Auxiliar 2: máquinas de soporte vectorial

MA5204 Aprendizaje de Máquinas

Arie Wortsman, Nelson Moreno, Víctor Faragi,
Francisco Vásquez, Fernando Fêtis.

Departamento de ingeniería matemática
Universidad de Chile

1 de junio de 2021

UNIVERSIDAD
DE CHILE

Dualidad lagrangiana

Un problema de optimización (P) tiene estructura de Karush-Kuhn-Tucker (KKT) si es de la forma

$$
\begin{aligned}
(P) \quad & \min_{s.a} f(x) \\
& g(x) \leq 0 \\
& h(x) = 0 \\
& x \in \mathbb{R}^n
\end{aligned}
$$

Donde $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ y $h : \mathbb{R}^n \to \mathbb{R}^p$ son funciones diferenciables.

Definición (Lagrangiano)

Se define el lagrangiano del problema (P) anterior como:

$$L(x, \lambda, \mu) := f(x) + \lambda^\top g(x) + \mu^\top h(x)$$

Donde $\lambda \in \mathbb{R}^m$ y $\mu \in \mathbb{R}^p$ se denominan ***multiplicadores de Lagrange*** *o variables duales.*

Dualidad lagrangiana

De la definición anterior, es directo el siguiente resultado:

Lemma (dualidad débil)

Sea f^* el valor óptimo de (P) y $x \in \mathbb{R}^n$ un punto factible de (P). Entonces, para todo $\lambda \in \mathbb{R}^m_+$ y $\mu \in \mathbb{R}^p$, se tiene que:

$$L(x, \lambda, \mu) \leq f^*$$

Fijando las variables duales λ y μ se puede optimizar de **forma irrestricta** sobre el lagrangiano, obteniendo el **lagrangiano dual**:

$$\theta(\lambda, \mu) := \inf_x L(x, \lambda, \mu)$$

Dualidad lagrangiana

De este modo, se tiene el problema lagrangiano dual:

$$(D) \quad \max_{\lambda \geq 0} \theta(\lambda, \mu)$$

Del teorema de dualidad débil, el valor óptimo de (D) será cota inferior del valor óptimo de (P). Bajo ciertas condiciones, se puede hablar de **dualidad fuerte**. En este caso:

- valor$(P) =$ valor(D).
- si valor$(D) \leq \infty$, entonces $\exists(\overline{\lambda}, \overline{\mu})$ que lo minimiza.
- si $\overline{x}$ es el punto que minimiza f, entonces $\overline{\lambda}_i g_i(\overline{x}) = 0$ (holgura complementaria).

SVM: motivación

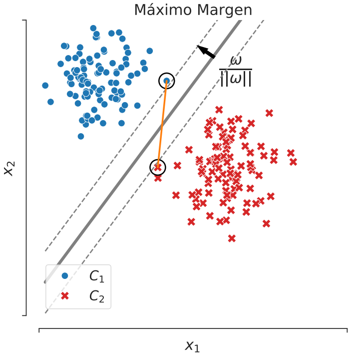
Se busca un hiperplano separador $w^\top x + b = 0$ que maximice el margen entre ambas clases (clase -1 y clase 1). Se tendrá en cuenta lo siguiente:

- Existen datos que limitan el margen (y la rotación). Dichos datos se llamarán **vectores soporte**.
- Para dichos vectores soporte, se impondrá que:

$$
\begin{aligned}
w^\top x_+ + b = 1, &\quad \forall x_+ \in C_1 \text{ vector soporte} \\
w^\top x_- + b = -1, &\quad \forall x_- \in C_{-1} \text{ vector soporte}
\end{aligned}
$$

- Dado que se busca un margen $m > 0$, será necesario que los datos seran **estrictamente separables** (mediante un hiperplano).

SVM: formulación primal



Sean $x_+ \in C_1$ y $x_- \in C_{-1}$ vectores soporte, luego:

$$m = \frac{1}{2}\,\|\mathrm{proy}_{\mathrm{w}}(\mathrm{x}_+ - \mathrm{x}_-)\| = \frac{1}{2}\frac{w^\top(x_+ - x_-)}{\|w\|}$$

Donde $w^\top(x_+ - x_-) = ((1-b) - (-1-b)) = 2.$

Luego:

$$m = \frac{1}{||w||}$$

Para evitar problemas de diferenciabilidad, en vez de maximizar m, se minimizará el siguiente problema equivalente:

$$\begin{aligned}(P) \quad & \min_{w,b} \quad \frac{1}{2}||w||^2 \\ & \text{s.a} \quad y_i(w^\top x_i + b) \geq 1,\ i \in \{1,\ldots,N\}\end{aligned}$$

SVM: formulación dual

El problema de optimización (P) cumple la restricción de Slater por lo que hay dualidad fuerte.

