Análisis Matemático

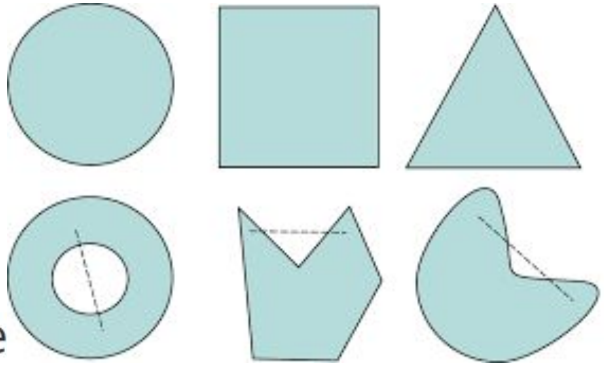
Clase 7

Optimización convexa

Conjuntos convexos

Def: Un conjunto $\mathcal{C}$ se dice convexo si para cualquier $x, y \in \mathcal{C}$, y todo escalar $\theta \in [0,1]$ vale que

$$\theta x + (1-\theta) y \in \mathcal{C}.$$



En criollo: la recta que une dos puntos cualesquiera de conjunto $\mathcal{C}$ está contenida en $\mathcal{C}$.

Funciones convexas

Def: Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función cuyo dominio es en un conjunto convexo. Diremos que f es una función convexa si para todo x, y en el dominio de f, y para todo escalar $\theta \in [0,1]$ vale que

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

Desigualdad de Jensen

Una forma gráfica de ver si una función es convexa es si la línea que une dos puntos cualesquiera $f(x)$ y y $f(y)$ queda por encima de la gráfica de f

Si f es **derivable**, diremos que f es convexa sii para dos puntos cualesquiera x, y satisface que $f(y) \geq f(x) + \nabla_x f(x)^T (y - x)$.

Optimización convexa

Dado el problema de optimización

$$
\begin{aligned}
\min_{x} \quad & f(x) \\
s.t. \quad & g_i(x) \leq 0 \quad i = 1, \ldots, m \\
& h_i(x) = 0 \quad i = 1, \ldots, k
\end{aligned}
$$

si función a optimizar $f(\cdot)$ es convexa y las restricciones $g(\cdot)$ y $h(\cdot)$ definen conjuntos convexos obtenemos lo que se conoce como problema de optimización convexo.

En estos casos hay dualidad fuerte y la solución del problema dual es la misma que la del problema primal.

Condiciones KKT

Supongamos $f,\ g_1,\ \ldots,\ g_m,\ h_1,\ \ldots,\ h_k$ son diferenciables. Sean x^* y $\boldsymbol{\lambda}^*,\ \boldsymbol{\mu}^*$ son puntos óptimos para los cuales hay dualidad fuerte (es decir que el $\mathcal{D}(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f(x^*)$. Como x^* debe ser un punto factible y además un mínimo, se obtiene lo que se conoce como las condiciones de Karush-Kuhn-Tucker (KKT):

$$\nabla_x f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla_x g_i(x^*) + \sum_{i=1}^{k} \mu_i \nabla_x h_i(x^*) = 0$$

$$\begin{aligned} g_i(x^*) leq 0, &\quad i = 1, \ldots, m \\ h_i(x^*) = 0, &\quad i = 1, \ldots, k \\ \lambda_i^* \geq 0, &\quad i = 1, \ldots, m \\ \boxed{\lambda_i^* f_i(x^*) = 0,} &\quad \boxed{i = 1, \ldots, m} \end{aligned}$$

Por dualidad fuerte

Si el problema a resolver es convexo, las condiciones KKT son necesarias y suficientes para que $x^*,\ \boldsymbol{\lambda}^*,\ \boldsymbol{\mu}^*$ sean óptimos y $\mathcal{D}(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f(x^*)$.

