TEMA 7: CLASIFICADORES LINEALES ÓPTIMOS

MÁRGENES

+ En casos lin. separables → ¿Podemos preferir cierta h^*?
  * Si elegimos h: ↑↑ margen → intuitivamente, seremos menos subceptibles al "ruido" en los datos. → ↓↓ E_{out}
+ Por tanto, buscaremos h: tenga el mayor margen posible.

CÁLCULO DEL HIPERPLANO SEPARADOR CON MÁRGEN MÁX.

PASOS PRELIMINARES

+ Normalización de w: Si multiplicamos todos los valores en w por una cte. "k" ⟶ Nos quedamos con el mismo plano.
  * Por simplicidad, buscaremos el w tal que $|w^T x_n| = 1$, siendo x_n el ejemplo más cercano al plano: $\forall x_i \in D \to |w^T x_i| \geq 1$.
+ Suponemos que los ejemplos son lin. separables: $\underbrace{y(w^T x_n)}_{\text{signo}(y) = \text{signo}(w^T x_n)} \geq 1$.
+ Cambiamos w_0 por b → Plano: $w^T x + b = 0$.

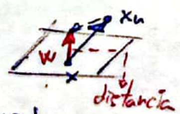
DISTANCIA PUNTO - PLANO. CÁLCULO DEL MÁRGEN.

+ El vector w es perpendicular al plano w.
  * Si tomamos 2 puntos del plano: x' y x'':

$$\left.\begin{matrix} w^T x' + b = 0 \\ w^T x'' + b = 0 \end{matrix}\right\} \; w^T \underbrace{(x' - x'')}_{\text{vector sobre el plano}} = 0 \to w^T \perp \overrightarrow{x'x''}$$

+ Distancia de x_n al plano:

$$\text{distancia}(w, x_n) = \underbrace{|\hat{w}^T (x_n - x)|}_{\text{Proyección de } \overrightarrow{xx_n} \text{ en } \vec{w}}, \quad \hat{w}^T = \frac{w}{\|w\|} \text{ normalizado (tam}=1).$$



$$\text{dist}(w, x_n) = \frac{1}{\|w\|} |w^T x_n - w^T x| = \frac{1}{\|w\|} |\underbrace{w^T x_n + b}_{\text{Punto a dist. 1}} - \underbrace{(w^T x + b)}_{\text{Punto en plano} \to = 0}|.$$

$$\boxed{\min(\text{dist}(w, x_n)) = \frac{1}{\|w\|}} \to \text{Margen de } w.$$

OPTIMIZACIÓN DEL MARGEN

+ Tenemos que maximizar el margen $= \frac{1}{\|w\|} = \frac{1}{\sqrt{w^T w}}$.
+ Este problema se traduce en minimizar $w^T w$.
  * Tenemos restricciones: $\left.\begin{matrix} y_n \underset{(+b)}{(w^T x_n)} \geq 1 \\ \min(y_n(w^T x_n + b)) = 1 \end{matrix}\right\}$ - Ejemplos bien clasificados - Mínima valoración $= 1$.

→ Podemos resolver el problema con "programación cuadrática".

* P.C: $\left\{\begin{matrix} \text{- Resuelve } \min_{u \in \mathbb{R}^d} \frac{1}{2} u^T Q u + p^T u \\ \text{- Con restricciones: } Au \geq c. \end{matrix}\right.$

* En nuestro caso
  - Minimizamos $\frac{1}{2} w^T Q w + p^T w$ — Q: Hacemos que sea $\approx I$ — p: No necesitamos esto, así que $p = 0^d$.
  - Sujeto a: $A \cdot w \geq C \to C = 1$

$$A = \begin{bmatrix} y_n^T x_n \end{bmatrix} \to Aw = y_n (w^T x_n) \quad \text{↳ Incluye } b.$$

→ Otra aproximación: Problema Dual de Lagrange.

