Máquinas de vectores de soporte

Hernán Felipe García M.Sc.

UTP

Contenido

Red RBF

- Recordemos que la salida de una red de base radial se modela como

$$y(\mathbf{x}) = \sum_{n=1}^{N} w_n h(\|\mathbf{x} - \mathbf{x}_n\|) + w_0 = \sum_{n=1}^{N} w_n \phi_n(\mathbf{x}) + w_0,$$

donde $h(\cdot)$ son funciones de base radial.

Regularización

- Para obtener los coeficientes $\mathbf{w} = [w_0, w_1, \ldots, w_N]^\top$ se pueden minimizar diferentes funciones

$$J(\mathbf{w}) = L(\mathbf{t}, \mathbf{y}),$$

Indice de desempeño

$$J_2(\mathbf{w}) = L(\mathbf{t}, \mathbf{y}) + \frac{\lambda}{2} \sum_{j=0}^{N} w_j^2$$

Norma L2

$$J_1(\mathbf{w}) = L(\mathbf{t}, \mathbf{y}) + \frac{\lambda}{2} \sum_{j=0}^{N} |w_j|,$$

LASSO, norma L1

donde $L(\mathbf{t}, \mathbf{y})$ es la función de pérdida.

- Para un problema de regresión, por ejemplo,

$$L(\mathbf{t}, \mathbf{y}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - y(\mathbf{x}_n)\}^2.$$

Kernel machines

- El tipo de regularización en J_2 se conoce como *regularización ridge*, y el tipo de regularización en J_1 se conoce como *regularización lasso*.

- Si en el modelo $y(\mathbf{x})$ de la red RBF la función $h(\mathbf{x}, \mathbf{x}')$ se reemplaza por un función *kernel* $k(\mathbf{x}, \mathbf{x}')$, el modelo para $y(\mathbf{x})$ se conoce como una máquina basada en kernel (*kernel machine*),

$$y(\mathbf{x}) = \sum_{n=1}^{N} w_n k(\mathbf{x}, \mathbf{x}_n) + w_0,$$

donde b es la tendencia.

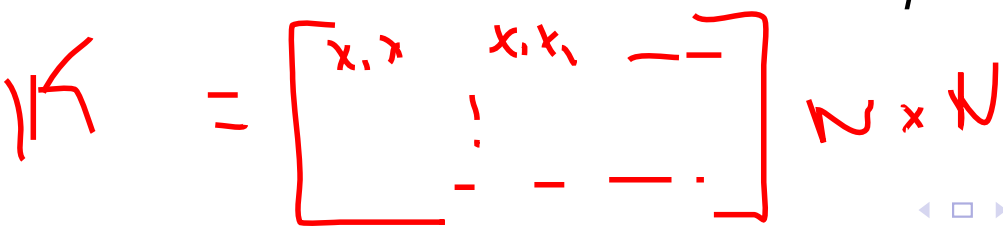
Kernel (fcn escalar)

- Un kernel $k(\mathbf{x}, \mathbf{x}')$ es una función real de dos argumentos.

- Típicamente la función es simétrica, es decir $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$, y no negativa, es decir $k(\mathbf{x}, \mathbf{x}') \geq 0$, así puede interpretarse como una medida de similaridad, aunque no se requiere.

Sparse kernel machines

- Si la regularización que se utiliza induce sparsity, la máquina basada en kernels se conoce como *sparse kernel machine*.
- Otra forma de introducir sparsity es usar regularización de norma dos, pero modificar la función de pérdida.
- Esta última forma conduce a las *máquinas de vectores de soporte*.
- Lo interesante es que al ser sparse, la predicción termina dependiendo sólo de cierto datos.
- Estos datos se conocen como *vectores de soporte*.



Contenido

Funciones kernel

- Un kernel $k(\mathbf{x}, \mathbf{x}')$ es una función real de dos argumentos.

- Típicamente la función es simétrica, es decir $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$, y no negativa, es decir $k(\mathbf{x}, \mathbf{x}') \geq 0$, así puede interpretarse como una medida de similaridad, aunque no se requiere.

