Departamento de Matemática

Multivariate Analysis

Mater in Eng. and Data Science & Master in Mathematics and Applications

2^{nd} Test | 1^{st} Semester – 2019/2020
Duration: 1.5 hours | 18/12/2019 – 18:00

Please justify conveniently your answers

Group I **9.0 points**

1. The pairwise distances (dissimilarities) between six objects are as follows:

$$\boldsymbol{D} = \begin{bmatrix} 0 & 3.2 & 2 & 0 & 2.8 & 2 \\ & 0 & 2.4 & 3.2 & 1.4 & 2.4 \\ & & 0 & 2 & 2.8 & 2 \\ & & & 0 & 2.8 & 2 \\ & & & & 0 & 2 \\ & & & & & 0 \end{bmatrix}.$$

   (a) Use complete-linkage (also known as furthest neighbour) cluster analysis on the dissimilarity matrix above, and draw the associated dendrogram for your analysis. (4.0)

   (b) How many clusters do you recommend to consider? Justify your choice. Indicate the chosen partition. (2.0)

2. Let $\boldsymbol{C} = [c_{ij}]$ be a similarity matrix, and $\boldsymbol{D} = [d_{ij}]$ such that $d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}$. Prove that $\boldsymbol{D}$ is a dissimilarity matrix. (3.0)

Group II **11.0 points**

1. An observation x comes from one of the two populations with prior probabilities $P(Y=0) = 0.4$ and $P(Y=1) = 0.6$ and probability density functions:

$$f_{X|Y=j}(x) = \begin{cases} \lambda_j^2 x \exp(-\lambda_j x), & x \geq 0, \\ 0, & x < 0, \end{cases}$$

   with $\lambda_1 > \lambda_0 > 0$, $j = 0, 1$.

   (a) Obtain the classification rule that minimizes the total probability of misclassification. (4.0)

   (b) Assuming $\lambda_1 = 2$, and $\lambda_0 = 1$, obtain:

      i. The recall of each class. (3.0)

      ii. The total probability of misclassification. (1.0)

      iii. The precision of each class. (2.0)

   Remark: The cumulative distribution function of $X|Y=j$ is

$$F_j(x) = 1 - e^{-\lambda_j x}(1 + \lambda_j x),$$

   for $x > 0$ and takes the value zero otherwise.

   (c) Comment your results. (1.0)

Multivariate Analysis - 2nd Test -
18 Dec 2019 - MECD/MMA

1. The pairwise distances (dissimilarities) between six objects are as follows:

$$
\boldsymbol{D} = \begin{bmatrix} 0 & 3.2 & 2 & 0 & 2.8 & 2 \\ & 0 & 2.4 & 3.2 & 1.4 & 2.4 \\ & & 0 & 2 & 2.8 & 2 \\ & & & 0 & 2.8 & 2 \\ & & & & 0 & 2 \\ & & & & & 0 \end{bmatrix}.
$$

(a) Use complete-linkage (also known as furthest neighbour) cluster analysis on the dissimilarity matrix above, and draw the associated dendrogram for your analysis. (4.0)

(b) How many clusters do you recommend to consider? Justify your choice. Indicate the chosen partition. (2.0)

1. (a) $\min(d_{ij}) = 0 = d_{14}$

Step 1: $d_{2(14)} = \max(d_{21}, d_{24}) = \max(3.2, 3.2) = 3.2$

$d_{3(14)} = \max(d_{31}, d_{34}) = \max(2, 2) = 2$

$d_{5(14)} = \max(d_{51}, d_{54}) = \max(2.8, 2.8) = 2.8$

$d_{6(14)} = \max(d_{61}, d_{64}) = \max(2, 2) = 2$

$$
\begin{array}{c|c}
 & \begin{matrix} 14 & 2 & 3 & 5 & 6 \end{matrix} \\
\begin{matrix} 14 \\ 2 \\ 3 \\ 5 \\ 6 \end{matrix} & \begin{bmatrix} 0 & 3.2 & 2 & 2.8 & 2 \\ & 0 & 2.4 & 1.4 & 2.4 \\ & & 0 & 2.8 & 2 \\ & & & 0 & 2 \\ & & & & 0 \end{bmatrix}
\end{array}
$$

Step2 : $\min d_{ij} = 1.4 = d(2,5)$

$d_{14(25)} = \max\left(d_{14(2)}, d_{14(5)}\right) = \max(3.2, 2.8) = 3.2$

$d_{3(25)} = \max\left(d_{32}, d_{35}\right) = \max(2.4, 2.8) = 2.8$

$d_{6(25)} = \max\left(d_{62}, d_{65}\right) = \max(2.4, 2) = 2$

$$\begin{array}{c|cccc} & 14 & 25 & 3 & 6 \\ \hline 14 & 0 & 3.2 & 2 & 2 \end{array}$$

$$\begin{array}{c} 143 \\ 25 \\ 6 \end{array} \begin{bmatrix} 0 & 3.2 & 2 \\ & 0 & 2.4 \\ & & 0 \end{bmatrix}$$

