Aprendizaje Automatizado

Tarea 2

PCIC - UNAM

18 de marzo de 2020

Diego de Jesús Isla López
(*dislalopez@gmail.com*)
(*diego.isla@comunidad.unam.mx*)

Ejercicio 1

Para ambos estimadores se tomarán distribuciones normales para los atributos **estatura** y **peso**, así como una distribución categórica para el atributo **nombre**. Para las clase **género** (M, F), se toma una distribución Bernoulli.

Estimador de máxima verosimilitud (EMV)

Dado que para los atributos **estatura** y **peso** se toma una distribución normal, tenemos:

$$\hat{\mu}_{EMV} = \frac{1}{n} \cdot \sum_{i=1}^{n} x^{(i)} \tag{1}$$

$$\sigma^2_{EMV} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x^{(i)} - \hat{\mu}_{EMV})^2 \tag{2}$$

Para el atributo **nombre** se calcula

$$\hat{q}_k = \frac{1}{n} \cdot c_k \tag{3}$$

Para la clase **M** tenemos:

- Estatura:

$$\hat{\mu}_{M_{estatura}} = \frac{1}{7} \cdot \sum_{i=1}^{7} x^{(i)} = \frac{1}{7} \cdot (12.37) = 1.7671 \tag{4}$$

$$\sigma^2_{M_{estatura}} = \frac{1}{7} \cdot \sum_{i=1}^{7} (x^{(i)} - \hat{\mu}_{M_{estatura}})^2 = \frac{1}{7} \cdot (0.0137) = 0.0019 \tag{5}$$

- Peso:

$$\hat{\mu}_{M_{peso}} = \frac{1}{7} \cdot \sum_{i=1}^{7} x^{(i)} = \frac{1}{7} \cdot (547.4) = 78.2 \tag{6}$$

$$\sigma^2_{M_{peso}} = \frac{1}{7} \cdot \sum_{i=1}^{7} (x^{(i)} - \hat{\mu}_{M_{peso}})^2 = \frac{1}{7} \cdot (110.3599) = 15.7657 \tag{7}$$

- Nombre:

$$\hat{q}_{Denis} = \frac{1}{7} \tag{8}$$

$$\hat{q}_{Alex} = \frac{2}{7} \tag{9}$$

$$\hat{q}_{Cris} = \frac{1}{7} \tag{10}$$

$$\hat{q}_{Juan} = \frac{2}{7} \tag{11}$$

$$\hat{q}_{Guadalupe} = \frac{1}{7} \tag{12}$$

$$\hat{q}_{Rene} = \frac{0}{7} \tag{13}$$

Para la clase **F** tenemos:

- Estatura:

$$\hat{\mu}_{F_{estatura}} = \frac{1}{6} \cdot \sum_{i=1}^{6} x^{(i)} = \frac{1}{6} \cdot (9.7098) = 1.6183 \tag{14}$$

$$\sigma^2_{F_{estatura}} = \frac{1}{6} \cdot \sum_{i=1}^{6} (x^{(i)} - \hat{\mu}_{F_{estatura}})^2 = \frac{1}{6} \cdot (0.1344) = 0.0224 \tag{15}$$

- Peso:

$$\hat{\mu}_{F_{peso}} = \frac{1}{6} \cdot \sum_{i=1}^{6} x^{(i)} = \frac{1}{6} \cdot (351.9) = 58.65 \tag{16}$$

$$\sigma^2_{F_{peso}} = \frac{1}{6} \cdot \sum_{i=1}^{6} (x^{(i)} - \hat{\mu}_{F_{peso}})^2 = \frac{1}{6} \cdot (426.0546) = 71.0091 \tag{17}$$