Programación lineal

Es el caso especial de optimización convexa donde todas las funciones y regiones son convexas:

$$\begin{aligned} \min_{x} \quad & c^T x \\ s.t. \quad & Ax \leq b \end{aligned}$$

donde $A \in \mathbb{R}^{m \times d}$, $x \in \mathbb{R}^d$. El lagrangiano queda dado por

$$\mathcal{L}(x, \lambda) = c^T x + \lambda^T (Ax - b) = (c + A^T \lambda)^T x - \lambda^T b\,.$$

Programación cuadrática

Es el caso en que la función objetivo es cuadrática y las restricciones afines:

$$\begin{aligned} \min_{x} \quad & \frac{1}{2}x^T Q x + c^T x \\ s.t. \quad & Ax \leq b, \end{aligned} \qquad \text{donde} \qquad \begin{aligned} & x \in \mathbb{R}^d,\ A \in \mathbb{R}^{m\times d} \\ & b \in \mathbb{R}^m,\ Q > 0 \in \mathbb{R}^{d\times d} \end{aligned}$$

$$\mathcal{L}(x,\lambda) = \tfrac{1}{2}x^T Q x + c^T x + \lambda^T(Ax - b) = \tfrac{1}{2}x^T Q x + (C + A^T\lambda)^T x - \lambda^T b$$

Derivando e igualando a cero recuperamos que $x = -Q^{-1}(c + A^T\lambda)$ y

$$\max_{\lambda > 0} \quad -\tfrac{1}{2}(x + A^T\lambda)^T Q^{-1}(c + A^T\lambda) - \lambda^T b$$

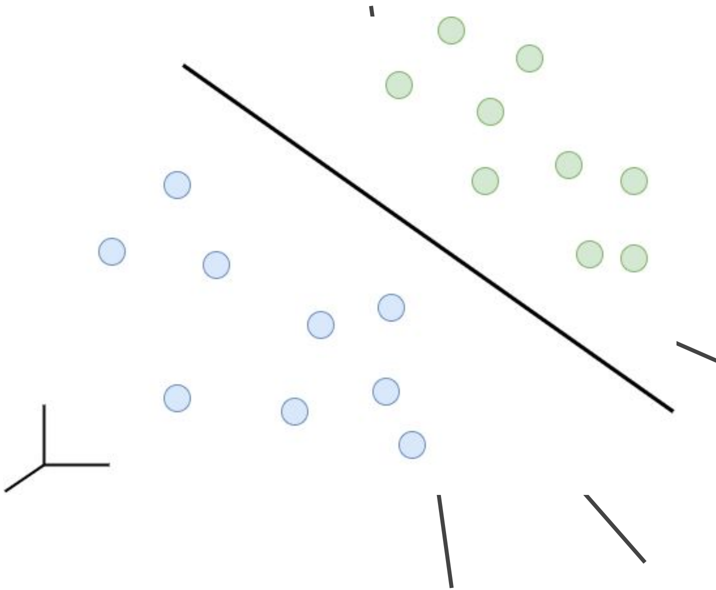
Support Vector Machine

Caso linealmente separable

Planteo del problema

Las SVM son un algoritmo de clasificación binaria supervisado, es decir que los predictores son funciones de la forma $f : \mathbb{R}^d \to \{-1, 1\}$.

Consideremos el conjunto de entrenamiento $\mathcal{T} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ de pares (x_k, y_k), donde $x_k \in \mathbb{R}^d$ son las observaciones y $y_k = \{-1, 1\}$ es la etiqueta correspondiente. El objetivo de la SVM es hallar $w \in \mathbb{R}^n$ y $b \in \mathbb{R}$ que definen el hiperplano $\langle w, x \rangle + b$ que mejor separa entre los casos positivos y negativos partir de $\mathcal{T}$.



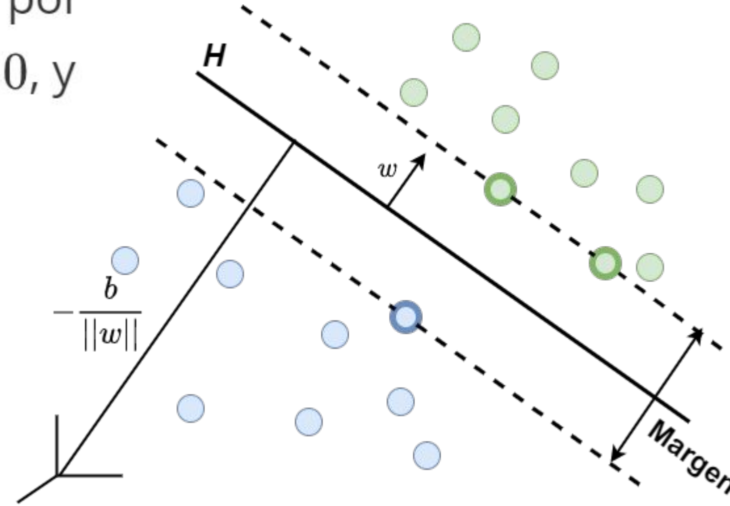
Planteo del problema

Lo que buscamos es que los casos positivos queden por encima del hiperplano, de forma que $\langle w, x_+ \rangle + b \geq 0$, y los negativos por debajo $\langle w, x_- \rangle + b \leq 0$. Esto se puede reescribir como $y_k(\langle w, x_k \rangle + b) \geq 0$.