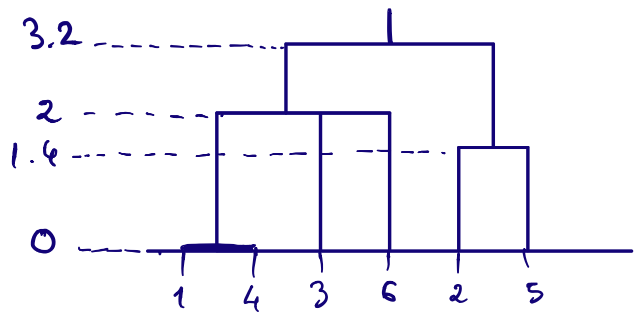
Step4: $\min d_{ij} = 2 = d_{143,6} = d_{25,6}$

$d_{1436,25} = \max\left(d_{143,25}, d_{6,25}\right) = \max(2.8, 2) = 2.8$

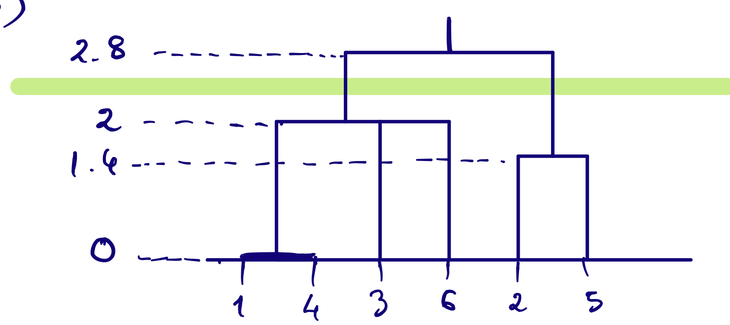
$$\begin{array}{c} 1436 \\ 25 \end{array} \begin{bmatrix} 0 & 3.2 \\ & 0 \end{bmatrix}$$

Step 5: when $d_{ij} = d_{1436,25} = 3.2$

Dendrogram:



(2.b)



The dendrogram does not suggest a very clear partition in clusters, but eventually two clusters:

$$c_1 = \{1, 3, 4, 6\} \text{ and } c_2 = \{2, 5\}$$

may be a good partition.

Indexes, like average silhouette may give a better indication of the quality of the partition.

Group I **9.0 points**

2. Let $\boldsymbol{C} = [c_{ij}]$ be a similarity matrix, and $\boldsymbol{D} = [d_{ij}]$ such that $d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}$. Prove that $\boldsymbol{D}$ is a dissimilarity matrix. (3.0)

We have to prove:

(i). $d_{ij} \geq 0 \quad \forall i, j$

(ii). $d_{ii} = 0$

(iii). $d_{ij} = d_{ji}$

(i). $d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}$, and we know that

$c_{ii} = \max_j c_{ij}$, thus $c_{ii} \geq c_{ij}, \forall j$ and $c_{jj} \geq c_{ij}, \forall i$

Being so, $c_{ii} - 2c_{ij} + c_{jj} \geq 0$ and $d_{ij} \geq 0$

(ii) $d_{ii} = \sqrt{c_{ii} - 2c_{ii} + c_{ii}} = 0$ ✓

(iii) $d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}} = \sqrt{c_{ii} - 2c_{ji} + c_{jj}}$

symmetry of c_{ij} given that it is a similarity

$= d_{ji}$, ✓

So we can conclude that d_{ij} is a dissimilarity

□

Group II **11.0 points**

1. An observation x comes from one of the two populations with prior probabilities $P(Y=0)=0.4$ and $P(Y=1)=0.6$ and probability density functions:

$$f_{X|Y=j}(x) = \begin{cases} \lambda_j^2 x \exp(-\lambda_j x), & x \geq 0, \\ 0, & x < 0, \end{cases}$$

with $\lambda_1 > \lambda_0 > 0$, $j = 0, 1$.

