Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN STATISTICAL SCIENCES
END OF SEMESTER EXAMINATION
STA 8316: MACHINE LEARNING AND PATTERN RECOGNITION

DATE: August 25, 2021 Time: **2 Hours**

Instructions

1. This examination consists of **FOUR** questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

Question 1 (20 Marks)

a) Explain what EDA is and in addition,
   i) Distinguish between EDA and summary, classical and Bayesian analysis
   ii) Enumerate the EDA assumptions
   iii) Identify and implement relevant EDA Techniques for given problems

(**6** Marks)

b) The Box-cox transform can be used to remove skewness in data. Describe this approach and also explain how maximum likelihood estimation can be used to estimate the transformation parameter λ.

(**7** Marks)

c) Bagging, boosting, random forests and stochastic gradient boosting algorithms are ensemble methods that are often used to enhance the quality of a decision tree. Briefly describe each approach, clearly highlighting the enhancement that they give.

(**7** Marks)

Question 2 (20 Marks)

a) Explain what you understand by the fixed distribution assumption in exploratory data analysis (EDA). In addition, answer the following question as they relate to EDA assumptions:
   i) What are the consequences of violating this assumption in statistical analysis?
   ii) Enumerate 3 approaches that can be used to test this assumption.
   iii) Explain how the modified power transformation can be used to remove kurtosis in a symmetric distribution.

(**10** Marks)

b) Outliers and anomalies occur quite often in data. In relations to these two issues, answer the following two questions:
   i) What are outliers and how might they arise in a practical setting?

ii) The Grubbs, Tietjen-Moore, Dixon and the generalized (extreme Studentized deviate) ESD (Rosner) tests are approaches used in exploratory data analysis. Distinguish between them and explain in (mathematical) detail how each approach works.

(**10** Marks)

Question 3 (20 Marks)

a) Briefly explain your understanding of each of the following techniques:
   i) Principal component analysis (PCA)
   ii) correspondence analysis (CA) and
   iii) multiple factor analysis (MFA)

(**12** Marks)

b) Eigenfaces and text mining are two common applications of principal components analysis and correspondence analysis. Explain how each approach works.

(8 Mark**s)**

Question 4 (20 Marks**)**

a) Decision trees are procedures employed extensively in machine learning and statistical literature. Briefly explain how these procedures work and also mention aspects of their efficiency and reliability.

(**7** Marks)

b) A major problem associated with regression trees is instability. Explain what you understand by this problem.

(**3** Marks)

c) Regression tree can be seen as a kind of additive model or a piecewise constant regression model. Mathematically or using an appropriate example, clearly explain this assertion.

(**5** Marks)

d) Describe the recursive partitioning algorithm and its utility in decision trees

(**5** Marks)