionalmente eficiente: se apela a una transformación no lineal del espacio en otro de mucha mayor dimensión en el cual los dos grupos transformados quedan separados como en el primer caso, es decir por una distancia positiva. Luego, es posible proceder como el caso linealmente separable. La transformación inversa deforma dicho hiperplano en una frontera no lineal que separa a los dos grupos en el espacio original.

Existe hasta aquí un grave problema: se produce un sobreajuste (el análogo de sobreajuste en regresión será tomar un polinomio interpolatorio que perfectamente ajuste los puntos). En efecto, la transformación que separa perfectamente ambos grupos no sólo contiene la información que distingue a un grupo del otro, sino que contiene todo el ruido como si fuese información relevante. Una manera de lograr que la frontera separe sólo por información relevante, ignorando el ruido, es tolerar que algunos puntos queden mal asignados a un grupo. Para ello se introducen variables de holgura que cuantifican la distancia de un punto que se encuentra del lado equivocado de la frontera de separación. De este modo, ambos problemas se formulan como problemas de programación cuadrática y sólo se necesita conocer el producto interno de los puntos transformados que, como función de los datos originales, se denomina función kernel.

2. METODOLOGÍA

a. CASO LINEALMENTE SEPARABLE

Sea $(x_1, y_1), \ldots, (x_n, y_n)$ una muestra de dos grupos separables por un hiperplano (linealmente separables), donde x_i corresponde a las coordenadas espaciales de cada punto y y_i asume los

valores $\{1, -1\}$ indicando la pertenencia a uno u otro grupo. La ecuación del hiperplano que separa ambos grupos tiene la forma:

$$\phi(x) = \beta' x + b_0$$

donde β es un vector normal al hiperplano y, por conveniencia en las interpretaciones geométricas, se lo supone de norma uno. Los coeficientes β y b_0 se eligen para maximizar el margen M que mide la distancia del hiperplano al punto más cercano de cada grupo. Este planteo corresponde a un problema de programación cuadrática, pues se tiene una función objetivo cuadrática con restricciones lineales:

$$\underset{\beta, b_0}{máx} M$$

sujeto a:

$$\| \beta \| = 1$$
$$y_i(\beta' x_i + b_0) \geq M \; para \; i = 1, \ldots, \text{n}$$

Aquí, el sentido del vector β indica el grupo rotulado por $y_i = 1$ y la expresión $y_i(\beta' x_i + b_0)$ mide la distancia del i-ésimo punto al plano separador; el factor y_i compensa con un menos cuando el punto x_i se encuentra del lado opuesto al que apunta β. El problema de arriba puede ser reformulado de una manera más conveniente, como sigue:

$$\underset{\beta, b_0}{mín} \| \beta \|$$

sujeto a:

$$y_i(\beta' x_i + b_0) \geq 1 \; para \; i = 1, \ldots, \text{n}$$

de este modo b_0 puede escribirse como $-\beta' x_0$ para cualquier punto x_0 que se encuentre en el hiperplano $\beta' x_i + b_0 = 0$. Así, las restricciones en el primer planteo pueden escribirse como: $\frac{\beta'}{M}(x_i - x_0) \geq 1$ y maximizar M resulta equivalente a minimizar $\frac{1}{M} = \left\| \frac{\beta}{M} \right\|$ con $\| \beta \| = 1$. Llamando β_1 a $\frac{\beta}{M}$, el problema es minimizar β_1 sujeto a las restricciones $\beta'_1(x_i - x_0) \geq 1$ que es el segundo planteo.

b. CASO NO SEPARABLE LINEALMENTE

Sea $(x_1, y_1), \ldots, (x_n, y_n)$ una muestra de dos grupos superpuestos espacialmente, es decir, que no son separables por un hiperplano a menos que se introduzcan variables de holgura que permitan a algunos puntos permanecer en el grupo equivocado.

Cada variable de holgura, una por punto, mide la distancia de un punto al margen (a distancia M del hiperplano separador) que corresponde al grupo al que pertenece el punto. Si el punto se encuentra del lado correcto, su variable holgura es cero, pero si sobrepasa el margen a distancia M del hiperplano separador, su variable de holgura es positiva.

