JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATION- 2023

B.Tech-VII Semester (CSE/IT/ECE/CE/BT/BI)

COURSE CODE (CREDITS): 19B1WCI731 MAX. MARKS: 35

COURSE NAME: Computational Data Analysis

COURSE INSTRUCTORS: Dr. Nishant Sharma MAX. TIME: 2 Hours

***Note:** (a)All questions are compulsory.*

(b)Marks are indicated against each question in square brackets.

(c) The candidate is allowed to make Suitable numeric assumptions wherever required for solving problems

JUIT TEST-3 EXAMINATION- DEC 2023

Q1. (a) Compare and contrast the Normal Equation method and Gradient Descent approach for solving linear regression problems. Discuss the advantages and disadvantages of each method, considering factors such as computational efficiency and applicability to large datasets.

[CO-1] [5 marks]

Q2. Consider a dataset of emails categorized as spam (S) or non-spam (NS), with the following information:

Training Data:

Total emails: 100

Spam emails (S): 30

Non-spam emails (NS): 70

Word Occurrences:

For the word "discount":

In spam emails: 25 occurrences

In non-spam emails: 5 occurrences

For the word "urgent":

In spam emails: 15 occurrences

In non-spam emails: 2 occurrences

Given a new email with the words "discount" and "urgent," calculate the probabilities of it being spam (P(S)) and non-spam (P(NS)) using the Naive Bayes classification. **[CO-2] [5 marks]**

Q3. Discuss the fundamental differences between Naive Bayes and K-NN algorithms in the context of generative algorithms. Explain the underlying assumptions, strengths, and weaknesses of each algorithm. Provide a hypothetical scenario where one algorithm might outperform the other and justify your choice. **[CO-2] [4 marks]**

Q4. (a) What is Lasso regularization, and how does it contribute to variable selection in regression models? Compare Lasso with Ridge regularization in terms of the penalty term.

(b) Explain the terms "Forward Regression" and "Backward Regression" in the context of feature selection. Provide a brief comparison of these methods, highlighting their strengths and limitations.

(c) Define Elastic-net regularization and discuss its motivation in the context of linear regression. How does Elastic-net combine L1 and L2 regularization, and what is its impact on the regression coefficients? **[CO-4] [3 + 3 + 3 marks]**

Q5. (a) Consider a dataset with 100 features (F1, F2, ..., F100) and a target variable Y. You are tasked with performing feature selection using Information Gain. The Information Gain for each feature is calculated, and the top 20 features with the highest Information Gain scores are selected. If the initial dataset had an entropy of 0.8 and the entropy of the selected subset is 0.2, calculate the Information Gain achieved through feature selection.

(b) You are working with a classification problem, and you decide to build an ensemble model using Random Forest. The original dataset has 1000 samples, and you use a Random Forest with 100 decision trees. Each tree is trained on a bootstrap sample of the data. If, on average, each decision tree in the Random Forest correctly classifies 80% of the samples it is exposed to, what is the overall accuracy of the Random Forest on the entire dataset? **[CO-5] [4 + 3 marks]**

Q6. Compare and contrast Bagging, Boosting, and Stacking ensemble learning techniques. Explain the underlying principles of each method and provide scenarios where one technique might outperform the others. Support your answer with examples. **[CO-4] [5 marks]**