TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV – Grupo 075

(ist1103860, ist1103883)

I. Pen-and-paper

I Pen-and-paper

$\left\{\begin{pmatrix}1\\0.6\\0.1\end{pmatrix}, \begin{pmatrix}0\\-0.4\\0.8\end{pmatrix}, \begin{pmatrix}0\\0.2\\0.5\end{pmatrix}, \begin{pmatrix}1\\0.4\\-0.1\end{pmatrix}\right\}$ ⇒

	Y_1	Y_2	Y_3
x_1	1	0.6	0.1
x_2	0	-0.4	0.8
x_3	0	0.2	0.5
x_4	1	0.4	-0.1

$\{Y_1\} \perp\!\!\!\perp \{Y_2, Y_3\}$,

- Cluster C_1 ($C=1$) → $\pi_1 = 0.5$ → $p_1 = P(Y_1 = 1) = 0.3$ → $N_1\left(\mu_1 = \begin{pmatrix}1\\1\end{pmatrix}, \Sigma_1 = \begin{pmatrix}2 & 0.5\\0.5 & 2\end{pmatrix}\right)$
- Cluster C_2 ($C=2$) → $\pi_2 = 0.5$ → $p_2 = P(Y_1 = 1) = 0.7$ → $N_2\left(\mu_2 = \begin{pmatrix}0\\0\end{pmatrix}, \Sigma_2 = \begin{pmatrix}1.5 & 1\\1 & 1.5\end{pmatrix}\right)$

1)

→ Atendendo que $\{Y_2, Y_3\}$ tem uma distribuição multivariada gaussiana:

$$P(\underbrace{Y_2 = a, Y_3 = b}_{w_i} \mid C_k) = N_k(\underbrace{Y_2 = a, Y_3 = b}_{w_i} \mid \mu_k, \Sigma_k) = \frac{1}{2\pi |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(w_i - \mu_k)^T \Sigma_k^{-1} (w_i - \mu_k)\right)$$

Temos também que a probabilidade de um posterior é dada por: $P(C_k \mid x_i) \overset{T.Bayes}{=} \frac{P(x_i \mid C_k) P(C_k)}{P(x_i)} =$

→ $P(x_i, C_k) = P(x_i \mid C_k) P(C_k) = P(x_i \mid C_k) \pi_k$

→ $P(x_i) = \sum_k P(x_i, C_k)$

$= \frac{P(x_i, C_k)}{P(x_i)}$

Expectation (E-Step)

(x_1) → $C=1$

→ Prior: $P(C=1) = \pi_1 = 0.5$

→ Likelihood: $P(x_1 \mid C=1) \overset{ind}{=} P(Y_1 = 1 \mid C=1)\, P(Y_2 = 0.6, Y_3 = 0.1 \mid C=1) =$

$= p_1 \times N_1(Y_2 = 0.6, Y_3 = 0.1 \mid \mu_1, \Sigma_1) = 0.3 \times 0.06658 \simeq 0.01997$

→ $P(x_1, C=1) = P(x_1 \mid C=1) P(C=1) = 0.01997 \times 0.5 \simeq 0.00999$

→ $C=2$

→ Prior: $P(C=2) = \pi_2 = 0.5$

→ Likelihood: $P(x_1 \mid C=2) \overset{ind}{=} P(Y_1 = 1 \mid C=2)\, P(Y_2 = 0.6, Y_3 = 0.1 \mid C=2) =$

$= p_2 \times N_2(Y_2 = 0.6, Y_3 = 0.1 \mid \mu_2, \Sigma_2) = 0.7 \times 0.11962 \simeq 0.08373$

→ $P(x_1, C=2) = P(x_1 \mid C=2) P(C=2) = 0.08373 \times 0.5 \simeq 0.041867$

(x_2) → $C=1$

→ Prior: $P(C=1) = \pi_1 = 0.5$

→ Likelihood: $P(x_2 \mid C=1) \overset{ind}{=} P(Y_1 = 0 \mid C=1)\, P(Y_2 = -0.4, Y_3 = 0.8 \mid C=1) =$

$= (1 - p_1) \times N_1(Y_2 = -0.4, Y_3 = 0.8 \mid \mu_1, \Sigma_1) = 0.7 \times 0.05005 \simeq 0.035035$

→ $P(x_2, C=1) = P(x_2 \mid C=1) P(C=1) = 0.035035 \times 0.5 = 0.017517$

→ C=2

→ Prior: $P(C=2) = \pi_2 = 0.5$

→ Likelihood: $P(x_2|C=2) \overset{ind}{=} P(y_1=0|C=2)\, P(y_2=-0.4, y_3=0.8|C=2) =$

$= (1-\rho_2)\, N_2(y_2=-0.4, y_3=0.8|\mu_2,\Sigma_2) = 0.3 \times 0.06819 \simeq 0.020457$

→ $P(x_2, C=2) = P(x_2|C=2)\, P(C=2) = 0.020457 \times 0.5 \simeq 0.01023$

(x_3) → C=1

→ Prior: $P(C=1) = \pi_1 = 0.5$

→ Likelihood: $P(x_3|C=1) \overset{ind}{=} P(y_1=0|C=1)\, P(y_2=0.2, y_3=0.5|C=1) =$

