FE-520 Assignment 4

Zhiyuan Yao, Zhi Chen

April 5, 2022

Submission Requirement:

For all the problems in this assignment you need to design and use Python 3, output and present the results in nicely format.

Please submit a written report (pdf), where you detail your results and copy your code into an Appendix. You are required to submit a single python file (.py) and a brief report. Your grade will be evaluated by combination of report and code.

You are strongly encouraged to write comment for your code, because it is a convention to have your code documented all the time.

Python script must be a '.py' script, Jupyter notebook '.ipynb is not allowed.

Do NOT copy and paste from others, all homework will be firstly checked by plagiarism detection tool.

1 Linear Regression (30 pts)

Continue with the last problem in assignment 3, in this problem you should do the same thing to implement linear regression in class.

Instead of building 1d linear model, we want to build linear regression model with input multiple dimension.

$$y^T = xm + c$$

The Algorithm of gradient decent to find m and c is :

- Set initial variable. m=[0,0, ..., 0] (m is numpy array with (M, 1), where x has size of (N,M)) and c=0, learning rate L=0.001, max number of iterations I=10000.
- Write a for loop to update the weights m and constant c, in this loop, all the calculations are based on matrix multiplication:
    1. calculate $\delta_m = -x^T(y^T - xm - c)$
    2. calculate $\delta_c = -\mathbb{1}(y^T - xm - c)$,
    3. calculate $m = m - L \times \delta_m$
    4. calculate $c = c - L \times \delta_c$

- repeat previous step until max iteration numbers is reached

1. Initialize this class with x (2D array) and y (1D array) using numpy array as input.
2. Constructor should also take m (array), epochs (number of iterations), L (learning rate) as input, and set default values.
3. Define a gradient descent function (method) to find m and c with given iterations number.
4. Write a predictive function to predict y based on new input x (where x is a 2D array).

2 Credit Transaction data (30 points)

This dataset is simulated individual credit card transactions by one company. Please use this dataset to answer following question. Please notice that you may need to observe the dataset and clean it before answering the following question.

1. What is total amount spending captured in this dataset?

   Hint: you may observe $ in front of the amount, which you need remove, and () stands for negative value, which you need deduct the amount.

2. How much was spend at WW GRAINGER?

   Hint: All 'WW GRAINGER' contained in the 'Vendor'. Find the item as long as the name contain 'WW GRAINGER'

3. How much was spend at WM SUPERCENTER?

   Hint: All 'WM SUPERCENTER' contained in the 'Vendor'. Find the item as long as the name contain 'WW SUPERCENTER'

4. How much was spend at GROCERY STORES?

   Hint: All 'GROCERY STORES' contained in the 'Merchant Category Code'. Find the item as long as the name contain 'GROCERY STORES'

3 Data Processing with Pandas (40 points)

In this practice, you are expected to play around Pandas and get familiar with it. The dataset is quarterly dataset downloading from WRDS. Please remember that you need to do data transformation based on the new dataset generated by previous step. Do not using other package other than numpy and pandas.

1. Read 'Energy.xlsx' and 'EnergyRating.xlsx' as BalanceSheet and Ratings(dataframe).
2. For BalanceSheet, drop the column if more than 30% value in this colnmn is missing value, see how many features are remaining.

3. For BalanceSheet, drop the column if more than 90% value in this colnmn is 0, see how many features are remaining.

4. For BalanceSheet, replace all None or NaN with average value of each column.

5. For BalanceSheet, Normalize the table (Only need to normalize columns from 1Accounting Changes - Cumulative Effect' to 'Selling, General and Administrative Expenses') (Ignore it when the feature is already dropped)

   Using pd.apply() to normalize the table, in this table, you need to implement follow formula to calculate the normalized value:

   $$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

   (Do not using any function like MinMax(), you need to write it by yourself)

6. Calculate the correlation matrix for variables = ['Current Assets - Other - Total', 'Current Assets - Total', 'Other Long-term Assets', 'Assets Netting & Other Adjustments']. (Ignore it when the feature is already dropped)

7. Merge (inner) Ratings and BalanceSheet based on 'datadate' and 'Global Company Key', and name merged dataset 'Matched'.

8. Mapping

   For dataset 'Matched', we have following mapping:
   AAA = 0
   AA+ = 1
   AA = 2
   AA- =3
   A+ = 4
   A = 5
   A- = 6
   BBB+ = 7
   BBB = 8
   BBB- = 9
   BB+ = 10
   BB = 11
   others = 12
   Using map function to create a new varible = 'Rate', which maps ratings (S&P Domestic Long Term Issuer Credit Rating) to numerical ratings.

9. Choose ten features as input x, their corresponding Rate as target y. Run the linear regression function from question 1, return the coefficients.