Máquinas de Soporte Vectorial: Dualidad de Lagrange

Lección 11

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

1. Dualidad de Lagrange
   - Optimización del Lagrangiano
2. Máquinas de Soporte Vectorial
   - Clasificadores de margen óptimo
   - Kernels

Problemas de optimización con restricciones

- Un problema de optimización con restricciones tiene la forma

$$\min_{\underline{\mathbf{w}}} f(\underline{\mathbf{w}})$$

$$\text{sujeto a } h_i(\underline{\mathbf{w}}) = 0, \quad i = 1, \ldots, l$$

- Esto se soluciona con **multiplicadores de Lagrange**
- Para ello definimos el **Lagrangiano** como

$$\mathcal{L}(\underline{\mathbf{w}}, \underline{\beta}) = f(\underline{\mathbf{w}}) + \sum_{i=1}^{l} \beta_i h_i(\underline{\mathbf{w}})$$

- Aquí los β_i son los **multiplicadores de Lagrange**
- Lo solucionamos buscando $\underline{\mathbf{w}}$, $\underline{\underline{\beta}}$ que haga que

$$\nabla_{\underline{\mathbf{w}}, \underline{\underline{\beta}}} \mathcal{L}(\underline{\mathbf{w}}, \underline{\beta}) = \underline{\mathbf{0}} \quad \Longrightarrow \quad \frac{\partial \mathcal{L}}{\partial w_j} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0$$

Restricciones con desigualdades (1)

- Consideremos problema de optimización **primal** con igualdades y desigualdades en restricciones:

$$\begin{aligned} &\min_{\underline{\mathbf{w}}} f(\underline{\mathbf{w}}) \\ \text{sujeto a } & g_i(\underline{\mathbf{w}}) \leq 0, \quad i = 1, \ldots, k \\ & h_i(\underline{\mathbf{w}}) = 0, \quad i = 1, \ldots, l \end{aligned}$$

- Lo resolvemos con el Lagrangiano generalizado:

$$\mathcal{L}(\underline{\mathbf{w}}, \underline{\alpha}, \underline{\beta}) = f(\underline{\mathbf{w}}) + \sum_{i=1}^{k} \alpha_i g_i(\underline{\mathbf{w}}) + \sum_{i=1}^{l} \beta_i h_i(\underline{\mathbf{w}})$$

con multiplicadores de Lagrange α_i y β_i.

Restricciones con desigualdades (2)

- Consideremos la cantidad

$$\theta_{\mathcal{P}}(\underline{\mathbf{w}}) = \max_{\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}:\alpha_i \geq 0} \mathcal{L}(\underline{\mathbf{w}},\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}})$$

- Si nos dan un $\underline{\mathbf{w}}$ que viola cualquiera de las restricciones, entonces:

$$\theta_{\mathcal{P}}(\underline{\mathbf{w}}) = \max_{\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}:\alpha_i \geq 0} \left(f(\underline{\mathbf{w}}) + \sum_{i=1}^{k} \alpha_i g_i(\underline{\mathbf{w}}) + \sum_{i=1}^{l} \beta_i h_i(\underline{\mathbf{w}}) \right) = \infty$$

- Si $\underline{\mathbf{w}}$ satisface las restricciones entonces $\theta_{\mathcal{P}}(\underline{\mathbf{w}}) = f(\underline{\mathbf{w}})$.
- Se cumple:

$$\theta_{\mathcal{P}}(\underline{\mathbf{w}}) = \begin{cases} f(\underline{\mathbf{w}}) & \text{si } \underline{\mathbf{w}} \text{ satisface restricciones primales} \\ \infty & \text{en otro caso} \end{cases}$$

Restricciones con desigualdades (3)

- Si consideramos el problema de minimización

$$\min_{\underline{\mathbf{w}}} \theta_{\mathcal{P}}(\underline{\mathbf{w}}) = \min_{\underline{\mathbf{w}}} \max_{\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}:\alpha_i \geq 0} \mathcal{L}(\underline{\mathbf{w}},\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}})$$

  entonces será el mismo problema primal (= soluciones).
- Denominaremos **valor óptimo** del problema primal a

$$p^* = \min_{\underline{\mathbf{w}}} \theta_{\mathcal{P}}(\underline{\mathbf{w}})$$