$= (1-\rho_1)\, N_1(y_2=0.2, y_3=0.5|\mu_1,\Sigma_1) = 0.7 \times 0.06837 \simeq 0.047859$

→ $P(x_3, C=1) = P(x_3|C=1)\, P(C=1) = 0.047859 \times 0.5 \simeq 0.02393$

→ C=2

→ Prior: $P(C=2) = \pi_2 = 0.5$

→ Likelihood: $P(x_3|C=2) \overset{ind}{=} P(y_1=0|C=2)\, P(y_2=0.2, y_3=0.5|C=2) =$

$= (1-\rho_2)\, N_2(y_2=0.2, y_3=0.5|\mu_2,\Sigma_2) = 0.3 \times 0.12958 \simeq 0.038874$

→ $P(x_3, C=2) = P(x_3|C=2)\, P(C=2) = 0.038874 \times 0.5 = 0.019437$

(x_4) → C=1

→ Prior: $P(C=1) = \pi_1 = 0.5$

→ Likelihood: $P(x_4|C=1) \overset{ind}{=} P(y_1=1|C=1)\, P(y_2=0.4, y_3=-0.1|C=1) =$

$= \rho_1 \times N_1(y_2=0.4, y_3=-0.1|\mu_1,\Sigma_1) = 0.3 \times 0.05905 \simeq 0.017715$

→ $P(x_4, C=1) = P(x_4|C=1)\, P(C=1) = 0.017715 \times 0.5 = 0.008857$

→ C=2

→ Prior: $P(C=2) = \pi_2 = 0.5$

→ Likelihood: $P(x_4|C=2) \overset{ind}{=} P(y_1=1|C=2)\, P(y_2=0.4, y_3=-0.1|C=2) =$

$= \rho_2 \times N_2(y_2=0.4, y_3=-0.1|\mu_2,\Sigma_2) = 0.7 \times 0.124500 \simeq 0.08715$

→ $P(x_4, C=2) = P(x_4|C=2)\, P(C=2) = 0.08715 \times 0.5 \simeq 0.043575$

Calculando agora $P(x_i)$:

- $P(x_1) = \sum_k P(x_1, C_k) = 0.00999 + 0.0481867 = 0.051857$
- $P(x_2) = \sum_k P(x_2, C_k) = 0.017517 + 0.01023 = 0.027747$
- $P(x_3) = \sum_k P(x_3, C_k) = 0.02393 + 0.019437 = 0.043367$
- $P(x_4) = \sum_k P(x_4, C_k) = 0.008857 + 0.043575 = 0.052432$

Agora já podemos calcular os posteriors para cada observação e cluster usando

$P(C_k|x_i) \overset{T.\,Bayes}{=} \frac{P(x_i|C_k)\,P(C_k)}{P(x_i)} = \frac{P(x_i, C_k)}{P(x_i)}$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV – Grupo 075

(ist1103860, ist1103883)

x_1

$\rightarrow P(C=1|x_1) = \frac{P(x_1, C=1)}{P(x_1)} = \frac{0.00999}{0.051857} \simeq 0.19259$

$\rightarrow P(C=2|x_1) = \frac{P(x_1|C=2)}{P(x_1)} = \frac{0.041867}{0.051857} \simeq 0.80741$

x_2

$\rightarrow P(C=1|x_2) = \frac{P(x_2, C=1)}{P(x_2)} = \frac{0.017517}{0.027747} \simeq 0.63135$

$\rightarrow P(C=2|x_2) = \frac{P(x_2, C=2)}{P(x_2)} = \frac{0.01023}{0.027747} \simeq 0.36865$

x_3

$\rightarrow P(C=1|x_3) = \frac{P(x_3, C=1)}{P(x_3)} = \frac{0.02393}{0.043367} \simeq 0.55181$

$\rightarrow P(C=2|x_3) = \frac{P(x_3, C=2)}{P(x_3)} = \frac{0.019437}{0.043367} \simeq 0.44819$

x_4

$\rightarrow P(C=1|x_4) = \frac{P(x_4, C=1)}{P(x_4)} = \frac{0.008857}{0.052432} \simeq 0.16892$

$\rightarrow P(C=2|x_4) = \frac{P(x_4, C=2)}{P(x_4)} = \frac{0.043575}{0.052432} \simeq 0.83108$

Maximization (M-Step)

	x_1	x_2	x_3	x_4
C=1	0.19259	0.63135	0.55181	0.16892
C=2	0.80741	0.36865	0.44819	0.83108

Para cada cluster C_k, vamos calcular o seguinte:

- $N_k = \sum_{i}^{N} P(C_k|x_i) = \sum_{i=1}^{N} \gamma_{ki}$
- $\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} P(C_k|x_i)\, w_i = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ki}\, w_i$
- $\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ki} \cdot (w_i - \mu_k)(w_i - \mu_k)^T$
- $\pi_k = P(C_k) = \frac{N_k}{N} = \frac{N_k}{\sum_k N}$