Kernel exponencial cuadrático

- El kernel exponencial cuadrático (SE - squared exponential) o kernel Gaussiano está definido como

$$k(\mathbf{x}, \mathbf{x}') = \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top \Sigma^{-1}(\mathbf{x} - \mathbf{x}')\right\}.$$

- Si Σ es diagonal, se puede escribir como

$$k(\mathbf{x}, \mathbf{x}') = \exp\left\{-\frac{1}{2}\sum_{j=1}^{D}\frac{(x_j - x_j')^2}{\sigma_j^2}\right\},$$

donde σ_j se define como la *longitud de escala característica* de la dimensión j.

- Si Σ es esférica, se obtiene el kernel isotrópico

$$k(\mathbf{x}, \mathbf{x}') = \exp\left\{-\frac{1}{2}\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2}\right\},$$

donde σ^2 se conoce como el *ancho de banda*.

Kernel Mercer (I)

- Algunos métodos de aprendizaje de máquina requieren que la función kernel satisfaga el requerimiento de que la matriz Gramo (Gram matrix) definida como

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \vdots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

  sea positiva definida para cualquier conjunto de entradas $\{\mathbf{x}_i\}_{i=1}^{N}$.

- A tal kernel se le conoce como *kernel Mercer* o *kernel positivo definido*.

Kernel Mercer (II)

- En general, si el kernel es Mercer, luego existe una función ϕ que mapea $\mathbf{x} \in \mathbb{R}^D$ tal que

$$k(\mathbf{x}, \mathbf{x}') = \phi^\top(\mathbf{x})\phi(\mathbf{x}'),$$

donde ϕ depende de las funciones propias de $k(\cdot, \cdot)$.

Producto punto entre funciones base

$\phi : \mathbb{R}^D \mapsto \mathbb{R}^M$

$k(x, x') : \mathbb{R}^D \mapsto \mathbb{R}$

Tipos de kernel

En la literatura se han definido kernels sobre objetos estructurados, por ejemplo

- texto (string kernels).
- imágenes (Pyramid match kernels).
- gráficos (graph kernels).
- modelos probabilísticos (Fisher kernels).

Ver el libro de Cristianini y Shawe-Taylor (2004): *Kernel Methods for Pattern Analysis*.

Contenido

Función objetivo

- En las máquinas de vectores de soporte para regresión, la función objetivo se escribe como

$$J(\mathbf{w}) = C \sum_{n=1}^{N} L_\epsilon(t_n, y_n) + \frac{1}{2}\|\mathbf{w}\|_2^2,$$

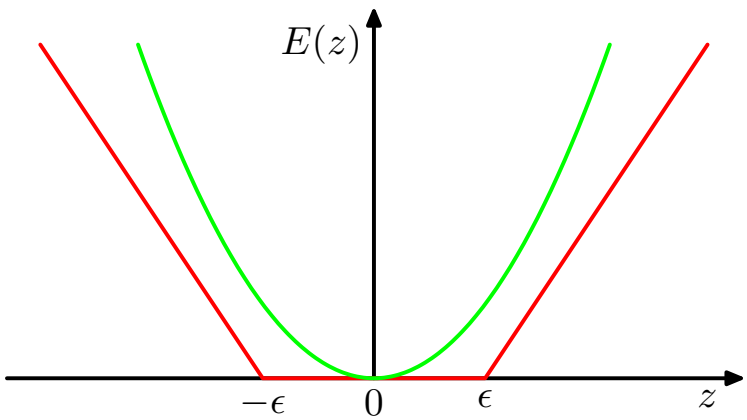
donde $y_n = y(\mathbf{x}_n) = \mathbf{w}^\top \phi_n + w_0$, y C es una constante de regularización.

