[NTU ESOE] 110 Data Mining Midterm Project Deadline: 2021.11.10.23:00

Student ID : R09631007　Name : 吳乙澤　Department : 生物機電工程學系

ProjectType : Regression

***please check your project type in "Data Mining 期中作業說明.pdf"**

1. 使用指定資料集於 InAnalysis 進行分析，並完成機器學習畫布。（機器學習畫布請使用 MLC.docx 為模板修改）

2. 請詳述你對資料作的前處理，以及調整參數的過程邏輯，可以參考 InAnalysis 上 data preview 以及 model preview 的結果來描述，必須超過 350 字。

首先，我不做任何前處理，先把 cnt 選為 Output，其他所有的特徵都選為 Input，並以 Linear Regression 對資料進行訓練。訓練結束後，透過 model preview 觀察訓練結果。見表一的第一個模型 LinearRegression_NoPreprocessing_AllFeatures，訓練集的 MAE 為 107.55，測試集的 MAE 為 103.77。為了改善訓練，我利用相關係數的熱度圖，選擇和 cnt 有關的特徵，也就是相關係數絕對值在 0.1 以上的特徵，作為模型的輸入進行訓練。訓練結果顯示，訓練集的 MAE 為 107.97，測試集的 MAE 為 102.95。此時測試集的誤差 102.95，比選取所有特徵作為輸入時的誤差 103.77 要低，表示選擇有相關性的特徵進行訓練，對降低測試集的誤差是有用的。

表一、使用 Linear Regression 對未經前處理的資料做訓練

模型名稱	訓練集的 MAE	測試集的 MAE
LinearRegression_NoPreprocessing_AllFeatures	107.55	103.77
LinearRegression_NoPreprocessing_SelectDataByCorr	107.97	102.95

但是，以上訓練出來的 MAE 誤差都在 100 以上，令我覺得模型的預測能力不佳。於是，我開始對數據作前處理，希望使測試集的 MAE 降低到 100 以下。在使用 1st Standard Deviation 對所有資料做前處理時，發現部分特徵失去大量資料。舉例來說，yr 特徵在經過 1st Standard Deviation 後，資料數量從原來的 1500 筆，變成 763 筆資料，幾乎只剩下一半的資料量。見表二的 LinearRegression_1SD_Filter_AllFeatures，在使用 Linear Regression 對所有特徵做訓練後，導致了 Overfitting 的現象。此時訓練結果的 MAE 為 67.67，測試時的 MAE 為 191.6，兩者相差 2.8 倍。推測是前處理時過濾太多資料，導致模型在預測測試集時失常。為了解決此問題，我改用能減少複雜度，避免 Overfitting 發生的 Ridge Regression 和 Lasso Regression 進行訓練。在多次調整 Normalize 的 Alpha 參數後，我仍然沒有解決 Overfitting 的問題，但是有減少 Overfitting 的程度。見表二的 LassoRegression_1SD_Filter_AllFeatures，這

是我最好的訓練結果。訓練集 MAE 為 86.55，測試集 MAE 為 137.48，兩者相差 1.6 倍，比 Linear Regression 的訓練結果(2.8 倍)好上不少。

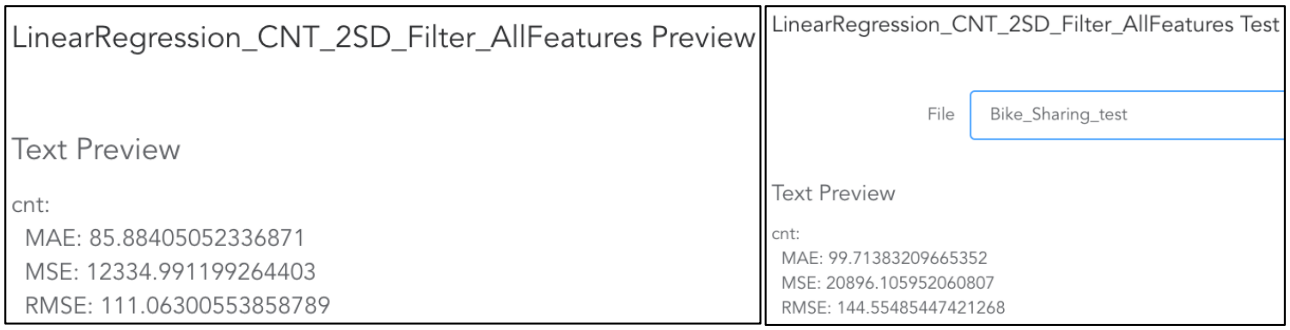
表二、使用 Linear/Ridge/Lasso Regression 對有經 1st Standard Deviation 處理的資料做訓練

模型名稱	訓練集的 MAE	測試集的 MAE
LinearRegression_1SD_Filter_AllFeatures	67.67	191.6
LassoRegression_1SD_Filter_AllFeatures	86.55	137.48

由上述結果可知，使用 1st Standard Deviation 會導致我無法解決的 Overfitting 問題。因此，只好使用 2nd Standard Deviation 和 3rd Standard Deviation 做前處理，或是只選擇部分特徵做 1st Standard Deviation 就好，看這樣的前處理策略，能不能使測試集的 MAE 誤差降至 100 以下。這部分的訓練結果會在題(3)中做說明。

3. 最終的訓練以及測試結果為何？請有邏輯的解釋你的結果。

我的最終訓練成功地使測試集的誤差降至 100 以下。該訓練的主要策略是以 2nd Standard Deviation 去除 cnt 的 Outlier，並利用 Linear Regression 演算法對所有特徵進行訓練。見圖一，此時訓練集的 MAE 為 85.88，測試集的 MAE 為 99.71，所用模型的名稱為 LinearRegression_CNT_2SD_Filter_AllFeatures。



圖一、最終訓練結果

4. 報告影片連結（請將報告影片連結附在這邊，並確保影片觀看權限開啟）

影片連結為：https://www.youtube.com/watch?v=F5_xq-htsdk