

Tarea 2: Clasificador bayesiano ingenuo

Saul Ivan Rivas Vega

Aprendizaje Automatizado

7 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 (<i>m</i>)	Peso (<i>Kg</i>)	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: $x_1 = (\text{Rene}, 1.68, 65)$, $x_2 = (\text{Guadalupe}, 1.75, 80)$, $x_3 = (\text{Denis}, 1.80, 79)$, $x_4 = (\text{Alex}, 1.90, 85)$ y $x_5 = (\text{Cris}, 1.65, 70)$. 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:

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

Donde las categorías son los nombres y los podemos enumerar:

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

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

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

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

$$\hat{q}_k = \frac{c_k}{n}$$

Donde c_k :

$$c_k = \sum_{i=1}^n [x_{nombre}^{(i)} = k] \quad (3)$$

Así podemos estimar el parámetro para la primer categoría:

Para la clase **Femenino** :

$$\begin{aligned} c_{1F} &= \sum_{i=1}^{13} [x_{nombre}^{(i)} = 1 \text{ y } x^{(i)} \text{ es de la clase Femenino}] \\ &= 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0 \\ &= 1 \\ \hat{q}_{(1|F)} &= \frac{1}{6} \end{aligned} \quad (4)$$

Para la clase **Masculino** :

$$\begin{aligned} c_{1M} &= \sum_{i=1}^{13} [x_{nombre}^{(i)} = 1 \text{ y } x^{(i)} \text{ es de la clase Masculino}] \\ &= 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 \\ &= 1 \\ \hat{q}_{(1|M)} &= \frac{1}{7} \end{aligned}$$

Y de la misma forma para las categorías restantes:

$$\begin{aligned}
c_{2F} = 2, \hat{q}_{(2|F)} = \frac{2}{6}, c_{2M} = 1, \hat{q}_{(2|M)} = \frac{1}{7} & \quad c_{3F} = 1, \hat{q}_{(3|F)} = \frac{1}{6}, c_{3M} = 2, \hat{q}_{(3|M)} = \frac{2}{7} \\
c_{4F} = 1, \hat{q}_{(4|F)} = \frac{1}{6}, c_{4M} = 1, \hat{q}_{(4|M)} = \frac{1}{7} & \quad c_{5F} = 0, \hat{q}_{(5|F)} = \frac{0}{6} = 0, c_{5M} = 2, \hat{q}_{(5|M)} = \frac{2}{7} \\
c_{6F} = 1, \hat{q}_{(6|F)} = \frac{1}{6}, c_{6M} = 0, \hat{q}_{(6|M)} = \frac{0}{7} = 0 &
\end{aligned} \tag{5}$$

1.1.2. Atributo Estatura

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

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

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

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

Donde podemos estimar a μ y a σ usando el estimador de máxima verosimilitud como:

Para la clase **Femenino**:

$$\begin{aligned}
\hat{\mu}_F &= \frac{1}{n} \sum_{i=1}^n x_{estatura}^{(i)} \\
&= \frac{1}{6}(1.50 + 1.52 + 1.62 + 1.67 + 1.65 + 1.75) \\
&= \frac{1}{6}(9.71) \\
&= 1.618\bar{3}
\end{aligned} \tag{8}$$

$$\begin{aligned}
\hat{\sigma}_F^2 &= \frac{1}{n} \sum_{i=1}^n (x_{estatura}^{(i)} - \hat{\mu})^2 \\
&= \frac{1}{6}[(1.50 - 1.618\bar{3})^2 + (1.52 - 1.618\bar{3})^2 + \dots + (1.75 - 1.618\bar{3})^2] \\
&= \frac{1}{6}[0.014003 + 0.009669 + 0.000002\bar{7} + \dots + 0.0173361111] \\
&= \frac{1}{6}[0.04468\bar{3}] \\
&= 0.007447\bar{2}
\end{aligned} \tag{9}$$

Para la clase **Masculino**:

$$\begin{aligned}
\hat{\mu}_M &= \frac{1}{n} \sum_{i=1}^n x_{estatura}^{(i)} \\
&= \frac{1}{7}(1.72 + 1.82 + 1.80 + 1.70 + 1.73 + 1.80 + 1.80) \\
&= \frac{1}{7}(12.37) \\
&= 1.7671428571428571
\end{aligned} \tag{10}$$

$$\begin{aligned}
\hat{\sigma}_M^2 &= \frac{1}{n} \sum_{i=1}^n (x_{estatura}^{(i)} - \hat{\mu})^2 \\
&= \frac{1}{7}[(1.72 - 1.76714)^2 + (1.82 - 1.76714)^2 + \dots + (1.80 - 1.76714)^2] \\
&= \frac{1}{7}[0.00222245 + 0.00279388 + 0.00107959 + \dots + 0.00107959] \\
&= \frac{1}{7}[0.014142857142857169] \\
&= 0.0020204081632653097
\end{aligned} \tag{11}$$

