Data Science 110
Homework 3

2021.11.05

Submission

- Deadline: 11/26 Fri. 23:59
  - Since the upcoming midterm week, you have 3 weeks to finish this homework.
- Submission delay: will get no points

Submission

- Upload: Ceiba homework section
- 上傳格式 : hw3_<student_id>.zip, e.g. hw3_r10921001.zip

⭐ 資料夾與檔案路徑如下 ⭐

- hw3_<student_id>/
  - hw3_Data1/
    - gene.txt
    - index.txt
    - label.txt
  - hw3_Data2/
    - -train.csv
    - -test.csv
  - report.pdf
  - hw3_p1.py
  - hw3_p2.py
  - hw3_p3.py

- 命名與規定不同or路徑底下檔案多or少 -> 均算檔案格式錯誤
- 檔案格式錯誤 -10%

Dataset for Problem 1 & 2

- Colon cancer data set (Alon et al. 1999).
- Please download the dataset hw3_Data1.zip (link). There are three files.
- (gene.txt) 62 samples with 2000 genes
- (index.txt) EST number and description of each of the 2000 genes
- (label.txt) The numbers correspond to patients, a positive sign to a normal tissue, and a negative sign to a tumor tissue.

Problem 1: One-by-one Feature Selection (30%)

- Do simple feature selection using **naive one-by-one selection** and answer the following questions in your report. (The formula in the lecture note is permitted.)

    1. (10%) Describe your feature selection method.

    2. (20%) Show your result and code of the feature selection. Which features are selected?

        - Code submission: hw3_p1.py

Problem 2: Subset-Based Feature Selection (40%) (1/2)

- Do a subset-based feature selection using **PSO (particle swarm optimization)**, **SA (simulated annealing)** or **GA (genetic algorithm)** heuristics and answer the following questions in your report.

  1. (20%) Describe the following details:

  (a) your algorithm, the metaheuristic you choose (PSO, SA or GA)

  (b) your objective function

  (c) the tunable parameters and tunable algorithm components (besides the objective function/cost function module) in your metaheuristic. What are the specific values/methods you use for your tunable parameters and algorithm component(s), if any.

  - Note: You can use the related algorithms covered in class, or you can reference any paper on these methods and implement their version. Add link to the original paper of the algorithm variation you use, if you use a variation different from the ones covered in class.

Problem 2 Subset-Based Feature Selection (40%) (2/2)

- Do a subset-based feature selection using **PSO (particle swarm optimization)**, **SA (simulated annealing)** or **GA (genetic algorithm)** heuristics and answer the following questions in your report.

  2. (20%) Show your result and code of the feature selection. (How many features are selected? Which features are selected? Etc.)

  - Code submission: hw3_p2.py

Problem 3 ARIMA Forecast (30%)(1/3)

- Write a Python code to perform an **ARIMA** analysis on Taiwan's Stock Exchange Index. Build an ARMIA model based on the "close" value data from 11/04/2020 up to 09/03/2021. And then forecast the "close" values for 09/06/2021 up to 11/04/2021. You can use the python package "pmdarima" to help with your implementation.
  - https://alkaline-ml.com/pmdarima/, https://github.com/alkaline-ml/pmdarima
  - Note:
    - You can perform any data transformation in the data preprocessing step if you see fit.
    - Also, you can use the "auto_arima()" call in your private experiments, but for homework submission, you are only allowed to use the "arima()" function call.

Problem 3 ARIMA Forecast (30%)(2/3)

- Please download the dataset hw3_Data2.zip (link). There are two files.
- (train.csv) stock info from 11/04/2020 to 09/03/2021 (for modeling)
- (test.csv) stock info from 09/06/2021 to 11/04/2021 (for prediction)

- Please use the column “Close” (which are exactly the close values) here.
- Please do not use the test data in modeling your ARIMA model.

*Data source: https://finance.yahoo.com/quote/%5ETWII?p=%5ETWII

Problem 3 ARIMA Forecast (30%)(3/3)

1. (15%) What are the ARIMA parameters (p, d, q, P, D, Q, s) that you use? And what is the mean square error (MSE) of your forecast?

2. (15%) Plot the whole stock (11/04/2020-11/04/2021) and your forecast data (09/06/2021~/11/04/2021) on the same figure. (x-axis: date, y-axis: close value)

- Code submission: hw3_p3.py