F74076027 林政傑 FDA_HW3-2

我選擇的資料集是 Electrical Grid Stability Simulated Data Data Set，這個資料集是參考一個論文，模擬 4 節點的分散式系統的電網所產出的，總共 14 個欄位。其中有三種特徵，分別是反應時間(tau1 ~ tau4)、用電狀況(p1 ~ p4)、跟電價有關的參數(g1 ~ g4)，每種特徵各有 4 份，總共 12 分。剩下的兩個欄位其一是特徵方程根的最大實部，如果為正，則電力系統不穩定，反之則穩定；另一欄位就是用字串 'stable' 和 'unstable' 紀錄其穩定與否。

分析資料時，先新增一個欄位，用 1 代表 stable；0 代表先 unstable，然後比較各項數據跟穩定度的關係，然而我比較所有 feature 之後，只能看出除了 p1，其他資料都是絕對值越大，電力系統越不穩定；p1 則是大於 5.7 時穩定度會大幅提升。而我也沒辦法讀完、讀懂那整篇論文，所以就沒有把這些數據拼湊起來做更深入的分析。在許多資訊不是這麼清楚的狀況下，我選擇使用隨機森林分類器，把這些屬性都納入考慮最後票選出最合適的結果。把 80%資料作為 train data，使用 5 次 kfold 進行驗證，得到:

```
average train accuracy: 0.9995625
min train accuracy: 0.99921875
max train accuracy: 0.99984375
average valid accuracy: 0.9026249999999999
min valid accuracy: 0.891875
max valid accuracy: 0.91
```

最後把 train data 都丟入模型訓練，拿剩下的 20% test data 來預測，得到正確率: 0.8895。最後的正確率比我預期的高很多，有可能是因為在論文中，這些參數都可以用一些方程式去互相轉換，彼此之間本來就有強烈相關性，而且資料量很充足(總共 10000 筆，用 8000 筆訓練)，所以可以簡單的套用隨機森林就達到 0.89 的準確度。