Problema de optimización dual

- Definamos el término **dual** al problema $\theta_{\mathcal{P}}(\underline{\mathbf{w}})$ anterior:

$$\theta_{\mathcal{D}}(\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}) = \min_{\underline{\mathbf{w}}} \mathcal{L}(\underline{\mathbf{w}},\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}})$$

- Mientras que en $\theta_{\mathcal{P}}$ **maximizamos** respecto a $\underline{\boldsymbol{\alpha}}$ y $\underline{\boldsymbol{\beta}}$, aquí **minimizamos** respecto a $\underline{\mathbf{w}}$
- El problema de optimización **dual** es ahora

$$\max_{\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}:\alpha_i \geq 0} \theta_{\mathcal{D}}(\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}) = \max_{\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}:\alpha_i \geq 0} \min_{\underline{\mathbf{w}}} \mathcal{L}(\underline{\mathbf{w}},\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}})$$

- Observe que el problema dual es similar al primal con los mín y máx invertidos
- El **valor óptimo** del problema dual es

$$d^* = \max_{\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}:\alpha_i \geq 0} \theta_{\mathcal{D}}(\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}})$$

Relación entre los problemas primal y dual

- Los problemas primal y dual tienen la siguiente relación:

$$d^* = \max_{\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}:\alpha_i \geq 0} \min_{\underline{\mathbf{w}}} \mathcal{L}(\underline{\mathbf{w}},\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}) \leq \min_{\underline{\mathbf{w}}} \max_{\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}:\alpha_i \geq 0} \mathcal{L}(\underline{\mathbf{w}},\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}}) = p^*$$

- Bajo *ciertas* condiciones $d^* = p^*$
- En ese caso podemos usar el planteo dual para resolver el problema primal.

Condiciones para equivalencia dual primal

- Si
  - f y g_i son convexas (hessianas positivas semidefinidas), y
  - Las restricciones g_i son estrictas (esto es $g_i(\underline{\mathbf{w}}) < 0, \forall i$)
  - h_i son afines ($h_i(\underline{\mathbf{w}}) = \underline{\mathbf{a}}_i^T \underline{\mathbf{w}} + b_i$)

  entonces existen $\underline{\mathbf{w}}^*$, $\underline{\boldsymbol{\alpha}}^*$, $\underline{\boldsymbol{\beta}}^*$, tales que
  - $\underline{\mathbf{w}}^*$ es la solución al problema primal
  - $\underline{\boldsymbol{\alpha}}^*$, $\underline{\boldsymbol{\beta}}^*$ son las soluciones al problema dual
  - $p^* = d^* = \mathcal{L}(\underline{\mathbf{w}}^*, \underline{\boldsymbol{\alpha}}^*, \underline{\boldsymbol{\beta}}^*)$

Condiciones de Karush-Kuhn-Tucker (KKT)

- Con las condiciones anteriores, también se cumple:

$$\frac{\partial}{\partial w_i}\mathcal{L}(\underline{\mathbf{w}}^*,\underline{\boldsymbol{\alpha}}^*,\underline{\boldsymbol{\beta}}^*) = 0, \quad i = 1,\ldots,n \tag{1}$$

$$\frac{\partial}{\partial \beta_i}\mathcal{L}(\underline{\mathbf{w}}^*,\underline{\boldsymbol{\alpha}}^*,\underline{\boldsymbol{\beta}}^*) = 0, \quad i = 1,\ldots,l \tag{2}$$

$$\alpha_i^* g_i(\underline{\mathbf{w}}^*) = 0, \quad i = 1,\ldots,k \tag{3}$$

$$g_i(\underline{\mathbf{w}}^*) \leq 0, \quad i = 1,\ldots,k \tag{4}$$

$$\alpha_i^* \geq 0, \quad i = 1,\ldots,k \tag{5}$$

conocidas como **condiciones de Karush-Kuhn-Tucker (KKT)**

- Si $\underline{\mathbf{w}}^*, \underline{\boldsymbol{\alpha}}^*, \underline{\boldsymbol{\beta}}^*$ satisfacen KKT, entonces también son soluciones de los problemas dual y primal