De todo los hiperplanos que sirven para separar las clases, una idea es buscar aquel que maximice el margen entre la región positiva y negativa.

$$\max_{w,b,r} \quad r$$

$$s.t. \quad y_k \left(\left\langle \frac{w}{\|w\|}, x_k \right\rangle + b \right) \geq r, \quad r > 0$$



Planteo del problema

Puedo operar sobre las restricciones y dividir todo por r,

obteniendo
$$y_k\left(\langle \underbrace{\frac{w}{\|w\|r}}_{w'}, x_k\rangle + \underbrace{\frac{b}{r}}_{b'}\right) \geq \underbrace{\frac{r}{r}}_{1}$$

Además $\|w'\| = \left\|\frac{w}{\|w\|r}\right\| = \frac{1}{r}\left\|\frac{w}{\|w\|}\right\| = \frac{1}{r} \Rightarrow r = \|w'\|$ y $b' = \frac{b}{\|w\|}$

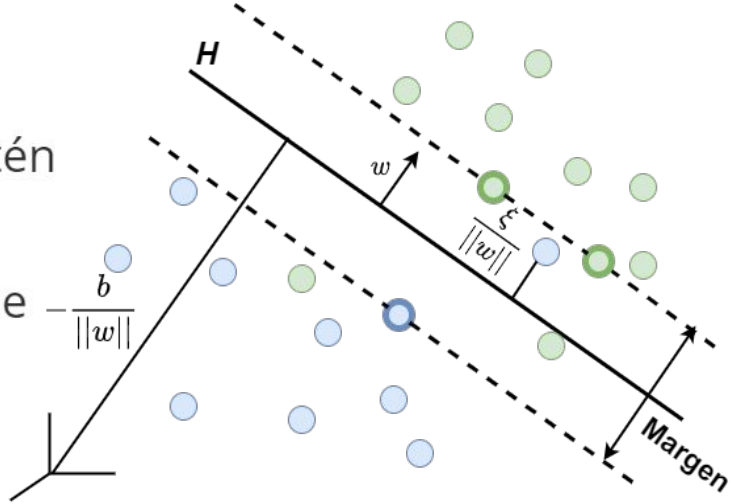
y podemos reescribir el problema de optimización como

$$\begin{aligned} &\min_{w,b} \quad \frac{1}{2}\|w\|^2 \\ &s.t. \quad y_k\left(\langle w, x_k\rangle + b\right) \geq 1 \quad k = 1, \ldots, n \end{aligned}$$

Caso no linealmente separable

Planteo del problema

Lo que vimos hasta ahora no admite que los puntos sean separables. Una forma de trabajar con esto es introduciendo una penalidad para los puntos que estén del lado incorrecto del hiperplano (o del margen) a través de las variables ξ_i (*slack variables*) de forma que el problema de optimización resulta



$$
\begin{aligned}
\min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{n} \xi_k \\
s.t. \quad & y_k\left(\langle w, x_k\rangle + b\right) \geq 1 - \xi_k \quad k = 1, \ldots, n \\
& \xi_k \geq 0
\end{aligned}
$$

Problema dual

Podemos construir el lagrangiano como

$$\mathcal{L}(w, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left(y_i(\langle w, x_i\rangle + b) - 1 + \xi_i\right) - \sum_{i=1}^{n}\gamma_i\xi_i$$

Derivando respecto de las variables primales e igualando a cero,

$$\frac{\partial \mathcal{L}}{\partial w} = w^T - \sum_{k=0}^{n}\alpha_k y_k x_k^T = 0,$$

obtenemos que $\frac{\partial \mathcal{L}}{\partial b} = -\sum_{k=1}^{n}\alpha_k y_k = 0,$ $\Rightarrow$ $\begin{aligned} w &= \sum_{k=1}^{n}\alpha_n y_n x_n \\ 0 &= \sum_{k=1}^{n}\alpha_k y_k \end{aligned}$