- $\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} P(C_k|x_i)\, z_i$

Considera-se z_i uma submatriz de x_i com a feature y_1

Considera-se w_i uma submatriz de x_i com as features y_2 e y_3

- $N_1 = \sum_{i=1}^{4} \gamma_{1,i} = 0.19259 + 0.63135 + 0.55181 + 0.16892 \simeq 1.54467$
- $N_2 = \sum_{i=1}^{4} \gamma_{2,i} = \gamma_{2,1} + \gamma_{2,2} + \gamma_{2,3} + \gamma_{2,4} = 0.80741 + 0.36865 + 0.44819 + 0.83108 = 2.45533$
- $N = N_1 + N_2 = 4.0$

Atualizando agora os valores de μ_k, Σ_k e π_k para cada um dos clusters:

$$\bullet\ \mu_1' = \frac{1}{N_1}\sum_i \gamma_{1,i}\, w_i = \frac{\gamma_{1,1} w_1 + \gamma_{1,2} w_2 + \gamma_{1,3} w_3 + \gamma_{1,4} w_4}{N_1} =$$

$$= \frac{0.19259 \begin{pmatrix}0.6\\0.1\end{pmatrix} + 0.63135 \begin{pmatrix}-0.4\\0.8\end{pmatrix} + 0.55181 \begin{pmatrix}0.2\\0.5\end{pmatrix} + 0.16892 \begin{pmatrix}0.4\\-0.1\end{pmatrix}}{1.54467} = \begin{pmatrix}0.026509\\0.507130\end{pmatrix}$$

$$\bullet\ \Sigma_1' = \frac{1}{N_1}\sum_i \gamma_{1,i}\,(w_i - \mu_1')(w_i - \mu_1')^T = \frac{1}{N_1}\Big(\gamma_{1,1}(w_1-\mu_1')(w_1-\mu_1')^T + \gamma_{1,2}(w_2-\mu_1')(w_2-\mu_1')^T +$$

$$+ \gamma_{1,3}(w_3-\mu_1')(w_3-\mu_1')^T + \gamma_{1,4}(w_4-\mu_1')(w_4-\mu_1')^T\Big) = \frac{1}{1.54467}\Bigg(0.19259 \left(\begin{pmatrix}0.6\\0.1\end{pmatrix} - \begin{pmatrix}0.026509\\0.507130\end{pmatrix}\right) \times \left(\begin{pmatrix}0.6\\0.1\end{pmatrix} - \begin{pmatrix}0.026509\\0.507130\end{pmatrix}\right)^T + 0.63135 \left(\begin{pmatrix}-0.4\\0.8\end{pmatrix} - \begin{pmatrix}0.026509\\0.507130\end{pmatrix}\right)\left(\begin{pmatrix}-0.4\\0.8\end{pmatrix} - \begin{pmatrix}0.026509\\0.507130\end{pmatrix}\right)^T + 0.55181 \left(\begin{pmatrix}0.2\\0.5\end{pmatrix} - \begin{pmatrix}0.026509\\0.507130\end{pmatrix}\right) \times \left(\begin{pmatrix}0.2\\0.5\end{pmatrix} - \begin{pmatrix}0.026509\\0.507130\end{pmatrix}\right)^T + 0.16892 \left(\begin{pmatrix}0.4\\-0.1\end{pmatrix} - \begin{pmatrix}0.026509\\0.507130\end{pmatrix}\right)\left(\begin{pmatrix}0.4\\-0.1\end{pmatrix} - \begin{pmatrix}0.026509\\0.507130\end{pmatrix}\right)^T\Bigg) = \begin{pmatrix}0.148137 & -0.10541\\ -0.10541 & 0.09605\end{pmatrix}$$

$$\bullet\ \pi_1' = \frac{N_1}{N} = \frac{1.54467}{4.0} = 0.38617$$

$$\bullet\ h_1' = \frac{1}{N_1}\sum_i \gamma_{1,i}\, z_i = \frac{1}{N_1}\left(\gamma_{1,1} z_1 + \gamma_{1,2} z_2 + \gamma_{1,3} z_3 + \gamma_{1,4} z_4\right) =$$

$$= \frac{0.19259 \times 1 + 0.63135 \times 0 + 0.55181 \times 0 + 0.16892 \times 1}{1.54467} = 0.2340$$

Repetindo agora para o cluster C_2:

$$\bullet\ \mu_2' = \frac{1}{N_2}\sum_i \gamma_{2,i}\, w_i = \frac{\gamma_{2,1} w_1 + \gamma_{2,2} w_2 + \gamma_{2,3} w_3 + \gamma_{2,4} w_4}{N_2} =$$

$$= \frac{0.80741 \times \begin{pmatrix}0.6\\0.1\end{pmatrix} + 0.36865 \times \begin{pmatrix}-0.4\\0.8\end{pmatrix} + 0.44819 \times \begin{pmatrix}0.2\\0.5\end{pmatrix} + 0.83108 \times \begin{pmatrix}0.4\\-0.1\end{pmatrix}}{2.45533} = \begin{pmatrix}0.30915\\0.21042\end{pmatrix}$$

