**DSE 2262 MACHINE LEARNING LABORATORY**

**SECTION A BATCH 2**
**WEEK 1 DATE: 15 MARCH 2022**

Exercise 1

1. Write a user defined function ‘myFnLinReg(x,y)’ to perform Simple Linear Regression given one predictor attribute and one response attribute. The function should return the coefficients of the straight line.
2. Use mtcars data set and consider the attributes mpg and weight. Split data into train and test sets (80 %,20%). Put training data set to ‘myFnLinReg(x,y)’ to build a linear regression model to predict mpg given the weight of the car.
3. What is the mpg of a car, whose weight is 5.5?
4. Compute and print accuracy measures such as RMSE and $R^2$ for the test set.
5. Apply the stochastic gradient descent and mini batch gradient descent algorithms to enhance the accuracy and visualize the cost function.

Exercise 2

1. Use the boston.csv dataset and determine the best 5 features to predict ‘MEDV’.
2. Using sklearn.linear_model, find the multiple regression model for the boston.csv dataset using the best 3 features. (from sklearn.linear_model import LinearRegression)
3. Find the accuracy of the model using appropriate metrics using 80, 20 split for training and test.

**DSE 2262 MACHINE LEARNING LABORATORY**

**SECTION A BATCH 1**
**WEEK 1 DATE: 17 MARCH 2022**

Exercise 1

6. Write a user defined function ‘myFnLinReg(x,y)’ to perform Simple Linear Regression given one predictor attribute and one response attribute. The function should return the coefficients of the straight line.
7. Use mtcars data set and consider the attributes mpg and weight. Split data into train and test sets (70 %,30%). Put training data set to ‘myFnLinReg(x,y)’ to build a linear regression model to predict mpg given the weight of the car.
8. What is the mpg of a car, whose weight is 6.7?
9. Compute and print accuracy measures such as RMSE and $R^2$ for the test set.
10. Apply the stochastic gradient descent and mini batch gradient descent algorithms to enhance the accuracy and visualize the cost function.

Exercise 2

4. Use the boston.csv dataset and determine the best 5 features to predict ‘MEDV’.
5. Using sklearn.linear_model, find the multiple regression model for the boston.csv dataset using the best 4 features. (from sklearn.linear_model import LinearRegression)
6. Find the accuracy of the model using appropriate metrics using 80, 20 split for training and test.

**DSE 2262 MACHINE LEARNING LABORATORY**

**SECTION A BATCH 2**

**WEEK 2 DATE: 22 MARCH 2022**

**EXER 1:**

1. Use the "pima-indians-diabetes.csv" dataset and note down the meta information.
2. Compute mean & standard deviation , tabulate and visualize the age of the patients.
3. Analyze and tabulate the relationship of age, BMI of patients with respect to the class.
4. Tabulate the class label and comment on whether the classes are balanced.
5. Use the data set to build a logistic regression model (using sklearn) and predict the class label. Divide the dataset into training and test set (70,30) using train_test_split method in sklearn.
6. Use the test data set and evaluate the performance using a confusion matrix. Visualize the confusion matrix using a heat map.
7. Compute accuracy rate, true positive and true negative rate and comment on the performance.
8. Visualize the ROC curve, and comment on the performance of the classifier.

**EXER 2:**

1. For the IRIS data set write down the meta information.
2. Visualize the class label against the predictor variable using appropriate plots.
3. Use the IRIS data set to build a logistic regression model (using sklearn) and predict the class label 'Species'. Divide the dataset into training and test set (70,30) using train_test_split method in sklearn.
4. Analysis and visualize the performance of the classifier using metrics, confusion matrix .
5. Use the IRIS data and KNeighborsClassifier (using sklearn) and predict the class label 'Species' for k value between 2 and 20. Divide the dataset into training and test set (70,30) using train_test_split method in sklearn.
6. Identify the best k ( for k between 2 and 20) for the model built.
7. Comment on the classifier (Logistic Regression or KNeighborsClassifier ) that has a better performance for the IRIS dataset.

**DSE 2262 MACHINE LEARNING LABORATORY**

**SECTION A BATCH 1**

**WEEK 2 DATE: 24 MARCH 2022**

**EXER 1:**

9. Use the "pima-indians-diabetes.csv" dataset and note down the meta information.
10. Compute mean & standard deviation , tabulate and visualize the age of the patients.
11. Analyze and tabulate the relationship of age, BMI of patients with respect to the class.
12. Tabulate the class label and comment on whether the classes are balanced.
13. Use the data set to build a logistic regression model (using sklearn) and predict the class label. Divide the dataset into training and test set (70,30) using train_test_split method in sklearn.
14. Use the test data set and evaluate the performance using a confusion matrix. Visualize the confusion matrix using a heat map.
15. Compute accuracy rate, true positive and true negative rate and comment on the performance.
16. Visualize the ROC curve, and comment on the performance of the classifier.

