Nanyang Technological University, SPMS, MAS

MH4501 Multivariate Analysis, Semester 2 AY 2022-23

Assignment 4 (Due date: 4:30pm, April 13, 2023)

Remark: NTU places very high importance on honesty in academic work submitted by students, and adopts a policy of **ZERO** tolerance on cheating and plagiarism.

1. Let $f_1(x) = \frac{1}{2}(1-|x|)$ for $|x| \le 1$ and $f_2(x) = \frac{1}{4}(2-|x-0.5|)$ for $-1.5 \le x \le 2.5$.
   (a) Sketch the two densities on the same plot.
   (b) Determine the classification regions when $p_1 = p_2$ and $c(1|2) = c(2|1)$.

2. Consider the classification problem of two normal populations with same covariance matrix Σ. A sample of 17 observations is got from the first population and we have

$$\bar{x}_1 = \begin{pmatrix} 12.5 \\ 7.3 \\ 10.1 \end{pmatrix}, \quad S_1 = \begin{pmatrix} 5 & 0.1 & 2 \\ 0.1 & 4 & -0.3 \\ 2 & -0.3 & 5 \end{pmatrix}$$

   A sample of 21 observations is got from the second population and we have

$$\bar{x}_2 = \begin{pmatrix} 10 \\ 8.1 \\ 10 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 6 & -0.08 & 2 \\ -0.08 & 5 & 0.24 \\ 2 & 0.24 & 4 \end{pmatrix}$$

   (a) Suppose $p_2 = 2 \times p_1$ and $c(1|2) = 2 \times c(2|1)$. Give the classification rule to allocate a new observation $x_0 = (x_1, x_2, x_3)^T$ to either of these two populations.
   (b) Suppose that we have a new observation $x_0 = (11, 8, 10)^T$, which population will you allocate it to?

3. The table below provides a data set containing 7 observations with 2 features:

Obs.	1	2	3	4	5	6	7
X_1	0.9	1.5	3.0	5.0	3.5	4.5	3.5
X_2	0.9	2.0	4.0	7.0	5.0	5.0	4.5

   We wish to identify two clusters of this data set using K-means clustering with $K = 2$. We use the Euclidean distance measure. Suppose that we initially assign the observations #1, #2, #3 as cluster 1 and the observations #4, #5, #6, #7 as cluster 2.

(a) What are the cluster centroids and cluster assignments after the first iteration of K-means clustering?

(b) What are the cluster centroids and cluster assignments after the second iteration of K-means clustering?

(c) Do we need to proceed to the third iteration? If yes, continue the K-means clustering algorithm until it converges. If no, explain why.

4. Suppose that we have five observations, for which we compute a dissimilarity (distance) matrix as follows:

$$\begin{pmatrix} 0 & 9 & 3 & 6 & 11 \\ 9 & 0 & 7 & 5 & 10 \\ 3 & 7 & 0 & 9 & 2 \\ 6 & 5 & 9 & 0 & 8 \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix}$$

(a) On the basis of the dissimilarity matrix, sketch the dendogram that results from hierarchically clustering these 5 observations using **complete linkage**. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

(b) Repeat (a), this time using **single linkage** clustering.

5. The (2×1) random vectors $X^{(1)}$ and $X^{(2)}$ have the joint mean vector and joint covariance matrix

$$\mu = \left[\begin{array}{c} \mu^{(1)} \\ \hline \mu^{(2)} \end{array}\right] = \left[\begin{array}{c} -3 \\ 2 \\ \hline 0 \\ 1 \end{array}\right]; \quad \Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 8 & 2 & 3 & 1 \\ 2 & 5 & -1 & 3 \\ \hline 3 & -1 & 6 & -2 \\ 1 & 3 & -2 & 7 \end{array}\right]$$

(a) Calculate the canonical correlations ρ_1, ρ_2.

(b) Determine the canonical variate pairs (U_1, V_1) and (U_2, V_2).

(c) Suppose that we have 100 samples and obtain the same canonical correlations as in part (a) based on sample covariance matrix. Construct a hypothesis testing for whether $X^{(1)}$ and $X^{(2)}$ are uncorrelated at 5% significance level (considering large sample size).

Thank you for your feedback that greatly improved the quality of this course.