資訊科學概論作業說明

2019.11.1

資料集 研究所入取率預測

主要欄位說明

1. GRE Scores (out of 340)
2. TOEFL Scores (out of 120)
3. University Rating (out of 5)
4. Statement of Purpose and Letter of Recommendation Strength (out of 5)
5. Undergraduate GPA (out of 10)
6. Research Experience (either 0 or 1)
7. Chance of Admit (ranging from 0 to 1)

目標

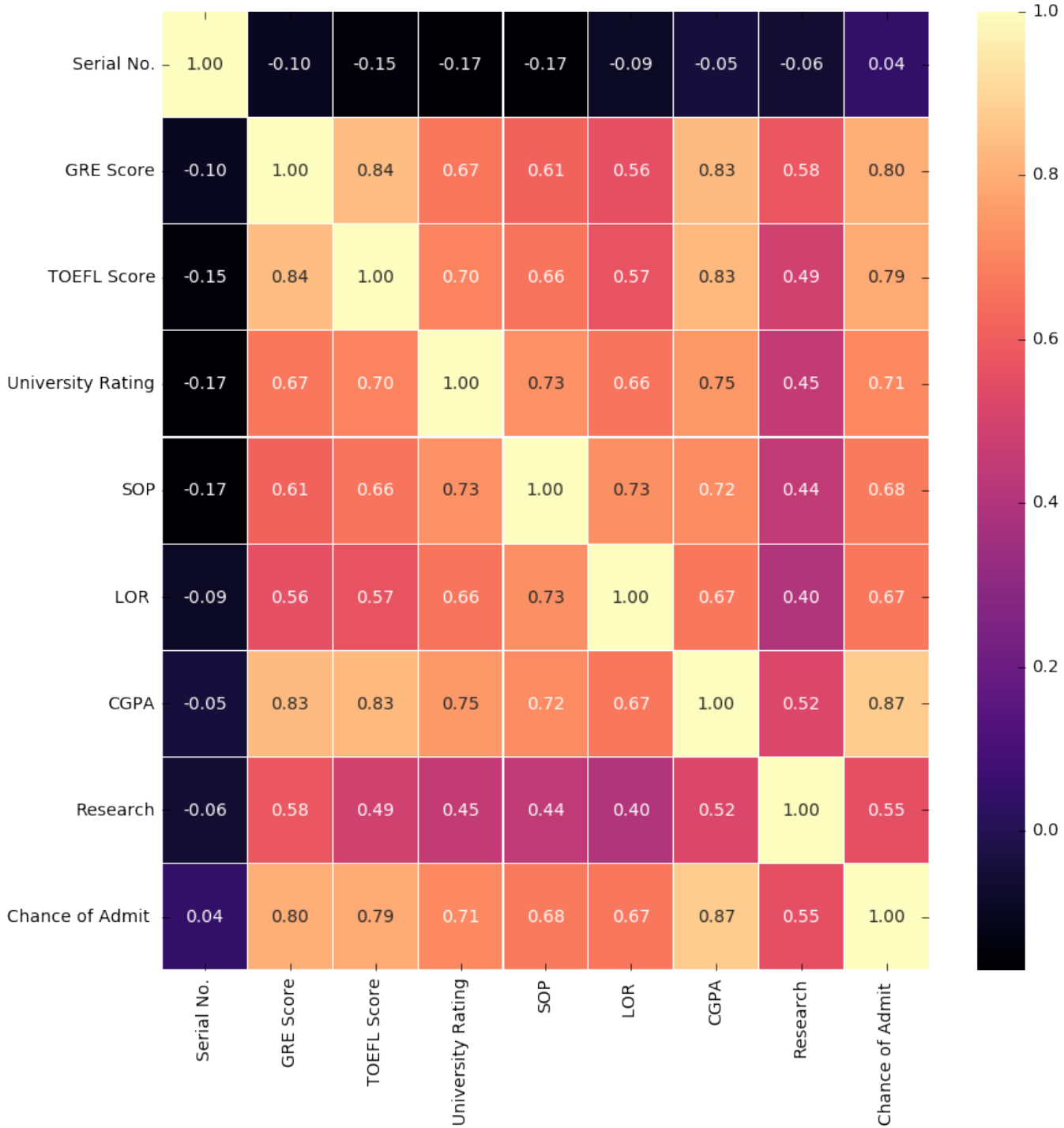
利用以下幾種機器學習方式去預測和分析