**EXER 2:**

8. For the IRIS data set write down the meta information.
9. Visualize the class label against the predictor variable using appropriate plots.
10. Use the IRIS data set to build a logistic regression model (using sklearn) and predict the class label 'Species'. Divide the dataset into training and test set (70,30) using train_test_split method in sklearn.
11. Analysis and visualize the performance of the classifier using metrics, confusion matrix .
12. Use the IRIS data and KNeighborsClassifier (using sklearn) and predict the class label 'Species' for k value between 2 and 20. Divide the dataset into training and test set (80,20) using train_test_split method in sklearn.
13. Identify the best k ( for k between 2 and 20) for the model built.
14. Comment on the classifier (Logistic Regression or KNeighborsClassifier ) that has a better performance for the IRIS dataset.

**DSE 2262 MACHINE LEARNING LABORATORY**

**SECTION A BATCH 2**
**WEEK 3 DATE: 29 MARCH 2022**
**EXER 1**

1. Use the titanic data set, perform preprocessing by deal with missing values, drop irrelevant attributes.
2. Use the scikit learn pipelines to perform the preprocessing - standardizing, encoding and model fitting in one step.
3. Perform Bayes classification using cross validation.
4. Tabulate using relevant measures of accuracy , Sensitivity and specificity.
5. Visualize the ROC curve and comment on performance

EXER 2

Download the "Womens Clothing E-Commerce Reviews.zip" file and answer the following:

1. Preprocessing:
    a. Find any null values are present or not, If present remove those data.
    b. Remove the data that have less than 5 reviews.
    c. Clean the data and remove the special characters and replace the contractions with its expansion. Convert the uppercase character to lower case. Also, remove the punctuations.
2. Separate the columns into dependent and independent variables (or features and labels). Then you split those variables into train and test sets (80:20).
3. Apply the Naïve Bayes Classification Algorithm on Sentiment category to predict if item is recommended
4. Tabulate accuracy in terms of precision, recall and F1 score.

**DSE 2262 MACHINE LEARNING LABORATORY**

**SECTION A BATCH 1**
**WEEK 3 DATE: 31 MARCH 2022**
**EXER 1**

6. Use the titanic data set, perform preprocessing by deal with missing values, drop irrelevant attributes.
7. Use the scikit learn pipelines to perform the preprocessing - standardizing, encoding and model fitting in one step.
8. Perform Bayes classification using cross validation.
9. Tabulate using relevant measures of accuracy , Sensitivity and specificity.
10. Visualize the ROC curve and comment on performance

EXER 2

Download the "Womens Clothing E-Commerce Reviews.zip" file and answer the following:

5. Preprocessing:
    a. Find any null values are present or not, If present remove those data.
    b. Remove the data that have less than 5 reviews.
    c. Clean the data and remove the special characters and replace the contractions with its expansion. Convert the uppercase character to lower case. Also, remove the punctuations.
6. Separate the columns into dependent and independent variables (or features and labels). Then you split those variables into train and test sets (80:20).
7. Apply the Naïve Bayes Classification Algorithm on Sentiment category to predict if item is recommended
8. Tabulate accuracy in terms of precision, recall and F1 score.

**DSE 2262 MACHINE LEARNING LABORATORY**

**SECTION A BATCH 2**

**WEEK 3 DATE: 5th APRIL 2022**

**EXER 1**

1. Use the German credit rating dataset "German Credit Data.csv" , Decision tree classifier to predict good or bad credit. Use "sklearn.model_selection" and GridSearchCV to search the hyperparameter values and report the most optimal one. Configure the grid search to search for optimal parameters:
    - Splitting criteria: gini or entropy.
    - Maximum depth of decision tree ranging from 2 to 10.
    - The searching of optimal parameter will be validated using 10-fold cross validation and the most optimal parameter will be chosen based on ROC AUC score.

2. Visualize the tree using graphviz software.

3. Display the text representation of the rules learnt.

**EXER 2**

Download fuel consumption dataset "FuelConsumption.csv", which contains model-specific fuel consumption ratings and estimated carbon dioxide emissions.

- Select the features 'ENGINESIZE', 'CYLINDERS', 'FUELCONSUMPTION_COMB', 'CO2EMISSIONS' to use for building the model. Plot Emission values with respect to Engine size.

- split the data into training and test sets (70:30) to create a model using training set, evaluate the model using test set, and use model to predict unknown value.

- Try to use a polynomial regression with the dataset of degree – 3, 4 & 5. Verify the accuracy by calculating Mean absolute error, Residual sum of squares, R2-score and comment on which model is the best.