Estadística y computación para metagenómica

Victor Muñiz Sánchez
victor_m@cimat.mx

Centro de Investigación en Matemáticas.
Unidad Monterrey.

Máquinas de soporte vectorial

Máquinas de Soporte Vectorial

Los dos artículos fundamentales.

A Training Algorithm for Optimal Margin Classifiers

Bernhard E. Boser*
EECS Department
University of California
Berkeley, CA 94720
boser@eecs.berkeley.edu

Isabelle M. Guyon
AT&T Bell Laboratories
50 Fremont Street, 6th Floor
San Francisco, CA 94105
isabelle@neural.att.com

Vladimir N. Vapnik
AT&T Bell Laboratories
Crawford Corner Road
Holmdel, NJ 07733
vlad@neural.att.com

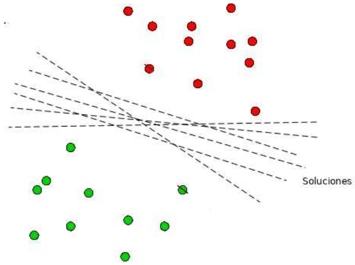
Machine Learning, 20, 273–297 (1995)

Support-Vector Networks

CORINNA CORTES corinna@neural.att.com
VLADIMIR VAPNIK vlad@neural.att.com
AT&T Bell Labs., Holmdel, NJ 07733, USA

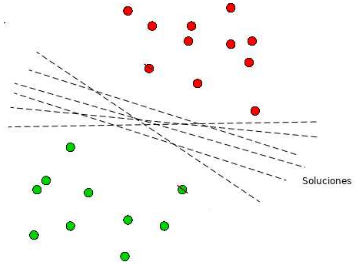
Máquinas de Soporte Vectorial

Recuerda el problema que teníamos anteriormente con clasificadores lineales



Máquinas de Soporte Vectorial

Recuerda el problema que teníamos anteriormente con clasificadores lineales



¿Cómo tener una solución única (óptima)?
Debemos definir un criterio.
Para SVM, éste criterio está dado por la solución óptima que involucra la distancia entre el hiperplano obtenido y un **subconjunto** de los datos de entrenamiento, llamados **vectores soporte**.

Máquinas de Soporte Vectorial

Veamos el caso más sencillo: dos categorías separables linealmente.

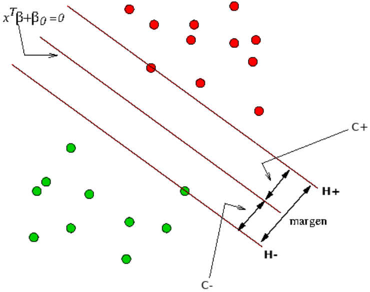
Sean $\{\mathbf{x}_i, y_i\}_{i=1}^n$, con $\mathbf{x}_i \in \mathbb{R}^d$ y $y \in \{-1, 1\}$, nuestros datos de entrenamiento.
Define un hiperplano separador

$$f(\mathbf{x}) = \beta_0 + \mathbf{x}'\boldsymbol{\beta} = 0.$$

Sea C_+ la distancia más corta de $f(\mathbf{x})$ al punto con clase 1 $(\mathbf{x}_+)$ más cercano.
Sea C_- la distancia más corta de $f(\mathbf{x})$ al punto con clase -1 $(\mathbf{x}_-)$ más cercano.
Definimos el **márgen** como

$$C_+ + C_-$$

Máquinas de Soporte Vectorial



Observa que:

$$\begin{aligned} \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} &\geq 1 \quad \text{si } y_i = 1 \\ \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} &\leq -1 \quad \text{si } y_i = -1. \end{aligned}$$

Por lo tanto,

$$\beta_0 + \mathbf{x}_+'\boldsymbol{\beta} = 1 \tag{1}$$

$$\beta_0 + \mathbf{x}_-'\boldsymbol{\beta} = -1 \tag{2}$$

para los puntos $\mathbf{x}_+$ y $\mathbf{x}_-$ que están en los hiperplanos H_+ y H_-, respectívamente.