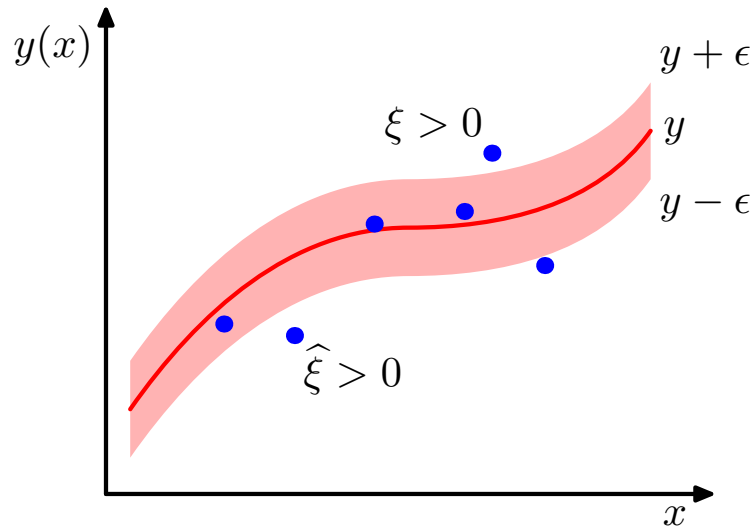
- La función $L_\epsilon(t, y)$ se conoce como la función de pérdida insensible épsilon (epsilon insensitive loss function, ϵ-insensitive loss function).

ϵ-insensitive loss function

La función L_ϵ está dada por

$$L_\epsilon(t, y) = \begin{cases} 0, & \text{si } |t - y| < \epsilon \\ |t - y| - \epsilon, & \text{de otra forma.} \end{cases}$$





Variables slack (I)

- La función $L_\epsilon(t, y)$ no es diferenciable, por el valor aboluto.
- El problema se formula como un problema de optimización con restricciones.
- En particular, se introducen *variables slack* para representar el grado para el cual cada punto está por fuera del tubo.
- Para cada dato $\mathbf{x}_n$ se necesitan dos variables slack $\xi_n \geq 0$, y $\hat{\xi}_n \geq 0$.
- $\xi_n > 0$ corresponde un punto para el cual $t_n > y(\mathbf{x}_n) + \epsilon$, y $\hat{\xi}_n > 0$ corresponde a un punto para el cual $t_n < y(\mathbf{x}_n) - \epsilon$.

Variables slack (II)

- La condición para que un punto objetivo se encuentre dentro de un ϵ-tube es que

$$y(\mathbf{x}_n) - \epsilon \leq t_n \leq y(\mathbf{x}_n) + \epsilon.$$

- Al introducir las variables slack se permite que los puntos estén por fuera del tubo (si las variables slack son diferentes de cero).

- Las condiciones correspondientes son

$$t_n \leq y(\mathbf{x}_n) + \epsilon + \xi_n$$
$$t_n \geq y(\mathbf{x}_n) - \epsilon - \hat{\xi}_n.$$

Función de error modificada

- La función de error para la regresión con vectores de soporte puede escribirse como

$$C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|_2^2,$$

que debe minimizarse sujeta a las restricciones

$$\begin{aligned}
\xi_n &\geq 0 \\
\hat{\xi}_n &\geq 0 \\
t_n &\leq y(\mathbf{x}_n) + \epsilon + \xi_n \\
t_n &\geq y(\mathbf{x}_n) - \epsilon - \hat{\xi}_n.
\end{aligned}$$

Multiplicadores de Lagrange

- Lo anterior puede lograrse introduciendo *multiplicadores de Lagrange*

$$
\begin{aligned}
a_n &\geq 0 \\
\hat{a}_n &\geq 0 \\
\mu_n &\geq 0 \\
\hat{\mu}_n &\geq 0,
\end{aligned}
$$

y optimizando el Lagrangiano

$$
\begin{aligned}
L = C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|_2^2 - \sum_{n=1}^{N}(\mu_n\xi_n + \hat{\mu}_n\hat{\xi}_n) \\
- \sum_{n=1}^{N} a_n(\epsilon + \xi_n + y(\mathbf{x}_n) - t_n) - \sum_{n=1}^{N} \hat{a}_n(\epsilon + \hat{\xi}_n - y(\mathbf{x}_n) + t_n).
\end{aligned}
$$

Derivadas del Lagrangiano

Reemplazando $y(\mathbf{x}_n) = \mathbf{w}^\top \phi(\mathbf{x}_n) + w_0$ en la expresión anterior, e igualando a cero las derivadas del Lagrangiano con respecto a $\mathbf{w}$, w_0, ξ_n, y $\hat{\xi}_n$, se tiene

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} (a_n - \hat{a}_n)\phi(\mathbf{x}_n)$$

