"Vector de Soporte y Métodos de Kernel"

Carlos Valle Vidal
cvalle@inf.utfsm.cl

Departamento de Informática -
Universidad Técnica Federico Santa María

Santiago, Abril 2009

Temario

Temario

Introducción

- Los Métodos de kernel son una familia de algoritmos relativamente nueva en el mundo del análisis inteligente de datos y reconocimiento de patrones.
- Combinan la simplicidad y eficiencia de algoritmos como el perceptrón y ridge regression con la flexibilidad de sistemas no-lineales y el rigor de los métodos de regularización.
- Sea X el espacio de entrada e Y el dominio de salida. $X \subseteq \mathbb{R}^n, Y = \{-1, 1\}$ clasificación, $Y \subseteq \mathbb{R}$ regresión

Introducción (2)

- A lo largo del capítulo usaremos como ejemplo una función de aprendizaje binaria $f : X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ de la siguiente forma
  - La entrada $x = (x_1, \ldots, x_n)^T$ es asignada a la clase positiva si $f(x) \geq 0$,
  - Sino es asignada a la clase negativa
- Estaremos interesados además cuando $f(x)$ es una función no-lineal de $x \in X$, y lo resolveremos encontrando una función lineal $f(x)$ que sea la imagen de un mapeo no-lineal.

Introducción (3)

- El conjunto de entrenamiento
  $S = \{(x_1, y_1), \ldots, (x_m, y_m) \subseteq (X \times Y)^m\}$
- m es el número de ejemplos, x_i es el ejemplo o instancia e y_i es su etiqueta
- Denotamos por $\langle x, w \rangle = x^T w = \sum_i x_i w_i$ al producto interno entre x y w

Métodos de Kernel

- Métodos de Kernel $\Rrightarrow$ llevar los datos a un espacio vectorial donde se pueda aplicar métodos lineales $\Rrightarrow$ identificar patrones
- Este mapeo es no-lineal $\Rrightarrow$ descubrir relaciones no-lineales usando métodos lineales.
- Consideremos el mapeo $\phi : x \in X \longrightarrow \phi(x) \in F$
- X: espacio de entrada. F : Espacio característico.

Temario

Representación Primal

- El algoritmo del perceptrón aprende una clasificación binaria usando una función lineal con valores reales $f : X \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$

$$\begin{aligned} f(x) &= \langle w, x \rangle + b \\ &= \sum_{i=1}^{n} w_i x_i + b \end{aligned}$$

- Donde $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ son parámetros que controlan la función y la regla de decisión está dada por $\text{sign}(f(x))$
- w, b se aprenden de los datos y su salida es el algoritmo del perceptrón

Interpretación Geométrica

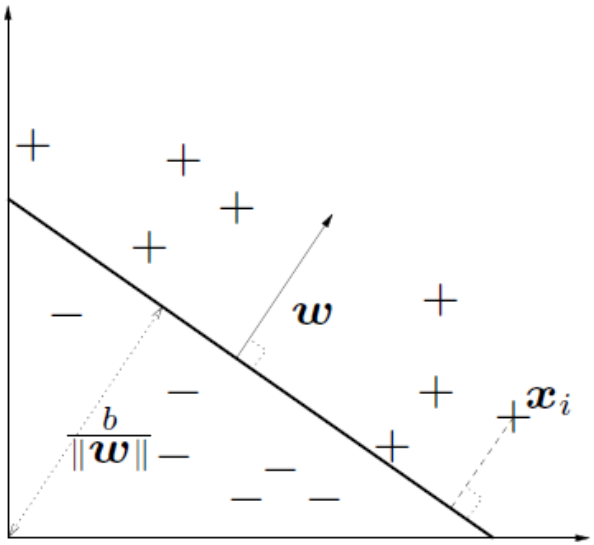
- El espacio de entradas X se divide en dos partes por el hiperplano definido por $\langle w, x \rangle + b = 0$
- Un hiperplano es un subespacio afín de dimensión $n-1$ que divide el espacio en dos partes correspondientes a dos clases.
- Esta representación de $n+1$ parámetros libres permite representar todos los posibles hiperplanos en $\mathbb{R}^n$

Interpretación Geométrica (2)

- La distancia euclidiana de un punto x_j al plano

$$\sum_{j=1}^{m} \frac{w_j x_j + b}{||w||} = \frac{\langle w, x_j \rangle + b}{||w||}$$

- $||w|| = \sqrt{\langle w \cdot w \rangle}$



Notaciones

- Los estadísticos usan este modelo y le llaman discriminante lineal (Fisher, 1936)
- En redes neuronales se le llama perceptrón (Rosemblatt, 1960)
- w es el vector de pesos y b el sesgo (bias).
- $-b = \theta \Rightarrow$ umbral
- Se desea minimizar

$$D(w,b) = -\sum_j y_j(x_j w + b)$$

- Derivando respecto de w y b tenemos

$$\frac{\delta D(w,b)}{\delta w} = \sum_j y_j x_j$$

$$\frac{\delta D(w,b)}{\delta b} = \sum_j y_j$$

El perceptrón: Algoritmo forma primal

Algoritmo 1 Algoritmo del Perceptrón (forma primal)

Dado un conjunto de entrenamiento S
$w_0 \leftarrow 0; b_0 \leftarrow 0; k \leftarrow 0$
$R \leftarrow \max_{1 \leq j \leq m} ||x_j||$
repeat
    for $j = 1$ to m **do**
        if $y_j(\langle w_k, x_j \rangle + b_k) \leq 0$ **then**
            $w_{k+1} \leftarrow w_k + \eta y_j x_j$
            $b_{k+1} \leftarrow b_k + \eta y_j R^2$
            $k \leftarrow k + 1$
        end if
    end for
until No haya errores dentro del loop
Salida: (w_k, b_k)

Observaciones

- Nótese que la contribución de x_j al cambio de peso es $\alpha_j \eta y_j x_j$
- α_j es el número de veces que x_j está mal clasificado
- El número de fallos totales k

$$k = \sum_{j=1}^{l} \alpha_j$$

-

$$w = \sum_{j=1}^{l} \alpha_j y_j x_j$$

- El algoritmo modifica directamente w y b
- Si existe un hiperplano que clasifique correctamente los datos $\Rightarrow$ datos linealmente separables
- Sino puede caer en infinitos ciclos

Representación dual

- Dada las observaciones anteriores, podemos representar la función de decisión de la siguiente manera

$$
\begin{aligned}
h(x) &= \text{sign}(\langle w, x\rangle + b) && (1)\\
&= \text{sign}\left(\left\langle \sum_{j=1}^{m} \alpha_j y_j x_j, x \right\rangle + b\right) && (2)\\
&= \text{sign}\left(\sum_{j=1}^{m} \alpha_j y_j \langle x_j, x\rangle + b\right) && (3)
\end{aligned}
$$

El perceptrón: Algoritmo forma dual

Algoritmo 2 Algoritmo del Perceptrón (forma primal)

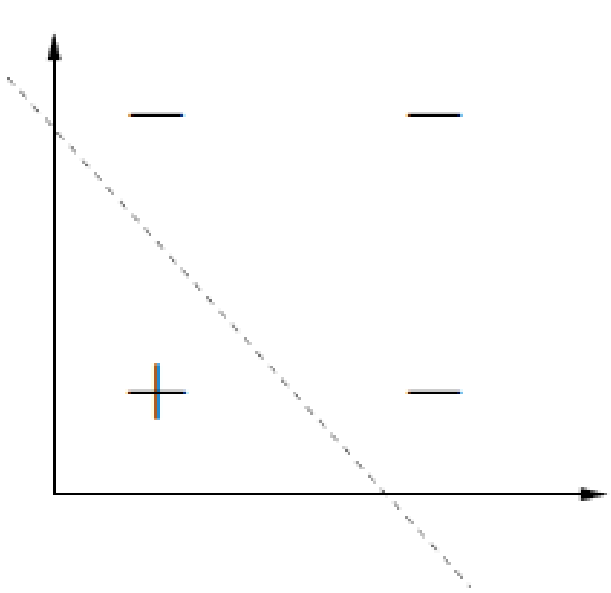
1: Dado un conjunto de entrenamiento S
2: $\alpha \leftarrow 0; b_0 \leftarrow 0$
3: $R \leftarrow \max_{1 \leq i \leq m} ||x_i||$
4: **repeat**
5: **for** $i = 1$ to m **do**
6: **if** $\left(\sum_{j=1}^{m} \alpha_j y_j \langle x_j, x \rangle + b\right) \leq 0$ **then**
7: $\alpha_i \leftarrow \alpha_i + 1$
8: $b \leftarrow b + \eta y_i R^2$
9: **end if**
10: **end for**
11: **until** No haya errores dentro del loop
12: Salida: (α, b) para definir $h(x)$ según ecuación (1)