$$\bullet\ \Sigma_2' = \frac{1}{N_2}\sum_i \gamma_{2,i}\,(w_i - \mu_2')(w_i - \mu_2')^T = \frac{1}{N_2}\Big(\gamma_{2,1}(w_1-\mu_2')(w_1-\mu_2')^T + \gamma_{2,2}(w_2-\mu_2')(w_2-\mu_2')^T +$$

$$\gamma_{2,3}(w_3-\mu_2')(w_3-\mu_2')^T + \gamma_{2,4}(w_4-\mu_2')(w_4-\mu_2')^T\Big) = \frac{1}{2.45533}\Bigg(0.80741 \times \left(\begin{pmatrix}0.6\\0.1\end{pmatrix} - \begin{pmatrix}0.30915\\0.21042\end{pmatrix}\right) \times \left(\begin{pmatrix}0.6\\0.1\end{pmatrix} - \begin{pmatrix}0.30915\\0.21042\end{pmatrix}\right)^T + 0.36865 \left(\begin{pmatrix}-0.4\\0.8\end{pmatrix} - \begin{pmatrix}0.30915\\0.21042\end{pmatrix}\right)\left(\begin{pmatrix}-0.4\\0.8\end{pmatrix} - \begin{pmatrix}0.30915\\0.21042\end{pmatrix}\right)^T + 0.44819 \left(\begin{pmatrix}0.2\\0.5\end{pmatrix} - \begin{pmatrix}0.30915\\0.21042\end{pmatrix}\right) \times \left(\begin{pmatrix}0.2\\0.5\end{pmatrix} - \begin{pmatrix}0.30915\\0.21042\end{pmatrix}\right)^T + 0.83108 \left(\begin{pmatrix}0.4\\-0.1\end{pmatrix} - \begin{pmatrix}0.30915\\0.21042\end{pmatrix}\right)\left(\begin{pmatrix}0.4\\-0.1\end{pmatrix} - \begin{pmatrix}0.30915\\0.21042\end{pmatrix}\right)^T\Bigg) = \begin{pmatrix}0.10829 & -0.08865\\ -0.08865 & 0.10412\end{pmatrix}$$

$$\bullet\ \pi_2' = \frac{N_2}{N} = \frac{2.45533}{4.0} = 0.61383$$

$$\bullet\ h_2' = \frac{1}{N_2}\sum_{i=1} \gamma_{2,i}\, z_i = \frac{1}{N_2}\left(\gamma_{2,1} z_1 + \gamma_{2,2} z_2 + \gamma_{2,3} z_3 + \gamma_{2,4} z_4\right) =$$

$$= \frac{0.80741 \times 1 + 0.36865 \times 0 + 0.44819 \times 0 + 0.83108 \times 1}{2.45533} = 0.6673$$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV – Grupo 075

(ist1103860, ist1103883)

Após aplicar 1 iteração do algoritmo EM, acabamos com os seguintes parâmetros para cada cluster:

Cluster	μ	Σ	π	π
Cluster 1	$\begin{pmatrix} 0.026509 \\ 0.507130 \end{pmatrix}$	$\begin{pmatrix} 0.148137 & -0.10541 \\ -0.10541 & 0.09695 \end{pmatrix}$	0.2340	0.38617
Cluster 2	$\begin{pmatrix} 0.30915 \\ 0.21042 \end{pmatrix}$	$\begin{pmatrix} 0.10829 & -0.08865 \\ -0.08865 & 0.10412 \end{pmatrix}$	0.6673	0.61383

2) $x_{new} = \begin{pmatrix} 1 \\ 0.3 \\ 0.7 \end{pmatrix}$

- $C=1$
  - Prior: $P(C=1) = \pi_1 = 0.38617$
  - Likelihood: $P(x_{new} \mid C=1) \overset{ind}{=} P(y_1=1 \mid C=1)\, P(y_2=0.3, y_3=0.7 \mid C=1) =$
  $= p_1 \times N_1(y_2=0.3, y_3=0.7 \mid \mu_1, \Sigma_1) = 0.2340 \times 0.02708 = 0.00634$
  - $P(x_{new}, C=1) = P(x_{new} \mid C=1)\, P(C=1) = 0.00634 \times 0.38617 = 0.002447$

- $C=2$
  - Prior: $P(C=2) = \pi_2 = 0.61383$
  - Likelihood: $P(x_{new} \mid C=2) \overset{ind}{=} P(y_1=1 \mid C=2)\, P(y_2=0.3, y_3=0.7 \mid C=2) =$
  $= p_2 \times N_2(y_2=0.3, y_3=0.7 \mid \mu_2, \Sigma_2) = 0.6673 \times 0.06843 = 0.04566$
  - $P(x_{new}, C=2) = P(x_{new} \mid C=2)\, P(C=2) = 0.04566 \times 0.61383 =$
  $= 0.02803$

- $P(x_{new}) = \sum_{k}^{2} P(x_{new}, C_k) = 0.002447 + 0.02803 = 0.030477$

Calculando agora os posteriors do x_{new} para cada um dos clusters, usando

$$P(C_k \mid x_i) \overset{T.\,Bayes}{=} \frac{P(x_i \mid C_k)\, P(C_k)}{P(x_i)} = \frac{P(x_i, C_k)}{P(x_i)}$$

$$\rightarrow P(C=1 \mid x_{new}) = \frac{P(x_{new}, C=1)}{P(x_{new})} = \frac{0.002447}{0.030477} = 0.08029 \Rightarrow P(C=2 \mid x_{new}) = \frac{P(x_{new}, C=2)}{P(x_{new})} = \frac{0.02803}{0.030477} = 0.91971$$

