Departamento de Matemática

Multivariate Analysis

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

2^{nd} Test — 1^{st} Semester – 2020/2021

Duration: 1.5 hours — 17/12/2020 – 18:30

Please justify conveniently your answers

Group I **9.0 points**

1. Let $\boldsymbol{x} = (x_1, \ldots, x_p)^t$ and $\boldsymbol{y} = (y_1, \ldots, y_p)^t$ represents the p continuous measurements characterizing two different objects, where for $\boldsymbol{x}$, $\boldsymbol{y} \in (\mathbb{R}^+)^p$, (3.0)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(\frac{x_i - y_i}{x_i + y_i}\right)^2} \quad \text{and} \quad d(\mathbf{0}, \mathbf{0}) = 0.$$

   Prove that the previous coefficient is a dissimilarity measure.

2. Show that the distance $d_{k(ij)}$ (between cluster k and the cluster formed by merging cluster i and j) used by average-linkage verifies: (2.0)

$$d_{k(ij)} = \frac{n_i}{n_i + n_j} d_{ki} + \frac{n_j}{n_i + n_j} d_{kj},$$

   where d_{ij} is the dissimilarity between i-th and j-th cluster and n_i is the number of objects belonging to the i-th cluster.

3. The pairwise dissimilarities between four objects are as follows:

$$\boldsymbol{D} = \begin{bmatrix} 0 & 1 & 11 & 5 \\ & 0 & 2 & 3 \\ & & 0 & 4 \\ & & & 0 \end{bmatrix}.$$

   Use average-linkage cluster analysis on the dissimilarity matrix above, and draw the associated dendrogram. (4.0)

   Suggestion: Use the result stated in Question 2.

Group II **11.0 points**

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

$$P(X = x|Y = j) = \begin{cases} \frac{\lambda_j^x e^{-\lambda_j}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise,} \end{cases}$$

   with $\lambda_1 > \lambda_0 > 0$, $j = 0, 1$, and $X|Y = j \sim \text{Poisson}(\lambda_j)$.

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

   (b) Assuming $\lambda_0 = 1$, and $\lambda_1 = 4$, obtain:

      i. The recall of class 1. (2.5)

      ii. The precision of class 1. (2.5)

      iii. The F_1 measure of class 1. (1.0)

   (c) Comment on the accuracy of the classifier. (1.0)

Departamento de Matemática

Multivariate Analysis

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

2^{nd} Test — 1^{st} Semester – 2020/2021

Duration: 1.5 hours — 17/12/2020 – 18:30

Please justify conveniently your answers

Group I — **9.0 points**

1. Let $\boldsymbol{x} = (x_1, \ldots, x_p)^t$ and $\boldsymbol{y} = (y_1, \ldots, y_p)^t$ represents the p continuous measurements characterizing two different objects, where for $\boldsymbol{x}, \boldsymbol{y} \in (I\!R^+)^p$, (3.0)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(\frac{x_i - y_i}{x_i + y_i}\right)^2} \quad \text{and} \quad d(\boldsymbol{0}, \boldsymbol{0}) = 0.$$

Prove that the previous coefficient is a dissimilarity measure.

We have to prove:

1. $d(\underset{\sim}{x}, \underset{\sim}{x}) = 0$
2. $d(\underset{\sim}{x}, \underset{\sim}{y}) \geq 0 \quad \forall \underset{\sim}{x}, \underset{\sim}{y} \in \mathbb{R}^p$
3. $d(\underset{\sim}{x}, \underset{\sim}{y}) = d(\underset{\sim}{y}, \underset{\sim}{x})$

Proof:

1. If $\underset{\sim}{x} \neq \underset{\sim}{0} \Rightarrow d(\underset{\sim}{x}, \underset{\sim}{x}) = \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(\frac{x_i - x_i}{x_i + x_i}\right)^2} = 0$

If $\underset{\sim}{x} = \underset{\sim}{0} \Rightarrow d(\underset{\sim}{0}, \underset{\sim}{0}) = 0$ so (1) is proved.

2. $d^2(x, y) = \frac{1}{p}\sum_{i=1}^{p} (x_i - y_i)^2 / (x_i + y_i)^2 \geq 0$

so $d(x, y) \geq 0$

And finally,

(3) $d(\underset{\sim}{x}, \underset{\sim}{y}) = \frac{1}{p}\sqrt{\sum_{i=1}^{p}\left(\frac{x_i - y_i}{x_i + y_i}\right)^2} =$

$= \frac{1}{p}\sqrt{\sum_{i=1}^{p}\left(\frac{y_i - x_i}{x_i + y_i}\right)^2} = d(\underset{\sim}{y}, \underset{\sim}{x})$ QED

2. Show that the distance $d_{k(ij)}$ (between cluster k and the cluster formed by merging cluster i and j) used by average linkage verifies: (2.0)

$$d_{k(ij)} = \frac{n_i}{n_i + n_j} d_{ki} + \frac{n_j}{n_i + n_j} d_{kj},$$

where d_{ij} is the dissimilarity between i-th and j-th cluster and n_i is the number of objects belonging to the i-th cluster.