Condición complementaria dual

- La tercera condición KKT $\alpha_i^* g_i(\underline{\mathbf{w}}^*) = 0$ se llama **condición complementaria dual**
- Implica que si $\alpha_i^* > 0$ entonces $g_i(\underline{\mathbf{w}}^*) = 0$
- Se dice en ese caso $g_i(\underline{\mathbf{w}}^*) = 0$ que la restricción está **activa** (esto es, que cumple $=$ y no $<$)
- Esto será clave luego para mostrar que solo hay un número pequeño de vectores de soporte
- También nos brindará la prueba de convergencia del algoritmo SMO

Clasificador de margen óptimo (1)

- Anteriormente planteamos el problema de clasificador de margen óptimo (primal):

$$\min_{\gamma,\underline{\mathbf{w}},b} \frac{1}{2}\|\underline{\mathbf{w}}\|^2$$
$$\text{sujeto a } y^{(i)}(\underline{\mathbf{w}}^T\underline{\mathbf{x}}^{(i)} + b) \geq 1, \quad i = 1,\ldots,m$$

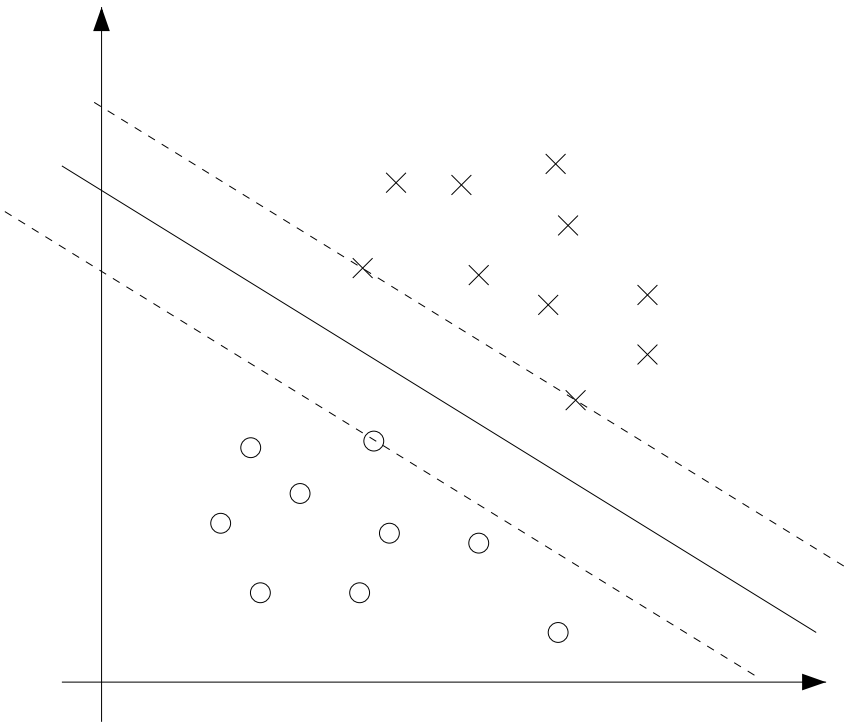
- Podemos escribir las restricciones como:

$$g_i(\underline{\mathbf{w}},b) = -y^{(i)}(\underline{\mathbf{w}}^T\underline{\mathbf{x}}^{(i)} + b) + 1 \leq 0$$

- Tenemos una restricción de esas para cada dato de entrenamiento
- No hay restricciones $h_i(\underline{\mathbf{w}}) = 0 \implies$ no hay β_i

Vectores de soporte (1)

- Por la condición complementaria KKT dual tendremos $\alpha_i > 0$ ¡solo para los datos de entrenamiento $(\underline{\mathbf{x}}^{(i)}, y^{(i)})$ que tienen margen funcional exactamente igual a uno! $(g_i(\underline{\mathbf{w}}) = 0)$



Vectores de soporte (2)

- Los puntos con el menor margen son aquellos más cercanos a la frontera de decisión
- En ejemplo, únicamente tres α_i son distintos de cero
- Esos datos de entrenamiento con $\alpha_i \neq 0$ se denominan **vectores de soporte**
- Usualmente solo unos pocos datos son vectores de soporte

Mapa de ruta...