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow \sum_{n=1}^{N} (a_n - \hat{a}_n) = 0$$

$$\frac{\partial L}{\partial \xi_n} = 0 \Rightarrow a_n + \mu_n = C$$

$$\frac{\partial L}{\partial \hat{\xi}_n} = 0 \Rightarrow \hat{a}_n + \hat{\mu}_n = C$$

Problema dual

- Haciendo los reemplazos necesarios, el problema dual implica maximizar

$$\widetilde{L}(\mathbf{a}, \hat{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m)$$
$$- \epsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n,$$

sujeto a las siguientes restricciones

$$0 \leq a_n \leq C,$$
$$0 \leq \hat{a}_n \leq C,$$
$$\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0.$$

- En la expresión anterior se usó $k(\mathbf{x}_n, \mathbf{x}_m) = \phi^{\top}(\mathbf{x}_n)\phi(\mathbf{x}_m)$.

Predicción

- Las predicciones se pueden realizar mediante

$$y(\mathbf{x}) = \sum_{n=1}^{N} (a_n - \hat{a}_n) k(\mathbf{x}, \mathbf{x}_n) + w_0.$$

- Los vectores de soporte son aquellos datos observados que contribuyen a las predicciones dadas por la ecuación anterior, en otras palabras, aquellos para los cuales $a_n = 0$, ó $\hat{a}_n = 0$.

- Estos son puntos que están en los límites del ϵ-tube o por fuera del tubo.

- Todos los puntos dentro del tubo tienen $a_n = \hat{a}_n = 0$.

- Para la predicción sólo se evalúan los términos que involucran vectores de soporte.

Encontrando w_0

- El parámetro w_0 puede encontrarse considerando un dato observado para el cual $0 < a_n < C$.
- Se puede demostrar que para ese dato

$$w_0 = t_n - \epsilon - \sum_{n=1}^{N} (a_n - \hat{a}_n) k(\mathbf{x}, \mathbf{x}_n).$$

- Se puede obtener un resultado análogo considerando un punto para el cual $0 < \hat{a}_n < C$.
- En la práctica es mejor promediar sobre los estimadores de w_0.

Implementación

- Se puede emplear programación cuadrática.

- Otros algoritmos
  - Sequential Minimal Optimization (SMO)
  - Método de punto interior

Contenido

Contenido

Modelo linealmente separable

- Se asume que el conjunto es linealmente separable en el espacio de características $\phi(\mathbf{x})$.

- Esto significa que para el modelo de los datos

$$y(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + w_0,$$

existe por lo menos una opción de parámetros $\mathbf{w}$, y w_0 tal que

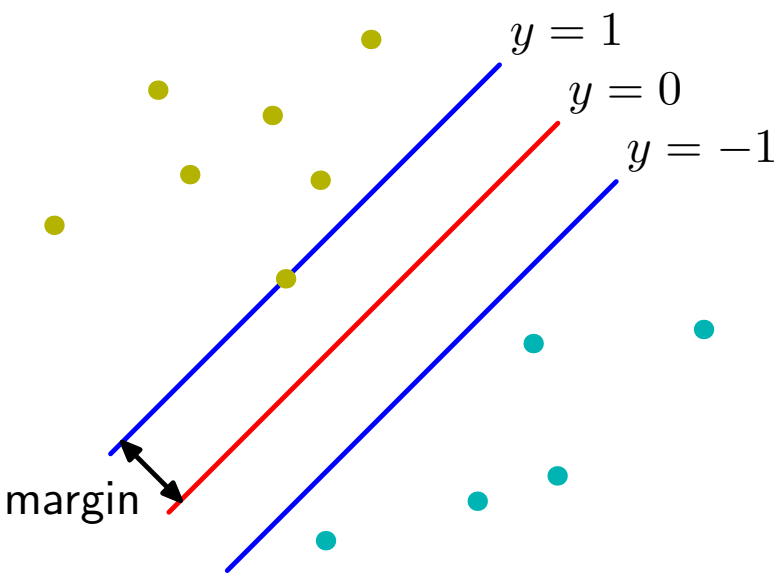
$$\begin{cases} y(\mathbf{x}_n) > 0 & \text{para puntos } t_n = +1 \\ y(\mathbf{x}_n) < 0 & \text{para puntos } t_n = -1, \end{cases}$$

