Aprendizaje automatizado

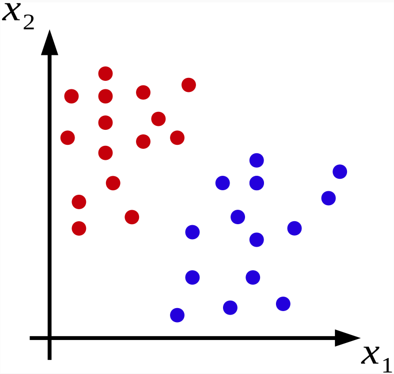
Máquinas de vectores de soporte y kernels

Gibran Fuentes-Pineda

Mayo-Junio 2021

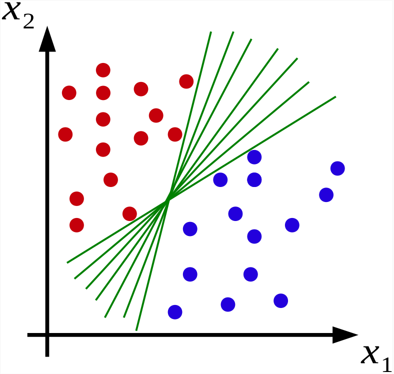
Caso 1: Clasificación binaria linealmente separable

- ¿Cómo separamos las clases?



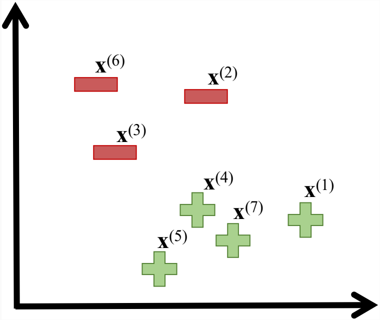
Caso 1: Clasificación binaria linealmente separable

- ¿Qué hiperplano elegimos?

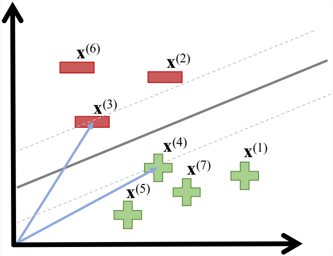


Clasificadores de margen máximo (1)

- El del margen más grande: hiperplanos paralelos a región de decisión que pasan por datos se llaman *vectores de soporte*



Clasificadores de margen máximo (2)



- Consideremos la frontera de decisión generada por $\mathbf{w}$ y una constante c. Dado un punto $\mathbf{x}^{(i)}$, la regla de decisión está definida por

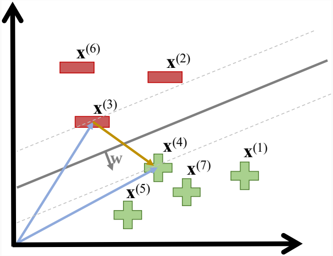
$$\mathbf{w}^\top \mathbf{x}^{(i)} \geq c$$

- La cual podemos reescribir como ($b = -c$)

$$\mathbf{w}^\top \mathbf{x}^{(i)} + b \geq 0$$

- $\mathbf{w}$ es perpendicular a la frontera de decisión

Clasificadores de margen máximo (3)



- Restricciones

$$\mathbf{w}^\top \mathbf{x}^{(i)} + b \geq 1, \text{ si } y^{(i)} = 1$$
$$\mathbf{w}^\top \mathbf{x}^{(i)} + b \geq -1, \text{ si } y^{(i)} = -1$$

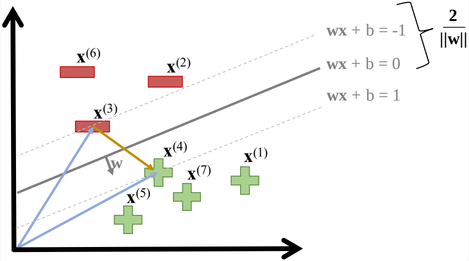
- Sea $y^{(i)} = 1$ para positivos y $y^{(i)} = -1$ para negativos, podemos reescribir

$$y^{(i)} \cdot (\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1$$
$$y^{(i)} \cdot (\mathbf{w}^\top \mathbf{x}^{(i)} + b) - 1 \geq 0$$

- Si $\mathbf{x}^{(i)}$ está exactamente en los hiperplanos de soporte

$$y^{(i)} \cdot (\mathbf{w}^\top \mathbf{x}^{(i)} + b) - 1 = 0$$

Clasificadores de margen máximo (4)



