Clase 6: Modelos lineales

Responsable: Carlos Enrique Lezama Jacinto

EST-25134, Primavera 2021
Dr. Alfredo Garbuno Iñigo
Febrero 2, 2021

1. Predictores lineales

En primer lugar, nos enfocaremos en predictores lineales[1] puesto que son fáciles de interpretar y se ajustan razonablemente bien a los datos de entrenamiento. Además, son muy intuitivos y presentan las bases para modelos más complejos.

Algunos de los modelos y algortimos que estudiaremos en las próximas clases son:

Modelos	**Algoritmos**	**Tarea**
Semiespacios	Programación lineal (LP), perceptrones	Clasificación
Regresión lineal	Mínimos cuadrados	Regresión
Regresión logística	Métodos iterativos	Clasificación

Definición 1.1 (Clase de transformaciones afines)**.** Definimos la clase de transformaciones afines como:

$$L_p = \{h_{\boldsymbol{w},b} \ : \ \boldsymbol{w} \in \mathbb{R}^p, \ b \in \mathbb{R}\},$$

donde

$$h_{\boldsymbol{w},b}(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = \left(\sum_{i=1}^{p} w_i x_i \right) + b.$$

Las clases que veremos a continuación son composiciones de una función $\varphi : \mathbb{R} \to \mathcal{Y}$ con L_p. En un caso de *clasificación binaria*, podemos proponer:

$$\varphi(h(x)) = \begin{cases} 1, & h(x) > 0 \\ 0, & \text{en otros casos} \end{cases}$$

Por otro lado, para problemas de *regresión*, nuestra φ fácilmente puede ser la función identidad, i.e. $\varphi(h(x)) = h(x)$.

1.1. Modelo de semiespacios

Definición 1.2 (Clase de semiespacios)**.** La clase de semiespacios, diseñada para problemas de clasificación binaria con $\mathcal{X} = \mathbb{R}^p$ y $\mathcal{Y} = \{-1, +1\}$, la definimos como:

$$HS_p = \text{sgn} \circ L_p = \{\text{sgn}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) : \boldsymbol{w} \in \mathbb{R}^p\}.$$

[1] Inicialmente, utilizaremos la *minimización de riesgo empírico* (ERM, por sus siglas en inglés) como método de aprendizaje preferido.

Dada la definición anterior, podemos considerar las siguientes tres situaciones:

1. Casos separables al asumir que se cumple la *hipótesis de realizabilidad*.
2. Casos no-separables en un sentido agnóstico.
3. Cuando una función lineal no es suficiente y necesitamos una transformación no lineal.

1.1.1. Programación lineal

Los programas lineales son problemas que pueden expresarse como la maximización de una función lineal sujeta a restricciones lineales. Es decir,

$$\begin{aligned} \max_{\boldsymbol{w}\in\mathbb{R}_p} \quad & \langle \boldsymbol{u}, \boldsymbol{w}\rangle \\ \text{sujeto a} \quad & A\boldsymbol{w} \geq \boldsymbol{v} \end{aligned}$$

Observación 1.3. Sea $S = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$ un conjunto de entrenamiento. Al asumir el caso 1, existe $\boldsymbol{w} \in \mathbb{R}^p$ tal que

$$y_i\langle \boldsymbol{w}, \boldsymbol{x}^{(i)}\rangle \geq 1, \qquad \forall i = 1, \ldots, m,$$

y $\boldsymbol{w}$ es un predictor ERM.

Por la observación anterior, si definimos $A \in \mathbb{R}^{m\times p}$ tal que $A_{i,j} = y^{(i)}x_j^{(i)}$, entonces podemos escribir $A\boldsymbol{w} \geq \boldsymbol{v}$.

1.1.2. Algoritmo de perceptrones

Este algortimo iterativo construye una secuencia de vectores $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots$. Inicialmente, establecemos $\boldsymbol{w}^{(1)} = \bar{0}$ (vector de ceros). En la iteración t, el Perceptrón encuentra un ejemplo i mal etiquetado por $\boldsymbol{w}^{(t)}$, es decir, $y^{(i)}\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}^{(i)}\rangle < 0$. Entonces, el Perceptrón actualiza $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + y^{(i)}\boldsymbol{x}^{(i)}$. Recordemos que nuestro objetivo es tener $y^{(i)}\langle \boldsymbol{w}, \boldsymbol{x}^{(i)}\rangle > 0$ para toda i. Nótese que

$$y^{(i)}\langle \boldsymbol{w}^{(t+1)}, \boldsymbol{x}^{(i)}\rangle = y^{(i)}\langle \boldsymbol{w}^{(t)} + y^{(i)}\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i)}\rangle = y^{(i)}\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}^{(i)}\rangle + \|\boldsymbol{w}^{(i)}\|^2.$$

Teorema 1.4. *Al asumir* $\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$ *separables, y sean* $B = \min\left\{\|\boldsymbol{w}\| : y^{(i)}\langle \boldsymbol{w}, \boldsymbol{x}^{(i)}\rangle \geq 1\right\}$, *y* $R = \max_i\left\{\|\boldsymbol{x}^{(i)}\|\right\}$. *Entonces, el algortimo de perceptrones se detiene a lo más* $(RB)^2$ *iteraciones después y devuelve* $\boldsymbol{w}^{(t)}$ *tal que* $y^{(i)}\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}^{(i)}\rangle > 0$.

Demostración. Bastante larga para estas notas. □

Agradecimientos

Este *template* se ha adaptado y traducido del provisto en la clase ACM 204 (Otoño 2017) por el profesor Joel Tropp.