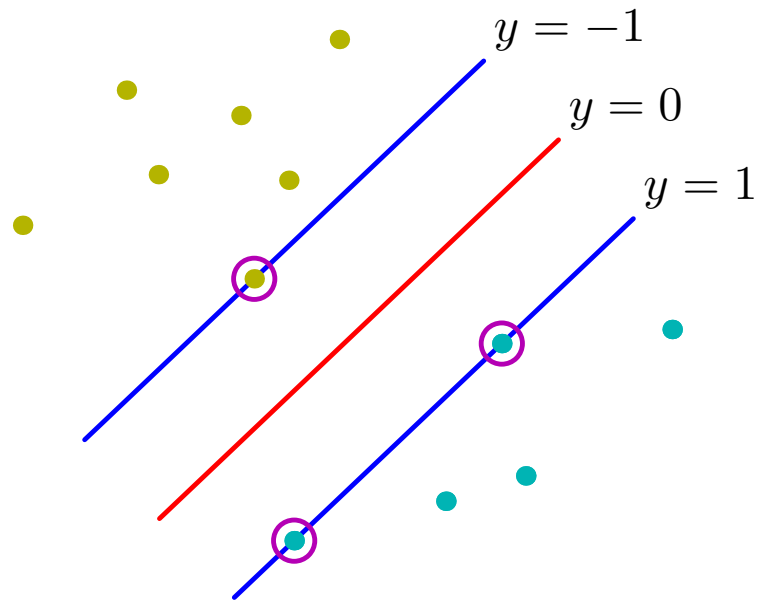
de forma tal que $t_n y(\mathbf{x}_n) > 0$ para todos los datos de entrenamiento.

Margen

- La SVM se acerca a este problema a través del concepto de *margen*.

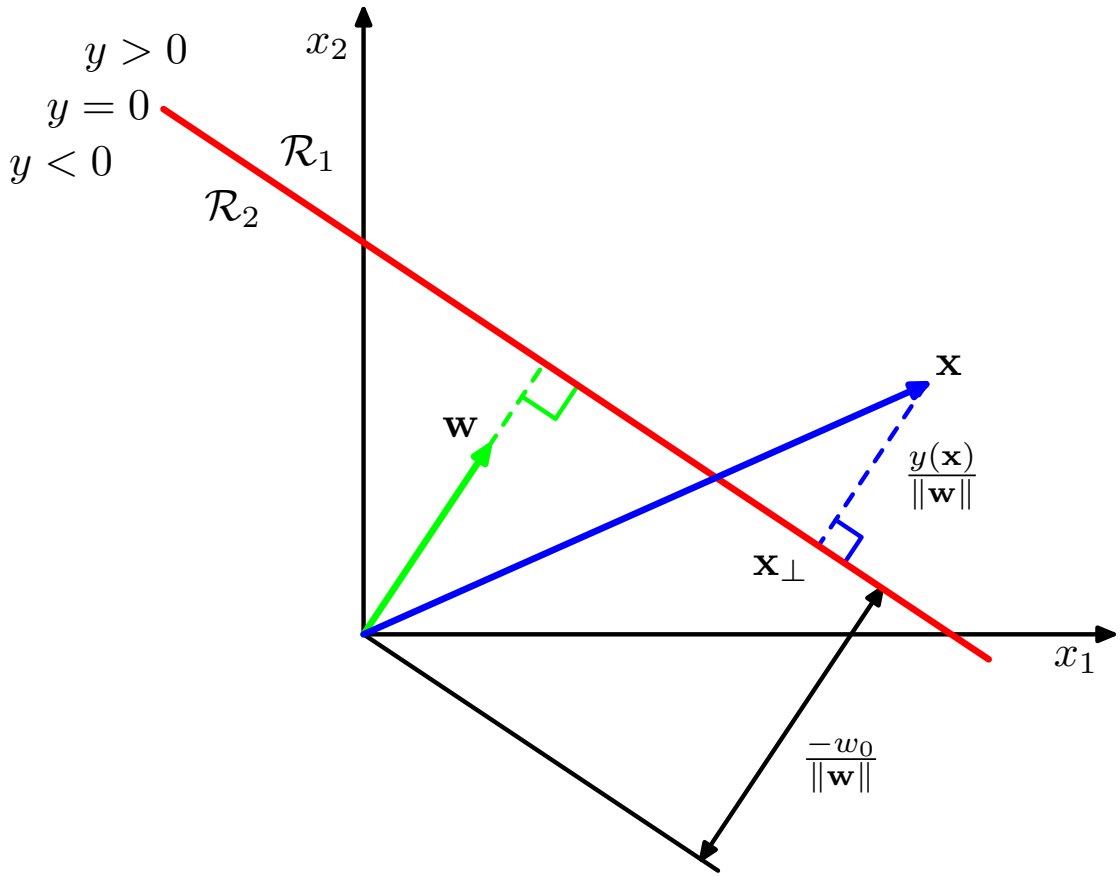
- El margen se define como la distancia más pequeña entre la frontera de decisión y cualquiera de las muestras.