- Nombre:

$$\hat{q}_{Denis} = \frac{1}{6} \tag{18}$$

$$\hat{q}_{Alex} = \frac{1}{6} \tag{19}$$

$$\hat{q}_{Cris} = \frac{1}{6} \tag{20}$$

$$\hat{q}_{Juan} = \frac{0}{6} \tag{21}$$

$$\hat{q}_{Guadalupe} = \frac{2}{6} \tag{22}$$

$$\hat{q}_{Rene} = \frac{1}{6} \tag{23}$$

Los parámetros de la clase **género** los obtenemos mediante:

$$\hat{q}_k = \frac{N_k}{N} \tag{24}$$

Entonces, para **M** tenemos:

$$\hat{q}_M = \frac{7}{13} \tag{25}$$

Para **F** tenemos:

$$\hat{q}_F = \frac{6}{13} \tag{26}$$

Dado que los atributos son independientes, la probabilidad en cada clase se calculará como:

$$P(F|x) = P(F) \cdot P(x_{nombre}|F) \cdot P(x_{estatura}|F) \cdot P(x_{peso}|F) \tag{27}$$

$$P(M|x) = P(M) \cdot P(x_{nombre}|M) \cdot P(x_{estatura}|M) \cdot P(x_{peso}|M) \tag{28}$$

La probabilidad de los atributos con distribución normal se calcula como:

$$L(\mu, \sigma^2|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x^{(i)}-\mu)^2}{2\sigma^2}} \tag{29}$$

Para los atributos con distribución categórica y Bernoulli, se utilizan sus respectivos valores de q_k.

Utilizado el estimador para el primer caso $x_1 =$ (Rene, 1.68, 65), tenemos:

- Probabilidad para **F**:

$$P(F|x_1) = \frac{6}{13} \cdot \frac{1}{6} \cdot (3.4113) \cdot (0.0341) = 0.00894 = 0.89\,\% \tag{30}$$

- Probabilidad para **M**: Dado que la probabilidad para $\hat{q}_{Rene} = 0$ para la clase **M**, la probabilidad es 0.

Entonces, el resultado de estimador para x_1 es **F**.

Para el caso $x_2 = $ (Guadalupe, 1.75, 80), tenemos:

- Probabilidad para **F**:

$$P(F|x_2) = \frac{6}{13} \cdot \frac{1}{6} \cdot (1.5998) \cdot (0.0229) = 0.0007 = 0.07\,\% \tag{31}$$

- Probabilidad para **M**:

$$P(M|x_2) = \frac{7}{13} \cdot \frac{1}{7} \cdot (7.7204) \cdot (0.0851) = 0.0505 = 5.05\,\% \tag{32}$$

Entonces, el resultado de estimador para x_2 es **M**.

Para el caso $x_3 = $ (Denis, 1.80, 79), tenemos:

- Probabilidad para **F**:

$$P(F|x_3) = \frac{6}{13} \cdot \frac{1}{6} \cdot (0.6658) \cdot (0.0038) = 0.0001 = 0.01\,\% \tag{33}$$

- Probabilidad para **M**:

$$P(M|x_3) = \frac{7}{13} \cdot \frac{1}{7} \cdot (6.5352) \cdot (0.0914) = 0.0459 = 4.59\,\% \tag{34}$$

Entonces, el resultado de estimador para x_3 es **M**.

Para el caso $x_4 = $ (Alex, 1.90, 85), tenemos:

- Probabilidad para **F**:

$$P(F|x_4) = \frac{6}{13} \cdot \frac{1}{6} \cdot (0.0498) \cdot (0.0007) = 0.000002 = 0.0002\,\% \tag{35}$$

- Probabilidad para **M**:

$$P(M|x_4) = \frac{7}{13} \cdot \frac{1}{7} \cdot (0.1943) \cdot (0.0264) = 0.0007 = 0.07\,\% \tag{36}$$

Entonces, el resultado de estimador para x_4 es **M**.

Para el caso $x_5 = $ (Cris, 1.65, 70), tenemos:

- Probabilidad para **F**:

$$P(F|x_5) = \frac{6}{13} \cdot \frac{1}{6} \cdot (3.9898) \cdot (0.0202) = 0.0062 = 0.62\,\% \tag{37}$$

- Probabilidad para **M**:

$$P(M|x_5) = \frac{7}{13} \cdot \frac{1}{7} \cdot (0.4472) \cdot (0.0149) = 0.0005 = 0.05\,\% \tag{38}$$

Entonces, el resultado de estimador para x_5 es **F**.

Estimador máximo a posteriori (MAP)

Para este estimador se tomará un valor de $\alpha = 2$ para el atributo **nombre** en ambas clases.