- Dados $x_{pos}^{(i)}$ (ejemplo positivo) y $x_{neg}^{(j)}$ (ejemplo negativo), el margen se puede calcular como

$$\frac{w^\top}{||w||} \cdot (x_{pos}^{(i)} - x_{neg}^{(j)}) = \frac{2}{||w||}$$

- Queremos encontrar la **w** que maximice el ancho o de forma equivalente minimizar

$$\frac{||w||}{2}$$

Optimización con restricciones

- Podemos convertir el problema a una optimización con restricciones

$$\begin{aligned} \text{minimiza } & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{sujeto a } & y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 \end{aligned}$$

donde $y^{(i)} \in \{-1, +1\}$

Optimización con restricciones

- Podemos convertir el problema a una optimización con restricciones

$$\begin{aligned} &\text{minimiza } \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{sujeto a } y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 \end{aligned}$$

  donde $y^{(i)} \in \{-1, +1\}$

- Optimización cuadrática con restricciones lineales y estrictamente convexa con solución única para problemas linealmente separables

Caso 2: No linealmente separables

- Penalizando suavemente clasificacionnes erróneas a través de *variables flojas*, $\xi_i \geq 0, i = 1, \ldots, n$

$$\text{minimiza } C \sum_{i=1}^{n} \xi_i + \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{sujeto a } y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 - \xi_i, i = 1, \ldots, n$$

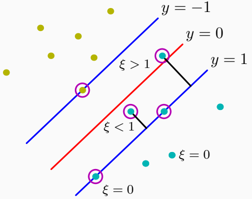


Imagen tomada de Bishop, PRML 2007

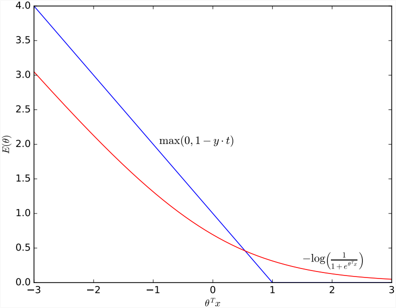
- $\xi^{(i)} = 0$, si están del lado correcto
- $\xi^{(i)} = |y^{(i)} - (\mathbf{w}^\top \mathbf{x}^{(i)} + b)|$ para otros puntos

Función de pérdida bisagra

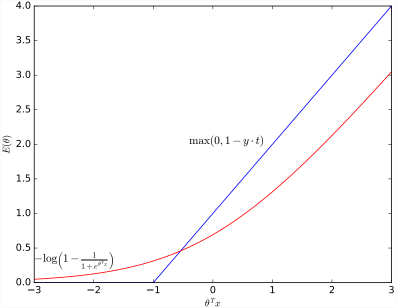
- Error respecto a parámetros está dado por función bisagra

$$B(\hat{y}, y) = \text{máx}(0, 1 - \hat{y} \cdot y)$$

$y = 1$



$y = -1$



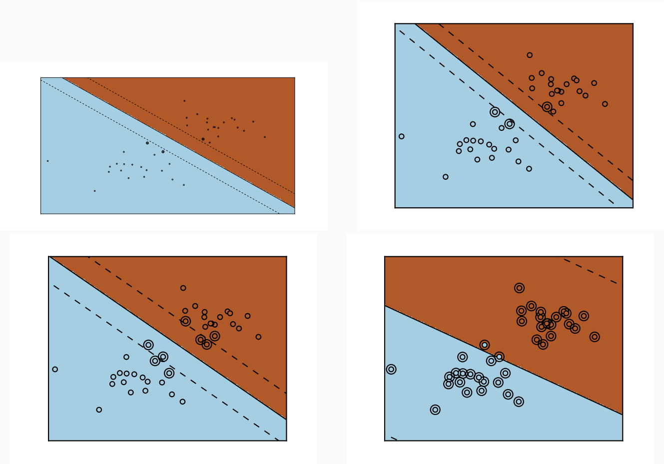
Encontrando el clasificador margen máximo

- El problema de optimización

$$\min_{\mathbf{w},b}\left[C\cdot\sum_{i=1}^{N}B(\hat{y}_i,y^{(i)})+\frac{1}{2}\|\mathbf{w}\|^2\right]$$