En este caso, el **lagrangiano** es el siguiente:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 + \sum_{i=1}^{N} \alpha_i \left(1 - y_i(w^\top x_i + b)\right)$$

Para el **lagrangiano dual** $\theta(\alpha) = \inf_{w,b} L(w, b, \alpha)$ se utiliza CPO (ya que L es convexo):

- $\frac{\partial L}{\partial w} = w^\top - \sum_{i=1}^{N} \alpha_i y_i x_i^\top = 0 \implies \overline{w} = \sum_{i=1}^{N} \alpha_i y_i x_i$.
- $\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \implies \sum_{i=1}^{N} \alpha_i y_i = 0$.

Sustituyendo:

$$\theta(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

Observación: falta encontrar el parámetro óptimo $\overline{b}$.

SVM: formulación dual

Luego, el **problema dual** $(D)\, \max_{\alpha \geq 0} \theta(\alpha)$ corresponde a:

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
\text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\
& \alpha_i \geq 0
\end{aligned}
$$

Por dualidad fuerte, si $\overline{\alpha}$ es solución de (D), entonces se tiene que $\overline{w} = \sum_{i=1}^{N} \overline{\alpha}_i y_i x_i$ es solución para (P).

Para el parámetro óptimo $\overline{b}$:

- $a := \min_{i:y_i=1} \overline{w}^\top x_i \to$ distancia a la muestra más cercana (clase 1).
- $b := \max_{i:y_i=-1} \overline{w}^\top x_i \to$ distancia a la muestra más cercana (clase -1).

Luego, se toma $\overline{b} = -\frac{a+b}{2}$.

SVM: predicción

Dado que se utiliza el clasificador $\overline{w}^\top x + \overline{b} = 0$, para una nueva observación x_0, se tiene que su clase predicha es:

$$\hat{y}_0 = \operatorname{sign}\left(\sum_{i=1}^{N} \overline{\alpha_i} y_i \langle x_i, x_0 \rangle + \overline{b}\right)$$

Donde $\overline{\alpha}$ representa la solución del problema dual.

Por otra parte, debido al teorema de holgura complementaria:

$$\alpha_i \left(1 - y_i \left(\overline{w}^\top x_i + \overline{b}\right)\right) = 0, \quad \forall i \in \{1, \dots, N\}$$

Por lo tanto, si x_i no es vector soporte (i.e., no está en el borde) $\implies \overline{\alpha}_i = 0$. Luego:

$$\hat{y}_0 = \operatorname{sign}\left(\sum_{x_i \text{ vector soporte}} \overline{\alpha_i} y_i \langle x_i, x_0 \rangle + \overline{b}\right)$$

Por qué usar el problema dual

Existen distintos argumentos de por qué usar el problema dual para SVM:

- El teorema de dualidad fuerte asegura que resolver el dual entregará una solución primal (en caso de existir).
- Las entradas x_i solo interactúan en forma de productos internos.
- Las variables duales son indicatrices de vectores soporte y simplifica el cálculo de una predicción.
- El problema dual no depende de la dimensión de los datos (excepto en el producto interno).
- La forma dual permitirá el uso de kernels.
- Existen algoritmos muy rápidos para resolver la forma dual (coordinate descent).

Soft margin SVM

El problema original es infactible cuando los datos no son linealmente separables. Una forma de solucionar este problema es agregar una penalización para puntos que están dentro o al otro lado del margen:

$$\begin{aligned}
(P) \quad \min_{w,b,\xi} \quad & \frac{1}{2}||w||^2 + c\sum_{i=1}^{N} \xi_i \\
\text{s.a} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i \\
& \xi_i \geq 0,\ i \in \{1, \ldots, N\}
\end{aligned}$$

Donde $\xi \geq 0$ se interpretan como variables de holgura y $c > 0$ es el peso entregado al regularizador $\sum_{i=1}^{N} \xi_i$. En este caso, el **lagrangiano** es el siguiente:

$$\begin{aligned}
L(w, b, \xi, \alpha, \beta) &= \frac{1}{2}||w||^2 + c\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i \left(1 - \xi_i - y_i(w^\top x_i + b)\right) + \sum_{i=1}^{N} \beta_i(-\xi_i) \\
&= \frac{1}{2}||w||^2 + \sum_{i=1}^{N} \xi_i \left(c - \alpha_i - \beta_i\right) + \sum_{i=1}^{N} \alpha_i \left(1 - y_i(w^\top x_i + b)\right)
\end{aligned}$$

