Probabilidades y Estadística - Clasificación

Federico Yulita

Primer cuatrimestre, 2019.

Índice

1. Introducción

El problema de clasificación es un ejemplo de aprendizaje supervisado. El problema consiste en clasificar ciertos datos $\boldsymbol{x}$ con distintos *labels* (o clases) $\boldsymbol{t}$ mediante una función lineal para luego poder predecir la clasificación de nuevos datos. La elección de la codificación para las clases - es decir el valor de $\boldsymbol{t}$ - tiene un gran impacto en la optimización del problema. En la Figura 1 se puede ver un ejemplo de clasificación. En el ejemplo vamos a clasificar datos que caigan por debajo de recta naranja como de clase roja y a los que caigan por encima de clase azul.

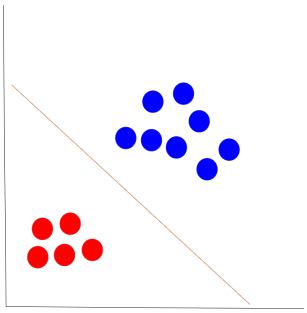

Figura 1: Ejemplo de clasificación.

Existen dos tipos de métodos para resolver este problema: El **Método Directo** y el **Método Probabilista**. En el método directo se va a buscar optimizar algo sin modelos y en el probabilista se usa un modelo estadístico de los datos.

2. *Generalized Linear Model*

En este modelo vamos a usar una función f de activación con parámetros $\boldsymbol{w}$ que vamos a intentar optimizar para que la función clasifique los datos. Consideremos:

$$y_k(x) = \boldsymbol{w}_k^\dagger \cdot \boldsymbol{x},$$

donde a $\boldsymbol{x} = \left(1, \boldsymbol{x}^\dagger\right)^\dagger$. Es decir, le agregamos un 1 a los datos para que el primer término proveniente del producto interno es una ordenada al origen. El subíndice k indica la clase. Entonces:

$$y(\boldsymbol{x}) = \bar{\bar{W}}^\dagger \cdot \boldsymbol{x}. \tag{1}$$

Definimos la **Función de Error** como:

$$Err\left(\bar{\bar{W}}\right) = \frac{1}{2}Tr\left(\left(\bar{\bar{X}}\cdot\bar{\bar{W}}-\bar{\bar{T}}\right)^{\dagger}\cdot\left(\bar{\bar{X}}\cdot\bar{\bar{W}}-\bar{\bar{T}}\right)\right) \tag{2}$$
$$= \frac{1}{2}\sum_{i,k}\left(\boldsymbol{x}_i\cdot\boldsymbol{w}_k^{\dagger}-t_i\right)^2.$$

Minimizando esta función de error (2) podemos encontrar los parámetros óptimos y una vez hallados podemos clasificar nuevos datos usando la función (1). Estaría bueno que esta última ecuación esté definida en el rango $[0,1]$ para usarla como una probabilidad.

Resulta que sale de la minimización analítica que los parámetros óptimos son:

$$\bar{\bar{W}} = \left(\bar{\bar{X}}^{\dagger}\cdot\bar{\bar{X}}\right)^{-1}\cdot\bar{\bar{X}}^{\dagger}\cdot\bar{\bar{T}}. \tag{3}$$

El problema de este método para clasificar es que los *outliers* tienen un gran peso sobre los parámetros hallados. Para solucionar esto lo que se hace es usar el **Discriminante de Fisher** para optimizar. Esto es, se toma el promedio de los datos de cada clase y entonces se trabajan con varios promedios de cada clase para cada set de datos. Luego se define:

$$s_k^2 = \sum_{n\in C_k}\left|y_k-\boldsymbol{w}^{\dagger}\cdot\boldsymbol{m}_k\right|^2,$$

donde C_k es la clase k y $\boldsymbol{m}_k$ es el promedio de la clase k. Entonces, s_k es como una medida de la distancia máxima entre los promedios para cada clase. Entonces se define el discriminante de Fisher (para dos clases) como:

$$J(\boldsymbol{w}) = \frac{\left(\boldsymbol{w}^{\dagger}\cdot(\boldsymbol{m}_1-\boldsymbol{m}_2)\right)^2}{s_1^2-s_2^2}. \tag{4}$$

Esta solución es mucho mejor que la otra ya que los outliers ya no pesan sobre la solución. Lo único malo es que no es una probabilidad.