Observaciones

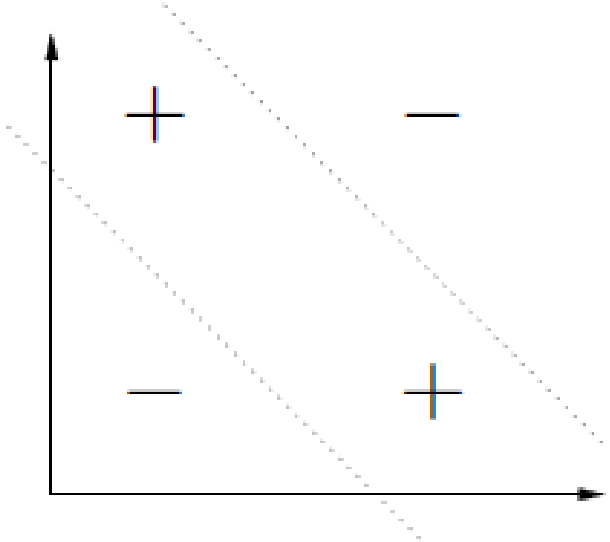
- Nótese que la información sobre los datos sólo llega $\langle x_i, x_j \rangle$
- Esto significa que no necesitamos los puntos, sólo sus productos internos.
- Veremos que esto se puede realizar de manera computacionalmente eficaz usando una función de kernel
- La información acerca de las posiciones relativas de los datos en el nuevo espacio está codificado en los productos internos entre ellos
- Los productos internos entre las proyecciones de los datos de entrada en el espacio nuevo altamente dimensional pueden solucionarse computacionalmente mediante una función de kernel

Problema

- Minsky 1969



AND



XOR

Temario

Mapeo implícito con kernels

- Podemos llevar los datos originales al espacio característico F donde los métodos lineales capturen características relevantes de los datos. Usamos el mapeo ϕ

$$\phi : x \in X \mapsto \phi(x) \in F$$

- Aprovechándonos de la representación dual, $f(x)$ en el espacio característico queda

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle + b$$

- Podemos reemplazar estos productos internos en el nuevo espacio, por una función de kernel que calcule directamente los productos internos como una función de las entradas x originales.

Función de Kernel

- Sea K función de kernel, tal que para todo $x, z \in X$

$$K(x,z) = \langle \phi(x), \phi(z) \rangle$$

- donde ϕ es el mapeo de X al espacio característico F
- La dimensión del espacio característico no afecta la computación del kernel.

Función de Kernel: Ejemplo

- Consideremos dos puntos $x = (x_1, x_2)$ y $z = (z_1, z_2)$ en un espacio bi-dimensional y la función $K(x,z) = \langle x, z \rangle^2$

$$\begin{aligned} \langle x, z \rangle^2 &= \langle (x_1, x_2), (z_1, z_2) \rangle^2 \\ &= (x_1 z_1 + x_2 z_2)^2 \\ &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 x_2 z_1 z_2 \\ &= \langle (x_1^2, x_2^2, \sqrt{2} x_1 x_2), (z_1^2, z_2^2, \sqrt{2} z_1 z_2) \rangle \end{aligned}$$

- El producto interno anterior en el espacio característico corresponde al mapeo

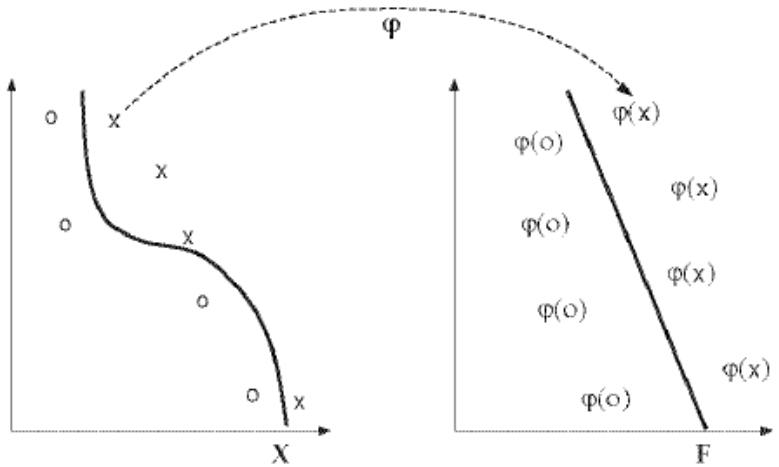
$$(x_1, x_2) \mapsto \phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$$

Función de Kernel: Ejemplo(2)

- Por lo tanto

$$K(x,z) = \langle x,z\rangle^2 = \langle \phi(x), \phi(z)\rangle$$

- Es fácil computar un kernel $K(x,z) = \langle x,z\rangle^d$, pero genera un espacio característico de $\binom{n+d-1}{d}$ dimensiones, por lo que computar $\phi(x)$ se vuelve infactible computacionalmente
- Podemos usar kernels sin conocer el mapeo ϕ asociado



Teorema de Mercer

- Sea X un subconjunto compacto de $\mathbb{R}^n$. Supongamos que K es una función continua simétrica tal que el operador integral $T_K : L_2(X) \rightarrow L_2(X)$

$$(T_k f)(\cdot) = \int_X K(\cdot, x) f(x) dx$$

- Es positivo, es decir

$$\int_{X \times X} K(x,z) f(x) f(z) dx dz \geq 0, \forall f \in L_2(X)$$

- Podemos expandir $K(x,z)$ en una serie con convergencia uniforme sobre $X \times X$ en términos de T_K funciones propias $\phi_j \in L_2(X)$, normalizado de tal forma que $||\phi_j||_{L_2} = 1$ y valores propios asociados positivos $\lambda_j \geq 0$

$$K(x,z) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x) \phi_j(z)$$

Teorema de Mercer (2)

- La imagen de un vector x vía el mapeo implícito definido por el kernel es $\sum_{j=1}^{\infty} \sqrt{\lambda_j}\phi_j(x)$.
- Kernel estándar gaussiano (radio basal)

$$K(x,z) = \exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)$$

- Kernel polinomial de grado d

$$K(x,z) = (\langle x,z\rangle + 1)^d$$

- Red neuronal

$$K(x,z) = \tanh\left(k_1\langle x,z\rangle + k_2\right)$$

Temario

Sobreajuste y cotas de generalización

- El principal supuesto que haremos, es que todos los puntos de los conjuntos de entrenamiento y prueba son independientes y vienen de la misma distribución (fija, pero desconocida) La capacidad de generalización depende del tamaño de la muestra y de la capacidad efectiva de la hipótesis usada en el sistema $(\widehat{f})$
- Esta cantidad es una medida de flexibilidad del algoritmo, contando de cuantas formas distintas el algoritmo puede etiquetar un conjunto de datos, separando correctamente al usar una función equivalente a la solución obtenida.

