Manufacturing Data Science 製造數據科學

Assignment 4

Due Date: Dec. 16, 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. (40%) Data Imbalance Problem

在 UCI Machine Learning Repository 開放數據中包含了一個半導體製造數據(semiconductor manufacturing dataset，https://archive.ics.uci.edu/ml/datasets/SECOM)，一共包含了 1,567 個觀測值，而每個觀測值具有 590 個特徵(感測值的量測結果)以及作為目標值的測試結果(為二元變數，良品為-1，不良品為 1)，其中不良品僅有 104 個樣本。試著參考網路資源學習並撰寫程式，使用此數據回答下列問題。

(a) (15%) 試在該數據分析流程中加入數據平衡的步驟，使用**三種**方法來進行數據平衡(e.g. 使用上抽樣、下抽樣或是代價敏感學習等)。

(b) (5%) 建議選用哪種方法最為合適？為什麼？

(c) (10%) 對於數據多數群與少數群的比例應當調整至多少？為什麼？試透過調整生成比率(i.e. 敏感度分析)來看模型分類結果。(提示：將敏感度分析以繪圖呈現兩條曲線，x 軸為不同生成比例、y 軸為偽陽性率(false positive rate)與偽陰性率(false negative rate))

(d) (10%) 試說明特徵挑選步驟應於數據平衡**前**或**後**，這對預測結果有何影響？

2. (30%) Programming Questions

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) **(0%)** 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".

3. (30%) Markov Decision Process

考慮一個沒有折損因子的機台維修保養的馬可夫決策過程，機台有四個狀態(健康, 可用, 耗損, 損壞)，其各別獎勵為(6, 3, 1, -15)。「損壞」狀態為吸收狀態(absorbing)，行動主要有兩種(加工, 保養)。

- 在狀態「健康」的情況下，採取行動「加工」，轉移到「健康」的機率為 ~~0.8~~；轉移到「可用」的機率為 0.2；轉移到「耗損」的機率為 0.1；轉移到「損壞」的機率為 0.0。
- 在狀態「健康」的情況下，採取行動「保養」，轉移到「健康」的機率為 1.0；轉移到其他狀態的機率為 0.0。
- 在狀態「可用」的情況下，採取行動「加工」，轉移到「健康」的機率為 0.0；轉移到「可用」的機率為 0.6；轉移到「耗損」的機率為 0.3；轉移到「損壞」的機率為 0.1。
- 在狀態「可用」的情況下，採取行動「保養」，轉移到「健康」的機率為 0.8；轉移到「可用」的機率為 0.2；轉移到其他狀態的機率為 0.0。
- 在狀態「耗損」的情況下，採取行動「加工」，轉移到「健康」的機率為 0.0；轉移到「可用」的機率為 0.1；轉移到「耗損」的機率為 0.5；轉移到「損壞」的機率為 0.4。
- 在狀態「耗損」的情況下，採取行動「保養」，轉移到「健康」的機率為 0.2；轉移到「可用」的機率為 0.5；轉移到「耗損」的機率為 0.3；轉移到「損壞」的機率為 0.0。

試著參考網路資源學習並撰寫程式，使用此數據回答下列問題。

(a) (5%) 試根據題目繪製轉移機率圖(transition probability diagram)；
(b) (10%) 使用價值迭代來決定最佳策略以及各個狀態的價值；
(c) (10%) 使用策略迭代來決定最佳策略以及各個狀態的價值，假設初始策略為在所有狀態皆採取行動「加工」；
(d) (5%) (d)承接(c)的答案，如果初始策略在所有狀態皆採取行動「保養」，策略迭代的計算過程與結果有什麼差異？

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!!