Deste modo, x_{new} é atribuído ao Cluster C_2 ($P(C=1 \mid x_{new}) < P(C=2 \mid x_{new})$)

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV – Grupo 075

(ist1103860, ist1103883)

3) x_1

→ $C=1$

→ $P(x_1|C=1) \overset{ind}{=} P(Y_1=1|C=1)\, P(Y_2=0.6, Y_3=0.1|C=1) = p_1 \times N_1(Y_2=0.6, Y_3=0.1|\mu_1,\Sigma_1) =$

$= 0.2340 \times 0.98904 = 0.23147$

→ $C=2$

→ $P(x_1|C=2) \overset{ind}{=} P(Y_1=1|C=2)\, P(Y_2=0.6, Y_3=0.1|C=2) = p_2 \times N_2(Y_2=0.6, Y_3=0.1|\mu_2,\Sigma_2) =$

$= 0.6673 \times 1.42292 = 0.94954$ //

x_2

→ $C=1$

→ $P(x_2|C=1) \overset{ind}{=} P(Y_1=0|C=1)\, P(Y_2=-0.4, Y_3=0.8|C=1) = (1-p_1)\, N_1(Y_2=-0.4, Y_3=0.8|\mu_1,\Sigma_1) =$

$= (1-0.2340) \times 1.65326 = 1.26633$ //

→ $C=2$

→ $P(x_2|C=2) \overset{ind}{=} P(Y_1=0|C=1)\, P(Y_2=-0.4, Y_3=0.8|\mu_2,\Sigma_2) = (1-p_2)\, N_2(Y_2=-0.4, Y_3=0.8|\mu_2,\Sigma_2) =$

$= (1-0.6673) \times 0.26673 = 0.08874$

x_3

→ $C=1$

→ $P(x_3|C=1) \overset{ind}{=} P(Y_1=0|C=1)\, P(Y_2=0.2, Y_3=0.5|C=1) = (1-p_1)\, N_1(Y_2=0.2, Y_3=0.5|\mu_1,\Sigma_1) =$

$= (1-0.2340) \times 1.87753 = 1.43811$ //

→ $C=2$

→ $P(x_3|C=2) \overset{ind}{=} P(Y_1=0|C=2)\, P(Y_2=0.2, Y_3=0.5|C=2) = (1-p_2)\, N_2(Y_2=0.2, Y_3=0.5|\mu_2,\Sigma_2) =$

$= (1-0.6673) \times 1.36519 = 0.45417$

x_4

→ $C=1$

→ $P(x_4|C=1) \overset{ind}{=} P(Y_1=1|C=1)\, P(Y_2=0.4, Y_3=-0.1|C=1) = p_1\, N_1(Y_2=0.4, Y_3=-0.1|\mu_1,\Sigma_1) =$

$= 0.2340 \times 0.08873 = 0.02077$

→ $C=2$

→ $P(x_4|C=2) \overset{ind}{=} P(Y_1=1|C=2)\, P(Y_2=0.4, Y_3=-0.1|C=2) = p_2\, N_2(Y_2=0.4, Y_3=-0.1|\mu_2,\Sigma_2) =$

$= 0.6673 \times 1.08391 = 0.72331$ //

Fazendo um hard assignment das observações aos clusters com ML assumption temos que $mL = \max_k P(x_i|C_k)$, ou seja, a observação x_i será atribuída ao cluster C_k com $P(x_i|C_k)$ máxima. Deste modo temos: Clusters = $\{C_1=\{x_2,x_3\}, C_2=\{x_1,x_4\}\}$

Atendendo à distância de Manhattan: $d(x_i,x_j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \dots + |x_{im}-x_{jm}|$

Para calcular as silhuetas dos clusters, primeiro temos de calcular as silhuetas das nossas observações usando: $s(x_i) = 1 - \frac{a(x_i)}{b(x_i)}$ se $a<b$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV – Grupo 075

(ist1103860, ist1103883)

Temos que $a(x_i) = \frac{1}{m_k - 1} \sum_{j \in C_k \setminus \{i\}} d(x_i, x_j)$ e $b(x_i) = \min_{k' \neq k} \frac{1}{m'_k} \sum_{j \in C_{k'}} d(x_i, x_j)$

x_1

$a(x_1) = \frac{\|x_1 - x_4\|_1}{2-1} = \|x_1 - x_4\|_1 =$

$= |1-1| + |0.6-0.4| + |0.1+0.1| = 0.2+0.2 = 0.4$

$b(x_1) = \frac{\|x_1 - x_2\|_1 + \|x_1 - x_3\|_1}{2} = \frac{(|1-0| + |0.6+0.4| + |0.1-0.8|) + (|1-0| + |0.6-0.2| + |0.1-0.5|)}{2} = \frac{(1+1+0.7) + (1+0.4+0.4)}{2} = \frac{4.5}{2} = 2.25$

