Regresión Logística

Ya conocemos el método de regresión lineal para predecir valores numéricos, pero ¿qué pasa si queremos hacer clasificación?

Uno pensaría que podemos ajustar una regresión lineal y definir un threshold. Así, los valores mayores al threshold son de una clase, mientras que los menores son de otra.

Sin embargo esta opción <u>no</u> nos gusta porque los valores que entrega la regresión **no están acotados**.

Afortunadamente podemos usar el siguiente truco!

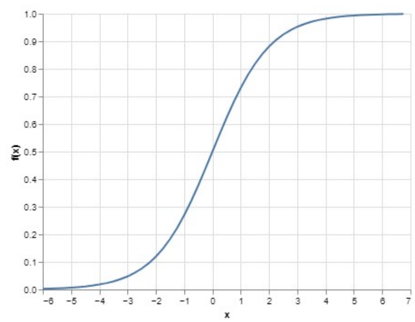
Supongamos una observación con k features x y un vector de coeficientes β. Un modelo útil para clasificar es:

$$\hat{p} = \sigma(x^T\beta) \quad \text{con} \quad \sigma(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{e^t+1}$$

$$\boxed{\hat{p} = \frac{1}{1+e^{-(x^T\beta)}}}$$

Esto es, intuitivamente, pasar el resultado de la regresión lineal por la función logística!

La función logística va de $]-\infty, \infty[\longmapsto [0,1]$



Luego para predecir:

$$\hat{y} = \begin{cases} 0 & \text{si } \hat{p} < 0.5 \\ 1 & \text{si } \hat{p} \geq 0.5 \end{cases}$$

Y como $\sigma(0) = 0.5$, el modelo predice 1 si $x^T\beta$ es mayor o igual que 0.

Ahora, ¿cómo encontramos los valores de β? Vamos a usar el método de máxima verosimilitud (Maximum likelihood).

Máxima verosimilitud

El método de máxima verosimilitud nos sirve para encontrar los parámetros de una distribución dado que conocemos ciertas observaciones (datos).

Sean $X_1, \dots, X_n$ valores IID con función de densidad $f(x;\theta)$, queremos encontrar θ!

→ Normalmente uno conoce θ (el parámetro) y pregunta la probabilidad de ver X_i.

Ej. Si sabemos que la estatura distribuye normal de parámetros μ, σ en una población, preguntamos la probabilidad de encontrar a alguien de 1.7 metros

En nuestro caso tenemos observaciones $X_1, \dots, X_n$ que vendrían a ser "estaturas" en el ejemplo anterior, y queremos aprender μ y σ (Ojo: aquí sabemos la distribución, nos faltan sus parámetros).

Ahora, la función de verosimilitud es:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

Y el estimador de máxima verosimilitud $\hat{\theta}$ es el valor de θ que maximiza $\mathcal{L}(\theta)$

Así encontramos el valor de θ que mejor se ajusta a nuestros datos. Pero, ¿cuál es la intuición detrás?

Sea $p_i = f(x_i;\theta)$, la probabilidad de ver $x_1, \dots, x_n$ es $p_1 \times \dots \times p_n = \prod_{i=1}^{n} f(x_i;\theta)$, así que intuitivamente, buscamos el parámetro que maximiza la probabilidad de "ver lo que vimos".

Entrenando una regresión logística

Recordemos que $\hat{p} = \sigma(x^T\beta)$ y queremos aprender β. Como esto es clasificación binaria, modelamos esto como una distribución Bernoulli:

$X \sim Be(p)$

$f(x) = p^x (1-p)^{1-x}$ con $x = \{0,1\}$

$$P(\underbrace{Y=y}_{\text{La respuesta}} \mid \underbrace{X=x}_{\text{Las features}}) = \sigma(x^T\beta)^y \cdot (1 - \sigma(x^T\beta))^{1-y}$$

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \left[\sigma(x_i^T \beta)^{y_i} \left(1 - \sigma(x_i^T \beta)\right)^{1-y_i} \right]$$

Usamos logaritmo:

$$\log(\mathcal{L}(\beta)) = \sum_{i=1}^{n} \left[y_i \log(\sigma(x_i^T \beta)) + (1-y_i) \log(1 - \sigma(x_i^T \beta)) \right]$$

Así que lo que hacemos es maximizar

$\log(\mathcal{L}(\beta))$ que es equivalente a maximizar $\mathcal{L}(\beta)$ o minimizar $-\log(\mathcal{L}(\beta))$.

También encontraremos en la literatura que buscamos minimizar:

$$-\frac{1}{n} \log(\mathcal{L}(\beta)) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(\sigma(x_i^T \beta)) + (1-y_i) \log(1 - \sigma(x_i^T \beta)) \right]$$

Pero lamentablemente no hay una fórmula explícita para encontrar el óptimo. Sin embargo, al ser una función convexa, podemos usar Gradient Descent:

$$\frac{\delta}{\delta \beta_j} -\frac{1}{n} \log(\mathcal{L}(\beta)) = \frac{1}{n} \sum_{i=1}^{n} \left(\sigma(x_i^T \beta) - y_i \right) x_{ij}$$

Feature j de la observación i

Así, podemos encontrar el óptimo para nuestra función objetivo. El valor de β en el óptimo son los parámetros que mejor se aproximan para los datos que tenemos.