CMPE 255-04, Fall 2023

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 (assignment3_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.

1. (6 pts) Binary classification with Logistic regression

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

Please download lending_club_data1a.csv file.
In this dataset, the following features are extracted from the original data.

loan_amnt	*loan amount of borrowers*
term	*the term of the loan*
int_rate	the interest rate of the loan
credit	numerical level of credit score of borrowers
emp_length	number of years of employment of borrowers
home_ownership	home_ownership status of borrowers: own, mortgage or rent
annual_inc	annual income of borrowers
dti	debt to income ratio of borrowers
bad_loans	1' means a risky (bad) loan, '0' means a safe loan

The target column (label column) of the dataset is called ‘bad_loans’.
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). Logistic regression using gradient descent method from scratch
Please build and train a logistic regression model to predict whether a loan is bad or not using the train dataset.
NOTE: To implement logistic regression, it requires only numerical values. Please turn categorical variables into binary features via one-hot encoding.

1-2 (2pts). Please build confusion matrix based on the trained model
Please calculate accuracy, precision, recall, F1 score for train/test dataset.

1-3 (2pts). Please plot ROC curve and calculate AUC for train/test dataset
NOTE: You might need to adjust threshold to get different data points for ROC curve.
Please 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.

2. (4 pts) Binary classification with Decision tree method

Please download lending_club_data2a.csv file.
In this dataset, the following features are extracted from the original data.

loan_amnt	*loan amount of borrowers*
term	*the term of the loan*
int_rate	the interest rate of the loan
credit	categorial level of credit score of borrowers
emp_length	number of years of employment of borrowers
home_ownership	home_ownership status of borrowers: own, mortgage or rent
annual_inc	annual income of borrowers
purpose	purpose of loan
dti	debt to income ratio
bad_loans	1' means a risky (bad) loan, '0' means a safe loan

The target column (label column) of the dataset is called ‘bad_loans’.
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

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

NOTE: You can use scikit-learn library for Decision tree.

1-2 (2pts). 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.