En el caso de los atributos con distribución normal (**estatura**, **peso**), el estimador se calcula como:

$$\hat{\mu} = \frac{\sigma_0^2(\sum_{i=1}^{n} x^{(i)} + \sigma^2 \mu_0)}{\sigma_0^2 \cdot n + \sigma^2} \tag{39}$$

donde σ_0^2 y σ^2 se conocen de antemano.

Para el atributo categórico **nombre**, el estimador se calcula como:

$$\hat{q}_k = \frac{c_k + a_k - 1}{n + \sum_{k=1}^{K} - K} \tag{40}$$

donde K es el número total de clases del atributo (6) y n es el número de elementos de la clase (para **M** o **F**).

Finalmente, el estimador para la clase **género** se obtiene mediante:

$$\hat{q}_k = \frac{N_k + \alpha_k - 1}{N + \beta_k + \alpha_k - 2} \tag{41}$$

donde β_k es el número de elementos de la clase.

Para la clase **M** tenemos:

- Estatura:
  - $\mu_0 = 1.7$
  - $\sigma_0^2 = 0.3$

- $\sigma^2 = 0.0020$

Entonces:

$$\hat{\mu}_{M_{estatura}} = \frac{(0.3)(12.37) + (0.0020)(1.7)}{(7)(0.3) + 0.0020} = 1.767 \tag{42}$$

- Peso:
  - $\mu_0 = 85.5$
  - $\sigma_0^2 = 17.0$
  - $\sigma^2 = 15.76$

  Entonces:

$$\hat{\mu}_{M_{peso}} = \frac{(17)(547.4) + (15.76)(85.5)}{(17)(7) + 15.76} = 79.0537 \tag{43}$$

- Nombre:

$$\hat{q}_{Denis} = \frac{1+2-1}{7-6+12} = \frac{2}{13} \tag{44}$$

$$\hat{q}_{Guadalupe} = \frac{1+2-1}{7-6+12} = \frac{2}{13} \tag{45}$$

$$\hat{q}_{Alex} = \frac{2+2-1}{7-6+12} = \frac{3}{13} \tag{46}$$

$$\hat{q}_{Cris} = \frac{1+2-1}{7-6+12} = \frac{2}{13} \tag{47}$$

$$\hat{q}_{Juan} = \frac{2+2-1}{7-6+12} = \frac{3}{13} \tag{48}$$

$$\hat{q}_{Rene} = \frac{0+2-1}{7-6+12} = \frac{1}{13} \tag{49}$$

Para la clase **F** tenemos:

- Estatura:
  - $\mu_0 = 1.5$
  - $\sigma_0^2 = 0.1$
  - $\sigma^2 = 0.0074$

  Entonces:

$$\hat{\mu}_{F_{estatura}} = \frac{(0.1)(9.7098) + (0.0074)(1.5)}{(6)(0.1) + 0.0074} = 1.6168 \tag{50}$$

- Peso:

- $\mu_0 = 70.3$
- $\sigma_0^2 = 85.0$
- $\sigma^2 = 71.0$

Entonces:

$$\hat{\mu}_{F_{peso}} = \frac{(85)(351.9) + (71)(70.3)}{(6)(85) + 71} = 60.0736 \tag{51}$$

- Nombre:

$$\hat{q}_{Denis} = \frac{1+2-1}{6-6+12} = \frac{1}{6} \tag{52}$$

$$\hat{q}_{Guadalupe} = \frac{2+2+1}{6-6+12} = \frac{1}{4} \tag{53}$$

$$\hat{q}_{Alex} = \frac{1+2-1}{6-6+12} = \frac{1}{6} \tag{54}$$

$$\hat{q}_{Cris} = \frac{1+2-1}{6-6+12} = \frac{1}{6} \tag{55}$$

$$\hat{q}_{Juan} = \frac{0+2-1}{6-6+12} = \frac{1}{12} \tag{56}$$

$$\hat{q}_{Rene} = \frac{1+2-1}{6-6+12} = \frac{1}{6} \tag{57}$$

Para la clase **género**, tenemos:

$$\hat{q}_{F} = \frac{6+2-1}{13+2+2-2} = \frac{7}{15} \tag{58}$$

$$\hat{q}_{F} = \frac{7+2-1}{13+2+2-2} = \frac{8}{15} \tag{59}$$

Utilizado el estimador para el primer caso $x_1 =$ (Rene, 1.68, 65), tenemos:

- Probabilidad para **F**:

$$P(F|x_1) = \frac{7}{15} \cdot \frac{1}{6} \cdot (3.5434) \cdot (0.0399) = 0.0109 = 1.09\,\% \tag{60}$$

- Probabilidad para **M**:

$$P(M|x_1) = \frac{8}{15} \cdot \frac{1}{13} \cdot (1.34) \cdot (0.0001) = 0.00001 = 0.001\,\% \tag{61}$$

Entonces, el resultado de estimador para x_1 es **F**.