3. Perceptrón

Consideremos un vector $\boldsymbol{x}$ con *features* $\boldsymbol{\Phi}(\boldsymbol{x})$. definimos $y(\boldsymbol{x}) = f\left(\boldsymbol{w}^{\dagger}\cdot\boldsymbol{x}\right)$, donde $f(a) = \begin{cases}1 & a\geqslant 0\\ 0 & a<0\end{cases}$. Vamos a usar una codificación para las clases de tipo $t_i\in\{-1,1\}$. Decimos que $\begin{cases}\boldsymbol{w}^{\dagger}\cdot\boldsymbol{\Phi}(x_n)<0 \implies t_n=-1\\ \boldsymbol{w}^{\dagger}\cdot\boldsymbol{\Phi}(x_n)>0 \implies t_n=1\end{cases}$. El objetivo entonces es hallar los parámetros que minimicen:

$$Err(\boldsymbol{w}) = -\sum_n \boldsymbol{w}^{\dagger}\cdot\boldsymbol{\Phi}(x_n)\,t_n. \tag{5}$$

Para esto puede usarse el algoritmo de ***Gradient Descent***. Este algoritmo consiste en iterar $\boldsymbol{w}^{\dagger}\overset{\text{def}}{:=}\boldsymbol{w}^{\dagger}-\eta\boldsymbol{\nabla}Err(\boldsymbol{w})$ hasta converger a un mínimo, donde η es algún *step-size*. El problema de este algoritmo es que dependiendo de tus valores iniciales del parámetro podes llegar a distintas soluciones de los parámetro óptimos. En la Figura 2 puede verse la red neuronal utilizada en este caso. Es de una capa ya que tiene una sola columna de neuronas (en rosa) donde cada una tiene un peso, en el caso de esta función f solo una neurona tiene un valor no nulo y depende del input dado.

4. Método Probabilista

Consideremos la fórmula de Bayes:

$$\mathbb{P}\left(C_1|x\right) = \frac{\mathbb{P}\left(x|C_1\right)\mathbb{P}\left(C_1\right)}{\mathbb{P}\left(x|C_1\right)\mathbb{P}\left(C_1\right) + \mathbb{P}\left(x|C_2\right)\mathbb{P}\left(C_2\right)} \tag{6}$$

Notemos que esto es igual a $\sigma\left(A\right)$, donde σ es la **Función Sigmoidal**:

$$\sigma\left(x\right) = \frac{1}{1+\exp\left(-x\right)}, \tag{7}$$

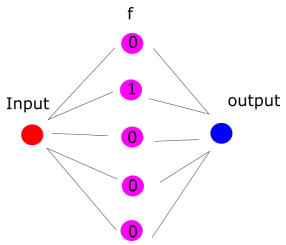

Figura 2: Red neuronal.

y $A = \ln\left(\frac{\mathbb{P}(x|C_1)\mathbb{P}(C_1)}{\mathbb{P}(x|C_2)\mathbb{P}(C_2)}\right)$. Si utilizamos el **Modelo Gaussiano** tenemos que:

$$\mathbb{P}\left(x|C_k\right) = \frac{1}{\left(2\pi\right)^d \left|\bar{\bar{\Sigma}}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)^\dagger \cdot \bar{\bar{\Sigma}}^{-1} \cdot \left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)\right),$$

donde $\bar{\bar{\Sigma}}$ es la matriz de covarianza. Entonces, usando el método de máxima verosimilitud podemos estimar $\boldsymbol{\mu}_k$, $\bar{\bar{\Sigma}}$ y $\mathbb{P}\left(C_k\right)$ y usando Bayes para predecir las clases. Vamos a usar que $\mathbb{P}\left(C_1|\boldsymbol{x}\right) = \sigma\left(\boldsymbol{w}^\dagger \cdot \boldsymbol{x} + w_o\right)$ y entonces hallamos que:

$$\begin{aligned}
\boldsymbol{w} &= \bar{\bar{\Sigma}}^{-1} \cdot \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^\dagger \cdot \bar{\bar{\Sigma}}^{-1} \cdot \boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\dagger \cdot \bar{\bar{\Sigma}}^{-1} \cdot \boldsymbol{\mu}_2 + \ln\left(\frac{\mathbb{P}\left(C_1\right)}{\mathbb{P}\left(C_2\right)}\right).
\end{aligned}$$

Glosario

D
Discriminante de Fisher, 2

F
Función de Error, 1
Función Sigmoidal, 3

G
Gradient Descent, 2

M
Método Directo, 1
Método Probabilista, 1
Modelo Gaussiano, 3