$S(x_1) = 1 - \frac{0.4}{2.25} = 0.82222$

x_2 $a(x_2) = \frac{\|x_2 - x_3\|_1}{2-1} = \|x_2 - x_3\|_1 = |0-0| + |-0.4-0.2| + |0.8-0.5| = 0.6+0.3 = 0.9$

$b(x_2) = \frac{\|x_2 - x_1\|_1 + \|x_2 - x_4\|_1}{2} = \frac{(|0-1| + |-0.4-0.6| + |0.8-0.1|) + (|0-1| + |-0.4-0.4| + |0.8+0.1|)}{2} = \frac{(1+1+0.7) + (1+0.8+0.9)}{2} = 2.7$

$S(x_2) = 1 - \frac{a(x_2)}{b(x_2)} = 1 - \frac{0.9}{2.7} = 0.66667$

x_3 $a(x_3) = \frac{\|x_3 - x_2\|_1}{2-1} = \|x_2 - x_3\|_1 = |0-0| + |-0.4-0.2| + |0.8-0.5| = 0.6+0.3 = 0.9$

$b(x_3) = \frac{\|x_3 - x_1\|_1 + \|x_3 - x_4\|_1}{2} = \frac{(|0-1| + |0.2-0.6| + |0.5-0.1|) + (|0-1| + |0.2-0.4| + |0.5+0.1|)}{2} = \frac{(1+0.4+0.4) + (1+0.2+0.6)}{2} = 1.8$

$S(x_3) = 1 - \frac{a(x_3)}{b(x_3)} = 1 - \frac{0.9}{1.8} = 0.500$

x_4 $a(x_4) = \frac{\|x_4 - x_1\|_1}{2-1} = \|x_1 - x_4\|_1 = |1-1| + |0.6-0.4| + |0.1+0.1| = 0.2+0.2 = 0.4$

$b(x_4) = \frac{\|x_4 - x_2\|_1 + \|x_4 - x_3\|_1}{2} = \frac{(|1-0| + |0.4+0.4| + |-0.1-0.8|) + (|1-0| + |0.4-0.2| + |-0.1-0.5|)}{2} = \frac{(1+0.8+0.9) + (1+0.2+0.6)}{2} = 2.25$

$S(x_4) = 1 - \frac{a(x_4)}{b(x_4)} = 1 - \frac{0.4}{2.25} = 0.82222$

Sendo a silhueta de um cluster a média das silhuetas das suas observações, temos:

$S(C_1) = \frac{S(x_2) + S(x_3)}{2} = \frac{0.66667 + 0.5}{2} = 0.58(3)$ $S(C_2) = \frac{S(x_1) + S(x_4)}{2} = \frac{0.82222 + 0.82222}{2} = 0.82222$

4) $Purity = \frac{1}{n}\sum_{k=1}^{k} \max_{j} (|C_k \cap L_j|)$ • $j = ?$

Temos que: • $Purity = 0.75 \iff \frac{1}{4}\sum_{k=1}^{2} \underbrace{\max_{j} (|C_k \cap L_j|)}_{\phi_k} = \frac{3}{4} \iff \phi_1 + \phi_2 = 3$

Sendo que os clusters contêm as seguintes observações: $Clusters = \{C_1 = \{x_3, x_3\}, C_2 = \{x_1, x_4\}\}$

Temos que $\phi_k = \max (|C_k \cap L_j|)$ é no máximo 2, pois cada cluster tem 2 observações, ou tem o valor 1. Ou seja, $\phi_1 + \phi_2 = 3$ sse $\phi_1 = 1$ e $\phi_2 = 2$ ou $\phi_1 = 2$ e $\phi_2 = 1$

• 1 classe (+): C_1 (++) C_2 (++) • $\phi_1 + \phi_2 = 2 + 2 = 4 \neq 3$ ✗

O número de classes não pode ser 1.

• 2 classes (+,-): C_1 (+ -) C_2 (+ -) • $\phi_1 + \phi_2 = \max(1,1) + \max(1,1) = 1 + 1 = 2 \neq 3$ ✗

C_1 (++) C_2 (+ -) • $\phi_1 + \phi_2 = \max(2,0) + \max(1,1) = 2 + 1 = 3 = 3$ ✓

Pelo que qualquer variação deste caso também funciona, desde que 3 observações pertençam à mesma classe.

C_1 (++) C_2 (- -) • $\phi_1 + \phi_2 = \max(2,0) + \max(0,2) = 2 + 2 = 4 \neq 3$ ✗

• 3 classes (+, -, X): C_1 (+ -) C_2 (X +) • $\phi_1 + \phi_2 = \max(1,1,0) + \max(1,0,1) = 1 + 1 = 2 \neq 3$ ✗

C_1 (+ -) C_2 (X X) • $\phi_1 + \phi_2 = \max(1,1,0) + \max(0,0,2) = 1 + 2 = 3 = 3$ ✓

Pelo que qualquer variação deste caso também funciona, desde que 2 observações de um mesmo cluster sejam da mesma classe e as restantes classes tenham uma observação cada.