Definición

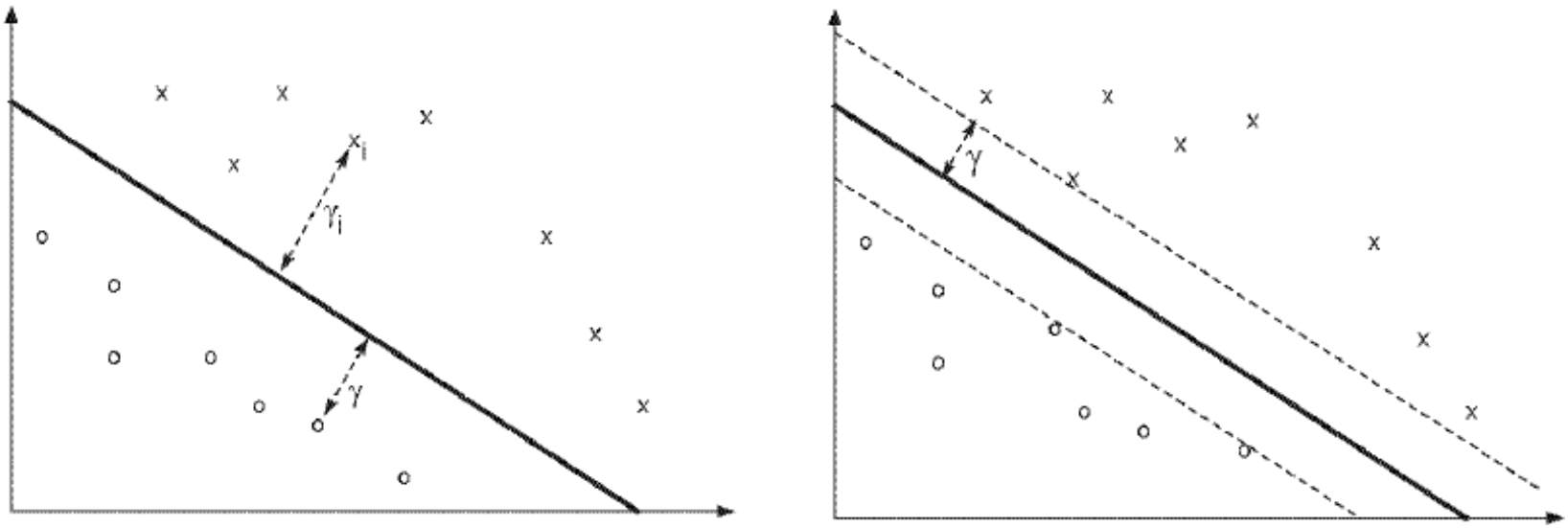
- Definamos el margen funcional de una ejemplo (x_i, y_i) con respecto al hiperplano, como

$$\gamma_i = y_i(\langle w, x_i \rangle + b)$$

- Notemos que $\gamma_i > 0$ implica una correcta clasificación
- Reemplazaremos la definición de margen funcional por el margen geométrico obtenemos una función lineal normalizada $\left(\frac{w}{||w||}, \frac{b}{||w||}\right)$ correspondiente a la distancia Euclidiana a la frontera de decisión en el espacio de entrada.
- El margen de un conjunto de entrenamiento S es el máximo margen geométrico sobre todos los hiperplanos
- El tamaño del margen será positivo si el conjunto de entrenamiento es linealmente separable

Margen geométrico

Figura: Márgen geométrico de dos puntos (a la izquierda), márgen de una muestra (a la derecha)



Teorema 1

- Considere una función de umbral $\mathcal{L}$ con vector de pesos unitarios sobre el producto interno en un espacio X y $\gamma \in \mathbb{R}$ fijo. Para cualquier distribución de probabilidad $\mathcal{D}$ sobre $X \times \{-1, 1\}$ con soporte en una bola de radio R entorno al origen, con probabilidad $1 - \delta$ sobre m ejemplos aleatorios S, cualquier hipótesis $f \in \mathcal{L}$ que tiene margen $m_S(f) \geq \gamma$ sobre S tiene errores no mayores que

$$\text{err}_{\mathcal{D}}(f) \leq \varepsilon(m, \mathcal{L}, \delta, \gamma) = \frac{2}{m}\left(\frac{64R^2}{\gamma^2}\ln\frac{em\gamma}{8R^2}\ln\frac{32m}{\gamma^2} + \ln\frac{4}{\delta}\right)$$

- $m > 2/\varepsilon$ y $64R^2/\gamma^2 < m$

Teorema 1 (2)

- Lo importante de este resultado, es que la dimensión de entrada de x no parece, esto significa que el resultado también puede aplicarse a espacios infinito dimensionales
- Curse of dimensionality (Richard Bellman)
- Este teorema no dice nada respecto de data no linealmente separable ni data con ruido (que puede causar un margen estrecho)

Definición

- Consideremos una clase $\mathcal{F}$ de funciones en el dominio de los reales sobre un espacio de entrada X para clasificación, con umbral 0. Definamos la variable de holgura del margen de un ejemplo $(x_i, y_i) \in X \times \{-1, 1\}$ con respecto a una función $f \in \mathcal{F}$ y el margen γ como la cantidad

$$\xi((x_i, y_i), f, \gamma) = \xi_i = \max(0, \gamma - y_i f(x_i))$$

- Notemos que $\xi_i > \gamma$ implica una mala clasificación de (x_i, y_i)

Definición (2)

- $\xi(S,f,\gamma)$ es el vector de holgura del margen de un conjunto de entrenamiento $S = \{(x_1,y_1),\ldots,(x_m,y_m)\}$
- Respecto de f y γ contiene a las variables de holgura

$$\xi = \xi(S,f,\gamma) = (\xi_1,\ldots,\xi_m)$$

- Las variables de holgura pueden medir ruido en los datos, causado por puntos individuales en los datos (con pequeños márgenes o valores negativos)

Teorema 2

- Consideremos funciones $\mathcal{L}$ con valores reales, umbral y vector de pesos sobre un producto interno X y $\gamma \in \mathcal{R}^+$ fijo. Existe una constante c tal que para cualquier distribución de probabilidad $\mathcal{D}$ sobre $X \times \{-1, 1\}$ con soporte en una bola de radio R entorno al origen, con probabilidad $1 - \delta$ sobre m ejemplos aleatorios S, cualquier hipótesis $f \in \mathcal{L}$ tiene un error no mayor que

$$\text{err}_{\mathcal{D}}(f) \leq= \frac{c}{m}\left(\frac{R^2 + ||\xi||_2^2}{\gamma^2}\ln^2 m + \ln\frac{1}{\delta}\right)$$

- Donde $\xi = \xi(f, S, \gamma)$ es el vector de holgura respecto de f y γ

Teorema 3

- Consideremos funciones $\mathcal{L}$ con valores reales, umbral y vector de pesos sobre un producto interno X y $\gamma \in \mathcal{R}^+$ fijo. Existe una constante c talque para cualquier distribución de probabilidad $\mathcal{D}$ sobre $X \times \{-1, 1\}$ con soporte en una bola de radio R entorno al origen, con probabilidad $1 - \delta$ sobre m ejemplos aleatorios S, cualquier hipótesis $f \in \mathcal{L}$ tiene un error no mayor que

$$\text{err}_{\mathcal{D}}(f) \leq= \frac{c}{m}\left(\frac{R^2}{\gamma^2}\ln^2 m + \frac{||\xi||_1}{\gamma}\ln m + \ln\frac{1}{\delta}\right)$$

- Donde $\xi = \xi(f, S, \gamma)$ es el vector de holgura respecto de f y γ

Observaciones

- El Teorema 2 es la generalización de la cota del error y toma en cuenta la cantidad en la que los puntos fallan respecto del margen γ
- La cota es en términos de ξ, lo que sugiere que si minizamos esta cantidad, mejoraremos el desempeño.
- La cota no descansa en que los puntos son linealmente separables, por lo que puede existir ruido o problemas con los datos
- Optimizar la norma de ξ no implica directamente minimizar el número de elementos mal clasificados, sin embargo es computacionalmente más barato.
- Soft margin v/s hard margin (vector de holgura tiene un efecto difuso sobre el margen)