Caso 2: No linealmente separables

- Clasificación con diferentes valores de C



Representación dual

- Reformulación para tener espacio de entrada dado por producto punto de entrada
- Problema de optimización

$$\text{maximiza } \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \mathbf{x}^{(i)\top} \mathbf{x}^{(j)}$$

$$\text{sujeto a } 0 \leq \alpha_i \leq C, \sum_{i=1}^{n} \alpha_i y^{(i)} = 0, \forall i$$

- Para predecir la clase de una nueva instancia $\tilde{\mathbf{x}}$

$$\tilde{y} = \left(\sum_{i=1}^{n} \alpha_i y^{(i)} k(\mathbf{x}^{(i)}, \tilde{\mathbf{x}}) + b \right)$$

¿Qué es una función de kernel?

- Función evaluada en los reales $k(\mathrm{x}^{(i)}, \mathrm{x}^{(j)}) \in \mathbb{R}$
    - Simétrica: $k(\mathrm{x}^{(i)}, \mathrm{x}^{(j)}) = k(\mathrm{x}^{(j)}, \mathrm{x}^{(i)})$
    - No negativa: $k(\mathrm{x}^{(i)}, \mathrm{x}^{(j)}) \geq 0$

¿Qué es una función de kernel?

- Función evaluada en los reales $k(\mathrm{x}^{(i)}, \mathrm{x}^{(j)}) \in \mathbb{R}$
  - Simétrica: $k(\mathrm{x}^{(i)}, \mathrm{x}^{(j)}) = k(\mathrm{x}^{(j)}, \mathrm{x}^{(i)})$
  - No negativa: $k(\mathrm{x}^{(i)}, \mathrm{x}^{(j)}) \geq 0$
- Puede ser vista como una medida de similitud (aunque no necesariamente debe ser una)

¿Qué es una función de kernel?

- Función evaluada en los reales $k(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}) \in \mathbb{R}$
  - Simétrica: $k(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}) = k(\mathsf{x}^{(j)}, \mathsf{x}^{(i)})$
  - No negativa: $k(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}) \geq 0$

- Puede ser vista como una medida de similitud (aunque no necesariamente debe ser una)

- Para mapeos a espacios no lineales $\phi(\mathsf{x}^{(i)})$, el kernel está dado por
$$k(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}) = \phi(\mathsf{x})^{(i)\top}\phi(\mathsf{x}^{(j)})$$

Ejemplos de funciones de kernel

- Lineal

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathbf{x}^{(i)\top}\mathbf{x}^{(j)}$$

Ejemplos de funciones de kernel

- Lineal

$$k(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}) = \mathsf{x}^{(i)\top}\mathsf{x}^{(j)}$$

- Gaussiana

$$k(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}) = \exp\left(-\frac{1}{2}(\mathsf{x}^{(i)} - \mathsf{x}^{(j)})^\top \Sigma^{-1}(\mathsf{x}^{(i)} - \mathsf{x}^{(j)})\right)$$

Ejemplos de funciones de kernel

- Lineal

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathbf{x}^{(i)\top}\mathbf{x}^{(j)}$$

- Gaussiana

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

- Función de base radial (RBF)

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp\left(-\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_2^2}{2\sigma^2}\right)$$

Ejemplos de funciones de kernel

- Lineal

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathbf{x}^{(i)\top}\mathbf{x}^{(j)}$$

- Gaussiana

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

- Función de base radial (RBF)

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp\left(-\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_2^2}{2\sigma^2}\right)$$

- Similitud coseno

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \frac{\mathbf{x}^{(i)\top}\mathbf{x}^{(j)}}{\|\mathbf{x}^{(i)}\| \cdot \|\mathbf{x}^{(j)}\|}$$

El truco del kernel

- Proyectamos el espacio de entrada a un espacio de más alta dimensionalidad en la que sea posible separar las clases linealmente
- Muchos algoritmos se pueden *kernelizar* usando la representación dual
    - Substituímos producto punto en representación dual por una llamada a un kernel
- Es necesario definir funciones de kernel válidas
    - Elegir un mapeo $\phi(\mathbf{x}^{(i)})$ y definir el kernel en base a este.
    - Definir directamente funciones, sin conocer $\phi(\mathbf{x}^{(i)})$
    - Ciertas operaciones sobre funciones válidas producen otras funciones válidas

Kernels positivos definidos (Mercer)