Máquinas de Soporte Vectorial

La diferencia es

$$\mathbf{x}'_+\boldsymbol{\beta} - \mathbf{x}'_-\boldsymbol{\beta} = 2,$$

y la suma

$$\beta_0 = -\frac{1}{2}(\mathbf{x}'_+\boldsymbol{\beta} + \mathbf{x}'_-\boldsymbol{\beta}).$$

También:

$$\begin{aligned} C_+ &= \frac{|\beta_0+\mathbf{x}'_+\boldsymbol{\beta}|}{\|\boldsymbol{\beta}\|} &= \frac{1}{\|\boldsymbol{\beta}\|} \\ C_- &= \frac{|\beta_0+\mathbf{x}'_-\boldsymbol{\beta}|}{\|\boldsymbol{\beta}\|} &= \frac{1}{\|\boldsymbol{\beta}\|}. \end{aligned}$$

Por lo tanto, el márgen es

$$M = \frac{2}{\|\boldsymbol{\beta}\|}.$$

Máquinas de Soporte Vectorial

Ahora, observa que, según (1) y (2), decimos que $\mathbf{x}_i$ **es un vector de soporte** si

$$y_i(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) = 1,$$

entonces, lo que queremos es:

$$\max_{\beta_0, \boldsymbol{\beta}, \|\boldsymbol{\beta}\|=1} M,$$

sujeto a

$$y_i(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) \geq M, \quad i = 1, 2, \ldots, n.$$

La restricción en $\boldsymbol{\beta}$ es para que no crezca arbitrariamente.

Máquinas de Soporte Vectorial

Ahora, observa que, según (1) y (2), decimos que $\mathbf{x}_i$ **es un vector de soporte** si

$$y_i(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) = 1,$$

entonces, lo que queremos es:

$$\max_{\beta_0, \boldsymbol{\beta}, \|\boldsymbol{\beta}\|=1} M,$$

sujeto a

$$y_i(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) \geq M, \quad i = 1, 2, \ldots, n.$$

La restricción en $\boldsymbol{\beta}$ es para que no crezca arbitrariamente. Otra forma de considerar ésta restricción es modificando las condiciones:

$$\frac{1}{\|\boldsymbol{\beta}\|} y_i(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) \geq M$$

Máquinas de Soporte Vectorial

O también:

$$y_i(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) \geq \|\boldsymbol{\beta}\| M,$$

y ésta desigualdad se cumple para cualquier escalamiento positivo de $\boldsymbol{\beta}$.
Podemos definir arbitrariamente

$$\|\boldsymbol{\beta}\| = \frac{1}{M},$$

que es un valor positivo.

Máquinas de Soporte Vectorial

O también:

$$y_i(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) \geq \|\boldsymbol{\beta}\| M,$$

y ésta desigualdad se cumple para cualquier escalamiento positivo de $\boldsymbol{\beta}$.
Podemos definir arbitrariamente

$$\|\boldsymbol{\beta}\| = \frac{1}{M},$$

que es un valor positivo.
Observa que, si M es grande, $\|\boldsymbol{\beta}\|$ será pequeño.

Máquinas de Soporte Vectorial

Entonces, podemos redefinir el problema de optimización como

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{\beta}\|^2$$

s.a.

$$y_i(\beta_0 + \mathbf{x}_i' \boldsymbol{\beta}) \geq 1, \quad i = 1, 2, \ldots, n.$$

Máquinas de Soporte Vectorial

Entonces, podemos redefinir el problema de optimización como

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{\beta}\|^2$$

s.a.

$$y_i(\beta_0 + \mathbf{x}_i' \boldsymbol{\beta}) \geq 1, \quad i = 1, 2, \ldots, n.$$

Tenemos entonces un problema de optimización muy agradable:

- cuadrático, por lo tanto es convexo y solución única
- restricciones lineales
- hay métodos muy eficientes para resolverlo.