Temario

Clasificadores SV

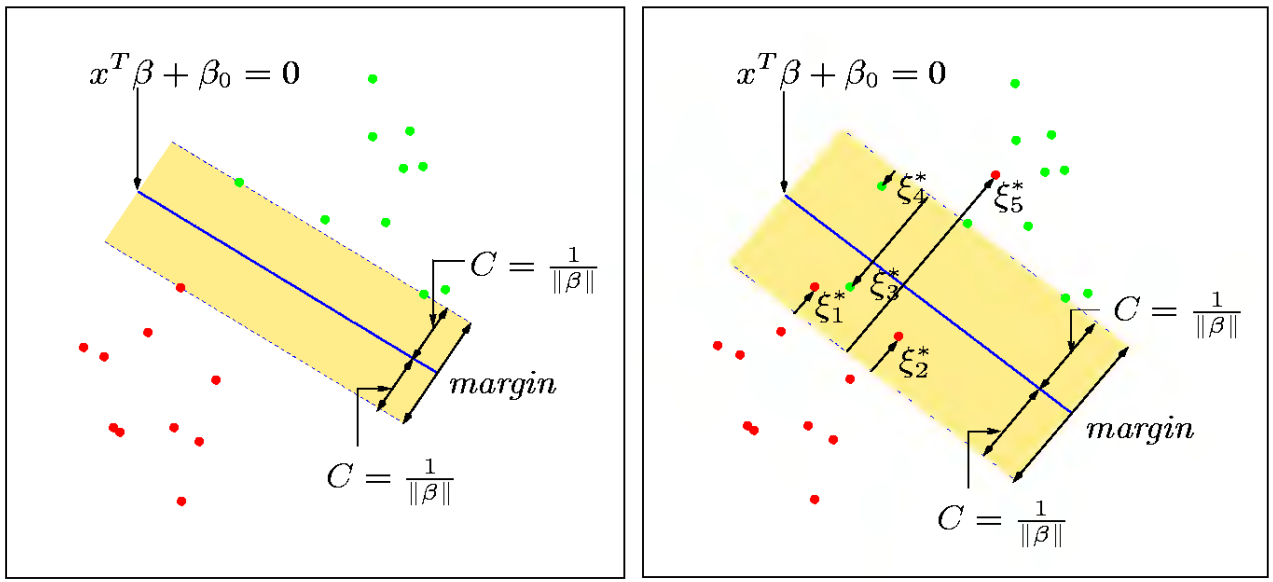


FIGURE 12.1. *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width* $2C = 2/\|\beta\|$. *The right panel shows the nonseparable (overlap) case. The points labeled* ξ_j^* *are on the wrong side of their margin by an amount* $\xi_j^* = C\xi_j$; *points on the correct side have* $\xi_j^* = 0$. *The margin is maximized subject to a total budget* $\sum \xi_i \leq$ *constant. Hence* $\sum \xi_j^*$ *is the total distance of points on the wrong side of their margin.*

Hiperplano Optimal

- Encontrar el hiperplano óptimo que separa dos clases equivale a resolver el problema de optimización

$$\max_{w,b} C$$
$$S.A. \frac{y_i(\langle w, x\rangle + b)}{||w||} \geq C, i = 1, \ldots, m$$

- Si elegimos $C = \frac{1}{||w||}$, el problema anterior es equivalente a

$$\min_{w,b} ||w||^2$$
$$S.A. \frac{y_i(\langle w, x\rangle + b)}{||w||} \geq 1, i = 1, \ldots, m$$

- Las restricciones definen un margen de decisión $\gamma = \frac{1}{||w||_2}$

Lagrangiano Primal

- Por lo tanto, es Lagrangiano primal es

$$L_P(w,b,\alpha) = \frac{1}{2}\langle w,w\rangle - \sum_{i=1}^{m} \alpha_i [y_i(\langle w,x_i\rangle + b) - 1]$$

- Donde los $\alpha_i \geq 0$ son los multiplicadores de Lagrange
- Este es un problema de programación no-lineal convexo $\Rightarrow$ función objetivo es convexa y los puntos que satisfacen las restricciones, forman un conjunto convexo (cualquier restricción lineal define un conjunto convexo, y un conjunto de K restricciones lineales simultáneas define la intersección de los K conjuntos convexos $\Rightarrow$ también es un conjunto convexo.

Lagrangiano dual

- Derivando respecto de w y b tenemos que

$$\frac{\delta L_P(w,b,\alpha)}{\delta w} = w - \sum_{i=1}^{m} y_i \alpha_i x_i = 0$$

$$\frac{\delta L_P(w,b,\alpha)}{\delta b} = \sum_{i=1}^{m} y_i \alpha_i = 0$$

- Utilizando estas relaciones en el Primal obtenemos

$$\begin{aligned} L_P(w,b,\alpha) &= \frac{1}{2}\langle w,w\rangle - \sum_{i=1}^{m} \alpha_i [y_i(\langle w,x_i\rangle + b) - 1] \\ &= \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i,x_j\rangle - \sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i,x_j\rangle + \sum_{i=1}^{m} \alpha_i \\ L_D(w,b,\alpha) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i,x_j\rangle \end{aligned}$$

- También se conoce como Wolfe dual
- Análogo al perceptrón, también se logra una solución en la que pueden usarse kernels.

Proposición 1

- Consideremos una muestra linealmente separable $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$
- Y supongamos que los parámetros α resuelven el problema de optimización cuadrática

$$\left.\begin{array}{ccc} \text{Maximizar } W(\alpha) & = & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle \\ S.A. & & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ \alpha_i & \geq 0 & , i = 1, \ldots, m \end{array}\right\}$$

- Entonces el vector de pesos $w^* = \sum_{i=1}^{m} y_i \alpha_i^* x_i$ forma parte del hiperplano de máximo margen $\gamma = 1/||w^*||_2$

Proposición 1 (2)

- b no aparece en el dual, por lo que b^* debe ser encontrado usando las restricciones del primal

$$b^* = \frac{\max_{y_i=-1}\left(\langle w^*, x_i\rangle\right) + \max_{y_i=1}\left(\langle w^*, x_i\rangle\right)}{2}$$

- Las condiciones de Karush-Kunh-Tucker(KKT) afirman que el óptimo α^*, w^*, b^* debe satisfacer

$$\alpha_i^*\left[y_i(\langle w_i^*, x_i\rangle + b^*) - 1\right] = 0, i = 1, \ldots, m$$

- Esto implica que solo para las entradas x_i donde el margen funcional es uno $\Rightarrow$ muy cerca del hiperplano clasificador $\Rightarrow$ Support Vectors (SV).

Proposición 1 (3)

- En otras palabras podemos expresar la representación dual en términos de este subconjunto de parámetros

$$\begin{aligned} f(x, \alpha^*, b^*) &= \sum_{i=1}^{m} y_i \alpha_i^* \langle x_i, x \rangle + b^* \\ &= \sum_{i \in SV} y_i \alpha_i^* \langle x_i, x \rangle + b^* \end{aligned}$$

- Los multiplicadores de Lagrange miden la importancia de cada ejemplo de entrenamiento en la solución final
- Perceptrón α_i media el número de equivocaciones del i-ésimo ejemplo

Proposición 1 (4)

- Otra importante consecuencia de las KKT para $j \in SV$

$$y_j f(x_j, \alpha^*, b^*) = y_j \left(\sum_{i \in SV} y_i \alpha_i^* \langle x_i, x \rangle + b^* \right) = 1$$

- Por lo tanto, podemos expresar $||w||$ y el margen como función de los multiplicadores α

$$\begin{aligned} \langle w^*, w^* \rangle &= \sum_{i,j}^{m} y_i y_j \alpha_i^* \alpha_j^* \langle x_i, x_j \rangle \\ &= \sum_{j \in SV} \alpha_j^* y_j \sum_{i \in SV} y_i \alpha_i^* \langle x_i, x_j \rangle \\ &= \sum_{j \in SV} \alpha_j^* (1 - y_j b^*) \\ &= \sum_{i \in SV} \alpha_i^* \end{aligned}$$

