CH5019 – Mathematical Foundations of Data Science

Project Details

General Instructions

- Use either MATLAB or Python for coding
- Report should be submitted in pdf format.
- Report should not exceed 10 A4 sized pages including figures and references. Use font size of 11 pts.
- Submit all your codes along with the report separately in a zipped folder with your Roll number as the file name.
- Upload the zipped file only in Moodle (.zip or .rar extensions are allowed).

Problem 01

Write your own code to fill the missing data in the given data set and answer the given questions:

Description of data set '***IPL_Twitter_missingData***':

This data set is collected from twitter posts across various regions of India for 3 different IPL seasons. Each data sample contains two binary variables and four integer variables. Each of these variables are described below

- Q_1 (first column): It is a binary variable indicating whether there exists a match for CSK on the particular day or not.
- Q_2 (second column): It is a binary variable indicating whether there exists a match for MI on the particular day or not.
- X_1 (third column): It is an integer value corresponds to the number of tweets specified David Warner on the particular day.
- X_2 (fourth column): It is an integer value corresponds to the number of tweets specified Mahendra Singh Dhoni on the particular day.
- X_3 (fifth column): It is an integer value corresponds to the number of tweets specified Rohit Sharma on the particular day.
- X_4 (sixth column): It is an integer value corresponds to the number of tweets specified Virat Kohli on the particular day.

i. How many data samples have missing values?

ii. How many categories can be identified in the data given by omitting the samples with missing values?

iii. Obtain the linear relationships between the variables (X_1, X_2, X_3 and X_4) for all the given scenarios.

   a. For the whole data
   b. If there is no match for both CSK and MI
   c. If there is no match for CSK but there is match for MI
   d. If there is no match for MI but there is match for CSK
   e. If there is match for both MI and CSK.

iv. Impute the missing values using your own ideas and also briefly explain those ideas in the report (scores will be proportional to the number of meaningful ways the data is filled). One has to use one unique idea to fill the complete data so submit multiple sets of filled data, if you have multiple ideas for filling the data.

Problem 02

Given csv files contain data for credit card fraud detection. Use SVM to classify the data

- q2_data_matrix.csv: This file contains a 100 × 5 data matrix. The 5 features and their corresponding ranges are described below:

1. Age: 18-100 years
2. Transaction Amount: $ 0-5000
3. Total Monthly Transactions: $ 0-50000
4. Annual Income: $ 30000-1000000
5. Gender: 0/1 (0 - Male, 1 - Female)

- q2_labels.csv: This file contains a 1000×1 vector of 0/1 labels for whether the transaction is fraudulent or not.
  – 0: The transaction is legitimate
  – 1: The transaction is fraudulent

i. Report the confusion matrix and the F1 Score for this data set.

ii. Use kernels in SVM and compare the accuracy against linear SVM. Which of the following kernels is better? Gaussian, Polynomial or linear kernel