* $$\underbrace{\mathcal{L}(w, b, \alpha)}_{\text{Lagrangiano}} = \frac{1}{2} w^T w - \sum_{n=1}^{N} \underbrace{\alpha_n}_{\text{Coeficientes de Lagrange.}} (y_n(w^T x_n + b) - 1)$$

* Minimizar $\mathcal{L}(w, b, \alpha)$ respecto a w y b, y después maximizar respecto a $\underset{(\alpha_n \geq 0)}{\alpha}$ resuelve nuestro problema.

1. Minimizamos $\left\{\begin{matrix} \nabla_w \mathcal{L} = w - \sum_{n=1}^{N} \alpha_n y_n x_n = 0 \\ \nabla_b \mathcal{L} = -\sum_{n=1}^{N} \alpha_n y_n = 0 \end{matrix}\right.$

   → Sustituyendo... $\mathcal{L}(\cancel{w, b,} \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m x_n^T x_m$

2. Maximizamos $\mathcal{L}(\alpha)$, con la restricción $\sum_{n=1} \alpha_n y_n = 0$. ($\nabla_b$). $\alpha_n \geq 0 \;\; \forall n$

* Para ello, utilizamos programación cuadrática.
* Pasamos $\mathcal{L}(\alpha)$ a representación matricial, y minimizamos $-\mathcal{L}(\alpha)$:

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - \underbrace{1^T \alpha}_{\to \sum_{n=1}^{N} \alpha_n}, \text{ sujeto a restricción } \underbrace{y^T \alpha}_{\sum \alpha_n y_n} = 0$$

$$0 < \alpha < \infty$$

$$Q = \begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \cdots \\ \vdots & \vdots & \end{bmatrix}$$

3. Calculamos, finalmente, w y b.

$$w^* = \sum_{n=1}^{\tilde{N}} \alpha_n y_n x_n = \sum_{\substack{n=1 \\ \text{no Vector de soporte}}}^{N} \alpha_n y_n x_n \dashrightarrow \text{Si no es vector de soporte, } \alpha = 0.$$

$$b^* = y_s - \sum y_n \alpha_n^* x_n^T x_n.$$

y_s de un punto tal que margen $= 1 \rightarrow$ Vector de soporte.

CONDICIONES DE LAS SOLUCIONES ÓPTIMAS (KKT).

* Son características que cumplen (u^*, α^*).
- u^* → Problema "primal"
- α^* → Problema dual.

1. Restricciones $\begin{cases} \text{- Primal: } a_m^T u^* \geq c_m. \\ \text{- Dual: } \alpha_m > 0. \end{cases}$

2. "Holgura" complementaria: $\alpha_m^* (\underbrace{y_m (x_m^T w^* + b^*) - 1}_{\text{Holgura (distancia)}}) = 0.$

* $dist(x, w) \geq 1$.
- Si es $= 1 \longrightarrow$ Parte derecha $= 0. \rightarrow \alpha_m^*$ puede ser $\neq 0$.
- Si es $> 1 \longrightarrow$ Parte derecha $> 0 \rightarrow \alpha_m^* = 0$.

* "Vector de soporte": aquellos x_n tales que $\begin{cases} dist(x, w) = 1 \\ \alpha_m^* > 0. \end{cases}$

3. Estacionaridad de $\underset{(w^T y\, b)}{u}$: $\nabla_u \mathcal{L}(u, \alpha) = 0$ si $u = u^*$ y $\alpha = \alpha^*$.

* Estas condiciones son las que garantizan que el problema dual es equivalente al primal.

TRANSFORMACIONES NO LINEALES

+ Podemos aplicar $\Phi(x) = z$, y simplemente utilizar z en vez de x.
+ ¿Cuánto perjudica esto en términos de eficiencia?
  → En vez de realizar $x_n^T x_n \rightarrow z_n^T z_n$ (producto escalar): no mucho, si z tiene una dimensión no muy grande.
+ Si visualizamos datos y separador → los vectores de soporte no tienen por qué estar a distancia cte. (sí lo están en Z-espacio).
+ ¿Cuánto afecta dim. Z en generalización? → Es INDEPENDIENTE...