Visto que $\phi_1 + \phi_2 = 3$ sse $(\phi_1 = 1 \wedge \phi_2 = 2) \vee (\phi_1 = 2 \wedge \phi_2 = 1)$, não faz sentido o nº de classes ser mais que 3 pois iríamos ficar com classes sem observações atribuídas e não conseguiríamos garantir que $\phi_1 + \phi_2 = 3$, ou seja, que a Purity fosse 0.75.

R: Concluímos, que sendo a Purity ≥ 0.75, o número de possíveis classes é 2 ou 3. Caso sejam 2 classes, garantimos Purity = 0.75 desde que 3 observações pertençam à mesma classe, e caso sejam 3 classes, garantimos Purity = 0.75 desde que as 2 observações de um mesmo cluster sejam da mesma classe e as restantes classes tenham uma observação cada.

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV – Grupo 075

(ist1103860, ist1103883)

II. Programming and critical analysis

1)

```
import pandas as pd
from sklearn import datasets, metrics, cluster, mixture, preprocessing
import numpy as np
from scipy.io.arff import loadarff

def purity_score(y_true, y_pred):
    # compute contingency/confusion matrix
    confusion_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    return np.sum(np.amax(confusion_matrix, axis=0)) / np.sum(confusion_matrix)

data = loadarff('column_diagnosis.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')

X = df.drop('class', axis=1)
y = df['class']

# MinMaxScaler
scaler = preprocessing.MinMaxScaler()
X_normalized = scaler.fit_transform(X)

# Values of k
k_values = [2, 3, 4, 5]

# Silhouette and purity scores
silhouette_scores = []
purity_scores = []

# Clustering
for k in k_values:
    kmeans = cluster.KMeans(n_clusters=k, random_state=0)

    # Isolate the k=3 values for question 3
    if k==3:
        labels3 = kmeans.fit_predict(X_normalized)
        cluster_labels = labels3
    else:
        cluster_labels = kmeans.fit_predict(X_normalized)

    silhouette_avg = metrics.silhouette_score(X_normalized, cluster_labels)
    purity = purity_score(y, cluster_labels)
```

```
    # Silhouette score
    silhouette_scores.append(silhouette_avg)
    purity_scores.append(purity)

# Print silhouette and purity scores for each k
for i in range(len(k_values)):
    k = k_values[i]
    silhouette = silhouette_scores[i]
    purity = purity_scores[i]
    print(f'k={k}: \nSilhouette Score = {silhouette} \nPurity = {purity}\n')
```

k=2:
Silhouette Score = 0.36044124340441114
Purity = 0.632258064516129

k=3:
Silhouette Score = 0.29579055730002257
Purity = 0.667741935483871

k=4:
Silhouette Score = 0.27442402122340176
Purity = 0.6612903225806451

k=5:
Silhouette Score = 0.23823928397844843
Purity = 0.6774193548387096

2) (i)

```
from sklearn.decomposition import PCA

# Create a PCA instance
pca = PCA(n_components=2)

# Fit the PCA model to the normalized data
X_pca = pca.fit_transform(X_normalized)

# Variability explained by the top two principal components
variability = pca.explained_variance_ratio_[0] + pca.explained_variance_ratio_[1]

# Print the explained variance ratio
print(f"Variability by the Top 2 Principal Components: {variability*100:.4f} %")
```

Variability by the Top 2 Principal Components: 77.1374 %

(ii)

```
# Get the absolute values
absolute_loadings = np.abs(pca.components_)

# DataFrame to associate the loadings with the input variables
loadings_df = pd.DataFrame(absolute_loadings, columns=X.columns, index=['PC1', 'PC2'])

# Sort the input variables by relevance for each principal component
sorted_loadings_pc1 = loadings_df.loc['PC1'].sort_values(ascending=False)
sorted_loadings_pc2 = loadings_df.loc['PC2'].sort_values(ascending=False)

# Prints
# Starting positions for each column, to organize the prints
variable_position = 0
loading1_position = 30
loading2_position = 60

# Prints
print("More relevant variables for Component 1:")
print("{0:41} {1:30} {2:30}".format("Variable", "Component 1 Loading", "Component 2 Loading\n"))

for attr in sorted_loadings_pc1.index:
    print("{0:30} {1:30} {2:30}".format(attr, sorted_loadings_pc1[attr], sorted_loadings_pc2[attr]))

print("\n\nMore relevant variables for Component 2:")
print("{0:41} {1:30} {2:30}".format("Variable", "Component 1 Loading", "Component 2 Loading\n"))

for attr in sorted_loadings_pc2.index:
    print("{0:30} {1:30} {2:30}".format(attr, sorted_loadings_pc1[attr], sorted_loadings_pc2[attr]))
```

More relevant variables for Component 1:

Variable	Component 1 Loading	Component 2 Loading
pelvic_incidence	0.5916206177372231	0.10003707489152218
lumbar_lordosis_angle	0.5150847620730923	0.08004745059088292
pelvic_tilt	0.4670394389672713	0.6703727595553627
sacral_slope	0.3256888625569193	0.4433029949470748
degree_spondylolisthesis	0.21692963450485395	0.004582909709400215
pelvic_radius	0.11582397626328882	0.581073837095359