Proposición 2

- Consideremos la muestra $S\{(x_1, y_1), \ldots, (x_m, y_m)\}$ y supongamos que α^* y b^* resuelven el problema de optimización dual. Entonces $w = \sum_{i=1}^{m} y_i \alpha_i^* x_i$ forma parte del hiperplano con margen geométrico

$$\gamma = 1/||w||_2 = \left(\sum_{i \in SV} \alpha_i^* \right)^{-1/2}$$

Proposición 1 usando kernels

- Consideremos una muestra $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ linealizable en el espacio característico definido por el kernel $K(x, z)$
- Y supongamos que los parámetros α resuelven el problema de optimización cuadrática

$$\left.\begin{array}{lll} \text{Maximizar } W(\alpha) & = & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ S.A. & & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ \alpha_i & \geq 0 & , i = 1, \ldots, m \end{array}\right\}$$

Proposición 1 usando kernels (2)

- Entonces la regla de decisión dada por $\text{sign}(f(x))$, donde $f(x) = \sum_{i=1}^{m} y_i \alpha_i^* K(x_i, x) + b^*$ es equivalente al hiperplano de máximo margen en el espacio característico implícito definido por el kernel $K(x,z)$ y que el hiperplano tiene margen geométrico

$$\gamma = \left(\sum_{i \in SV} \alpha_i^* \right)^{-1/2}$$

- Nótese que un Kernel que satisface las condiciones de Mercer, es equivalente a que la matriz $(K(x_i, x_j))_{i,j}^m$ sea definida positiva.
- Esto transforma el problema a convexo ya que la matriz $(y_i y_j K(x_i, x_j))_{i,j}^m$ también es definida positiva $\Rightarrow$ solución única
- El problema se puede comprimir sólo usando los SV

Teorema

- Consideremos funciones $\mathcal{L}$ reales con umbral con vector de pesos unitario sobre un espacio X con producto interno. Para cualquier distribución de probabilidad $\mathcal{D}$ sobre $X \times \{-1, 1\}$, con probabilidad $1 - \delta$ sobre m ejemplos de S, el hiperplano maximal tiene un error menor que

$$\text{err}_{\mathcal{D}}(f) \leq \frac{1}{m-d}\left(d \ln \frac{em}{d} + \ln \frac{m}{\delta}\right)$$

- Donde d es el número de SV.
- Elección del Kernel

Temario

SMO (Soft margin Optimization)

- Intenta mejorar el problema de generalización de Hard Margin
- Cuyo problema de optimización es

$$\begin{aligned} &\text{Minimizar}_{w,b}\langle w, w\rangle \\ &\text{S.A: } y_i(\langle w, x_i\rangle + b) \quad \geq \quad 1, i = 1, \ldots, m \end{aligned}$$

- Introduciremos variables de holgura permitiendo violar las restricciones

$$\begin{aligned} \text{S.A: } y_i(\langle w, x_i\rangle + b) &\geq 1 - \xi_i, i = 1, \ldots, m \\ \xi_i &\geq 0, i = 1, \ldots, m \end{aligned}$$

- De la misma forma necesitamos controlar el tamaño de las violaciones ξ_i

Soft Margin norma 2

- Cuyo problema de optimización es

$$\left.\begin{array}{lcc} \text{Minimizar}_{w,b}\langle w,w\rangle & + & C\sum_{i=1}^{m}\xi_i^2 \\ \text{S.A: } y_i(\langle w,x_i\rangle+b) & \geq & 1-\xi_i, i=1,\ldots,m \\ & & \xi_i \geq 0, i=1,\ldots,m \end{array}\right\} \quad (4)$$

- C controla el equilibrio entre los errores de clasificación y el margen.
- El Lagrangiano primal de este problema es

$$L(w,b,\xi,\alpha) = \frac{1}{2}\langle w,w\rangle + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i[y_i(\langle w,x_i\rangle+b)-1+\xi_i]$$

- donde $\alpha_i \geq 0$ son los multiplicadores de Lagrange

Soft Margin norma 2 (2)

- Para encontrar la forma dual debemos derivar respecto a w, ξ y b

$$\begin{aligned}
\frac{\partial L(w,b,\xi,\alpha)}{\partial w} &= w - \sum_{i=1}^{m} y_i \alpha_i x_i = 0 \\
\frac{\partial L(w,b,\xi,\alpha)}{\partial \xi} &= C\xi - \alpha = 0 \\
\frac{\partial L(w,b,\xi,\alpha)}{\partial b} &= \sum_{i=1}^{m} y_i \alpha_i = 0
\end{aligned}$$

- Reemplazando estas relaciones en el primal obtenemos la siguiente función objetivo

$$\begin{aligned}
L(w,b,\xi,\alpha) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle + \frac{1}{2C}\langle \alpha, \alpha \rangle - \frac{1}{2C}\langle \alpha, \alpha \rangle \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \frac{1}{2C}\langle \alpha, \alpha \rangle
\end{aligned}$$

Soft Margin norma 2 (3)

- Maximizar sobre α es equivalente a maximizar

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \left(\langle x_i, x_j \rangle + \frac{1}{C} \delta_{ij} \right)$$

- Donde δ_{ij} es la δ de Kronecker

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{e.t.o.c.} \end{cases}$$

- Usando Kernels obtenemos la siguiente proposición

Proposición

- Considerando la muestra de entrenamiento S
- Usando el espacio característico inducido por el kernel $K(x,z)$ y supongamos que α^* resuelve el problema de optimización cuadrático:

$$\begin{aligned} W(\alpha) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \left(\langle x_i, x_j \rangle + \frac{1}{C} \delta_{ij} \right) \\ S.A: & \quad \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & \quad \alpha_i \geq 0, i = 1, \ldots, m \end{aligned}$$

Proposición (2)

- Sea $f(x) = \sum_{i=1}^{m} y_i \alpha_i^* K(x_i, x) + b^*$, donde b^* es escogido de manera tal que $f(x_i) = 1 - \alpha_i^*/C$ para cualquier i con $\alpha_i^* \neq 0$. Entonces la regla de decisión definida por $\text{sign}(f(x))$ es equivalente al hiperplano en el espacio característico definido implícitamente por $K(x, z)$ el cual resuelve el problema de optimización (4), donde las variables de holgura están definidas respecto del margen geométrico

$$\gamma = \left(\sum_{i \in SV} \alpha_i^* - \frac{1}{C} \langle \alpha^*, \alpha^* \rangle \right)^{-1/2}$$

Proposición (3)

- Demostración: El valor de b^* es escogido usando la relación $\alpha_i = C\xi_i$ y por referencia a las restricciones del primal generadas por las condiciones complementarias de Karush Kuhn Tucker

$$\alpha_i[y_i(\langle w_i, x_i\rangle + b) - 1 + \xi_i] = 0, i = 1, \ldots, m$$

- Esto se mantiene al computar la norma de w^* la cual define el tamaño del margen geométrico

$$\begin{aligned}
\langle w^*, w^*\rangle &= \sum_{i,j}^{m} y_i y_j \alpha_i^* \alpha_j^* K(x_i, x_j) \\
&= \sum_{j \in SV} \alpha_j^* y_j \sum_{i \in SV} y_i \alpha_i^* K(x_i, x_j) \\
&= \sum_{j \in SV} \alpha_j^* (1 - \xi_j^* - y_j b^*) \\
&= \sum_{i \in SV} \alpha_i^* - \sum_{i \in SV} \alpha_i^* \xi_i^* \\
&= \sum_{i \in SV} \alpha_i^* - \frac{1}{C}\langle \alpha_i^*, \alpha_i^*\rangle
\end{aligned}$$

Proposición (4)

- El problema es equivalente a usar hardmargin, solamente hay que adicionar $1/C$ en la diagonal de la matriz de kernel.
- Esto provoca el efecto de agregar $1/C$ a los valores propios de la matriz, mejorando su condicionamiento
- Usar un suavizamiento con norma 2, equivale a cambiar el kernel

$$K(x,z) = K(x,z) + \frac{1}{C}\delta_x(z)$$

Soft Margin norma 1

- Cuyo problema de optimización es

$$\left.\begin{array}{lcc} \text{Minimizar}_{w,b}\langle w,w\rangle & + & C\sum_{i=1}^{m}\xi_i \\ \text{S.A: } y_i(\langle w,x_i\rangle+b) & \geq & 1-\xi_i, i=1,\ldots,m \\ & & \xi_i \geq 0, i=1,\ldots,m \end{array}\right\} \quad (5)$$