Frontera de decisión

La frontera de decisión se escoge como aquella para la cual se maximiza el margen.



Problema de optimización

- La distancia perpendicular de un punto $\mathbf{x}$ al hiperplano definido por $y(\mathbf{x}) = 0$ está dada por $|y(\mathbf{x})|/\|\mathbf{w}\|$.
- Se está interesado en soluciones para las cuales $t_n y(\mathbf{x}_n) > 0$, para todo n.
- Así, la distancia de un punto $\mathbf{x}_n$ a la superficie de decisión está dada por

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n \left(\mathbf{w}^\top \phi(\mathbf{x}_n) + w_0\right)}{\|\mathbf{w}\|}.$$

- El margen está dado por la distancia perpendicular al punto $\mathbf{x}_n$ más cercano, y se desea optimizar $\mathbf{w}$, y w_0 para maximizar esa distancia

$$\arg\max_{\mathbf{w}, w_0} \left\{ \min_n \left[\frac{t_n \left(\mathbf{w}^\top \phi(\mathbf{x}_n) + w_0\right)}{\|\mathbf{w}\|} \right] \right\}$$

$$\arg\max_{\mathbf{w}, w_0} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n \left[t_n \left(\mathbf{w}^\top \phi(\mathbf{x}_n) + w_0\right) \right] \right\}$$

Problema equivalente (I)

- La solución directa del problema anterior es extremadamente compleja.
- Se convierte a un problema equivalente, más sencillo de resolver.
- Nótese que si $\mathbf{w} \to k\mathbf{w}$, y $w_0 \to kw_0$, la distancia desde cualquier punto $\mathbf{x}_n$ a la superficie de decisión, dada por $t_n y(\mathbf{x}_n)/\|\mathbf{w}\|$, no cambia.
- Se puede escoger una constante k tal que

$$t_n \left(\mathbf{w}^\top \phi(\mathbf{x}_n) + w_0\right) = 1,$$

  para el punto $\mathbf{x}_n$ que está más cercano a la superficie.
- Se conoce como la representación canónica de la superficie de decisión.

Problema equivalente (II)

- En este caso todos los puntos satisfacen las siguientes condiciones

$$t_n \left(\mathbf{w}^\top \phi(\mathbf{x}_n) + w_0 \right) \geq 1, \quad n = 1, \ldots, N.$$

- El problema de optimización requiere maximizar $\|\mathbf{w}\|^{-1}$, o minimizar $\|\mathbf{w}\|^2$,

$$\arg \min_{\mathbf{w}, w_0} \frac{1}{2} \|\mathbf{w}\|^2,$$

sujeto a las restricciones $t_n \left(\mathbf{w}^\top \phi(\mathbf{x}_n) + w_0 \right) \geq 1, \ \forall n.$

Multiplicadores de Lagrange

- Se introducen multiplicadores de Lagrange,

$$L(\mathbf{w}, w_0, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n\{t_n\left(\mathbf{w}^\top \phi(\mathbf{x}_n) + w_0\right) - 1\},$$

donde $\mathbf{a} = [a_1, \ldots, a_N]^\top$.

Problema dual

- Haciendo los reemplazos correspondientes, se obtiene la representación dual del problema de margen máximo.

- Se busca maximizar

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m),$$

con respecto a $\mathbf{a}$, sujeto a las restricciones

$$a_n \geq 0, \quad n = 1, \ldots, N,$$

$$\sum_{n=1}^{N} a_n t_n = 0.$$

Vectores de soporte

- Las predicciones están dadas por

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + w_0.$$

- El signo de $y(\mathbf{x})$ indica la clase.