El problema consiste en encontrar un hiperplano que maximizando su distancia al margen (*M*), se minimice al mismo tiempo la totalidad de las holguras. De manera equivalente, el problema consiste en minimizar la inversa de *M* más la totalidad de las holguras penalizadas por una constante *C* que torna el procedimiento rígido a medida que se aumenta el valor de *C*. Con el objetivo de que el problema resultante sea convexo, se utiliza una holgura relativa a *M*:

$$\min_{M,\beta,b_0} \frac{1}{2}\frac{\| \beta \|^2}{M} + C\sum_{i=1}^{n}\xi_i$$

sujeto a:

$$\| \beta \| = 1$$
$$\zeta_i = 1 \; para \; i = 1,\ldots,\mathrm{n}$$
$$y_i(\beta' x_i + b_0) \geq M(1-\zeta_i) \; para \; i = 1,\ldots,\mathrm{n}$$

El problema se formula en términos de multiplicadores de Lagrange y con el fin de hacerlo más eficiente numéricamente, se computa el problema dual equivalente, obteniéndose la misma solución que en el problema primal. El Lagrangiano Dual a computar es:

$$L_{Dual} = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle \Phi(x_i), \Phi(x_j)\rangle$$

Esta función establece una cota inferior de la función objetivo primal en cada punto factible del problema primal y se maximiza sujeto a las restricciones:

$$0 \leq \alpha_i \leq C$$
$$\sum \alpha_i y_i = 0$$

La idea detrás de ajustar una frontera de clasificación no lineal que separe mejor los puntos de dos grupos espacialmente dispersos es usar una transformación biyectiva, no lineal Φ, del espacio $\mathbb{R}^n$ en otro $\mathbb{R}^m$ con $m>n$, que transforme los puntos $x_i \; en \; \Phi(x_i)$ en donde, debido a la curvatura que introduce y al aumento de dimensión, ambos grupos sean linealmente separables. No es necesario conocer la función Φ sino el producto interno $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$ que se denomina función kernel; las más utilizadas en SVM son lineal, polinomial, Gausseano de base radial (RBF) y Laplace de base radial.

3. RESULTADOS

La primera implementación de SVM en R (R Development Core Team) se introdujo en la librería *e1071* (Dimitriadou, et al. 2005) a través de la función *svm*. Además, la librería *kernlab* (Karatzoglou, et al. 2010) ofrece una variedad de métodos basados en funciones de kernel por medio de la función *ksvm*. Ambas funciones de R se aplican a dos ejemplos. El primero de ellos se trata de dos grupos no separables linealmente mientras que el segundo presenta, a modo ilustrativo, la generalización de SVM al un caso de tres grupos.

a. Caso linealmente no separable: dos poblaciones normales bivariadas

Para ejemplificar la aplicación de SVM en R, se simulan dos poblaciones normales bivariadas no separables linealmente. Una de las poblaciones consta de dos variables independientes con media 0 y varianza 1. La otra población consta de dos variables independientes con media 3 y varianza 3. La librería *e1071*, por medio de la función tune.svm permite seleccionar los valores óptimos de los parámetros de la función kernel elegida y del costo. Para estos valores se obtiene una tasa de error de entrenamiento es de 8.67%. En cuanto a la predicción de la población a la cual pertenece cada una de las observaciones en la muestra de validación, el error de validación para del modelo ajustado es de 7%. Utilizando la librería *kernlab*, se obtienen resultados similares (error de validación del 8%).

b. Extensión a más de dos grupos: datos Iris

Se consideran los datos de Fisher *iris* que consisten de cuatro mediciones realizadas a flores: largo de sépalo (sepal length), ancho de sépalo (sepal width), largo de pétalo (petal length), ancho de pétalo (petal width) y una variable clasificadora (Species) que corresponde a la especie de iris: setosa, versicolor, o virginica. Las primeras cuatro variables numéricas constituyen el vector de variables de entrada x y la variable categórica Species es la correspondiente respuesta y. Hay 150 flores en total, de los cuales 50 corresponden a la especie setosa, 50 a la especie versicolor y 50 a virginica. Aplicando la metodología de SVM, solo 3 individuos resultaron clasificados erróneamente.

4. REFERENCIAS

• **Dimitriadou, Hornik, Leisch, Meyer, Weingessel (2011)** e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.

• **Hastie, Tibshirani, Friedman (2009)** The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics.

• **Huang, Davis, TownShend (2002)** An assessment to support vector machines for land cover classification. Remote Sensing Vol. 23, No. 4.

• **Karatzoglou, Smola, Hornik, Zeileis (2004)** kernlab: An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20.

• **R Development Core Team (2010)** R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.