Máquinas de Soporte Vectorial

Generalmente, se resuelve el problema de optimización dual:

$$\text{máx}\, \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{M}\boldsymbol{\alpha},$$

s.a.

$$\boldsymbol{\alpha}'\mathbf{y} = 0, \qquad \boldsymbol{\alpha} \geq 0,$$

donde $\boldsymbol{\alpha}$ son los multiplicadores de Lagrange del problema primal (variables del problema dual) $\mathbf{M}$ una matriz cuadrada, simétrica, semidefinida positiva y entradas $M_{ij} = y_i y_j (\mathbf{x}_i' \mathbf{x}_j)$.

Máquinas de Soporte Vectorial

En el óptimo,

$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i,$$

pero, por las condiciones KKT de complementariedad del problema de optimización:

$$\boldsymbol{\beta}^* = \sum_{i \in SV} \alpha_i^* y_i \mathbf{x}_i,$$

donde SV es el conjunto de vectores soporte, es decir, aquellos donde

$$\alpha_i^* > 0$$

Generalmente, usamos un promedio de los α_i^* que son SV.

Máquinas de Soporte Vectorial

También, obtenemos

$$\beta_0^* = -\frac{1}{2}\left(\mathbf{x}_+'\boldsymbol{\beta}^* + \mathbf{x}_-'\boldsymbol{\beta}^*\right),$$

con $\mathbf{x}_+$ y $\mathbf{x}_-$ cualesquiera vectores de soporte.

Máquinas de Soporte Vectorial

También, obtenemos

$$\beta_0^* = -\frac{1}{2}\left(\mathbf{x}_+'\boldsymbol{\beta}^* + \mathbf{x}_-'\boldsymbol{\beta}^*\right),$$

con $\mathbf{x}_+$ y $\mathbf{x}_-$ cualesquiera vectores de soporte.

Finalmente, nuestro clasificador queda entonces como

$$f^*(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}^* + \beta_0^*,$$

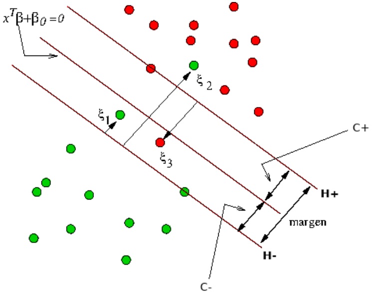
y la función de decisión será:

$$G(\mathbf{x}) = \text{signo}(f^*(\mathbf{x})).$$

Máquinas de Soporte Vectorial

Caso no separable (soft margin).

Aquí, permitimos que cada dato pueda aparecer en el lado "equivocado" del hiperplano separador.



Máquinas de Soporte Vectorial

Definimos las variables de holgura

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)' \geq 0.$$

El problema de optimización es

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{\beta}\|^2$$

s.a.

$$\begin{aligned} y_i(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) + \xi_i &\geq 1 \\ \xi_i &\geq 0 \\ \sum \xi_i &\leq \text{constante}, \end{aligned}$$

para $i = 1, 2, \ldots, n$.

Máquinas de Soporte Vectorial

El "soft margin" lo formulamos como un problema de regularización sobre los pesos $\boldsymbol{\xi}$ del margen:$\lambda\|\boldsymbol{\xi}\|_L$, $\lambda \geq 0$.

Máquinas de Soporte Vectorial

El "soft margin" lo formulamos como un problema de regularización sobre los pesos $\boldsymbol{\xi}$ del margen:$\lambda\|\boldsymbol{\xi}\|_L$, $\lambda \geq 0$. Generalmente usamos la norma L_1 para "activar" o permitir que solo algunos datos estén del lado incorrecto de $f(\mathbf{x})$. Entonces, el problema de optimización queda:

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^{n} \xi_i$$

s.a.

$$\begin{array}{rcl} y_i(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) & \geq & 1 - \xi_i \\ \xi_i & \geq & 0, \end{array}$$

para $i = 1, 2, \ldots, n$.