- Tenemos un problema primal: "clasificación de margen óptimo"
- Lo transformamos en un problema dual
- Queremos ahora desarrollar problema dual en términos del producto interno $\left\langle \underline{\mathbf{x}}^{(i)}, \underline{\mathbf{x}}^{(j)} \right\rangle$ (o $\underline{\mathbf{x}}^{(i)\,T} \underline{\mathbf{x}}^{(j)}$ o $\underline{\mathbf{x}}^{(i)} \cdot \underline{\mathbf{x}}^{(j)}$)
- En esta reformulación se basa el **truco del kernel**

Problema dual del clasificador de margen óptimo (1)

- El Lagrangiano de nuestro problema de optimización es:

$$\mathcal{L}(\underline{\mathbf{w}},b,\underline{\boldsymbol{\alpha}}) = \frac{1}{2}\|\underline{\mathbf{w}}\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(\underline{\mathbf{w}}^T \underline{\mathbf{x}}^{(i)} + b) - 1 \right]$$

- Note que solo tenemos α_i y ningún β_i (no hay restricciones de igualdad)
- Para encontrar las soluciones del problema dual, minimizamos primero $\mathcal{L}(\underline{\mathbf{w}},b,\underline{\boldsymbol{\alpha}})$ respecto a $\underline{\mathbf{w}}$ y b (para un $\underline{\boldsymbol{\alpha}}$ fijo) y así obtenemos $\theta_{\mathcal{D}}$:

$$\nabla_{\underline{\mathbf{w}}} \mathcal{L}(\underline{\mathbf{w}},b,\underline{\boldsymbol{\alpha}}) = \underline{\mathbf{w}} - \sum_{i=1}^{m} \alpha_i y^{(i)} \underline{\mathbf{x}}^{(i)} = 0$$

Problema dual del clasificador de margen óptimo (2)

de donde se deriva directamente que:

$$\underline{\mathbf{w}} = \sum_{i=1}^{m} \alpha_i y^{(i)} \underline{\mathbf{x}}^{(i)}$$

- La derivada respecto a b da:

$$\frac{\partial}{\partial b} \mathcal{L}(\underline{\mathbf{w}}, b, \underline{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Problema dual del clasificador de margen óptimo (3)

- Si tomamos el valor de $\underline{\mathbf{w}}$ y lo sustituimos en el Lagrangiano y simplificamos:

$$\mathcal{L}(\underline{\mathbf{w}},b,\underline{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (\underline{\mathbf{x}}^{(i)})^T \underline{\mathbf{x}}^{(j)} - b \sum_{i=1}^{m} \alpha_i y^{(i)}$$

  donde por la derivada respecto a b sabemos que lo último es cero, así que:

$$\mathcal{L}(\underline{\mathbf{w}},b,\underline{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (\underline{\mathbf{x}}^{(i)})^T \underline{\mathbf{x}}^{(j)}$$

- Esto lo obtuvimos minimizando $\mathcal{L}$ respecto a $\underline{\mathbf{w}}$ y b.

Problema dual del clasificador de margen óptimo (4)

- Agregando las restricciones $\alpha_i \geq 0$ y la derivada respecto a b obtenemos el problema de optimización dual:

$$\max_{\underline{\boldsymbol{\alpha}}} (W(\underline{\boldsymbol{\alpha}})) = \max_{\underline{\boldsymbol{\alpha}}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \left\langle \underline{\mathbf{x}}^{(i)}, \underline{\mathbf{x}}^{(j)} \right\rangle \right)$$

sujeto a $\alpha_i \geq 0, \quad i = 1, \ldots, m$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- En este caso se satisfacen las condiciones KKT $\Longrightarrow$ podemos resolver problema dual en vez del problema primal
- En este problema los parámetros buscados son α_i
- Observe que entradas usadas solo a través de $\langle \cdot, \cdot \rangle$

Problema dual del clasificador de margen óptimo (5)

- Una vez encontrado el $\underline{\boldsymbol{\alpha}}^*$ que maximiza $W(\underline{\boldsymbol{\alpha}})$, usamos

$$\underline{\mathbf{w}}^* = \sum_{i=1}^{m} \alpha_i^* y^{(i)} \underline{\mathbf{x}}^{(i)}$$