- El Lagrangiano primal de este problema es

$$L(w,b,\xi,\alpha) = \frac{1}{2}\langle w,w\rangle + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\langle w,x_i\rangle+b)-1+\xi_i] - \sum_{i=1}^{m} r_i\xi_i$$

- donde $\alpha_i \geq 0$ y $r_i \geq 0$

Soft Margin norma 1 (2)

- Para encontrar la forma dual debemos derivar respecto a w, ξ y b

$$\begin{aligned}
\frac{\partial L(w,b,\xi,\alpha,r)}{\partial w} &= w - \sum_{i=1}^{m} y_i \alpha_i x_i = 0 \\
\frac{\partial L(w,b,\xi,\alpha,r)}{\partial \xi} &= C - \alpha_i r_i = 0 \\
\frac{\partial L(w,b,\xi,\alpha,r)}{\partial b} &= \sum_{i=1}^{m} y_i \alpha_i = 0
\end{aligned}$$

- Reemplazando estas relaciones en el primal obtenemos la siguiente función objetivo

$$L(w,b,\xi,\alpha,r) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

Soft Margin norma 1 (3)

- Observemos que esta es idéntica al problema de máximo margen a excepción de la restricción $C - \alpha_i - r_i = 0$ junto con $r_i \geq 0$, lo que fuerza $\alpha_i \leq C$
- Las condiciones KKT son

$$\alpha_i [y_i (\langle x_i, w \rangle + b) - 1 + \xi_i] = 0, i = 1, \ldots, m$$
$$\xi_i (\alpha_i - C) = 0, i = 1, \ldots, m$$

- Observemos que $\xi_i \neq 0 \Rightarrow \alpha_i = C$
- Los puntos con $\xi_i \neq 0$ tienen un margen menor a $1/||w||$.

Proposición

- Considerando la muestra de entrenamiento S
- Usando el espacio característico inducido por el kernel $K(x,z)$ y supongamos que α^* resuelve el problema de optimización cuadrático:

$$\begin{aligned} \text{Maximizar } W(\alpha) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ S.A: & \quad \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & \quad C \geq \alpha_i \geq 0, i = 1, \ldots, m \end{aligned}$$

Proposición (2)

- Sea $f(x))\sum_{i=1}^{m} y_i \alpha_i^* K(x_i, x) + b^*$, donde b^* es escogido de manera tal que $y_i f(x_i) = 1$ para cualquier i con $C > \alpha_i^* > 0$. Entonces la regla de decisión definida por $\text{sign}(f(x))$ es equivalente al hiperplano en el espacio característico definido implícitamente por $K(x, z)$ el cual resuelve el problema de optimización (5), donde las variables de holgura están definidas respecto del margen geométrico

$$\gamma = \left(\sum_{i,j \in SV} y_i y_j \alpha_i^* \alpha_j^* K(x_i, x_j) \right)^{-1/2}$$

- b^* se obtiene a partir de las KKT las que si $C > \alpha_i^* > 0$, $\xi_i^* = 0$ y

$$y_i(\langle x_i, w^* \rangle + b^*) - 1 + \xi_i^* = 0$$

Proposición (3)

- La norma de w^* está dada por la expresión

$$\begin{aligned} \langle w^*, w^* \rangle &= \sum_{i,j}^{m} y_i y_j \alpha_i^* \alpha_j^* K(x_i, x_j) \\ &= \sum_{j \in SV} \sum_{i \in SV} y_i y_j \alpha_i^* \alpha_j^* K(x_i, x_j) \end{aligned}$$

- Restricción para $\alpha_i \Rrightarrow$ box constraint
- Limitar influencia de outliers
- Ambas formas de soft margin obtienen soluciones basadas en hard margin

ν-SVM

- Un problema es determinar la constante C
- Este problema equivale a encontrar $0 \leq \nu \leq 1$ en el problema de optimización

$$\begin{aligned} \text{Maximizar } W(\alpha) &= -\frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ S.A: & \quad \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & \quad \sum_{i=1}^{m} \alpha_i \geq \nu \\ & \quad 1/m \geq \alpha_i \geq 0, i = 1, \ldots, m \end{aligned}$$

- Podemos ver esto como la proporción del conjunto de entrenamiento cuyo margen es mayor que ν
- ν no depende de la escala del espacio característico, sino sólo del nivel de ruido en los datos.

SVM como método de penalización

- Usando $f(x) = \langle w, \phi(x) \rangle + b$ consideremos el problema de optimización

$$\min_{b,w} \sum_{i=1}^{m} [1 - y_i f(x_i)]_+ + \lambda ||w||^2 \tag{6}$$

- Donde $+$ indica la parte positiva. Se puede demostrar que con $lambda = 1/(2C)$ la solución es la misma que (5)

Tabla: Comparación de funciones de pérdida

Función de pérdida	L(y,f(x))	Función a minimizar
Log verosimilitud	$\log(1+e^{-yf(x)})$	$f(x) = \log \frac{Pr(y=+1\|X)}{Pr(y=-1\|X)}$
MSE	$(y-f(x))^2$	$f(x) = Pr(y=+1\|X) - Pr(y=-1\|X)$
SVM	$[1-y_i f(x_i)]_+$	$f(x) = \begin{cases} +1 & \text{si } Pr(y=+1\|X) \geq \frac{1}{2} \\ -1 & \text{e.t.o.c.} \end{cases}$

SVM como método de penalización(2)

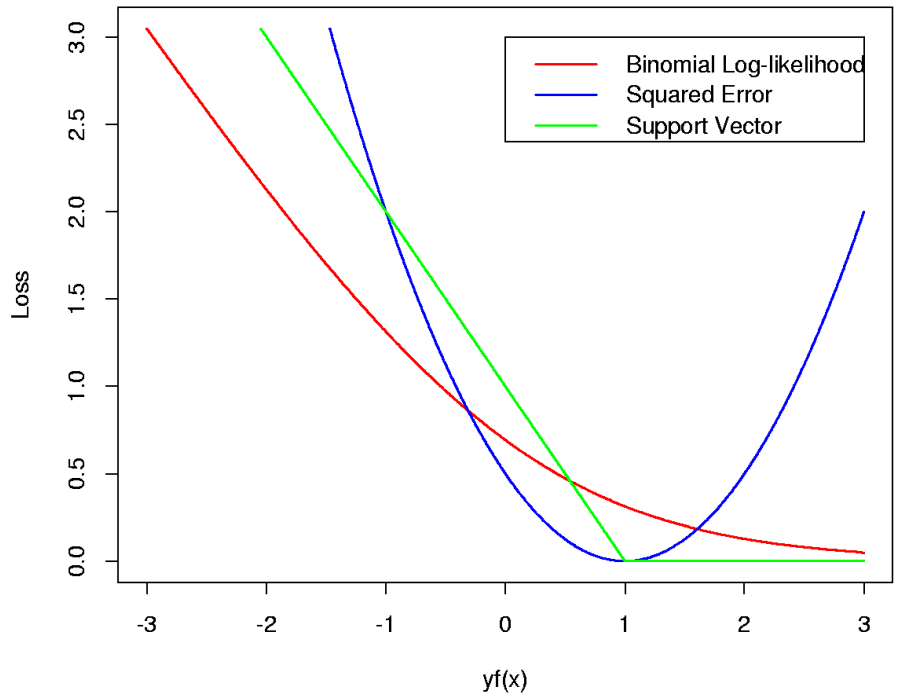


FIGURE 12.4. *The support vector loss function, compared to the (negative) log-likelihood loss for logistic regression, and squared-error loss. All are shown as a function of yf rather than f, because of the symmetry in all three between the $y = +1$ and $y = -1$ case. The log-likelihood has the same asymptotes as the SVM loss, but is rounded in the interior.*

