BINDURA UNIVERSITY OF SCIENCE EDUCATION SFM411

BScSFM

MODERN APPLIED STATISTICS

Time : 3 hours

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. (a) Suppose that objects are to be clustered on the basis of an observation vector that consists of p continuous measurements. Define the following terms:

(i) Euclidean distance, [2]

(ii) Minkowski metric. [2]

(b) Contrast each of the following pairs of terms:

(i) Cluster Analysis and Discriminant Analysis. [2]

(ii) Supervised and unsupervised machine learning? [2]

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. [8]

A3. (a) What is the difference between single linkage and complete linkage methods of hierarchical clustering? [4]

(b) What is a dendrogram? Why is it useful? [4]

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

A4. The examination marks of 100 students in five different courses in BSc Statistics and Financial Mathematics-Level 4.1 have been recorded. Each examination was marked out of 100 marks. An analyst has extracted the principal components and eigenvalues from the correlation matrix for the marks.

The coefficients for the first three principal components and the eigenvalues are given in the table below:

		Component		
	variable	1	2	3
	SFM411	-0.55	0.62	0.73
	SFM412	-0.34	0.48	-0.75
	SFM414	-0.51	-0.14	0.03
	SFM415	-0.47	-0.38	-0.12
	SFM416	-0.40	-0.50	0.35
Eigenvalue		3.24	0.76	0.44

(a) Write down the second principal component. [2]

(b) How much of the total variation in the data is explained by the first principal component? [2]

(c) How much of the total variation in the data is explained by the last two principal components? [3]

SECTION B (60 marks)

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

B5. (a) Suppose that $n_1 = 11$ and $n_2 = 12$ observations are made on two random variables X_1 and X_2, where X_1 and X_2 are assumed to have a bivariate normal distribution with common covariance matrix $\sum$, but possibly different mean vectors μ_1 and μ_2 for the two samples. The sample mean vectors and pooled covariance matrix are:

$$\bar{x}_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \bar{x}_2 = \begin{bmatrix} -4 \\ -2 \end{bmatrix}, \quad S_{pooled} = \begin{bmatrix} 7.4 & -1.2 \\ -1.2 & 4.8 \end{bmatrix}$$

(i) Construct Fisher's (sample) linear discriminant function correct to four decimal places. [9]

(ii) Assign observation $X_0^T = \begin{bmatrix} -1 & 2 \end{bmatrix}$ to either population π_1 or π_2. Assume equal costs and equal prior probability. [6]

(b) Suppose that the random variables X_1, X_2 and X_3 have the variance-covariance matrix

$$\sum = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

(i) Find the eigenvalue-eigenvector pair for the covariance matrix $\sum$, [11]

(ii) Find the principal components of $\sum$, [2]

(iii) Give the variances of the principal components in (ii) above. [2]

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

Object	1	2	3	4
1	0			
2	1	0		
3	11	2	0	
4	5	3	4	0

Cluster and draw the dendrogram for the four 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) Consider the following eight points where (x, y) represent location:

$$A_1(2,10),\ A_2(2,5),\ A_3(8,4),\ A_4(5,8),\ A_5(7,5),\ A_6(6,4),\ A_7(1,2),\ A_8(4,9)$$

(i) Graph the items in terms of their (x, y) and comment. [3,1]

(ii) Use the K-means clustering technique and perform two iterations to divide the points into K = 3 clusters.
Start with the initial cluster centres $A_1(2,10)$, $A_4(5,8)$, and $A_7(1,2)$ and use the distance function between two points $a = (x_1, y_1)$ and $b = (x_2, y_2)$ defined as:

$$P(a,b) = \mid x_2 - x_1 \mid + \mid y_2 - y_1 \mid .$$

[20]

(iii) Hence, find the three cluster centres after the second iteration. [6]

END OF QUESTION PAPER