1.資料特性分析 ex:相關係數..
2.線性回歸分析
3.分類
4.分群
5.類神經網路

資料特性分析

相關係數分析
各欄位數值分佈

註：
讀取資料時注意欄位的名稱



線性回歸分析

驗證：決定係數 R2 or 均方誤差 MSE

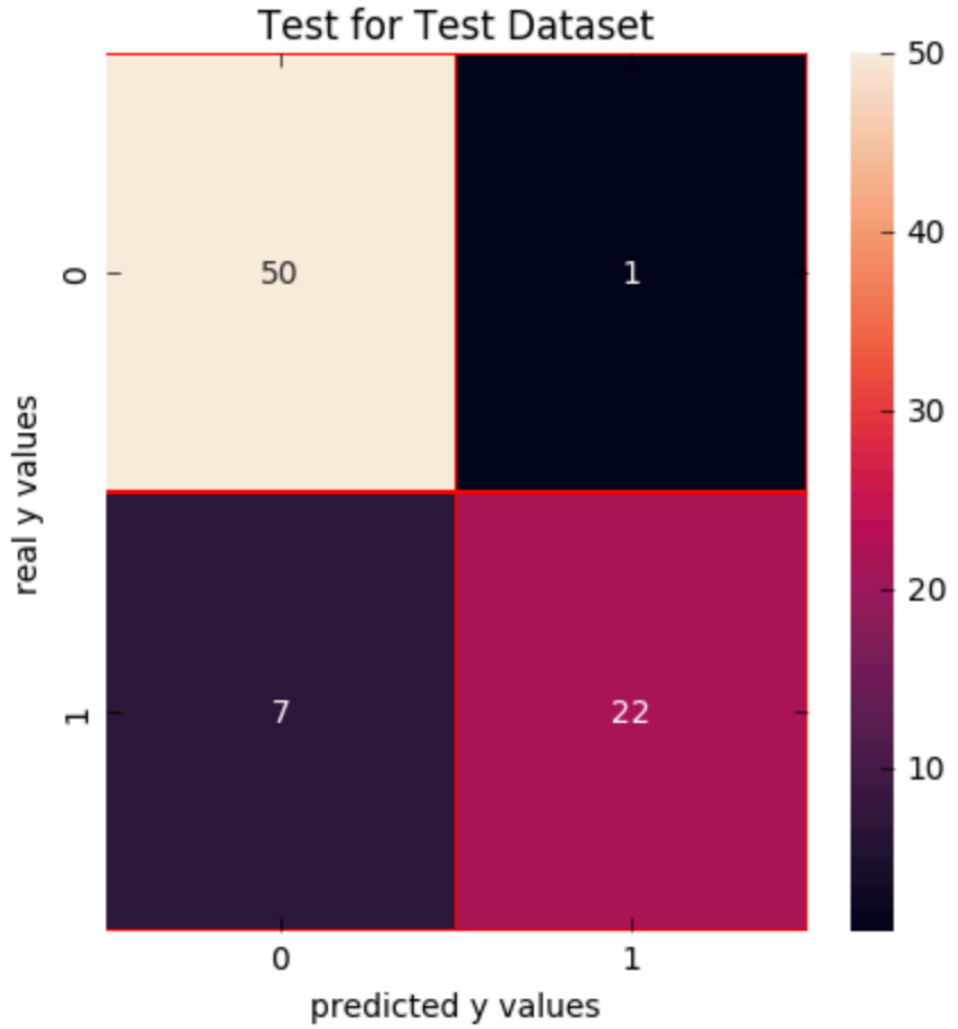
```
real value of y_test[1]: 0.68 -> the predict: [0.72368741]
real value of y_test[2]: 0.9 -> the predict: [0.93536809]
r_square score:  0.8212082591486991
r_square score (train dataset):  0.7951946003191086
```

