MTH 552: STATISTICAL & AI TECHNIQUES IN DATA MINING
Mid semester Examination: Full Marks 60

[1] (a) From a dataset of track records of 100 athletes in 5 track and field events; 100m, 200m, 400m, 5000m, 10000m; calculations based on sample correlation matrix gives the following:

$$\hat{\lambda}_1 = 3.75;\ \hat{f}_1' = (0.469, 0.532, 0.465, 0.387, 0.361)$$

$$\hat{\lambda}_2 = 1.01;\ \hat{f}_2' = (-0.368, -0.236, -0.315, 0.585, 0.606)$$

$\hat{\lambda}_1$ is the largest eigen value of the sample correlation matrix, $\hat{f}_1$ is the corresponding orthonormalized eigen vector; $\hat{\lambda}_2$ is the second largest eigen value of the sample correlation matrix, $\hat{f}_2$ is the corresponding orthonormalized eigen vector.

The standardized observation vectors of two athletes of interest are

Athlete 1 : $(0.01, 0.06, 0.8, 0.65, 0.75)'$

Athlete 2 : $(0.8, 0.7, 0.3, 0.12, 0.01)'$

(i) How would you interpret the 1st 2 principal components?

(ii) What proportion of total (standardized) sample variation does the 1st prinicipal component explain?

(iii) With the given information, can you suggest a ranking of the 2 athletes?

(b) Let $\underset{\sim}{X} = (X_1, \ldots, X_p)'$ be a random vector with $E(\underset{\sim}{X}) = \underset{\sim}{\mu} = (\mu_1, \ldots, \mu_p)'$ and $Cov(\underset{\sim}{X}) = \Sigma = ((\sigma_{ij})), (\Sigma > 0)$ and $\underset{\sim}{Y} = (Y_1, \ldots, Y_p)'$ denote the vector of principal components derived from standardized variables $\underset{\sim}{Z} = (Z_1, \ldots, Z_p)'$; for $i = 1, \ldots, p$, $Z_i = (X_i - \mu_i)/\sqrt{\sigma_{ii}}$. Find the covariance matrix, $Cov(\underset{\sim}{X}, \underset{\sim}{Y})$.

12 (6+6) Marks

[2] The distance matrix corresponding to 6 multidimensional cases $C_1, C_2, C_3, C_4, C_5, C_6$ is given by

$$D = \begin{pmatrix} 0 & 4 & 13 & 24 & 12 & 8 \\ & 0 & 10 & 22 & 11 & 10 \\ & & 0 & 7 & 3 & 9 \\ & & & 0 & 6 & 18 \\ & & & & 0 & 8.5 \\ & & & & & 0 \end{pmatrix}$$

(a) Construct the dendogram tree corresponding to an agglomerative complete linkage hierarchical clustering algorithm.

(b) Identify the clusters at a merger level 8.

(c) Find the merger level at which we get 4 clusters and list the objects in the clusters.

10 Marks

Scanned by CamScanner

[3] Consider the divergence distance measure between two multidimensional (p-dimensional) populations (π_1 and π_2)

$$J_D = \int \cdots \int \left(f(\underset{\sim}{x} \mid \pi_1) - f(\underset{\sim}{x} \mid \pi_2) \right) \log\left(f(\underset{\sim}{x} \mid \pi_1) / f(\underset{\sim}{x} \mid \pi_2) \right) d\underset{\sim}{x}$$

$f(\underset{\sim}{x} \mid \pi_i)$ denote the joint density under population π_i, $i = 1,2$. Prove or disprove the following statement "If the p-components of the underlying random vector are independent then $J_D = \sum_{i=1}^{p} J_D^i$, J_D^i is the divergence distance for the i^{th} component of the random vector".

6 Marks

[4] Let $(X_1, \ldots, X_n)$ be random sample from a population having a mixture exponential model density

$$p(x) = \sum_{j=1}^{g} \pi_j \, p(x \mid \theta_j),$$

$$\text{where, } p(x \mid \theta_j) = \begin{cases} \theta_j e^{-\theta_j x}, & x > 0 \\ 0, & \text{o/w} \end{cases}, \; j = 1, \ldots, g.$$

(a) Formulate the maximum likelihood estimation of the parameters involved in the E-M algorithm framework.

(b) Derive the E-M algorithm update equations for π_j and θ_j and hence outline the density estimation procedure based on a given set of observations.

12 Marks

[5] Let $\underset{\sim}{x}_1 = (1,2)'$, $\underset{\sim}{x}_2 = (3,2)'$, $\underset{\sim}{x}_3 = (1,6)'$ & $\underset{\sim}{x}_4 = (5,4)'$ be observed feature vectors of 4 cases. 3 different clustering algorithms gave the following 3 partitions:

Algorithm I Partition: (Case 1, Case 3), (Case 2, Case 4)

Algorithm II Partition: (Case 1, Case 2), (Case 3, Case 4)

Algorithm III Partition: (Case 1, Case 4), (Case 2, Case 3)

Let $S_W = \frac{1}{n} \sum_{j=1}^{g} \sum_{i=1}^{n} Z_{ji} \left(\underset{\sim}{x}_i - \underset{\sim}{m}_j \right) \left(\underset{\sim}{x}_i - \underset{\sim}{m}_j \right)'$ be the pooled within cluster scatter matrix for a fixed number, g, of clusters obtained from n cases. $Z_{ji} = 1$, if $\underset{\sim}{x}_i \in$ cluster j; 0, otherwise. $\underset{\sim}{m}_j$ is the mean of cluster j. Which of the above partition(s) would you prefer if clustering criterion based on $trace(S_W)$ is to be used?

10 Marks

[6]. Let $(20,10,16,2,3,4,4,8,1,12,11,19,18,21,5,11,11,12,19,2)$ be a sample from an univariate population with unknown probability density function $f(x)$.

(a) Find non-parametric, rectangular kernel based, density estimate at $x = 7, 11, 24, 30$ with kernel bandwidth, h, equal to 4 for the rectangular kernel.

(b) Find non-parametric k-nearest neighbor (with a symmetric region having center at the point x) density estimates at the same points as in (a) using $k = 4$.

10 Marks

Scanned by CamScanner