Kernels y la función de estimación

- Recordemos que

$$K(x,z) = \sum_{j=1}^{\infty} \lambda_j h_j(x) h_j(z)$$

- Y $\phi_j(x) = \sqrt{\lambda_j} h_j(x)$ entonces con $\theta_j = \sqrt{\lambda_j} w_j$
- Podemos reescribir (6) como

$$\min_{b,\theta} \sum_{i=1}^{m} \left[1 - y_i (b + \sum_{j=1}^{\infty} \theta_j h_j(x_i)) \right]_+ + C \sum_{j=1}^{\infty} \frac{\theta^2}{\lambda_j}$$

Kernels y la función de estimación (2)

- La teoría garantiza que existe una solución finito dimensional de la forma

$$f(x) = b + \sum_{i=1}^{m} \alpha_i K(x, x_i)$$

- En particular existe un criterio de optimización equivalente

$$\min_{\alpha} \sum_{i=1}^{m} (1 - y_i f(x_i))_+ + C\alpha^T K \alpha$$

- Donde K es una matriz $m \times m$ con entradas $K_{ij} = K(x_i, x_j)$
- Estos modelos se pueden expresar de manera más general

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{m} (1 - y_i f(x_i))_+ + CJ(f)$$

- Donde $\mathcal{H}$ es el espacio de funciones y $J(f)$ es el regularizador del espacio

SVM y la maldición de la dimensionalidad

TABLE 12.2. *Skin of the orange: shown are mean (standard error of the mean) of the test error over five simulations. BRUTO fits an additive spline model adaptively, while MARS fits a low-order interaction model adaptively.*

	Method	Test Error (SE)	
		No Noise Features	Six Noise Features
1	SV Classifier	0.423 (0.006)	0.466 (0.008)
2	SVM/poly 2	0.081 (0.016)	0.172 (0.015)
3	SVM/poly 5	0.212 (0.008)	0.393 (0.004)
4	SVM/poly 10	0.265 (0.011)	0.438 (0.006)
5	BRUTO	0.082 (0.009)	0.080 (0.009)
6	MARS	0.138 (0.016)	0.116 (0.004)
	Bayes	0.061	0.061

Temario

Kernel ridge regression

- Ridge regression + Kernels $\Rrightarrow$ Proceso Gaussiano
- Necesitamos encontrar una función de regresión lineal $f(x) = \langle w, x \rangle$ que entrene la muestra S, donde las etiquetas son numeros reales, es decir, $y_i \in \mathbb{R}$.
- La calidad del modelo se mide por el cuadrado de las desviaciones $y_i - \langle w, x_i \rangle$ junto con tratar de mantener la norma de la función lo más pequeña posible.

Kernel ridge regression (2)

- El resultado se describe en el siguiente problema de optimización

$$\left.\begin{array}{cc} \text{Minimizar} & \lambda||w||^2 + \sum_{i=1}^m \xi_i^2 \\ \text{S.A:} & y_i - \langle w, x_i \rangle = \xi_i, i = 1, \ldots, m \end{array}\right\} \quad (7)$$

- Del cual se deriva el Lagrangiano

$$\text{Minimizar} L(w, \xi, \alpha) = \lambda||w||^2 + \sum_{i=1}^m \xi_i^2 + \sum_{i=1}^m \alpha_i (y_i - \langle w, x_i \rangle - \xi_i)$$

- Derivando tenemos que

$$w = \frac{1}{2\lambda \sum_{i=1}^m \alpha_i x_i} \text{ y } \xi_i = \frac{\alpha_i}{2}$$

Kernel ridge regression (3)

- Reemplazando estas relaciones obtenemos el dual

$$\text{Maximizar } W(\alpha) = \sum_{i=1}^{m} y_i \alpha_i - \frac{1}{4\lambda} \sum_{i,j}^{m} \alpha_i \alpha_j \langle x_i, x_j \rangle - \frac{1}{4} \sum \alpha_i^2$$

- Reescribiendo de manera vectorial tenemos

$$W(\alpha) = y^T \alpha - \frac{1}{4\lambda} \alpha^T K \alpha - \frac{1}{4} \alpha^T \alpha$$

- Donde K es la matriz Gram $K_{ij} = \langle x_i, x_j \rangle$ o la matriz de Kernels $K_{ij} = K(x_i, x_j)$ si trabajamos en el espacio característico

Kernel ridge regression (4)

- Derivando respecto de α la siguiente condición

$$-\frac{1}{2\lambda}K\alpha - \frac{1}{2}\alpha + y = 0$$

- Obteniendo la solución

$$\alpha = 2\lambda(K + \lambda I)^{-1}y$$

- Y la función de regresión

$$f(x) = \langle w, x \rangle = y^T(K + \lambda I)^{-1}k$$

- Donde k es el vector con entradas $k_i = \langle x_i, x \rangle, i = 1, \ldots, m$

Proposición

- Supongamos que queremos hacer una regresión sobre la muestra de entrenamiento S
- Usando el espacio característico definido implícitamente por el kernel $K(x,z)$, y sea $f(x) = y^T(K+\lambda I)^{-1}k$ donde K es una matriz $m \times m$ con entradas $K_{ij} = K(x_i, x_j)$ y k es el vector con entradas $k_i = K(x_i, x)$.
- Entonces la función $f(x)$ es equivalente al hiperplano en el espacio característico implícitamente definido por el kernel $K(x,z)$ que resuelve el problema de optimización ridge regression (7)

Temario

Kernel PCA y CCA

Veremos dos técnicas que pueden ser adaptadas para ser usadas con Kernels

- Análisis de componentes principales (PCA), para reducir la dimensión de los datos.
- Análisis de correlación Canónica (CCA), que ayuda a encontrar correlaciones entre los datos formados por dos pares de vectores o entre dos dataset cuyos elementos forman una biyección.

Análisis de componentes principales (PCA)

- Es una técnica clásica para analizar data de alta dimensión $\Rrightarrow$ extraer su estructura oculta $\Rrightarrow$ Encontrar un subconjunto (pequeño) de coordenadas que describen la mayor parte de la información de los datos
- Se puede demostrar que las direcciones que proveen más información de los datos son los k vectores propios principales de los datos $\Rrightarrow$ Para un k fijo, minimizan la distancia cuadrática entre los datos originales y los datos transformados.
- Los vectores propios se llaman ejes principales de los datos y las nuevas coordenadas de cada punto se obtienen mediante la proyección de este en los k ejes principales.

Análisis de componentes principales (PCA)

- Como antes, representábamos los vectores en la forma dual, como combinaciones de los datos $v = \sum_i \alpha_i x_i$ y necesitamos encontrar los parámetros α_i
- Dado un conjunto de datos no etiquetados $S = \{x_1, \ldots, x_m\}$ que están centrados $\sum_i x_i = 0$
- La matriz de covarianza empírica se define como

$$C = \frac{1}{m-1} \sum_i x_i x_i^T$$

- Y es una matriz semi definida positiva.

Análisis de componentes principales (PCA) (2)

- Sus vectores y valores propios pueden ser escritos como

$$\lambda v = Cv = \sum_i \langle x_i, v \rangle x_i$$

- Cada vector propio puede escribirse como combinación lineal de los ejemplos de entrenamiento

$$v = \sum_i \alpha_i x_i$$

- Para algún α y así permitir una representación dual que utilice kernels.
- Usando los k primeros vectores propios genera la mejor aproximación que minimiza la suma de las 2-normas de los residuos de los ejemplos de entrenamiento.

Análisis de componentes principales (PCA) (3)

- Realizando la misma operación sobre el espacio característico $\Rightarrow$ usar imágenes de los puntos $\phi(x_i)$, con simples manipulaciones podemos encontrar los coeficientes α^n de los n vectores propios pueden obtenerse al resolver el problema de valores propios

$$m\lambda\alpha = K\alpha$$

- E imponiendo la normalización $1 = \lambda_n \langle \alpha^n, \alpha^n \rangle, n = 1, \ldots, n$

Análisis de componentes principales (PCA) (3)

- Aunque no tengamos las coordenados explícitas de los vectores propios, podemos calcular las proyecciones de los ejemplos sobre los n vectores propios v^n de la siguiente manera

$$\langle \phi(x), v^n \rangle = \sum_{i=1}^{m} \alpha_i^n K(x_i, x)$$

- Esta información es todo lo que necesitamos de nuestros datos para extraer regularidades.