- Si la matriz de Gram es positiva definida, se conoce como kernel de Mercer

$$\mathsf{K} = \begin{pmatrix} k(\mathsf{x}^{(1)}, \mathsf{x}^{(1)}) & \cdots & k(\mathsf{x}^{(1)}, \mathsf{x}^{(n)}) \\ & \vdots & \\ k(\mathsf{x}^{(n)}, \mathsf{x}^{(1)}) & \cdots & k(\mathsf{x}^{(n)}, \mathsf{x}^{(n)}) \end{pmatrix}$$

- La eigendescomposición de K está dada por

$$\mathsf{K} = \mathsf{U}^\top \Lambda \mathsf{U}$$

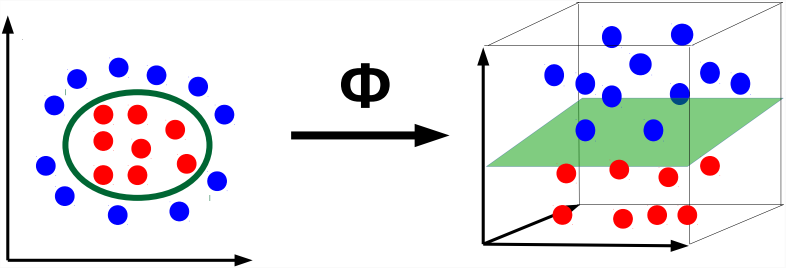
donde

$$k_{ij} = \left(\Lambda^{\frac{1}{2}} \mathsf{U}_{:\,,i}\right)^\top \left(\Lambda^{\frac{1}{2}} \mathsf{U}_{:\,,j}\right)$$

- Si un kernel es Mercer, existe un mapeo $\phi(\mathsf{x}^{(i)})$ tal que

$$k(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}) = \phi(\mathsf{x})^{(i)\top} \phi(\mathsf{x}^{(j)})$$