Máquinas de Soporte Vectorial

Como antes, podemos definir un problema de optimización dual que es muy parecido al caso de SVM separable:

$$\max \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{M}\boldsymbol{\alpha}, \tag{3}$$

s.a.

$$\boldsymbol{\alpha}'\mathbf{y} = 0, \qquad 0 \leq \boldsymbol{\alpha} \leq \lambda\mathbf{1},$$

con $\mathbf{M}$ una matriz cuadrada, simétrica, semidefinida positiva y entradas $M_{ij} = y_i y_j (\mathbf{x}_i'\mathbf{x}_j)$.

Máquinas de Soporte Vectorial

En el óptimo,

$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i,$$

y como en el caso anterior, por las condiciones de complementariedad:

$$\boldsymbol{\beta}^* = \sum_{i \in SV} \alpha_i^* y_i \mathbf{x}_i,$$

donde SV es el conjunto de vectores soporte, es decir, aquellos donde

$$0 < \alpha_i^* \quad \text{y} \quad \xi_i = 0.$$

También se usa un promedio de los α_i^* que son SV.

Máquinas de Soporte Vectorial

También, obtenemos

$$\beta_0^* = -\frac{1}{2}\left(\mathbf{x}'_+\boldsymbol{\beta}^* + \mathbf{x}'_-\boldsymbol{\beta}^*\right),$$

con $\mathbf{x}_+$ y $\mathbf{x}_-$ cualesquiera vectores de soporte.

Máquinas de Soporte Vectorial

También, obtenemos

$$\beta_0^* = -\frac{1}{2}\left(\mathbf{x}_+'\boldsymbol{\beta}^* + \mathbf{x}_-'\boldsymbol{\beta}^*\right),$$

con $\mathbf{x}_+$ y $\mathbf{x}_-$ cualesquiera vectores de soporte.

Finalmente, nuestro clasificador queda entonces como

$$f^*(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}^* + \beta_0^*,$$

y la función de decisión será:

$$G(\mathbf{x}) = \text{signo}(f^*(\mathbf{x})).$$

Máquinas de Soporte Vectorial

SVM no lineal.

Considera nuevamente el problema de optimización para el caso general (3).
Para generar fronteras de clasificación no-lineales, se utiliza el truco del kernel, considerando que la formulación de la solución de SVM está dada en términos de productos punto.
En este caso:

$$\begin{aligned} M_{ij} &= y_i y_j (\mathbf{x}_i' \mathbf{x}_j) \\ &= y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \\ &= y_i y_j k(\mathbf{x}_i' \mathbf{x}_j), \end{aligned}$$

y la solución está dada por:

$$\begin{aligned} f^*(\mathbf{x}) &= \boldsymbol{\beta}^{*T} \mathbf{x} + \beta_0^* \\ &= \sum_{i \in SV} \alpha_i^* y_i k(\mathbf{x}, \mathbf{x}_i) + \beta_0^*, \end{aligned}$$

con $k(\cdot, \cdot)$, un kernel válido.

Máquinas de Soporte Vectorial

SVM multiclase.

Generalmente hay dos opciones:

- One vs All.
  Resuelve K problemas binarios de clasificación, y toma $\hat{y} = f_k(\mathbf{x})$ con el valor más grande positivo, donde $f_k(\mathbf{x})$ es la solución óptima para el problema de clasificación binaria de la clase k contra el resto.
- One vs One.
  Construye $\binom{K}{2}$ clasificadores binarios, y $f_k(\mathbf{x})$ es la solución k que recibe más votaciones.

Máquinas de Soporte Vectorial

`SVM.ipynb`