Tarea 2: Clasificador bayesiano ingenuo

Saul Ivan Rivas Vega

Aprendizaje Automatizado

5 de marzo de 2020

1. Géneros

Un programa de salud gubernamental desea clasificar los registros de las personas en géneros femenino (F) o masculino (M) a partir de los atributos nombre, estatura y peso. Se cuentan con los siguientes registros:

Nombre	Estatura (m)	Peso (Kg)	Género
Denis	1.72	75.3	M
Guadalupe	1.82	81.6	M
Alex	1.80	86.1	M
Alex	1.70	77.1	M
Cris	1.73	78.2	M
Juan	1.80	74.8	M
Juan	1.80	74.3	M
Denis	1.50	50.5	F
Alex	1.52	45.3	F
Cris	1.62	61.2	F
Rene	1.67	68.0	F
Guadalupe	1.65	58.9	F
Guadalupe	1.75	68.0	F

Entrena un clasificador bayesiano ingenuo usando estimación por máxima verosimilitud y otro usando estimación por máximo a posteriori. Reporta los parámetros que obtuviste en ambos casos y usa los clasificadores entrenados para predecir la clase de los siguientes vectores: x1 = (Rene, 1.68, 65), x2 = (Guadalupe, 1.75, 80), x3 = (Denis, 1.80, 79), x4 = (Alex, 190, 85) y x5 = (Cris, 165, 70). Describe de forma detallada el procedimiento que seguiste tanto en el entrenamiento como en la predicción y discute los resultados obtenidos. Para el entrenamiento del clasificador por máximo a posteriori considera los siguientes valores para las distribuciones correspondientes:

Género	Nombre	Estatura			Peso		
	α_k	μ_0	σ_0^2	σ^2	μ_0	σ_0^2	σ^2
M	$1, \forall k$	1.7	0.3	0.0020	85.5	17.0	15.76
F	$1, \forall k$	1.5	0.1	0.0074	70.3	85.0	71.00

1.1. Estimador por máxima verosimilitud

Los atributos son: **nombre**, **estatura** y **peso**, y la clase es **género**.

1.1.1. Atributo *Nombre*

Para el **nombre** podemos asumir una distribución categórica:

$$X_{nombre}^{(i)} \sim Cat(X_{nombre}^{(i)}; q) \tag{1}$$

Donde las categorías son los nombres y los podemos enumerar:

1. Denis
2. Guadalupe
3. Alex
4. Cris
5. Juan
6. Rene

Y así con los nombres de 1 a 6 podemos definir a $Cat(X_{nombre}^{(i)}; q)$ como:

$$Cat(X_{nombre}^{(i)}; q) = \prod_{k=1}^{6} q_k^{[x_{nombre}^{(i)}=k]} \tag{2}$$

Donde podemos estimar a q_k usando el estimador de máxima verosimilitud como:

$$\begin{gathered} \hat{q}_k = \frac{c_k}{n} \\ \text{Donde } c_k\text{:} \\ c_k = \sum_{i=1}^{n} [x_{nombre}^{(i)} = k] \end{gathered} \tag{3}$$

Así podemos estimar el parámetro para las primer categoría:

$$\begin{aligned} c_1 &= \sum_{i=1}^{13} [x_{nombre}^{(i)} = 1] \\ &= 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0 \\ &= 2 \\ \hat{q}_1 &= \frac{2}{13} \end{aligned} \tag{4}$$

Y de la misma forma para las categorías restantes:

$$\begin{gathered} c_2 = 3, \quad \hat{q}_2 = \frac{3}{13} \qquad c_3 = 3, \quad \hat{q}_3 = \frac{3}{13} \\ c_4 = 2, \quad \hat{q}_4 = \frac{2}{13} \qquad c_5 = 2, \quad \hat{q}_5 = \frac{2}{13} \\ c_6 = 1, \quad \hat{q}_6 = \frac{1}{13} \end{gathered} \tag{5}$$

1.1.2. Atributo Estatura

Para la **estatura** podemos asumir una distribución normal:

$$X^{(i)}_{estatura} \sim \mathcal{N}(X^{(i)}_{estatura}; \mu; \sigma^2) \tag{6}$$

Donde $\mathcal{N}(X^{(i)}_{estatura}; \mu; \sigma^2)$ se define como:

$$\mathcal{N}(X^{(i)}_{estatura}; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x^{(i)}-\mu)^2}{2\sigma^2}} \tag{7}$$

2. Spam