BINDURA UNIVERSITY OF SCIENCE EDUCATION

SFM 411: MODERN APPLIED STATISTICS

NOV 2024

Time: 3 hours

Candidates may attempt ALL questions in Section A and at most two questions in Section B. Each question should start on a fresh page.

Section A (40 marks)

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

A1. Distinguish between the following terms;

(a) supervised and unsupervised learning, [4]

(b) agglomerative and divisive clustering, and [4]

(c) data reduction and data interpretation. [4]

A2. Suppose we measure two variables X_1 and X_2 for items A, B, C and D. The data are as follows:

	Observations	
Item	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

(a.) Graph the items in terms of their (X_1, X_2) and comment. [4]

(b.) Use the K-means clustering technique to divide the items into K = 2 clusters. Start with the initial groups (AB) and (CD). [7]

A3. (a) Define data mining. [2]

(b) List the sequence of steps in data mining process. [8]

A4. Compute the correlation Matrix from the covariance matrix

$$\Sigma = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}$$

Obtain $\mathbf{V}^{1/2}$ and $\boldsymbol{\rho}$ [7]

Section B (60 marks)

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

B5.(a) Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ be a random sample with covariance matrix $\mathbf{\Sigma}$, with eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_p \geq 0$ and corresponding eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$. For each $i = 1, 2, \cdots, p$, let $Y_i = \mathbf{e}_i'\mathbf{X}$. Prove that,

(i) $Var(Y_i) - \lambda_i$, and [5]

(ii) $Cov(Y_i, Y_j) = 0$ for $i \neq j$. [5]

(b) A random vector $\mathbf{X}' = [X_1, X_2]$ has the following variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$$

Find the:

(i) eigenvalue-eigenvector pairs for the covariance matrix. [10]

(ii) principal components of $\mathbf{\Sigma}$ and their variances. [6]

(iii) proportion of the total variance accounted for by the first principal component. [4]

B6. a) Consider two samples X_1 from population 1 and X_2 from population 2

$$\mathbf{X}_1 = \begin{pmatrix} 56 & 70 & 65 & 54 & 70 \\ 50 & 55 & 62 & 52 & 51 \end{pmatrix}$$

and

$$\mathbf{X}_2 = \begin{pmatrix} 55 & 80 & 73 & 64 & 73 & 81 \\ 53 & 75 & 72 & 61 & 74 & 73 \end{pmatrix}$$

Given that the pooled variance matrix for the data sets is

$$\mathbf{S}_{pooled} = \begin{bmatrix} 90.75 & 58.75 \\ 58.75 & 61.75 \end{bmatrix}$$

(i) Construct Fisher's (sample) linear discriminant function. [10]

(ii) Assign observation $\mathbf{X}_0^T = \begin{bmatrix} 50 & 55 \end{bmatrix}$ to either population $\pi1$ or $\pi2$. Assume equal costs and equal prior probability [5]

(b) (i) Given that the orthogonal factor model is given by $\mathbf{X} - \mu = \mathbf{LF} + \epsilon$, where matrix L is the matrix of loadings, F is an m × 1 vector of common factors and L is p × 1 vector of specific factors. Show that covariance matrix $\Sigma = \mathbf{LL}' + \Psi$ [8] .

(ii) Given that for a 4×1 observation vector X a 2 orthogonal factor model, with the matrix of loadings, L is given by

$$\mathbf{L} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix},$$

and specific variances of 2,4,1 and 3 for X_1, X_2, X_3 and X_4 respectively. Derive the covariance matrix Σ. [7]

B7. a)Explain the problems of data mining? [10]

b) Explain the purpose of cluster analysis and discuss briefly the decisions that need to be made when carrying out a cluster analysis. [20]

THE END