$$d_{k(ij)} = \frac{1}{(n_k)(n_i + n_j)} \sum_{u=1}^{n_k} \sum_{v=1}^{n_i + n_j} d_{uv} =$$

$$= \frac{1}{n_k (n_i + n_j)} \sum_{u=1}^{n_k} \left[\sum_{v=1}^{n_i} d_{uv} + \sum_{v=1}^{n_j} d_{uv} \right]$$

$$= \frac{n_i}{n_i + n_j} \sum_{u=1}^{n_k} \sum_{v=1}^{n_i} \frac{d_{uv}}{n_k n_i} + \frac{n_j}{n_i + n_j} \sum_{u=1}^{n_k} \sum_{v=1}^{n_j} d_{uv}$$

$$= \frac{n_i}{n_i + n_j} d_{ki} + \frac{n_j}{n_i + n_j} d_{kj}$$

QED

3. The pairwise dissimilarities between four objects are as follows:

$$\boldsymbol{D} = \begin{bmatrix} 0 & 1 & 11 & 5 \\ & 0 & 2 & 3 \\ & & 0 & 4 \\ & & & 0 \end{bmatrix}.$$

Use average-linkage cluster analysis on the dissimilarity matrix above, and draw the associated dendrogram. (4.0)

Suggestion: Use the result state in Question 2.

1st Step: min $d_{ij} = d_{12} = 1$

$$\boldsymbol{D} = \begin{bmatrix} 0 & 1 & 11 & 5 \\ & 0 & 2 & 3 \\ & & 0 & 4 \\ & & & 0 \end{bmatrix}.$$

$d(3,\{1,2\}) = \frac{1}{1+1} d(1,3) + d(2,3) \frac{1}{1+1}$

$= \frac{11+2}{2} = 6.5$

$d(4,\{1,2\}) = \frac{1}{1+1} d(4,1) + d(4,2) \frac{1}{2} =$

$= \frac{5+3}{2} = 4$

Thus, the new dissimilarity matrix is:

$$\begin{array}{c} \\ \{1,2\} \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} \{1,2\} & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 6.5 & 0 & \\ 4 & 4 & 0 \end{bmatrix} \end{array}$$

2nd Step:

min $d_{ij} = d(\{1,2\},4) = d(3,4) = 4$

$d(\{1,2\},\{3,4\}) = d(\{1,2\},3) \frac{1}{1+1} +$

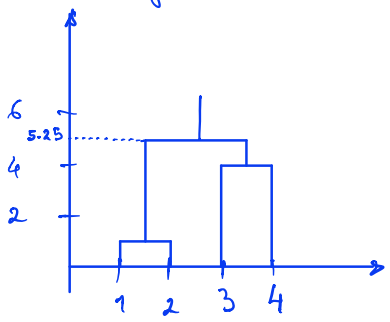
$+ d(\{1,2\},4) = \frac{1}{1+1} = \frac{6.5+4}{2} = \frac{10.5}{2}$

$= 5.25$

$$D = \begin{array}{c} \\ \{1,2\} \\ \{3,4\} \end{array} \begin{array}{c} \begin{array}{cc} & \end{array} \\ \begin{bmatrix} 0 & \\ 5.25 & 0 \end{bmatrix} \end{array}$$

3rd step : $\min d_{ij} = d(\{1,2\}, \{3,4\}) = 5.25$

Dendrogram :



2nd Possible Solution

1st Step: min $d_{ij} = \boxed{d_{12} = 1}$

$$d(3,\{1,2\}) = \frac{1}{1+1} d(1,3) + d(2,3) \frac{1}{1+1}$$

$$= \frac{11+2}{2} = 6.5$$

$$d(4,\{1,2\}) = \frac{1}{1+1} d(4,1) + d(4,2) \frac{1}{2} =$$

$$= \frac{5+3}{2} = 4$$

Thus, the new dissimilarity matrix is:

$$\begin{array}{c} \\ \{1,2\} \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} \{1,2\} & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 6.5 & 0 & \\ 4 & 4 & 0 \end{bmatrix} \end{array}$$

2nd Step:

min $d_{ij} = \boxed{d(\{1,2\},4) = d(3,4) = 4}$

$$d(3,\{1,2,4\}) = d(3,4)\frac{1}{1+2} + d(3,\{1,2\})\frac{2}{1+2}$$

$$= \frac{4 + 6.5*2}{3} = \frac{17}{3} = 5.67$$

$$D = \begin{array}{c} \\ 3 \\ \{1,2,4\} \end{array} \begin{array}{c} \begin{array}{cc} & \end{array} \\ \begin{bmatrix} 0 & \\ 5.67 & 0 \end{bmatrix} \end{array}; \quad \text{min } d_{ij} = 5.67 = d(3,\{1,2,4\})$$

