Manufacturing Data Science 製造數據科學

Assignment 1

Due Date: 5pm, Oct. 6, 2023

Please solve the following questions and justify your answer by using Python. Show all your analysis result including Python code in your report. **Upload your "zip" file including (1) MS Word/LaTeX pdf report (answering each question and its sub-questions) and Python code; or (2) notebook (including answer and code), with file name: "MDS_Assignment1_ID_Name.zip" to NTU COOL** by due. The late submission is not allowed.

1.* (35%) *Linear Regression Analysis for Wine Quality

For the attached red wine dataset (**MDS_Assignment1_winequality.xlsx**), please use "multiple regression" to find the potential linear pattern (i.e., linear regression equation) for 1599 observations with 11 input variables and 1 output variable (label variable is regarded as continuous variable 反應變數請視為連續變數). Please answer the following questions by using Python software and package:

(a) **(10%)** Show the results of regression analysis as follows.

	estimate	std. error	t value	p-value
Intercept				
Fixed acidity				
Volatile acidity				
…				
Alcohol				

R-squared: 0.xxxx, Adjusted R-squared: 0.xxxx

(b) **(5%)** The fitting of the linear regression is a good idea? If yes, why? If no, why? What's the possible reason of poor fitting?

(c) **(5%)** Based on the results, rank the independent variables by p-values and which one are statistically significant variables with p-values<0.01? (i.e. 重要變數挑選)

(d) **(15%)** Testify the underlying assumptions of regression (1) Normality, (2) Independence, and (3) Homogeneity of Variance with respect to residual.

Dataset is related to red vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests. See Cortez et al. (2009) for more information.

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

Source: Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez

A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

2. (35%) Association Rule- Market Basket Analysis

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. The dataset includes 541,909 instances and 8 attributes. Attribute information is described as follows.

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

That is exactly what the “market basket analysis” contains: a collection of invoices with each line representing 1 item purchased with respect to its corresponding invoice. A collection of several lines with the same InvoiceNo is called a transaction. You can see the online retail data set (MDS_Assignment1_OnlineRetail.xlsx). Note that you ONLY need the first column and second column, and transform into the “format” for extracting association rules (we don’t consider the purchasing quantity here). Use “association rule” to find the potential patterns which satisfy the following criterion:

- Set the minimum support to 0.001
- Set the minimum confidence of 0.6

Please answer the following questions:

(1) **(10%)** How to handle the raw dataset via data preprocessing?

(2) **(10%)** Define and what's the top 10 association rules? Show the support, confidence, and lift to each specific rule, respectively?

(3) **(5%)** Please provide/guess the "story" to interpret **one** of top-10 rules you are interested in (Third column in dataset shows item Description).

(4) **(10%)** Give a visualization graph of your association rules.

Source:
https://archive.ics.uci.edu/dataset/352/online+retail

3. (30%) Manufacturing System Analysis

針對一流水線產線，有四個工作站，每個工作站的平行等效機台數(只需經過其中一台)與其每機台的加工時間如下表。

(a) **(10%)**根據 Little's Law，試計算各工作站的產出率 TH 於下表；試問瓶頸站的產出率 r_b、最小生產週期時間(總加工時間，T_0)、關鍵在製品水準(W_0)各為多少？

(b) **(10%)**試給出最佳績效(best case)下，最大的產出率(THbest)與最小生產週期時間(CTbest)的計算公式(提示：參閱講義)

工作站編號	平行等效機台數	加工時間(小時)	工作站的產能 TH (個/小時)
1	5	7	
2	2	3	
3	6	15	
4	3	5	

(c) **(10%)**根據該問題的產線，試程式撰寫建立一**模擬**模型(或用套裝軟體、數值分析)來驗證，當在製品 WIP 數量超過工廠產能時，其生產週期將嚴重惡化。也就是當產線的投料速度(投產量)大於產線的產出率，此時生產系統將處於非穩態的狀態(non-steady state)。試用圖表呈現 WIP、CT 與 TH 之間惡化的關係。(提示：參閱講義)

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.