Attempt All Questions

Question 1 : Cluster Analysis

i. (2 Marks) What is the purpose of a cluster analysis?

ii. (2 Marks) A discriminant analysis is similar to a cluster analysis; however, there is one fundamental difference. Explain this difference.

iii. (2 Marks) Compare and contrast any two linkage methods.

iv. (2 Marks) What is the difference between a linkage method and a distance measure?

iv. (2 Marks) Compare and contrast any two linkage methods.

v. (2 Marks) How do we determine the appropriate number of clusters? Give two different visualization methods that are used to display the outcome of a cluster analysis.

vi. (2 Marks) Compute the Euclidean distance between the following points.

$$A = (4, 7, 8, 2)$$

$$B = (5, 6, 2, 6)$$

vii. (2 Marks) Give one reason why the squared Euclidean Distance may be used in preference to the Euclidean distance.

viii. (2 Marks) In the context of hierarchical cluster analysis, distinguish between agglomerative clustering and divisive clustering.

ix. (2 Marks) What is a vertical icicle plot used for? Give a brief description, supporting your answer with sketches.

x. (2 Marks) Explain the difference between Ward's method and k means clustering.

Question 2a : Dimensionality Reduction

The following questions relate to Principal Component Analysis and Factor Analysis

i. (2 Marks) What is the purpose of a principal component analysis?

ii. (3 Mark) Varimax, quartimax and equamax are the commonly used methods in a certain procedure. What is this procedure? What is the purpose of the procedure. Which method is the most commonly used?

iii. (5 Marks) What is meant by the "true" dimension of the data? How does an analyst determine the appropriate number of principal components to retain, making reference to three different approaches.

iv. (2 Marks) What problems occur if a principal component analysis is carried out on a data matrix where the columns contain measurements on very difference scales? What can be done to overcome this problem?

v. (2 Marks) The Kaiser-Meyer-Olkin (KMO) statistic is used to measure a certain characteristic of the data. What is this characteristic? Explain how the KMO statistic should be interpreted.

vi. (1 Marks) Briefly describe the Bartlett Test for Sphericity, with reference to the null and alternative hypotheses, and how those statements relate to the purpose of the test.

Question 2b : Multicollinearity

The following questions relate to multicollinearity in the context of multiple regression analysis.

i. (1 Mark) Define multicollinearity.

ii. (2 Marks) State two ways in which a multiple regression analysis could be affected by severe multicollinearity.

iii. (2 Marks) State two ways of formally diagnosing the severity of multicollinearity, making reference to how both should be used to make decisions about the data.

Question 3a : Discriminant Analysis

i. (2 Marks) What is the purpose of a discriminant analysis?

ii. (3 Marks) How does discriminant analysis differ from MANOVA?

Question 3b : Linear Models

The following questions relate to model selection and validation in the context of multiple regression analysis.

i. (1 Marks) What is meant by over

tting?

ii. (3 Marks) Compare and contrast the following variable selection procedures (*1 Mark for each*).

   a. Forward Selection

   b. Backward Elimination

   c. Stepwise Regression

iii. (1 Mark) Briefly describe how the Akaike Information Criterion would be used in the context of model selection.

Question 3c : Appraisal of Analytical Systems

	Predicted Negative	Predicted Positive
Observed Negative	True Negative	False Positive
Observed Positive	False Negative	True Positive

i. (3 Marks) With reference to the table above, define each of the following appraisal metrics (*1 Mark for each*).

   a. Accuracy

   b. Precision

   c. Recall

ii. (2 Marks) What is the F-score? Explain its function and how it is computed.

iii. (2 Marks) Define Specificity and Sensitivity. You make reference to previous answers.

iv. (3 Marks) What is a ROC curve? Explain its function, how it is determined, and the means of interpreting the curve. Support your answer with a sketch.

Question 4a : Missing Data

i. (2 Marks) What is Missing Data? Discuss the implications of Missing Data in the context of a statistical analysis.

ii. (3 Marks) Compare and contrast the following types of missing data: Missing At Random, Missing Not At Random, Missing Completely at Random.

iii. (3 Marks) Discuss some of the traditional techniques for dealing with Missing Data, making reference to the limitations of each.

Question 4b : Logistic Regression

i. (3 Marks) Under what circumstances would you use Logistic Regression?

ii. (2 Marks) Suppose that, out of a sample of 100 men and 100 women, 75 men drank alcohol in the last week, while 30 women drank alcohol in the past week. Compute the odds ratio for women to men.

iii. (3 Marks) What is a logit? How can you transform a logit into a probability?

iv. (2 Marks) What is a dummy variable? Explain how it is used in Logistic Regression. Support your answer with an example.

v. (2 Marks) Describe how the Likelihood Ratio Test is used for variable selection in Logistic Regression.