BT3017

Feature Engineering for Machine Learning

Course Project

Due date: 15th April 2022 (Friday) 2359 hrs

Semester 2, AY21/22, School of Computing, National University of Singapore

IMPORTANT:

For this project, you are supposed to submit your project file to LUMINUS.

Instruction for submission:

- *Create a folder using the following naming convention:*

  StudentNumber_yourName_Project

- *Put all your code and your report (.pdf) in this folder.*

- *Zip your folder. Name your zip file using the following convention:*

  StudentNumber_yourName_Project.zip

  For example, if your student number is A1234567B, and your name is Chow Yuen Fatt, for this tutorial, your file name should be A1234567B_ChowYuenFatt_Project.zip

  Note: Do not submit your file in .rar format

- *Submit the zip file in the "Project Submit Here" folder in Luminus.*

The following files containing stock market information were downloaded from Kaggle:

"Dow Jones Industrial Average Historical Data.csv"
"NASDAQ_100_Data_From_2010.csv"

You are to perform the following tasks using PANDAS, Seaborn, and python codes:

1. From the NASDAQ csv file,
   (a) create a Pandas dataframe (name the dataframe "com12") containing all data in the csv file corresponding to the following 12 companies:
      - AAPL
      - ADBE
      - AMD
      - AMZN
      - ASML
      - FB
      - GOOG
      - INTC
      - MSFT
      - NVDA
      - PYPL
      - TSLA

   (b) From com12, and with reference to the Dow Jones Industrial Average (DJIA) data in *"Dow Jones Industrial Average Historical Data.csv"*, plot the correlation of the "close" stock prices of each of the 12 companies with the *next day* "open" price of DJIA.

   (c) From com12, perform the following:
      i. create another Pandas dataframe of "close" stock prices with date resolution changed from day to month i.e. reduce all prices of the same

month with one record containing the average “close” price of the month. Name this dataframe “com12month”.

ii. Using KMEANS clustering technique, cluster the 12 com12month data vectors corresponding to the 12 companies into clusters based on the company’s monthly average price. You may decide on a reasonable/meaningful number of clusters.

2. From the NASDAQ csv file,
   (a) Create a Pandas dataframe (name the dataframe “Nas2020”) containing the “close” stock prices in the csv file of all the records in the year 2020.

   (b) Create another Pandas dataframe of “close” stock prices with date resolution changed from daily to weekly i.e. reduce all prices of the same week with one record containing the average “close” price of the week. Name this dataframe “Nas2020week”.

   (c) Perform Principal Component Analysis on the “close” stock price data vectors (one vector per company) in Nas2020week.

   (d) Supposing you wish to reduce the dimension of the data vectors so that the machine learning input can take a reduced dimension feature vector. You wish to do so by projecting the data vectors onto the principal components and represent the data points using the projection coefficients. Explain how you would determine how many eigenvectors to use to reduce the data dimension.

3. From the NASDAQ csv file, create a Pandas dataframe containing all data of the company AAPL.

   You wish to use Fourier Transform to study the fluctuations in the daily “close” of AAPL. You should decide on the length of the Fourier Transform (i.e. how may days of data for every batch of Fourier analysis), and decide on how you would overlap the batches (eg. first batch day1 to dayN, second batch dayN/2 to day3N/2, etc).

4. From the experience gained in Questions 1,2,3 above, and optionally including other techniques, devise a set of useful features to predict the next day “close” price of a company given the previous history. Explain in detail how you would pre-treat the data (eg. normalization, etc). There is no need to implement the machine learning system. The focus should be on feature engineering.

******** IMPORTANT *********

Submit your work in a report format (.pdf), and zip it together with your project codes. Use the naming convention stated on the first page of this document.

It is ok to learn from internet, but you must acknowledge the source. If you use anything from internet or from others including your peers, and fail to acknowledge or mislead people into thinking it is your own work, that constitutes plagiarism. As this is a substantial assignment comprising 30% of the total grade, students who are caught committing plagiarism or any other form of cheating will get a “F” grade for this module, and also will possibly face other disciplinary actions.