1.1.3. Atributo Peso

Para el **peso** podemos asumir una distribución normal:

$$X_{peso}^{(i)} \sim \mathcal{N}(X_{peso}^{(i)}; \mu; \sigma^2) \tag{12}$$

Donde $\mathcal{N}(X_{peso}^{(i)}; \mu; \sigma^2)$ se define como:

$$\mathcal{N}(X_{peso}^{(i)}; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x^{(i)} - \mu)^2}{2\sigma^2}} \tag{13}$$

Donde podemos estimar a μ y a σ usando el estimador de máxima verosimilitud como:

Para la clase **Femenino**:

$$\begin{aligned}
\hat{\mu}_F &= \frac{1}{n} \sum_{i=1}^n x_{peso}^{(i)} \\
&= \frac{1}{6}(50.5 + 45.3 + 61.2 + 68.0 + 58.9 + 68.0) \\
&= \frac{1}{6}(351.9) \\
&= 58.65
\end{aligned} \tag{14}$$

$$\begin{aligned}
\hat{\sigma}_F^2 &= \frac{1}{n} \sum_{i=1}^n (x_{peso}^{(i)} - \hat{\mu})^2 \\
&= \frac{1}{6} [(50.5 - 58.65)^2 + (45.3 - 58.65)^2 + \dots + (68.0 - 58.65)^2] \\
&= \frac{1}{6} [66.42250 + 178.2225 + 6.502500 + \dots + 87.42250] \\
&= \frac{1}{6} [426.055] \\
&= 71.00916\bar{7}
\end{aligned} \tag{15}$$

Para la clase **Masculino**:

$$\begin{aligned}
\hat{\mu}_M &= \frac{1}{n} \sum_{i=1}^n x_{peso}^{(i)} \\
&= \frac{1}{7} (75.3 + 81.6 + 86.1 + 77.1 + 78.2 + 74.8 + 74.3) \\
&= \frac{1}{7} (547.4) \\
&= 78.2
\end{aligned} \tag{16}$$

$$\begin{aligned}
\hat{\sigma}_M^2 &= \frac{1}{n} \sum_{i=1}^n (x_{peso}^{(i)} - \hat{\mu})^2 \\
&= \frac{1}{7} [(75.3 - 78.2)^2 + (81.6 - 78.2)^2 + \dots + (74.3 - 78.2)^2] \\
&= \frac{1}{7} [8.41 + 11.56 + 62.41 + \dots + 15.21] \\
&= \frac{1}{7} [110.36] \\
&= 15.7657142857142
\end{aligned} \tag{17}$$

1.1.4. Género

Para la clase (Género) podemos asumir una distribución Bernoulli:

$$Y^{(i)} \sim Ber(Y^{(i)}; q) \quad (18)$$

Donde $Ber(Y^{(i)}; q)$ se define como:

$$\begin{aligned} Ber(Y^{(i)}; q) &= q^C (1 - q)^{1-C} \\ C &= [y = \text{clase}] \end{aligned} \quad (19)$$

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

Para la clase **Femenino**:

$$\begin{aligned} \hat{q}_F &= \frac{N_F}{N} \\ &= \frac{6}{13} \end{aligned} \quad (20)$$

Para la clase **Masculino**:

$$\begin{aligned} \hat{q}_M &= \frac{N_M}{N} \\ &= \frac{7}{13} \end{aligned} \quad (21)$$

1.1.5. Uso del Estimador por Máxima Verosimilitud

Clasificaremos al vector de entrada con base en la siguiente ecuación:

$$\begin{aligned} C &= \max \arg_{C \in \{F, M\}} \{P(X|C)P(C)\} \\ \text{Donde las probabilidades de cada parametro al ser independientes se multiplicaran} \\ &= \max \arg_{C \in \{F, M\}} \{P(C)P(X_{\text{nombre}}|C)P(X_{\text{estatura}}|C)P(X_{\text{peso}}|C)\} \end{aligned} \quad (22)$$

2. Spam