Para el **lagrangiano dual** $\theta(\alpha, \beta) = \inf_{w,b,\xi} L(w, b, \xi, \alpha, \beta)$ se vuelve a usar CPO:

- $\frac{\partial L}{\partial w} = w^\top - \sum_{i=1}^N \alpha_i y_i x_i^\top = 0 \implies \overline{w} = \sum_{i=1}^N \alpha_i y_i x_i.$
- $\frac{\partial L}{\partial b} = -\sum_{i=1}^N \alpha_i y_i = 0 \implies \sum_{i=1}^N \alpha_i y_i = 0.$
- $\frac{\partial L}{\partial \xi_i} = c - \alpha_i - \beta_i = 0 \implies \beta_i = c - \alpha_i$

Por lo tanto, el lagrangiano dual es el mismo que para hard-margin. De esta forma, el problema dual es el siguiente:

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
\text{s.a} \quad & \sum_{i=1}^N \alpha_i y_i = 0 \\
& \alpha_i, \beta_i \geq 0, \quad \forall i \in \{1, \ldots, N\}
\end{aligned}
$$

Y dado que $\beta_i = c - \alpha_i$, la última restricción puede escribirse como:

$$0 \leq \lambda_i \leq c, \quad \forall i \in \{1, \ldots, N\}$$

Truco del kernel

En general, muchos algoritmos de ML, al igual que en SVM, trabajan los datos de entrada únicamente en forma de productos internos de la forma $\langle x_i, x_j \rangle$.

De esta forma, si $\phi : \mathbb{R}^M \to \mathbb{R}^D$ es un mapa de características, los datos de entrenamiento solo aparecerán de la forma $\langle \phi(x_i), \phi(x_j) \rangle$.

Estos productos internos pueden ser almacenados en una matriz (Gram matrix) $M \in \mathbb{R}^{D \times D}$ mediante $m_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$ y así, se puede trabajar únicamente con la matriz M.
Por lo tanto, nace la siguiente pregunta:

Dada una función $K : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$, ¿bajo qué condiciones $K(x_i, x_j)$ representa una matriz de productos internos?

Truco del kernel: teorema de Mercer

Theorem (teorema de Mercer)

Una función continua $K : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ se dirá Mercer kernel si

- *Es simétrica: $K(x_1, x_2) = K(x_2, x_1)$*
- *Es definida positiva: para toda función $g : \mathbb{R}^M \to \mathbb{R}$ continua,*

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} K(x_1, x_2) g(x_1) g(x_2) dx_1 dx_2 \geq 0$$

Luego, si $K : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ es un Mercer kernel, entonces existe un espacio de Hilbert $(\mathcal{H}, \langle , \rangle)$ y una función $\phi : X \to \mathcal{H}$ tal que:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

De esta forma, pueden sustituirse los productos $\langle x_i, x_j \rangle$ por $K(x_i, x_j)$ ya que el teorema anterior asegura que dicha evaluación representa el producto interno de los datos en algún espacio de características.

Kernel SVM

Al sustituir los productos internos en la forma dual de SVM se obtiene el problema de **Kernel SVM**, el cual es mucho más general ya que permite separar los datos en un espacio de mayor dimensionalidad:

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\
\text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq c
\end{aligned}
$$

Donde $K : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ es un Mercer kernel asociado a algún mapa de características ϕ no necesariamente de dimensión finita.

Ejemplos de kernels

- **Kernel polinomial:**
$$K_{pol}(x,y) = (c + x^\top y)^d$$
donde $c \geq 0$ es un parámetro libre y $d \in \mathbb{N}$ es el orden del polinomio.
  - Tiene la desventaja de ser numéricamente inestable dependiendo del orden d.
  - Es muy usado en NLP con $d = 2$.
- **Kernel gaussiano (RBF):**
$$K_{RBF}(x,y) = \sigma^2 \exp\left(-\frac{\|x-y\|^2}{2l^2}\right)$$
  - Tiene asociado un mapa de características infinitodimensional.
  - Está asociado a KNN debido a que suaviza un diagrama de Voronoi.
  - Es el kernel por defecto en SVM.
- **Kernel periódico:**
$$K_{per}(x,y) = \sigma^2 \exp\left(-\frac{2\,\mathrm{sen}^2\left(\frac{\pi|x-y|}{p}\right)}{l^2}\right).$$

Además, se pueden combinar nuevos kernels sumando y/o multiplicando kernels usuales.