More relevant variables for Component 2:

Variable	Component 1 Loading	Component 2 Loading
pelvic_tilt	0.4670394389672713	0.6703727595553627
pelvic_radius	0.11582397626328882	0.581073837095359
sacral_slope	0.3256888625569193	0.4433029949470748
pelvic_incidence	0.5916206177372231	0.10003707489152218
lumbar_lordosis_angle	0.5150847620730923	0.08004745059088292
degree_spondylolisthesis	0.21692963450485395	0.004582909709400215

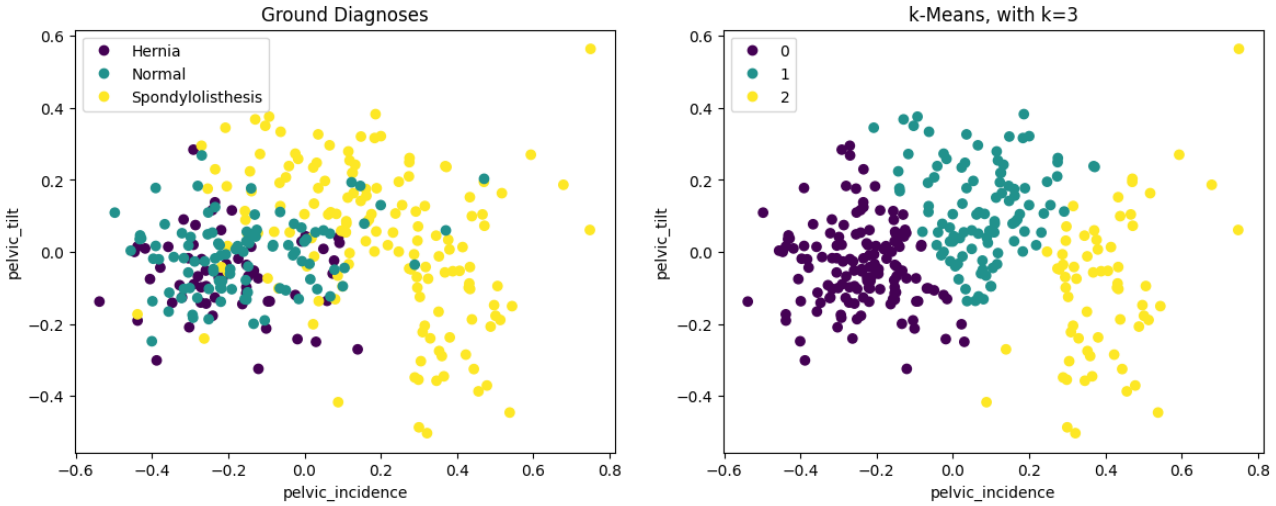
3)

```
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder

# Create a label encoder
label_encoder = LabelEncoder()

# Encode the labels to numeric values
c_original = label_encoder.fit_transform(y)
c_kmeans = label_encoder.fit_transform(labels3)

# Change the labels from the default to desired labels
labels = np.unique(y)
label_mapping = {str(i): label for i, label in enumerate(labels)}


# Plot of the ground diagnosis
plt.figure(figsize=(14, 5))
plt.subplot(121)
scatter = plt.scatter(X_pca[:, 0], X_pca[:, 1], c=c_original)
plt.title("Ground Diagnoses")
plt.xlabel(X.columns[0])
plt.ylabel(X.columns[1])

# Change the labels to desired ones
handles, _ = scatter.legend_elements()
custom_labels = [label_mapping[str(i)] for i in range(len(handles))]
plt.legend(handles, custom_labels)

# Plot for k=3
plt.subplot(122)
```

Homework IV – Grupo 075

(ist1103860, ist1103883)

```
scatter = plt.scatter(X_pca[:, 0], X_pca[:, 1], c=c_kmeans)
plt.title("k-Means, with k=3")
plt.xlabel(X.columns[0])
plt.ylabel(X.columns[1])
handles, labels = scatter.legend_elements()
plt.legend(handles, labels)

plt.show()
```



4)

1 - Atendendo aos resultados do exercício 1, e sabendo que uma silhueta mais elevada indica que os clusters estão bem separados e que a pureza mede o grau em que os elementos de um agrupamento pertencem à mesma classe, com k = 3 os resultados da silhueta e da pureza são mais equilibrados, podendo neste caso dividir o grupo de doentes em Hérnia e Spondylolisthesis com mais precisão. Deste modo, com o número adequado de clusters, podemos dividir grupos em subgrupos, ou seja, identificar diferentes tipos de doença dentro de um dataset.

2 - Os clusters podem associar certas categorias de risco a problemas específicos, permitindo estabelecer relações preditivas entre determinados atributos do paciente e a probabilidade de sofrer de certas doenças. Por exemplo, ao identificar que um valor de pelvic_tilt superior a 20 está consistentemente presente em um cluster onde a incidência de hérnias é mais prevalente, é possível estabelecer uma associação preditiva que sugere um maior risco de hérnia para pacientes com essa característica específica no conjunto de atributos. Assim, o clustering permite-nos identificar as features com maior impacto na separação da população em indivíduos saudáveis e doentes.

END