**Manufacturing Data Science 製造數據科學**

**Assignment 2**

Due Date: 5pm, Oct. 27, 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_Assignment2_ID_Name.zip" to NTU COOL** by due. The late submission is not allowed.

ps: For some questions, you may ask ChatGPT and copy-and-paste its answer. Then, based on its responses, you can provide your further response with more insights or mentioning some aspects which ChatGPT didn't take into account.

1. (20%) (a)試簡述何謂維度的詛咒？試列舉一案例說明。(b)避免維度詛咒的方法有哪些？(c)試找一個開放數據(e.g. Kaggle 開放數據)並選一種方法(e.g. 線性迴歸或決策樹)，用模擬方法固定樣本數但逐步增加變數個數，試著重新繪製圖 3.12，呈現維度與預測(或分類)績效間的關係。(提示：模擬方法可思考如下：(1)先做線性迴歸；(2)重要變數依 p-value 排序；(3)將重要的變數一個個依序放入迴歸並計算 adjusted-R2 作為預測準確度)。(d)若準確度有或沒有明顯下降，請試著說明為什麼？

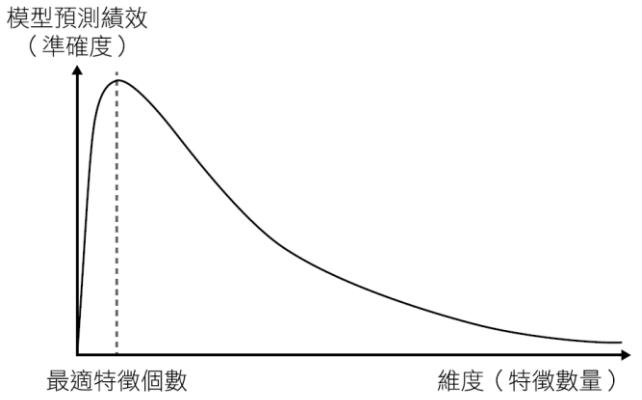


圖 3.12　數據維度與模型的預測的準確度

2. (20%) (a)試說明損失函數與模型評估指標有何不同？(b)試使用網際網路(internet)學習，損失函數的設計有哪些？試列舉兩種，並說明其各自的優缺點或可建議的使用時機。(c)如何根據不同情況選擇損失函數？試舉例或用開放數據說明之。

3. (30%) This dataset can be used to predict the chronic kidney disease and it can be collected from the hospital nearly 2 months of period. Data set is **MDS_Assignment2_kidney.xlsx** and data source is https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease#. The dataset includes 400 observations and 24+1 attributes ( 11 numeric, 14 nominal). The last attribute is the "Class" label. We use the following representation to collect the dataset.

1.Age(numerical)
age in years
2.Blood Pressure(numerical)
bp - in mm/Hg
3.Specific Gravity(nominal)
sg - (1.005,1.010,1.015,1.020,1.025)
4.Albumin(nominal)
al - (0,1,2,3,4,5)
5.Sugar(nominal)
su - (0,1,2,3,4,5)
6.Red Blood Cells(nominal)
rbc - (normal,abnormal)
7.Pus Cell (nominal)
pc - (normal,abnormal)
8.Pus Cell clumps(nominal)
pcc - (present,notpresent)
9.Bacteria(nominal)
ba - (present,notpresent)
10.Blood Glucose Random(numerical)
bgr in mgs/dl
11.Blood Urea(numerical)
bu in mgs/dl
12.Serum Creatinine(numerical)
sc in mgs/dl
13.Sodium(numerical)
sod - in mEq/L
14.Potassium(numerical)
pot- in mEq/L
15.Hemoglobin(numerical)
hemo - in gms
16.Packed Cell Volume(numerical)
pcv
17.White Blood Cell Count(numerical)
wc - in cells/cumm
18.Red Blood Cell Count(numerical)
rc - in millions/cmm
19.Hypertension(nominal)
htn - (yes,no)
20.Diabetes Mellitus(nominal)
dm - (yes,no)

21.Coronary Artery Disease(nominal)
cad - (yes,no)
22.Appetite(nominal)
appet - (good,poor)
23.Pedal Edema(nominal)
pe - (yes,no)
24.Anemia(nominal)
ane - (yes,no)
25.Class (nominal)
class - (ckd,notckd)

(a)根據此開放數據，您會用什麼方法來確認資料品質的好壞？試操作一次並說明其細節。(b)試建議三個可能衡量數據品質的量化指標(i.e. KPIs)。(c)如何處理遺漏值(missing values)？又或某些欄位不打算遺漏值處理的理由為何？

4. (30%) 根據上題 Chronic_Kidney_Disease 的數據集，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) 若要建構線性迴歸或羅吉斯迴歸分析，如何處理某些類別或名目尺度的欄位？
   (b) 試將羅吉斯迴歸分析的結果呈現如下表，並試著解釋任一特徵與目標值之間的關係。

| | estimate | std. error | t value | p-value |
|---|---|---|---|---|
| intercept | | | | |
| age | | | | |
| bp | | | | |
| … | | | | |
| ane | | | | |

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

   (c) 基於上述(b)的結果，將上述特徵以 t value 進行排序後，哪些特徵的迴歸係數在統計上是顯著的呢(p-value<0.01)？
   (d) 試問配適羅吉斯迴歸模型是否合適？試若配適不佳，試說明其可能的原因為何？
   (e) 試問配適線性判別分析模型是否合適？若配適不佳，試說明其可能的原因為何？
   (f) 試問配適二次判別分析模型是否合適？若配適不佳，試說明其可能的原因為何？

***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.