(a) Obtain the classification rule that minimizes the total probability of misclassification. (4.0)

$p = P(Y=1) = 0.6 \quad , \quad 1-p = 0.4$

$$TPM = P(\text{assign } X \text{ to } \{Y=1\} \mid Y=0)(1-p) + P(\text{assign } X \text{ to } \{Y=0\} \mid Y=1)\, p$$

$$= P(X \in R_1 \mid Y=0)(1-p) + p\, P(X \in R_0 \mid Y=1)$$

$$= P(X \in R_1 \mid Y=0)(1-p) + p - p\, P(X \in R_1 \mid Y=1)$$

$$= \int_{R_1} \left[(1-p) f_{X|Y=0}(x) - p\, f_{X|Y=1}(x)\right] dx$$

R_1 should be chosen as the values of $x \geq 0$ such that

$$(1-p) f_{X|Y=0}(x) - p\, f_{X|Y=1}(x) \leq 0$$

$$\iff \frac{f_{X|Y=1}(x)}{f_{X|Y=0}(x)} \geq \frac{1-p}{p}$$

$$\iff \frac{\lambda_1^2 \cancel{x} \exp(-\lambda_1 x)}{\lambda_0^2 \cancel{x} \exp(-\lambda_0 x)} \geq \frac{1-p}{p}$$

$$\Leftrightarrow \quad \exp\left(-(\lambda_1-\lambda_0)x\right) \geq \frac{1-p}{p}\left(\frac{\lambda_0}{\lambda_1}\right)^2$$

$$\Leftrightarrow \quad -(\lambda_1-\lambda_0)\,x \geq \log\left(\frac{(1-p)}{p}\left(\frac{\lambda_0}{\lambda_1}\right)^2\right)$$

$$\Leftrightarrow \quad x \leq \frac{1}{(\lambda_1-\lambda_0)} \log\left(\frac{p}{1-p}\left(\frac{\lambda_1}{\lambda_0}\right)^2\right)$$

Optimal classification Rule:

$$\begin{cases} \text{Assign } x \text{ to } \{Y=1\} \text{ iff } x \leq \dfrac{1}{(\lambda_1-\lambda_0)} \log\left(\dfrac{p}{1-p}\left(\dfrac{\lambda_1}{\lambda_0}\right)^2\right) \\ \text{Assign } x \text{ to } \{Y=0\} \quad \text{otherwise} \end{cases}$$

(b) Assuming $\lambda_1 = 2$, and $\lambda_0 = 1$, obtain:

i. The recall of each class. (3.0)

ii. The total probability of misclassification. (1.0)

iii. The precision of each class. (2.0)

Remark: The cumulative distribution function of $X|Y = j$ is

$$F_j(x) = 1 - e^{-\lambda_j x}(1 + \lambda_j x),$$

for $x > 0$ and takes the value zero otherwise.

b(i) $\quad \xi = \dfrac{1}{2-1} \log\left(\dfrac{6}{4}\, 2^2\right) = \log(6)$

$$Re(1) = P(X \leq \log(6) \mid Y=1) = 1 - e^{-2\log(6)}(1+2\log 6)$$

$$= 1 - \frac{1}{36}(1+\log 36) = 0.8727$$

$$Re(0) = P(X > \log 6 \mid Y=0) = 1 - P(X \leq \log 6 \mid Y=0)$$

$$= 1 - \left[1 - e^{-\log 6}(1+\log 6)\right]$$

$$= \frac{1}{6}\left(1+\log 6\right) = 0.4653$$

b(ii) $\quad TPR = P(X \in R_1 | Y=0)(1-p) + p\,P(X \in R_0 | Y=1)$

$$= 0.4\,(1-0.4653) + 0.6\,(1-0.8727)$$

$$= 0.2903$$

b(iii)

$$Pr(i) = P(Y=i \mid X \in R_i) =$$

$$= \frac{P(X \in R_i | Y=i)\,P(Y=i)}{P(X \in R_i | Y=i)\,P(Y=i) + P(X \in R_i | Y \neq i)\,P(Y \neq i)}$$

$$= \frac{Re(i)\,P(Y=i)}{Re(i)\,P(Y=i) + (1-Re(j))\,P(Y=j)}, \quad i \neq j$$

$$Pr(1) = \frac{Re(1)\,p}{Re(1)\,p + (1-Re(0))\,(1-p)}$$

$$= \frac{0.8727 \times 0.6}{0.8727 \times 0.6 + 0.4\,(1-0.4653)}$$

$$= 0.7100$$

$$Pr(0) = \frac{Re(0)\,(1-p)}{Re(0)(1-p) + p\,(1-Re(1))}$$

$$= \frac{0.4653 * 0.4}{0.4653 * 0.4 + 0.6 * (1 - 0.8727)}$$

$$= \boxed{0.7090}$$

(c) Comment your results.

Around 70% of the observations are well classify. the recall of class 1 is high but it is the obs. of class 3,4 & 6 that are not so well correctly assign to the right class by the classifier.