- Cualquier punto para el cual $a_n = 0$, no aparecerá en la suma de predicción, y así estos puntos no juegan ningún rol en la predicción.

- Los puntos restantes se conoce como *vectores de soporte*.

- Los vectores de soporte son puntos que subyacen sobre los hiperplanos del margen máximo en el espacio de características.

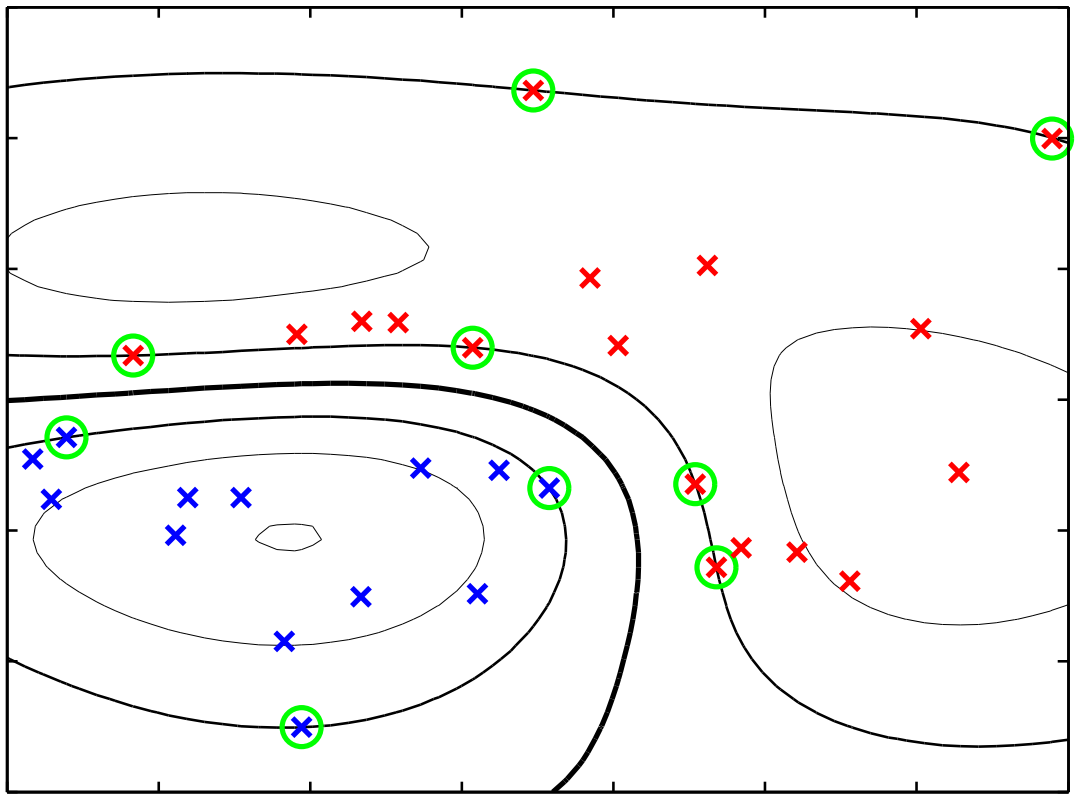
Estimando w_0

- El valor de w_0 se puede obtener a partir de

$$w_0 = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right),$$

donde S denota el conjunto de los índices de los vectores de soporte, y N_S es el número total de vectores de soporte.