Análisis de correlación Canónica (CCA)

- Asumamos una biyección entre los elementos de dos conjuntos, que posiblemente corresponden a diferentes descripciones del mismo objeto (Ej: dos vistas del mismo objeto tridimensional, o dos versiones del mismo documento en idiomas distintos)
- Dado dos pares $S = \{(x^1, x^2)_i\}$ CCA busca combinaciones lineales de variables en cada conjunto que con máxima correlación

$$r = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2 \sum_i b_i^2}}$$

- Donde $\{a_i\}$ y $\{b_i\}$ son realizaciones de la variable aleatoria con media cero

Análisis de correlación Canónica (CCA) (2)

- Dado dos conjuntos de vectores $x_i^1 \in X_1$ y $x_i^2 \in X_2, i = 1, \ldots, m$
- Encontraremos vectores $w^1 \in X_1$ y $w^2 \in X_2$ tal que la proyección de los datos en esos vectores tenga máxima correlación
- Resolver esto es equivalente a transformar el problema en

$$\begin{bmatrix} 0 & C_{12} \\ C_{21} & 0 \end{bmatrix} \begin{bmatrix} w^1 \\ w^2 \end{bmatrix} = \lambda \begin{bmatrix} C_{11} & 0 \\ 0 & C_{22} \end{bmatrix} \begin{bmatrix} w^1 \\ w^2 \end{bmatrix}$$

- Donde

$$C_{jk} = \sum_{i=1}^{m} x_i^j (x_i^k)^T, j, k = 1, 2$$

Análisis de correlación Canónica (CCA) (2)

- Se pueden introducir kernels usando los mismos procedimientos vistos anteriormente
- $w^1 = \sum_i \alpha_i^1 \phi(x_i^1)$ y $w^2 = \sum_i \alpha_i^2 \phi(x_i^2)$ lo que conduce al problema dual en α

$$\begin{bmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{bmatrix} \begin{bmatrix} \alpha^1 \\ \alpha^2 \end{bmatrix} = \lambda \begin{bmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{bmatrix} \begin{bmatrix} \alpha^1 \\ \alpha^2 \end{bmatrix}$$

- Donde K_1 y K_2 son las matrices de kernel para los vectores $x_i^1 \in X_1$ y $x_i^2 \in X_2, i = 1, \ldots, m$ asumiendo que están en biyección, esto es, la entrada ij en cada matriz corresponde al mismo par de puntos.
- Resolviendo este problema podemos encontrar transformaciones no-lineales de los datos que maximicen la correlación entre ellos.

Temario

Support vector regression

- Recordemos el modelo de regresión

$$f(x) = x^T \beta + \beta_0$$

- Para estimar β consideremos la generalización no lineal

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}||\beta||^2$$

- Donde

$$V_\varepsilon(t) = \begin{cases} 0 & \text{si } |t| < \varepsilon \\ |t| - \varepsilon & \text{e.t.o.c.} \end{cases}$$

Support vector regression (2)

- Medida de error ε-no sensible $\Rightarrow$ ignorar pequeños residuos
- Estimación robusta

$$v_H(r) = \begin{cases} r^2/2 & \text{si } r \leq c \\ c|r| - c^2/2 & \text{si } r > c \end{cases}$$

- Función de Huber menos sensible a outliers

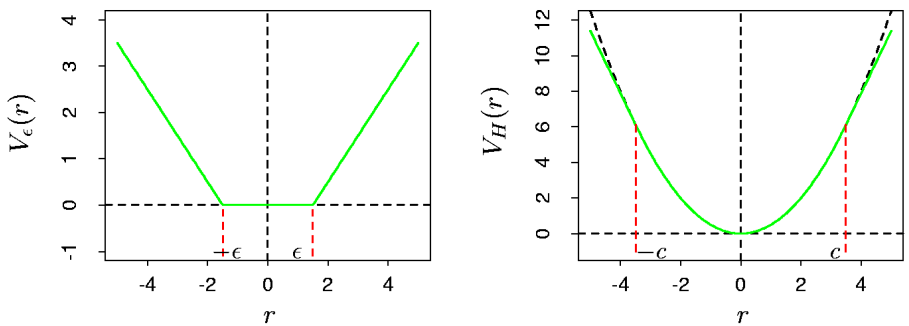


FIGURE 12.6. *The left panel shows the ϵ-insensitive error function used by the support vector regression machine. The right panel shows the error function used in Huber's robust regression (green curve). Beyond $|c|$, the function changes from quadratic to linear.*

Support vector regression (3)

- Si $\widehat{\beta}, \widehat{\beta}_0$ minimizan H la solución es de la forma

$$\widehat{\beta} = \sum_{i=1}^{N} (\widehat{\alpha}_i^* - \widehat{\alpha}_i) x_i \tag{8}$$

$$\widehat{f}(x) = \sum_{i=1}^{N} (\widehat{\alpha}_i^* - \widehat{\alpha}_i) \langle x, x_i \rangle + \beta_0 \tag{9}$$

- Donde $\alpha_i, \widehat{\alpha}_i^*$ son positivos y resuelven el problema de programación cuadrática

$$\min_{\alpha_i, \widehat{\alpha}_i^*} \varepsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{N} y_i (\alpha_i^* + \alpha_i) + \sum_{i,j}^{N} (\alpha_i^* + \alpha_i)(\alpha_i^* + \alpha_i) \langle x_i, x_j \rangle$$

$$\text{S.A. } 0 \leq \alpha_i, \alpha_i^* \leq 1/\lambda$$

$$\sum_{i=1}^{N} (\alpha_i^* + \alpha_i) = 0$$

$$\alpha_i \alpha_i^* = 0$$

- Usar $V_\varepsilon(r/\sigma)$

Regresión y kernels

- Supongamos que aproximamos una función de regresión en términos de funciones base $\{\phi_m(x)\}, m = 1, 2, \ldots, M$

$$f(x) \sum_{m=1}^{M} \beta_m \phi_m(x) + \beta_0$$

- Para estimar β y β_0 minimizamos

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \sum \beta_m^2$$

- Para una medida de error general $V(r)$ la solución de $\widehat{f}(x) = \sum \beta_m \phi_m(x) + \beta_0$ tiene la forma

$$\widehat{f}(x) = \sum_{i=1}^{N} \widehat{a}_i K(x, x_i)$$

- Donde $K(x, y) = \sum_{m=1}^{M} \phi_m(x) \phi_m(y)$

Regresión y kernels (2)

- Por ejemplo, para $V(r) = r^2$. Sea H una matriz de funciones base de $n \times m$ donde el im - ésimo elemento $\phi_m(x_i)$ y supongamos que $M \gg N$
- Por simplicidad asumamos $\beta_0 = 0$, o que la constante es absorbida en ϕ
- Estimemos β minimizando el error cuadrático

$$H(\beta) = (y - H\beta)^T (y - H\beta) + \lambda ||\beta||^2$$

- La solución es

$$\widehat{y} = H\widehat{\beta}$$

- Donde $\widehat{\beta}$ está determinado por

$$-H^T (y - H\widehat{\beta}) + \lambda\widehat{\beta} = 0$$

Regresión y kernels (3)

-
- Sin embargo
$$H\widehat{\beta} = (HH^T + \lambda I)^{-1} HH^T y$$
- Donde la matriz HH^T de $N \times N$ consiste en $\{HH^T\}_{ij} = K(x_i, x_j)$
- Se puede probar que la predicción de un x arbitrario debe satisfacer
$$\begin{aligned} \widehat{f}(x) &= \phi(x)^T \widehat{\beta} \\ &= \sum_{i=1}^{N} \widehat{\alpha}_i K(x, x_i) \end{aligned}$$
- Donde $\widehat{\alpha} = (HH^T + I)^{-1} y$

Consultas y Comentarios