GENERALIZACIÓN DEL MODELO

+ ¿Es tener un mayor margen realmente mejor?
  * Mientras que en regularización fijamos $w^T w < C$ y minim. E_{in}...
  * En SVM, fijamos $E_{in} = 0$ y minimizamos $w^T w$ (equiv. maximizar $\frac{1}{\|w\|}$).

* Cuanto mayor es el margen que exigimos → ↓ dicotomías son separables → ↓ d_{VC}. Concretamente: $d_{VC}(\underset{\text{Margen exigido}}{\rho}) = \min(\lceil R^2/\rho^2 \rceil, d) + 1$.

* Usando validación cruzada:
→ Si no quitamos un vector soporte, la solución no cambia: $e_n = 0$.
→ Si quitamos un vector soporte $0 \leq e_n \leq 1$.
* Por tanto, $E_{out}(SVM) \leq E_{cv}(SVM) = \frac{1}{N} \sum_{n=1}^{N} e_n \leq \frac{\#\text{vectores soporte}}{N}$.
→ Entonces, cuantos menos vectores soporte obtengamos, mejor será la generalización (¡independientemente de d_{VC}!).

EL TRUCO DEL KERNEL EN SVM DUAL

+ Estamos en el caso en que usamos $\Phi: X \rightarrow Z$ sobre x:

$$\mathcal{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{\tilde{N}} \sum_{m=1}^{\tilde{N}} y_n y_m \alpha_n \alpha_m \boxed{z_n^T z_m} \rightarrow \text{Producto escalar}$$

$$g(x) = \text{signo}(w^T z + b) = \text{signo}[(\sum \alpha_n y_n \boxed{z_n^T z}) + b]$$

$w = \sum \alpha_n y_n z_n$; $z_n^T z$ → Producto escalar

* Solo utilizamos z para realizar el producto escalar y obtener nº.
* ¿Somos capaces de calcular $z^T z$ sin calcular $\Phi(x)$?
→ $z_n^T z = K(x, x') \longleftarrow$ kernel (función que calcula, en algún espacio Z, $\Phi(x)^T \Phi(x')$).

+ Ejemplo: $K(x, x') = (1 + x^T x')^Q =$ (kernel polinómico).
$= (1 + x_1 x_1' + \ldots + x_d x_d')^Q =$ { si $Q = 100$, $d = 10$ → expansión enorme. coeficientes ctes.
$=$ producto escalar en Z de dimensión altísima. ¿Qué $\Phi(x)$ es?
↳ Nos da igual, existe

TEMA 7: CLASIFICADORES LINEALES ÓPTIMOS (2)

KERNEL GAUSSIANO - RBF (RADIAL BASIS FUNCTION)

+ $k(x,x') = e^{-\gamma \|x-x'\|^2}$
  γ → Parámetro ajustable.

* Ejemplo: $d=1$, $\gamma=1$ → $k(x,x') = e^{-(x-x')^2}$
  → Desarrollo de Taylor: $= \exp(-x^2)\exp(-x'^2) \underbrace{\sum_{k=0}^{\infty} \frac{2^k (x)^k (x')^k}{k!}}_{\exp(2xx')}$

* Podemos ver que { es $f(x)$; es $f(x')$; es cte.
* Si agrupamos, tenemos un producto escalar entre $f(x)$ y $f(x')$ de ∞ términos.
* Este kernel tiene dimensión infinita → $\Phi(x)$ es NLT infinita.

