BINDURA UNIVERSITY OF SCIENCE EDUCATION SFM411

BScSFM

MODERN APPLIED STATISTICS

Time : 3 hours MAR 2024

Candidates should attempt ALL questions in section A and at most TWO questions in section B.

SECTION A (40 marks)

Candidates may attempt ALL questions being careful to number them A1 to A4

A1. Distinguish between the following terms;

(a) data interpretation and data reduction, [3]

(b) single-linkage and complete-linkage, [4]

(c) supervised and unsupervised learning. [4]

A2. (a) Define data mining, [2]

(b) Give any three reasons why we mine data, [3]

(c) List the sequence of steps in the data mining process. [7]

A3. Consider the following design matrix, representing four sample points, $X_i \in \mathbb{R}^2$

$$X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$$

We want to represent the data in only one dimension, so we turn to principal components analysis (PCA).

(a) Compute the unit-length principal component directions of X, and [5]

(b) State which one of the PCA algorithmn would you choose if you request just one principal component. [1]

(c) Calculate the projected principal coordinate values for each sample points from X. [4]

A4. (a) What is the difference between a hierarchical and a non-hierarchical method of clustering? Give an example of a non-hierarchical method. [4]

(b) Define a dendrogram and outline why it is useful? [3]

SECTION B (60 marks)

Candidates may attempt TWO questions being careful to number them B5 to B7.

B5. (a) The pairwise distances (dissimilarities) between four objects are as follows:

Object	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Cluster and draw the dendrograms for the four items using each of the following procedures:

(i) Single linkage agglomerative clustering, [5,3]

(ii) Complete linkage agglomerative clustering. [5,3]

(b) Consider the eight data points

$$A1(2;10), A2(2;5), A3(8;4), A4(5;8), A5(7;5), A6(6;4), A7(1;2), A8(4;9).$$

The distance matrix based on the Euclidean distance between the points is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Suppose the initial centres of each cluster are $A1$, $A4$, and $A7$.

(i) Use the K-means clustering algorithmn to divide the data points into K = 3 clusters and write down the new clusters. Perform only one epoch(iteration). [7]

(ii) Find the centers of the new clusters obtained in (i) above. [3]

(iii) Graph the data points in terms of their (x_1, x_2) and on the graph show the clusters after the first iteration. [1,3]

B6. The pairwise distances (dissimilarities) between five objects are as follows:

Object	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Cluster and draw the dendrogram for the five items using each of the following procedures:

(a) Single linkage hierarchical procedure, [10]

(b) Complete linkage hierarchical procedure, [10]

(c) Average linkage hierarchical procedure. [10]

B7. (a) For what purposes would you use a linear discriminant function in a two-group problem? Discuss the conditions required for this method of analysis to be appropriate. [2,2]

(b) In a study of psychology students, aimed at predicting their academic success, two variables were considered: x_1, the GRE-Q score (a measure of quantitative ability), and x_2, the number of hours of psychology courses taken previously. Random samples of 30 successful students and 20 unsuccessful students were taken. The results were summarised as follows. For the successful group, the mean vector, $\bar{\mathbf{x}}_{\mathbf{S}}$, and sample covariance matrix, $\hat{\mathbf{\Sigma}}_{\mathbf{S}}$, were as follows.

$$\bar{\mathbf{x}}_{\mathbf{S}} = \begin{bmatrix} 548.5 \\ 28.7 \end{bmatrix} \quad \hat{\mathbf{\Sigma}}_{\mathbf{S}} = \begin{bmatrix} 8955 & -13 \\ -13 & 127 \end{bmatrix}$$

For the successful group, the mean vector, $\bar{\mathbf{x}}_{\mathbf{U}}$, and sample covariance matrix, $\hat{\mathbf{\Sigma}}_{\mathbf{U}}$, were as follows.

$$\bar{\mathbf{x}}_{\mathbf{U}} = \begin{bmatrix} 498.8 \\ 14.4 \end{bmatrix} \quad \hat{\mathbf{\Sigma}}_{\mathbf{S}} = \begin{bmatrix} 9323 & 292 \\ 292 & 72 \end{bmatrix}$$

(i) Use this information to provide a linear discriminant function to distinguish between these two groups. You may assume that the two types of possible error are considered to be equally important. [13]

(ii) Stating any assumptions you make, estimate the probability of classifying a future psychology student to the wrong group. [4]

(iii) Consider two new psychology students, A and B, where A has $x_1 = 525$ and $x_2 = 22$, while B has $x_1 = 530$ and $x_2 = 23$. Use your results to predict whether A will be successful or unsuccessful. [6]

(iv) If you were now asked to predict the outcome for B, would you be more or less confident about your prediction? Explain your answer. [1,2]

END OF QUESTION PAPER