CMPE 255-04, Fall 2023

Bonus Assignment #3

Release on Nov 1st , 2023
Due 11:59pm on Sunday, Nov 19th, 2023

Notes

This assignment should be submitted in Canvas as a format of ipython notebook (bonus3_yourFirstName_LastnName.ipynb).
No late assignments will be accepted.
You may collaborate on homework but must write independent code/solutions. Copying and other forms of cheating will not be tolerated and will result in a zero score for the homework (minimal penalty) or a failing grade for the course.
Your work will be graded in terms of correctness, completeness, and clarity, not just the answer. Thus, correct answers with no or poorly written supporting steps may receive very little credit.

The LendingClub is a peer-to-peer leading company that directly connects borrowers and potential lenders/investors. In this assignment, you will build a classification model to predict whether or not a loan provided by LendingClub is likely to default.

NOTE: There are two options for this bonus assignment. Please choose one of them. If you submit both cases, the highest-grade part will be counted.

CASE A:
1. (3 pts) Binary classification with Decision tree based on classification error from scratch

Please download lending_club_data3a.csv file.
In this dataset, the following features are extracted from the original data.

credit	categorial level of credit score of borrowers
term	*the term of the loan*
income	categorial level of income of borrowers
y	risky or safe loan

The target column (label column) of the dataset is called 'y'.
You need to split the data into train/test dataset with the ratio of 80% / 20% to check overfitting.
Please use random state=123 for splitting the data.

1-1 (2pts). Build a decision tree classifier of a loan prediction based on classification error.

Build a decision tree classifier of a loan prediction and train the model on the train dataset.

NOTE: Please do not use any package/library including scikit-learn library except NumPy, Pandas, and Matplotlib.

1-2 (1pt). Please build confusion matrix based on the decision tree model
Please calculate accuracy, precision, recall, F1 score for train/test dataset.
Please plot ROC curve and calculate AUC for train/test dataset
NOTE: Please use bagging method (random sampling for subset of dataset) to get more than 10 data points for ROC curve.
Please do not use any package for ROC curve and AUC calculation. You need visualize ROC curve using Matplotlib and calculate AUC from scratch.

CASE B:

1. (2 pts) Binary classification with Decision tree based on Entropy or Gini index from scratch

Please download lending_club_data3a.csv file.
In this dataset, the following features are extracted from the original data.

credit	categorial level of credit score of borrowers
term	*the term of the loan*
income	categorial level of income of borrowers
y	risky or safe loan

The target column (label column) of the dataset is called 'y'.
You need to split the data into train/test dataset with the ratio of 80% / 20% to check overfitting.
Please use random state=123 for splitting the data.

1-1 (1pt). Build a decision tree classifier of a loan prediction based on Entropy or Gini index.

Build a decision tree classifier of a loan prediction and train the model on the train dataset.

NOTE: Please do not use any package/library including scikit-learn library except NumPy, Pandas, and Matplotlib.

1-2 (1pt). Please build confusion matrix based on the decision tree model
Please calculate accuracy, precision, recall, F1 score for train/test dataset.
Please plot ROC curve and calculate AUC for train/test dataset
NOTE: Please use bagging method (random sampling for subset of dataset) to get more than 10 data points for ROC curve.
Please do not use any package for ROC curve and AUC calculation. You need visualize ROC curve using Matplotlib and calculate AUC from scratch.