CONSTRUCCIÓN DE KERNELS

+ Podemos formar nuevos kernels combinando varios de ellos.
+ Si queremos saber si k es un kernel válido → $\boxed{\text{Condición de Mercer.}}$

Sii {
* $k(x,x') = k(x',x)$ (simetría).
* $K = \begin{bmatrix} k(x_1,x_1') & \cdots & k(x_1,x_n') \\ \vdots & & \vdots \\ k(x_n,x_1') & \cdots & k(x_n,x_n') \end{bmatrix}$ es positiva semidefinida. $(|K|>0)$.

SVM SUAVE

+ SVM, pero ahora toleramos errores. → ¿Cuánto?
  * Nuestro error ahora lo medimos como "violación del margen":

$$\xi_n \dashrightarrow y_n(w^T x_n + b) \geq 1 - \xi$$

"xi"

+ Minimizamos$_{b,w,\xi}$ $\frac{1}{2} w^Tw + \underbrace{C\sum_{n=1}^{N} \xi_n}_{\text{"regularizador"}}$, sujeto a $\begin{cases} y_n(w^Tx_n+b) \geq 1-\xi_n; \\ \xi_n \geq 0 \ \forall n \in [1,N] \end{cases}$

  "regularizador" { $C \to \infty$ ≈ Hard SVM; $C \to 0$ ≈ No tenemos error en cuenta.

* Este problema es muy similar a clasificación con regularización.
  → w^Tw funciona como reg. // → $C \cdot$ Error funciona como guía.
  Lo cambia de lugar la cte

DUAL DE SVM SUAVE

+ Usando Lagrangianos:

$$\mathcal{L}(w,b,\xi,\alpha,\beta) = \frac{1}{2} w^Tw + \underbrace{C\sum_{n=1}^{N} \xi}_{\text{Añadimos el error}} - \sum_{n=1}^{N} \alpha_n (y_n(w^Tx_n+b) - 1 + \xi_n) - \underbrace{\sum_{n=1}^{N} \beta_n \xi_n}_{\text{Como } \xi \text{ tiene restricciones, surge.}}$$

* Minimizamos con w, b, ξ; maximizamos con α, β.

$$\nabla_w \mathcal{L} = w - \sum_{n=1}^{N} \alpha_n y_n x_n = 0.$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0 \ ; \quad \frac{\partial \mathcal{L}}{\partial \xi} = C - \alpha_n - \beta_n = 0 \ ^{*}$$

* Usando la fórmula anterior, tenemos que $C\sum^{N}\xi - \alpha\sum^{N}\xi - \beta\sum^{N}\xi = 0$
  → $\mathcal{L}(w,b,\xi,\alpha,\beta) = \frac{1}{2} w^Tw - \sum_{n=1}^{N} \alpha_n (y_n(w^Tx_n+b)-1)$, con las mismas condiciones que con SVM duro, excepto: $0 \leq \alpha_n \leq C$ (es lo único nuevo).

+ Ahora, tenemos dos tipos de vectores de soporte:
  → En el margen { $(0 < \alpha_n^* < C)$; $y_n(w^Tx_n)+b = 1$ (no tienen error).
  → No en el margen { $\alpha_n^* = C$; $y_n(w^Tx_n)+b < 1$ (hay error y viola el margen).

+ En cuanto a b^*, pasa algo interesante:
  * Si solo hay un vector soporte en el margen → $b^* = 1 - y_s(w^Tx_s)$
  * Pueden haber, si hay más de uno, un rango de soluciones para b.

CONCLUSIONES SOBRE SVM SUAVE

+ Puede ser visto como un caso especial de clasif. regularizada { - $E_{svm}(b,w)$ - weight decay
+ Controla la complejidad del modelo maximizando el margen.
+ Puede lidiar con transformaciones infinitas/muy grandes → kernel.
+ Expresa la solución con { - Vectores de soporte - Multiplicadores de Lagrange - Kernel
+ Puede controlar la sensibilidad a errores con C.
* Si C y $k(x,x')$ se eligen correctamente → ↓↓ E_{out} y define uno de los mejores modelos de clasificación.