Manufacturing Data Science 製造數據科學

Assignment 4

Due Date: Jan. 10, 2022

Please solve the following questions and justify your answer. Show all your analysis result including equation/calculation or Python code in your report. **Upload your "zip" file including MSWord/PDF report and Python code with 檔名: MDS_Assignment4_ID_Name.zip" to NTU COOL** by due. The late submission is not allowed.

1. **(30%)** Statistical Process Control (Problem in 15.8.5)

Exercise 15.8.5 in Montgomery and Runger (Applied Statistics and Probability for Engineers, 7th edition, 2018)

Heart rate (in counts/minute) is measured every 30 minutes. The results of 20 consecutive measurements are as follows.

Sample No.	Heart Rate	Sample No.	Heart Rate
1	68	11	79
2	71	12	79
3	67	13	78
4	69	14	78
5	71	15	78
6	70	16	79
7	69	17	79
8	67	18	82
9	70	19	82
10	70	20	81

Use $\mu = 70$ and $\sigma = 3$.

(a) **(5%)** Construct an EWMA control chart with $\lambda = 0.1$. Use $L = 2.81$. Does the process appear to be in control?

(b) **(5%)** Construct an EWMA control chart with $\lambda = 0.5$. Use $L = 3.07$. Compare your results to those in part (a).

(c) **(5%)** If the heart rate mean shifts to 76, approximate the ARLs for the charts in parts (a) and (b).

(d) **(5%)** What's the probability that the control chart detects a shift to 76 on the first sample following the shift.

(e) **(10%)** What is the probability that the control chart does not detect a shift to 76 on the first sample following the shift, but does detect it on the second sample?

2. (30%) Prognostics and Health Management (PHM)

This dataset was used for the prognostics challenge competition at the International Conference on Prognostics and Health Management (PHM2008).

Data sets consist of multiple multivariate time series. Each data set is further divided into training and test subsets. Each time series is from a different engine – i.e., the data can be considered to be from a fleet of engines of the same type. There are 218 engines. Each engine starts with different degrees of initial wear and manufacturing variation which is unknown to the user. This wear and variation is considered normal, i.e., it is not considered a fault condition. There are three operational settings that have a substantial effect on engine performance. These settings are also included in the data. The data are contaminated with sensor noise.

The engine is operating normally at the start of each time series, and starts to degrade at some point during the series. In the training set, the degradation grows in magnitude until a predefined threshold is reached beyond which it is not preferable to operate the engine. In the test set, the time series ends some time prior to complete degradation. The objective of the competition is to predict the number of remaining operational cycles before in the test set, i.e., the number of operational cycles after the last cycle that the engine will continue to operate properly.

The data are provided as a zip-compressed text file with 26 columns of numbers, separated by spaces. Each row is a snapshot of data taken during a single operational cycle; each column is a different variable. The columns correspond to:

1) unit number
2) time, in cycles
3) operational setting 1
4) operational setting 2
5) operational setting 3
6) sensor measurement 1
7) sensor measurement 2

...

26) sensor measurement 21

Users are expected to train their algorithms using data in the file named train.txt. You must then evaluate the RUL prediction performance on data provided in file test.txt. You may download the dataset source here: https://ti.arc.nasa.gov/c/13/. For model evaluation, the final score is a weighted sum of RUL errors. The scoring function is an asymmetric function that penalizes late predictions more than the early predictions. (Please download the dataset from the linkage and see attached documentation for details)

Answer all the following questions with respect to **unit number 1 (i.e. engine #1) ONLY** for **TRAINING** dataset.

(a) **(5%)** RUL Calculation: calculate remaining useful life (RUL) for each engine (different unit number) (hint: use the max time (in cycles) minus the current time, and create the remain useful life column). In fact, this is the task of "labelling" for supervised learning.

(b) **(5%)** Variation Analysis: calculate the coefficient of variation (the ratio of the standard deviation to the mean) of each sensor. Which sensor shows maximum and minimum coefficient of variation, respectively? What's the insight you can provide? (hint: variance implies information content)

(c) **(5%)** Feature Engineering: feature engineering is used to derive more features for prediction. In time domain, calculate the "moving" average/ variance/ peak value (max value) of each sensor respectively, by using predetermined length of the time window (eg. In unit number 1：calculate (0 to 10), (1 to 11), (2 to 13)…if the time window equal to 11 cycles). You don't need to consider the end of time series which does not have enough cycles in the window.

(d) **(5%)** Feature Selection 1: do the similar moving average work to RUL column, then calculate the correlation coefficients between each generated feature (i.e. average, variance, and max value) and RUL column. Identify the Top 10 features which have high **absolute value** of correlation coefficients with RUL.

(e) **(5%)** Feature Selection 2: Use random forest and identify the Top 10 important features (i.e. average, variance, and max value) with respect to the RUL column. Plot the line plot which x axis is cycle and y axis is these 10 important features.

(f) **(5%送分題)** Please "feel free" to read the two solutions shown in the following linkages when you are available. You don't need to do any work about this question (f).

Python solution: https://github.com/mustafashabbir10/Prognosis/blob/master/Prognosis.ipynb

R solution: http://mkalikatzarakis.eu/wp-content/uploads/2018/12/code.html

3. Programming Questions (35%)

Please use ***Python*** to answer the following questions. Provide your code and justify your answer. Show all your work in detail including specific algorithm and parameter design. You should hand in ***TWO files (one for Tabu and one for Genetic Algorithm)*** regarding to each meta-heuristic algorithm, respectively. The result should include ***optimal solution (i.e., job sequence), optimal function (i.e. fitness) value, running time, number of tardy jobs.*** For the parameter settings (eg. tabu size, crossover rate, mutation rate, etc.), please give a simple ***trial-and-error*** or ***design of experiment*** for sensitivity analysis.

Single-Machine Scheduling Problem

Please answer following single-machine total weighted tardiness problem. The objective function is to minimize the total weighted tardiness.

Jobs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Processing Time	10	10	13	4	9	4	8	15	7	1	9	3	15	9	11	6	5	14	18	3
Due Date	50	38	49	12	20	105	73	45	6	64	15	6	92	43	78	21	15	50	150	99
Weights	10	5	1	5	10	1	5	10	5	1	5	10	10	5	1	10	5	5	1	5

(a) **(5% 送分題)** Learn Genetic Algorithm (GA) from the internet video https://www.youtube.com/watch?v=kHyNqSnzP8Y or https://www.youtube.com/watch?v=Fdk7ZKJHFcI .

(b) **(15%)** Develop Tabu Search (TS) algorithm to solve the problem. Show your design and the "result".

(c) **(15%)** Develop Genetic Algorithm (GA) to solve the problem. Show your design and the "result".

(d) **(5%)** Please give a comparison between Tabu and GA. You may try different parameters to see the change of the results (i.e., sensitivity analysis) in your developed algorithm. What's the "insight" or interesting things you found?

Note

1. Show all your work in detail. **Innovative** idea is encouraged.
2. If your answer refers to any external source, please "must" give an academic citation. Any "plagiarism" is not allowed.

Merry Christmas and Happy New Year!!