Ejemplo: Clasificador bayesiano ingenuo

Febrero 2023

Supongamos que tenemos una colección de 11 textos, 5 de ellos pertenecen a la categoría de informática y 6 de ellos a la categoría de deportes. También tenemos el siguiente diccionario y bolsas para cada uno de los textos:

$$V = \begin{vmatrix} P_1 = \text{Gol} \\ P_2 = \text{Maestro} \\ P_3 = \text{Velocidad} \\ P_4 = \text{Defensa} \\ P_5 = \text{Rendimiento} \\ P_6 = \text{Campo} \\ P_7 = \text{Movimiento} \\ P_8 = \text{Ataque} \end{vmatrix}$$

$$B_D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \begin{array}{c} \begin{array}{cccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \end{pmatrix} \end{array} \tag{1}$$

$$B_I = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{cccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \tag{2}$$

A partir de los datos anteriores entrene un clasificador bayesiano ingenuo y calcule la predicción para los nuevos textos:

$$\tilde{x_1} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{3}$$

$$\tilde{x_2} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{4}$$

- Entrenamiento:
  La fase de entrenamiento consiste principalmente en encontrar los parámetros $(\hat{q_1}, \hat{q_2}, \ldots, \hat{q_8})$ condicionadas a las clases informática y deporte, suponiendo que cada $p_i \sim Ber(q_i)$ y que cada parámetro lo podemos estimar por máxima verosimilitud.

- ($Y = Deportes$)
  Recordemos que el estimador máximo verosimil del parámetro q de una variable aleatoria bernoulli es el promedio, así que para estimar cada $\hat{q}_i$, tomamos la columna i de la bolsa de deportes y obtenemos el promedio.

$$\hat{q}_1 = \frac{1}{2}, \hat{q}_2 = \frac{1}{6}, \hat{q}_3 = \frac{2}{3}, \hat{q}_4 = \frac{1}{3}, \hat{q}_5 = \frac{2}{3}, \hat{q}_6 = \frac{5}{6}, \hat{q}_7 = \frac{1}{6}, \hat{q}_8 = \frac{2}{3}$$

- ($Y = Informtica$)
  Análogamente para estimar cada $\hat{q}_i$, tomamos la columna i de la bolsa de informática y obtenemos el promedio.

$$\hat{q}_1 = \frac{1}{5}, \hat{q}_2 = \frac{3}{5}, \hat{q}_3 = \frac{1}{5}, \hat{q}_4 = 1, \hat{q}_5 = \frac{1}{5}, \hat{q}_6 = \frac{1}{5}, \hat{q}_7 = 1, \hat{q}_8 = \frac{3}{5}$$

- (Y)
  Asiganando el valor 1 cuando el documento es informática.

$$\hat{q} = \frac{5}{11}$$

- Predicción
  Usando teorema de Bayes:

$$P(y|\tilde{x}) = P(y|p_1)P(y|p_2)\dots P(y|p_8)P(y) \tag{5}$$

  Así, $P(deportes|\tilde{x_1}) = \frac{1}{2}\frac{5}{6}\frac{1}{3}\frac{1}{3}\frac{2}{3}\frac{5}{6}\frac{1}{6}\frac{2}{3}\frac{6}{11} = 0{,}001558798$ y $P(informatica|\tilde{x_1}) = \frac{4}{5}\frac{2}{5}\frac{4}{5}1\frac{1}{5}\frac{1}{5}1\frac{3}{5}\frac{5}{11} = 0{,}002792727$