Intuición de clasificación con kernels



SVM con kernel lineal

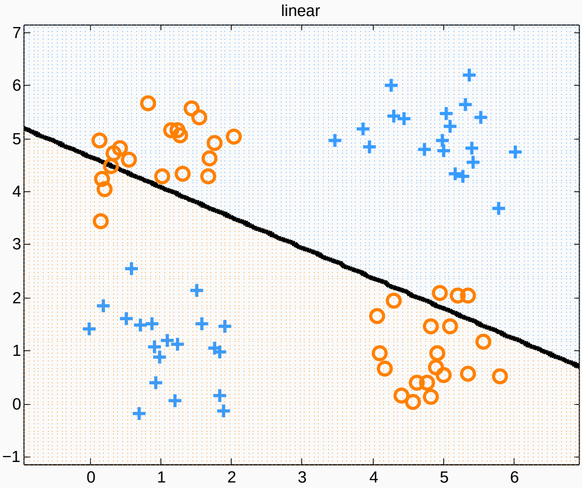


Imagen generada usando ejemplo de `https://github.com/probml/pmtk3`

SVM con función de base radial

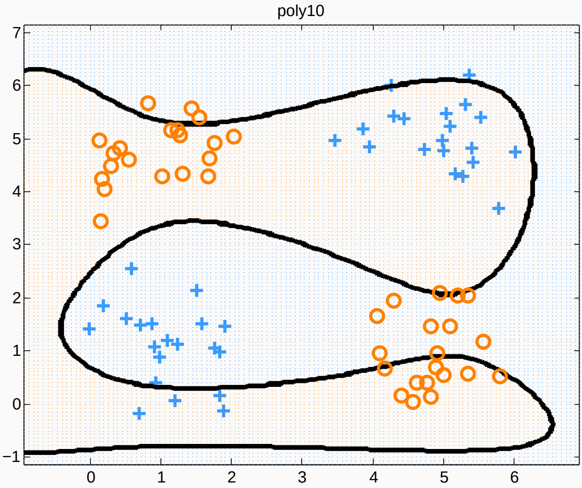


Imagen generada usando ejemplo de `https://github.com/probml/pmtk3`

SVM con kernel polinomial

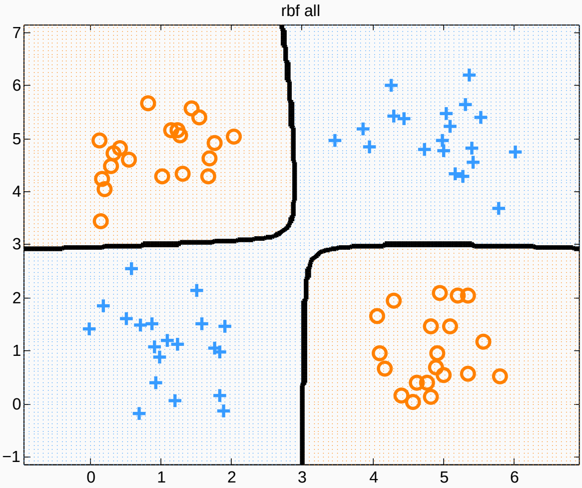


Imagen generada usando ejemplo de `https://github.com/probml/pmtk3`

Máquinas de vectores de soporte para regresión

- Extensión que preserva dispersidad en datos para regresión

Máquinas de vectores de soporte para regresión

- Extensión que preserva dispersidad en datos para regresión

- Usa función de pérdida ϵ-*sensible*

$$E(\hat{y}^{(i)}, y^{(i)}) = \begin{cases} 0 & \text{si } |\hat{y}^{(i)} - y^{(i)}| < \epsilon \\ |\hat{y}^{(i)} - y^{(i)}| - \epsilon & \text{en caso contrario} \end{cases}$$

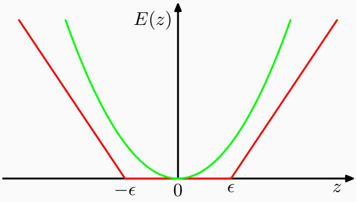


Imagen tomada de Bishop, PRML 2006

Problema de optimización para regresión

- Se busca resolver

$$\min_{\mathbf{w},b}\left[C\sum_{i=1}^{n}E(\hat{y}^{(i)},y^{(i)})+\frac{1}{2}\|\mathbf{w}\|^2\right]$$

- Expresado con variables flojas ξ

$$\min_{\mathbf{w},b}\left[C\sum_{i=1}^{n}(\xi^{(i)}+\hat{\xi}^{(i)})+\frac{1}{2}\|\mathbf{w}\|^2\right]$$

$$\text{sujeto a } \hat{y}^{(i)}+\epsilon+\xi^{(i)}\geq y^{(i)}$$

$$\hat{y}^{(i)}-\epsilon-\hat{\xi}^{(i)}\leq y^{(i)}$$

Tubo ϵ

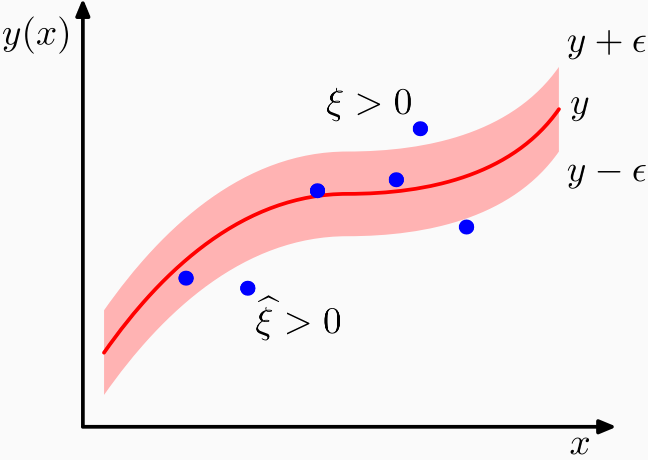


Imagen tomada de Bishop, PRML 2006

Algoritmo de optimización mínima secuencial (SMO)

- Divide el problema de optimización en una serie de subproblemas mínimos (con 2 multiplicadores de Lagrange debido a las restricciones)

- Es posible resolver cada subproblema de forma analítica

$$0 \leq \alpha_1, \alpha_2 \leq C$$
$$y^{(1)} \cdot \alpha_1 + y^{(2)} \cdot \alpha_2 = k$$

donde k es el negativo de la suma del resto de los términos de la restricción de igualdad

Algoritmo de descenso por subgradiente (PEGASOS)

- La función bisagra no es diferenciable
- Podemos usar el subgradiente

$$\tilde{\nabla} E(\mathbf{w}, b) = \begin{cases} 0, & y^i \cdot (\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 \\ y^i \cdot \mathbf{x}^{(i)}, & y^i \cdot (\mathbf{w}^\top \mathbf{x}^{(i)} + b) < 1 \end{cases}$$