MAHATMA GANDHI UNIVERSITY, KOTTAYAM

MGU-UGP (HONOURS) FIFTH SEMESTER EXAMINATION

(2024 ADMISSION ONWARDS)

MG5DSESTA300– INTRODUCTION TO MULTIVARIATE ANALYSIS

Duration: 1Hour Maximum: 30 Marks

PRACTICAL EXAMINATION

Answer any five questions. Each question carries 6 marks.

Solve the questions using R in the computer lab. The R codes along with output must be written in the answer sheets.

1. An automotive researcher is evaluating whether vehicle efficiency and physical build characteristics differ systematically based on transmission type. Specifically, the researcher wants to test if the combined mean vector of Miles per Gallon (mpg) and Weight (wt) is significantly different between cars with automatic transmissions and those with manual transmissions. Using the built-in mtcars dataset in R, perform a Two-Sample Hotelling's T-squared Test to determine if there is a statistically significant multivariate difference between the two transmission types. [A] [1]

2. In the iris dataset, consider the two sepal measurements (Sepal.Length, Sepal.Width) as the first set of variables X, and the two petal measurements (Petal.Length, Petal.Width) as the second set of variables Y. Write the R code to compute the canonical correlations and extract the canonical coefficients for the both variable sets. [A] [1]

3. The built-in object eurodist contains road distances (km) between 21 European cities.
   (a) Obtain a two-dimensional configuration using classical MDS (cmdscale).
   (b) Plot the cities.
   (c) Compute the goodness-of-fit measures and the proportion of variation explained by each dimension. [A] [2]

4. Use the variables mpg, cyl, disp, hp, drat, wt and qsec of the mtcars data set.
   (a) Examine the correlation matrix and test its suitability for factor analysis.
   (b) Fit a two – factor maximum- likelihood model with varimax rotation. Report the loadings, uniqueness's and communalities.
   (c) Compute regression factor scores and plot the cars in the factor space. Which cars have the highest and lowest scores on Factor 1? [A] [3]

5. Two samples of sizes 70 and 100 are given from two normal populations $N_p(\mu_1,\Sigma)$ and $N_p(\mu_2,\Sigma)$.
   Let $\bar{X}^{(1)}=\begin{bmatrix}3846\\5040\\2940\end{bmatrix}$ $\bar{X}^{(2)}=\begin{bmatrix}6160\\8080\\5200\end{bmatrix}$ and S = $\begin{bmatrix}50910 & 8825 & 3925\\8825 & 11085 & 5070\\3925 & 5070 & 2710\end{bmatrix}$.
   Obtain the best discriminant function to discriminate between two populations and classify the observations X=$\begin{bmatrix}56\\78\\52\end{bmatrix}$. [A] [4]

6. For the distance matrix, apply single linkage and complete linkage methods and draw corresponding dendrograms. [A] [5]

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

7. Use the knn () function of the package class (supplied with R) on the iris data.
   (a) Standardise the four measurements. Use 100 randomly selected flowers (set. seed (123)) as the training set and the remaining 50 as the test set. Classify the test flowers with k = 5 and give the confusion matrix and accuracy.
   (b) Choose k using leave-one-out cross-validation (knn.cv) for k = 1, …, 25 and plot the error against k. [A] [4]

8. Cross-classify the 32 cars of mtcars by number of cylinders (cyl) and number of forward gears (gear).
   (a) Construct the contingency table and test for association. Why does R give a warning, and how can it be avoided?
   (b) Perform correspondence analysis. How much of the inertia is explained by each dimension?
   (c) Draw and interpret the correspondence map. [A] [5]

MAHATMA GANDHI UNIVERSITY, KOTTAYAM

MGU-UGP (HONOURS) FIFTH SEMESTER EXAMINATION

(2024 ADMISSION ONWARDS)

MG5DSESTA302– STATISTICAL COMPUTING USING PYTHON

Duration: 1Hour Maximum: 30 Marks

PRACTICAL EXAMINATION

Answer any five questions. Each question carries 6 marks.

Solve the questions using Jupyter notebook in **the computer lab. The python codes along with output must be written in the answer sheets.**

1. Load the iris data set from sklearn.datasets. Standardize the four features using StandardScaler. Apply PCA with 2 components. Print the explained variance ratio of each component and the cumulative explained variance. Plot the first two principal component scores as a scatter plot coloured by the species. Write the codes used and output obtained. [A][4]

2. The table below gives a distance matrix of perceived dissimilarities between 5 air freshener brands measured during a market study. Conduct Classical Metric Multidimensional Scaling on this precomputed distance matrix with n_components=2, and random_state=42. [A][4]

Brand	**Godrej**	**Airwick**	**Odonil**	**Ambi Pure**	**Lia**
Godrej	0	206	963	2976	1949
Airwick	206	0	802	2815	1771
Odonil	963	802	0	2013	996
Ambi Pure	2976	2815	2013	0	1307
Lia	1949	1771	996	1307	0

3. Load the diabetes dataset from sklearn. Standardize the 10 predictor variables and fit factor analysis with n_components=3. Print the first five factor scores. [A][4]

4. Using the fetch_california_housing dataset available in sklearn.datasets, fit a multiple linear regression model using all predictor variables. Split the data using test_size=0.20 and random_state=501, and evaluate the model using MSE, RMSE, MAE and R^2. [A][5]

5. Train a logistic regression model on the load_breast_cancer dataset. Use 80% of the sample as training data and split with random_state = 42. Use max_iter = 100. Measure the accuracy using accuracy_score print the confusion matrix and display the confusion matrix. Write the codes used and the output obtained. [A][5]

6. Load the load_iris dataset from sklearn.datasets. Split the data into training and testing sets in the ratio 80:20 using random_state=101. Standardize the features using StandardScaler and train an SVM classifier using an RBF kernel with C=1 and gamma= 'scale'; Report the classification accuracy and predict the class of a new flower having the following measurements: sepal length = 5.8 cm, sepal width = 2.7 cm, petal length = 4.1cm, and petal width = 1.0 cm. [A][5]

7. Train a decision tree classifier using data set load_iris from sklearn.datasets. Using 75% of the sample as training data. Split the data with random_state = 88. Use gini as criterion, max_depth = 3 and random_state =11. Print the test accuracy and confusion matrix. Also predict the species of a flower with measurements [5.9, 3.0, 5.1, 1.8]. Write the codes used and output obtained. [A][5]

8. Using the Iris dataset, apply K-Means to the four measurements. Use the Elbow Method to choose k, then form k clusters with random_state=404. Display the cluster labels and centroids, and plot petal length against petal width. [A][5]