Assignment 3

CS 484

Submission Deadline: 03/24 11:59 pm

Problem 1: Linear Regression (13 marks)

The dataset lab03_dataset_1.csv has 6,435 rows of data pertaining to Walmart sales and employment. The input features are Store, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI and the output is Unemployment. Perform the following tasks:

1. Create a heatmap for the entire dataset. (1 mark)
2. The input features should be subjected to feature scaling, specifically the min-max scaling. (2 marks)
3. Once the scaled input features are ready, learn a model using sklearn's linear regression module. Use a 90-10 train-test split for the learning process. (2 marks)
4. After you generate the linear regression model, output the regression score, coefficients, intercept and mean squared error (over the test set). (5 marks)
5. Create a scatter plot which showcases the true output and the predicted output for the test case. Make sure to display a single plot which should contain both the data points. Use two different colors to represent the two types of data. Don't forget to add a legend to the plot. (3 marks)

Problem 2: *k* – Nearest Neighbors (12 marks)

The dataset lab03_dataset_2.csv has the results of fraud investigations of 5,960 cases. The binary variable FRAUD indicates the result (output class) with 1 = Fraud, 0 = Not Fraud. The other quantitative variables contain information about the cases.

- DOCTOR_VISITS: Number of visits to a doctor.
- MEMBER_DURATION: Membership duration in number of months.
- NUM_CLAIMS: Number of claims made recently.
- NUM_MEMBERS: Number of members covered.

- OPTOM_PRESC: Number of optical examinations.
- TOTAL_SPEND: Total amount of claims in dollars.

Use the first 20% of the dataset i.e., the first 20% of the rows as the test set, while the remaining bottom 80% rows will be your training set. During majority voting, if both the classes have equal distribution within the nearest neighborhood, choose class = 1 (Fraud).

1. The input features used during training should be subjected to feature scaling, specifically the min-max scaling. (2 marks)
2. You will use sklearn's *k* – nearest neighbors module to learn a classification model with multiple nearest neighbors ranging from 2 to 5. Apply the learned *k*–NN model to classify the test set. Compute the misclassification rates for *k* ranging from 2 to 5. Use Euclidean distance as the similarity measure. (5 marks)
3. Next, apply sklearn's *k*–d tree module to classify the test set. In a similar manner to the above scenario, compute the misclassification rates for *k* ranging from 2 to 5. Use Manhattan distance as the similarity measure. (5 marks)