Ejemplo



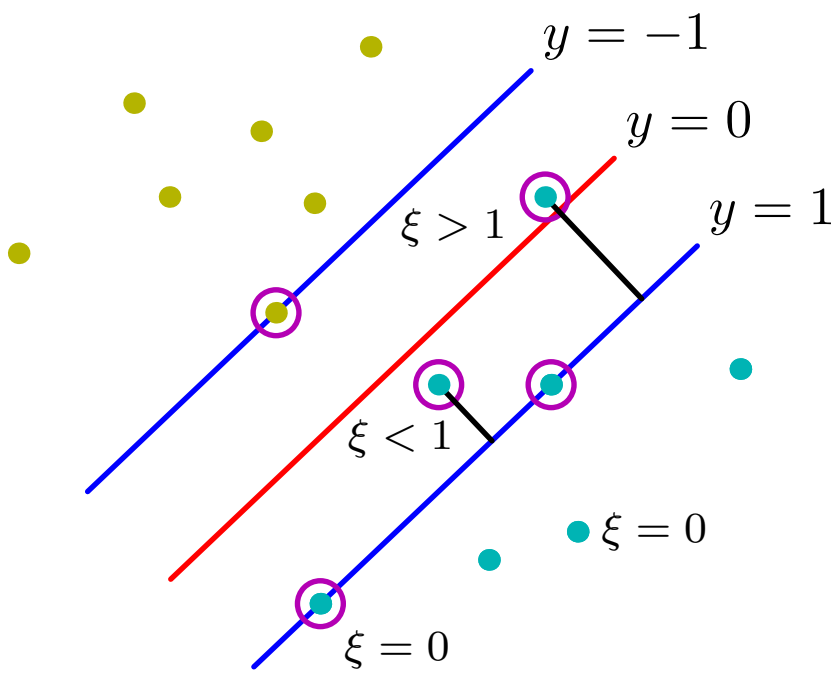
Contenido

Distribuciones de clase con traslape

- La SVM anterior entrega una separación exacta de los datos de entrenamiento en el espacio original **x**.
- En la práctica, sin embargo, las distribuciones de clase condicional pueden traslaparse, y la separación exacta de los datos de entrenamiento puede conducir a una pobre generalización.
- A continuación se modifica la SVM anterior, permitiendo que algunos datos estén en el “lado equivocado” del límite del margen, pero con una penalidad que incrementa con la distancia desde ese límite.

Variables slack ξ_n

- Se definen como $\xi_n = 0$ para puntos que están sobre o dentro del límite del margen correcto, y $\xi_n = |t_n - y(\mathbf{x}_n)|$, para otros puntos.
- Así, un punto que está sobre la superficie de decisión $y(\mathbf{x}) = 0$, tendrá $\xi_n = 1$, y puntos con $\xi_n \geq 1$ serán clasificados de forma incorrecta.



Problema de optimización

- El objetivo es maximizar el margen mientras se penalizan los puntos que están en el lado equivocado del margen.

- En particular, se desea minimizar

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n,$$

donde $C > 0$, y sujeto a las restricciones

$$t_n y(\mathbf{x}_n) \geq 1 - \xi_n,$$

con $\xi_n \geq 0$.

- El parámetro C permite hacer un compromiso entre la penalización por variables slack y el margen.

Multiplicadores de Lagrange

- Se introducen multiplicadores de Lagrange,

$$L(\mathbf{w}, w_0, \boldsymbol{\xi}, \mathbf{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n, - \sum_{n=1}^{N} a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n,$$

donde $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^\top$, $\mathbf{a} = [a_1, \ldots, a_N]^\top$, y $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]^\top$.

Problema dual

- Haciendo los reemplazos correspondientes, el problema dual consiste en maximizar

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m),$$

con respecto a $\mathbf{a}$, sujeto a las restricciones

$$0 \leq a_n \leq C, \quad n = 1, \ldots, N,$$

$$\sum_{n=1}^{N} a_n t_n = 0.$$

Vectores de soporte

- Las predicciones están dadas por

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + w_0.$$

- El signo de $y(\mathbf{x})$ indica la clase.
- Cualquier punto para el cual $a_n = 0$, no aparecerá en la suma de predicción, y así estos puntos no juegan ningún rol en la predicción.
- Los puntos para los cuales $a_n > 0$ son los vectores de soporte.
- Si $a_n < C$, entonces $\xi_n = 0$, y los puntos subyacen en el margen.
- Si $a_n = C$, entonces $\xi_n > 0$, y los puntos o está bien clasificados (si $\xi_n \leq 1$), o mal clasificados si $\xi_n > 1$.

Estimando w_0

- El valor de w_0 se puede obtener a partir de

$$w_0 = \frac{1}{N_M} \sum_{n \in M} \left(t_n - \sum_{m \in S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right),$$

donde S denota el conjunto de los índices de los vectores de soporte, M denota el conjunto de los índices de los datos que tienen $0 < a_n < C$, y N_M es la cardinalidad de M.