- Con $\underline{\mathbf{w}}^*$ podemos encontrar ahora b con

$$b^* = -\frac{\max\limits_{i:y^{(i)}=-1} \underline{\mathbf{w}}^{*T} \underline{\mathbf{x}}^{(i)} + \min\limits_{i:y^{(i)}=1} \underline{\mathbf{w}}^{*T} \underline{\mathbf{x}}^{(i)}}{2}$$

Producto interno en predicción (1)

- Retomando resultado de minimizar $\mathcal{L}$:

$$\underline{\mathbf{w}} = \sum_{i=1}^{m} \alpha_i y^{(i)} \underline{\mathbf{x}}^{(i)}$$

- Supongamos que ajustamos los parámetros del modelo a un conjunto de entrenamiento y queremos predecir y para $\underline{\mathbf{x}}$.
- Podemos predecir $y = 1$ si $(\underline{\mathbf{w}}^T \underline{\mathbf{x}} + b) > 0$
- Pero con valor anterior:

$$\underline{\mathbf{w}}^T \underline{\mathbf{x}} + b = \left(\sum_{i=1}^{m} \alpha_i y^{(i)} \underline{\mathbf{x}}^{(i)} \right)^T \underline{\mathbf{x}} + b = \sum_{i=1}^{m} \alpha_i y^{(i)} \left\langle \underline{\mathbf{x}}^{(i)}, \underline{\mathbf{x}} \right\rangle + b$$

- Como solo los vectores de soporte tienen $\alpha_i \neq 0$, solo ellos intervienen en predicción

Producto interno en predicción (2)

- Cálculo en términos de productos internos entre los vectores de soporte y $\underline{\mathbf{x}}$
- Nos falta un paso para llegar a las máquinas de soporte vectorial: el truco del kernel
- También nos falta corregir suposición de que los datos son linealmente separables

Concepto de kernel

- Cuando vimos el problema de predicción de precio en función del área de la casa, mejoramos el modelo extendiendo la entrada; es decir, no utilizamos como entrada a $\underline{\mathbf{x}} = [\text{área}]$ sino una extensión: $\underline{\hat{\mathbf{x}}} = \begin{bmatrix} 1 & x_1 & x_1^2 \end{bmatrix}$
- Denominamos a la entrada real del problema $\underline{\mathbf{x}}$ los **atributos** de entrada. Las entradas *viven* en el **espacio de entrada**.
- Denominamos a los componentes usados por el clasificador $\underline{\hat{\mathbf{x}}}$ **características** de entrada. Estos vectores *viven* en el **espacio de características**.

Mapeo de características

- El mapeo ϕ transforma los atributos en características.
- Por ejemplo $\phi : \mathbb{R} \rightarrow \mathbb{R}^3$:

$$\phi(\underline{\mathbf{x}}) = \begin{bmatrix} 1 \\ x_1 \\ x_1^2 \end{bmatrix}$$

- Podemos en los SVM usar $\phi(\underline{\mathbf{x}})$ en vez de $\underline{\mathbf{x}}$.
- Puesto que el problema está planteado en términos de $\langle \underline{\mathbf{x}}, \underline{\mathbf{z}} \rangle$, solo debe reemplazare eso por $\langle \phi(\underline{\mathbf{x}}), \phi(\underline{\mathbf{z}}) \rangle$

Kernels

- Dado el mapeo ϕ, se define el **kernel** como:

$$K(\underline{\mathbf{x}},\underline{\mathbf{z}}) = \phi(\underline{\mathbf{x}})^T \phi(\underline{\mathbf{z}})$$

- Todos los $\langle \phi(\underline{\mathbf{x}}), \phi(\underline{\mathbf{z}}) \rangle$ se reemplazan entonces por $K(\underline{\mathbf{x}},\underline{\mathbf{z}})$
- Con frecuencia evaluar K es poco costoso aun cuando ϕ sea muy caro de calcular (¡representando incluso vectores en infinitas dimensiones!)
- En otras palabras SVM aprende en el espacio de características de muchas dimensiones, ¡aun sin tener **nunca** que evaluar nada en dicho espacio, ni tan siquiera tener que calcular $\phi(\underline{\mathbf{x}})$!

Resumen

1. Dualidad de Lagrange
   - Optimización del Lagrangiano
2. Máquinas de Soporte Vectorial
   - Clasificadores de margen óptimo
   - Kernels

Este documento ha sido elaborado con software libre incluyendo LaTeX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux