Manufacturing Data Science 製造數據科學

Assignment 2

Due Date: 5pm, Oct. 28, 2022

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.

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 開放數據或第一次作業紅酒數據集)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)

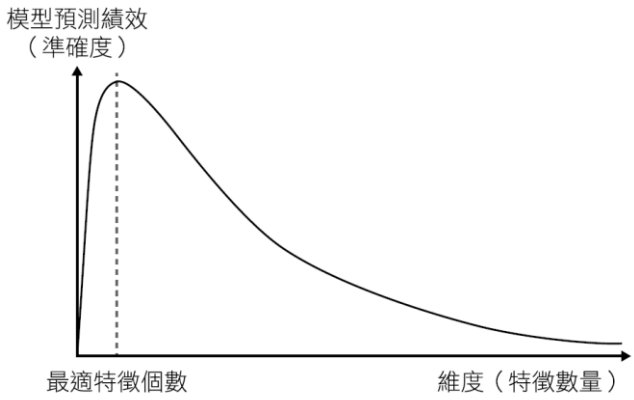

圖 3.12　數據維度與模型的預測的準確度

2. (20%) (a)試找一個開放數據(e.g. Kaggle 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)公司或您是否有現存方法來進行資料品質的確認？如果有(或沒有)，試依您的角度說明(或建議)確認資料品質的標準作業流程(i.e. SOP)。(c)試建議三個可能衡量數據品質的量化指標(i.e. KPIs)。

3. (20%) 在數據科學分析架構中的決策支援階段：(a)什麼是模型的適應性與擴充性？(b)在 AI 專案中(可根據第一題的開放數據與模型)，就您所使用的數據與建構的預測模型是否具備適用性與擴充性？為什麼？又該如何調整與改善呢？

4. (10%) 遺漏值填補的方法包括了統計量填補、預測式填補與生成式填補(a)試說明這些方法分別適用於什麼樣情形；(b)為什麼某特徵存在大量遺漏值不宜直接刪除？

5. (30%) 在 UCI Machine Learning Repository 開放數據中包含了一個鋼板缺陷數據(steel plates faults dataset，https://archive.ics.uci.edu/ml/datasets/steel+plates+faults)，一共包含了 1,941 個觀測值，而每個觀測值具有 27 個特徵以及作為目標值的 7 種缺陷。試挑選出凹凸不平(Bumps)以及刮痕(K_Scratch)兩種缺陷進行分析。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。

   (1) 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				
X_Minimum				
X_Maximum				
…				
SigmoidOfAreas				

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

   (2) 基於上述(a)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)？

   (3) 試問配適一個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？

   (4) 試問配適一個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？

   (5) 試問配適一個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.