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = C - \alpha_n - \gamma_n = 0$$

Problema dual

Reemplazando este w en el lagrangiano obtenemos el lagrangiano dual

$$\begin{aligned}\mathcal{D}(\xi, \alpha, \gamma) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} y_i y_k \alpha_i \alpha_k \langle x_i, x_k\rangle - \sum_{i=1}^{n} y_i \alpha_i \left\langle \sum_{k=1}^{n} y_k \alpha_k x_k, x_i \right\rangle \\ &\quad + C\sum_{i=1}^{n}\xi_i - b\underbrace{\sum_{i=1}^{n} y_i \alpha_i}_{=0} + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i \xi_i - \sum_{i=1}^{n}\gamma_i \xi_i \\ &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} y_i y_k \alpha_i \alpha_k \langle x_i, x_k\rangle + \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\underbrace{(C - \alpha_i - \gamma_i)}_{=0}\xi_i \\ &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} y_i y_k \alpha_i \alpha_k \langle x_i, x_k\rangle + \sum_{i=1}^{n}\alpha_i\end{aligned}$$

El problema dual resulta

$$\begin{aligned}\max_{\boldsymbol{\alpha}, \boldsymbol{\gamma}} \quad & \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \\ s.t. \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C\end{aligned}$$

Condiciones KKT

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} = w^T - \sum_{k=0}^{n} \alpha_k y_k x_k^T &= 0, \\
\frac{\partial \mathcal{L}}{\partial b} = -\sum_{k=1}^{n} \alpha_k y_k &= 0, \\
\frac{\partial \mathcal{L}}{\partial \xi_n} = C - \alpha_n - \gamma_n &= 0, \\
y_i(\langle w, x \rangle + b) - 1 + \xi_i &\geq 0, \\
\xi_i &\geq 0, \\
\alpha_i &\geq 0, \\
\mu_i &\geq 0, \\
\alpha_i [y_i(\langle w, x_i \rangle + b) - 1 + \xi_i] &= 0, \\
\mu_i \xi_i &= 0
\end{aligned}
$$

$$
\begin{aligned}
& w && = \textstyle\sum_{k=1}^{n} \alpha_k y_k x_k = \sum_{i|\alpha_i > 0} \alpha_i y_i x_i \\
\Rightarrow \; & 0 && = \textstyle\sum_{k=1}^{n} \alpha_k y_k \\
& 0 && \leq \alpha_i \leq C
\end{aligned}
$$

Luego, el w óptimo del problema primal se obtiene como combinación lineal de un subconjunto de las observaciones. A este subconjunto se lo llama vectores soporte

Etapa de test

Una vez hallados w y b óptimos, la función de predicción para una nueva observación x resulta

$$\begin{aligned}\hat{y} = f(x) &= sgn(\langle w, x\rangle + b)\\ &= sgn\left(\left\langle \sum_{i:\alpha_i\neq 0} \alpha_i y_i x_i, x \right\rangle + b\right)\\ &= sgn\left(\sum_{i:\alpha_i\neq 0} \alpha_i y_i \langle x_i, x\rangle + b\right)\end{aligned}$$

Observaciones

Finalmente, el modelo de clasificación depende únicamente de los vectores soporte. Esto presenta algunas ventajas:

- Es relativamente 'barato' en etapa de test (no hay que comparar contra todas las muestras)
- Es robusto a los outliers

Interpretación alternativa: función de pérdida

El problema primar puede reescribirse en términos de una función de pérdidas que se desea minimizar:

$$J(w,b) = \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{n} \underbrace{\max\{0, 1 - y_i(\langle w, x\rangle + b)\}}_{hinge\ loss}$$

Dado que la $hinge\ loss$ es derivable en todo punto (salvo donde $y_i(\langle w, x\rangle + b) = 1$), esta función de pérdidas puede resolverse por los métodos de optimización sin restricciones vistos en la clase 6.

Bibliografía

Bibliografía

- “A Tutorial on Support Vector Machines for Pattern Recognition”, Christopher J.C. Burges, Data Mining and Knowledge Discovery, 1998.
- “Convex Optimization”, Stephen Boyd and Lieven Vandenberghe, Cambridge University Press, 2004