Para el caso $x_2 =$ (Guadalupe, 1.75, 80), tenemos:

- Probabilidad para **F**:

$$P(F|x_2) = \frac{7}{15} \cdot \frac{1}{4} \cdot (1.4008) \cdot (0.0028) = 0.0004 = 0.04\,\% \tag{62}$$

- Probabilidad para **M**:

$$P(M|x_2) = \frac{8}{15} \cdot \frac{2}{13} \cdot (8.2932) \cdot (0.0976) = 0.0646 = 6.46\,\% \tag{63}$$

Entonces, el resultado de estimador para x_2 es **M**.

Para el caso $x_3 =$ (Denis, 1.80, 79), tenemos:

- Probabilidad para **F**:

$$P(F|x_3) = \frac{7}{15} \cdot \frac{1}{6} \cdot (0.4813) \cdot (0.0037) = 0.0001 = 0.01\,\% \tag{64}$$

- Probabilidad para **M**:

$$P(M|x_3) = \frac{8}{15} \cdot \frac{2}{13} \cdot (6.8033) \cdot (0.1004) = 0.056 = 5.6\,\% \tag{65}$$

Entonces, el resultado de estimador para x_3 es **M**.

Para el caso $x_4 =$ (Alex, 1.90, 85), tenemos:

- Probabilidad para **F**:

$$P(F|x_4) = \frac{7}{15} \cdot \frac{1}{6} \cdot (0.0206) \cdot (0.0005) = 0.0000009 = 0.00009\,\% \tag{66}$$

- Probabilidad para **M**:

$$P(M|x_4) = \frac{8}{15} \cdot \frac{3}{13} \cdot (0.1076) \cdot (0.0327) = 0.0004 = 0.04\,\% \tag{67}$$

Entonces, el resultado de estimador para x_4 es **M**.

Para el caso $x_5 =$ (Cris, 1.65, 70), tenemos:

- Probabilidad para **F**:

$$P(F|x_5) = \frac{7}{15} \cdot \frac{1}{6} \cdot (4.3065) \cdot (0.0236) = 0.0079 = 0.79\,\% \tag{68}$$

- Probabilidad para **M**:

$$P(M|x_5) = \frac{8}{15} \cdot \frac{2}{13} \cdot (0.2898) \cdot (0.0074) = 0.0001 = 0.01\,\% \tag{69}$$

Entonces, el resultado de estimador para x_5 es **F**.

Ejercicio 2

Los resultados obtenidos fueron:

- Reportados como spam: 1500 (29 %)
- Reportados como no spam: 3672 (71 %)

Se utilizaron dos clasificadores bayesianos: el primero usando una distribución multinomial para los regisros y el segundo usando una distribución Bernoulli, manejando los datos como incidencia de palabras en lugar de número de apariciones.

El clasificador multinomial obtuvo un 95 % de precisión en la predicción sobre el conjunto de entrenamiento y un 94 % sobre el conjunto de pruebas. A su vez, el clasificador Bernoulli obtuvo un 86 % de precisión en la predicción sobre el conjunto de entrenamiento y un 84 % sobre el conjunto de pruebas. Esto puede indicarnos utilizar ambos enfoques para un análisis de textos pudiera llegar a ser adecuado; sin embargo, es clara la ventaja que conlleva el utilizar una distribución multinomial.

Ejercicio 3

Se utilizaron tres conjuntos de prueba. Cada uno de ellos fue manipulado para completar los datos faltantes en tres formas: utilizando la media, la mediana y la moda. En los tres casos el clasificador obtuvo resultados de precisión del 100 % tanto con el conjunto de prueba como con el conjunto de entrenamiento. Esto pudiera indicar que el total de datos pudiera ser pequeño para el problema que se quiere resolver.