Group II **11.0 points**

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

$$P(X = x|Y = j) = \begin{cases} \frac{\lambda_j^x e^{-\lambda_j}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise,} \end{cases}$$

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

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

$$TPM = P(\text{classify } x \text{ as } Y=1 \mid Y=0) + P(\text{classify } x \text{ as } Y=0 \mid Y=1)$$

$$= \sum_{x \in R_1} P(X=x \mid Y=0) P(Y=0) + \sum_{x \in R_0} P(X=x \mid Y=1) P(Y=1)$$

$$= \left[1 - \sum_{x \in R_0} P(X=x \mid Y=0)\right] P(Y=0) + \sum_{x \in R_0} P(X=x \mid Y=1) P(Y=1)$$

mean TPM we define R_1 as the set of $x \in \mathbb{N}_0$ such that

$$P(X=x \mid Y=1) P(Y=1) - P(X=x \mid Y=0) P(Y=0) \geq 0$$

where in our case $P(Y=1) = P(Y=0) = \frac{1}{2}$, thus

Assign x_0 to $Y=1$ iff (since $p_1=p_0$)

$$P(X=x_0 \mid Y=1) \geq P(X=x_0 \mid Y=0)$$

$$\Leftrightarrow \quad \frac{e^{-d_1} d_1^{x_0}}{x_0!} \geq \frac{e^{-d_0} d_0^{x_0}}{x_0!}$$

$$\Leftrightarrow \quad e^{-(d_1-d_0)} \left(\frac{d_1}{d_0}\right)^{x_0} \geq 1 \qquad \text{where } d_1 > d_0 \Leftrightarrow \frac{d_1}{d_0} > 1$$

$$\Leftrightarrow \quad \left(\frac{d_1}{d_0}\right)^{x_0} \geq e^{(d_1-d_0)}$$

$$\Leftrightarrow \quad x_0 \log\left(\frac{d_1}{d_0}\right) \geq d_1 - d_0$$

$$\Leftrightarrow \quad x_0 \geq \frac{d_1-d_0}{\log(d_1)-\log(d_0)}, \text{ since } d_1 > d_0$$

Thus, the optimal classification rule is:

$$\begin{cases} \text{If } x_0 \geq \dfrac{d_1-d_0}{\log(d_1)-\log(d_0)} & \Rightarrow \text{assign } x_0 \text{ to } Y=1 \\ \text{otherwise} & \Rightarrow \text{assign } x_0 \text{ to } Y=0 \end{cases}$$

Group II **11.0 points**

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

$$P(X = x|Y = j) = \begin{cases} \frac{\lambda_j^x e^{-\lambda_j}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise,} \end{cases}$$

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

(b) Assuming $\lambda_0 = 1$, and $\lambda_1 = 10$, obtain:

i. The recall of class 1. (3.0)

ii. The precision of class 1. (2.0)

iii. The F_1 measure of class 1. (1.0)

If $d_0 = 1$ and $d_1 = 10$ then

$$m = \frac{d_1 - d_0}{\log(d_1) - \log(d_0)} = \frac{4-1}{\log(4) - \log(1)} \simeq 2.164$$

(i) $Rec(1) = P(\text{classify } X \text{ as } Y=1 \mid Y=1) =$

$= P(X \geq 2.164 \mid Y=1) = 1 - P(X \leq 2 \mid Y=1)$

$= 1 - F_{Po(4)}(2) = 1 - 0.2381$

$= 0.7619$

(ii) $Pr(1) = P(Y=1 \mid \text{classify as } Y=1) =$

$$= \frac{P(\text{classify as } Y=1 \mid Y=1)\, p_1}{P(\text{classify as } Y=1 \mid Y=1)\, p_1 + P(\text{classify as } Y=1 \mid Y=0)\, p_0}$$

$$Pr(1) = \frac{0.7619}{0.7619 + (1 - 0.9197)} = 0.90465$$

$P(\text{Classify } x \text{ as } y=1 \mid y=0) =$

$= 1 - P(X < 2.165 \mid y=0) = 1 - P(X \leq 2 \mid y=0)$

$= 1 - F_{Po(1)}(2) = 1 - 0.9197 = 0.0803$

(iii) $F_1(1) = \dfrac{2 \; Re(1) * Pr(1)}{Re(1) + Pr(1)} = 0.8272$

(c) Comment your results. (1.0)

The Precision, and F_1-measure of class $y=1$ are very high but Re(1) is not so high (0.7619)

thus, to be sure about the classifier's performance it is important to consider an overall measure of performance, like:

Overall Accuracy $= Re(1) f_1 + Re(0) f_0$

$= \dfrac{0.7619 + (1 - 0.0803)}{2}$

$= 0.8408$

is fairly high, meaning that there is a good separation between the groups leading to a high overall accuracy. nevertheless the probability of correctly assign an observation of $y=1$ is not as high as we could like.