分類

方法：決策樹分類、羅吉斯回歸

驗證：混淆矩陣、ROC曲線

註：
由於目標值是個連續的隨機變量
因此需設替目標值設閾值threshold

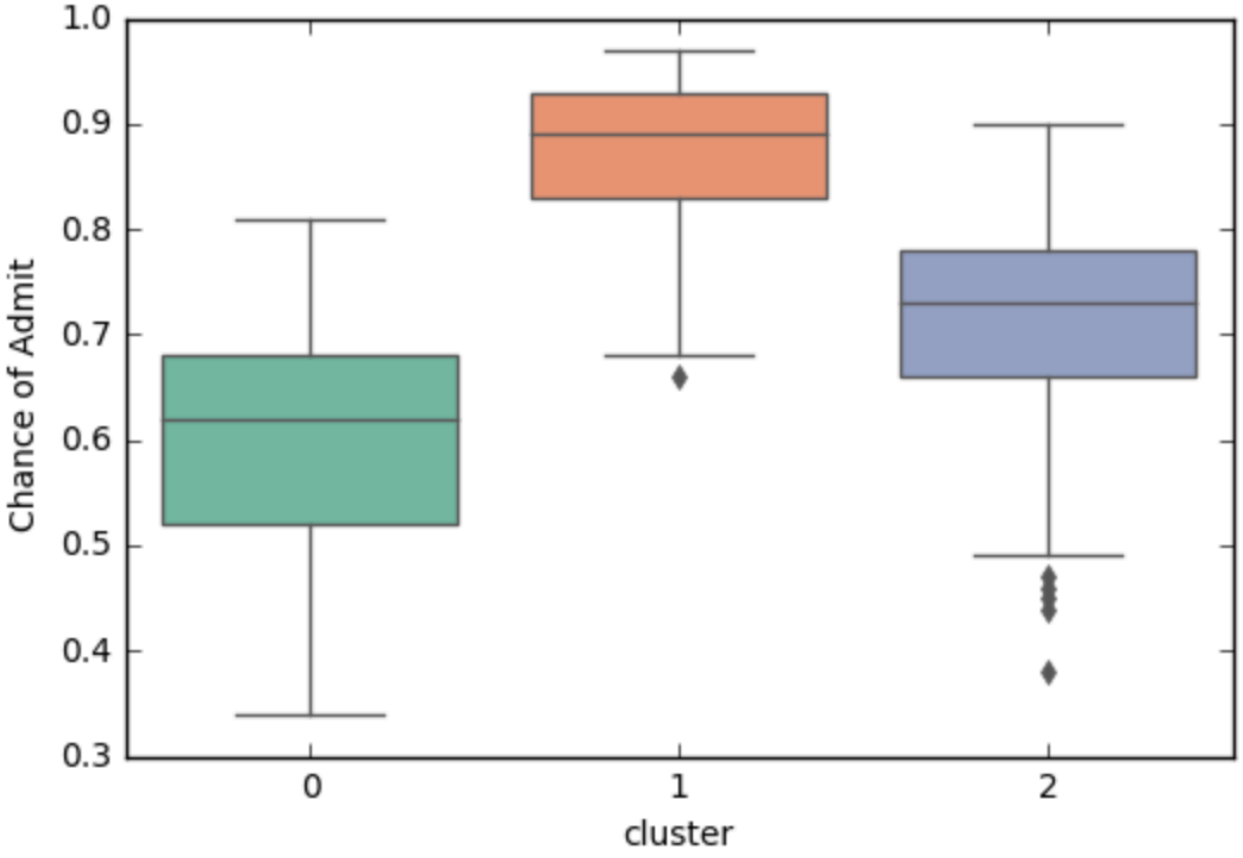



```
precision_score:  0.9565217391304348
recall_score:  0.7586206896551724
f1_score:  0.8461538461538461
```

分群

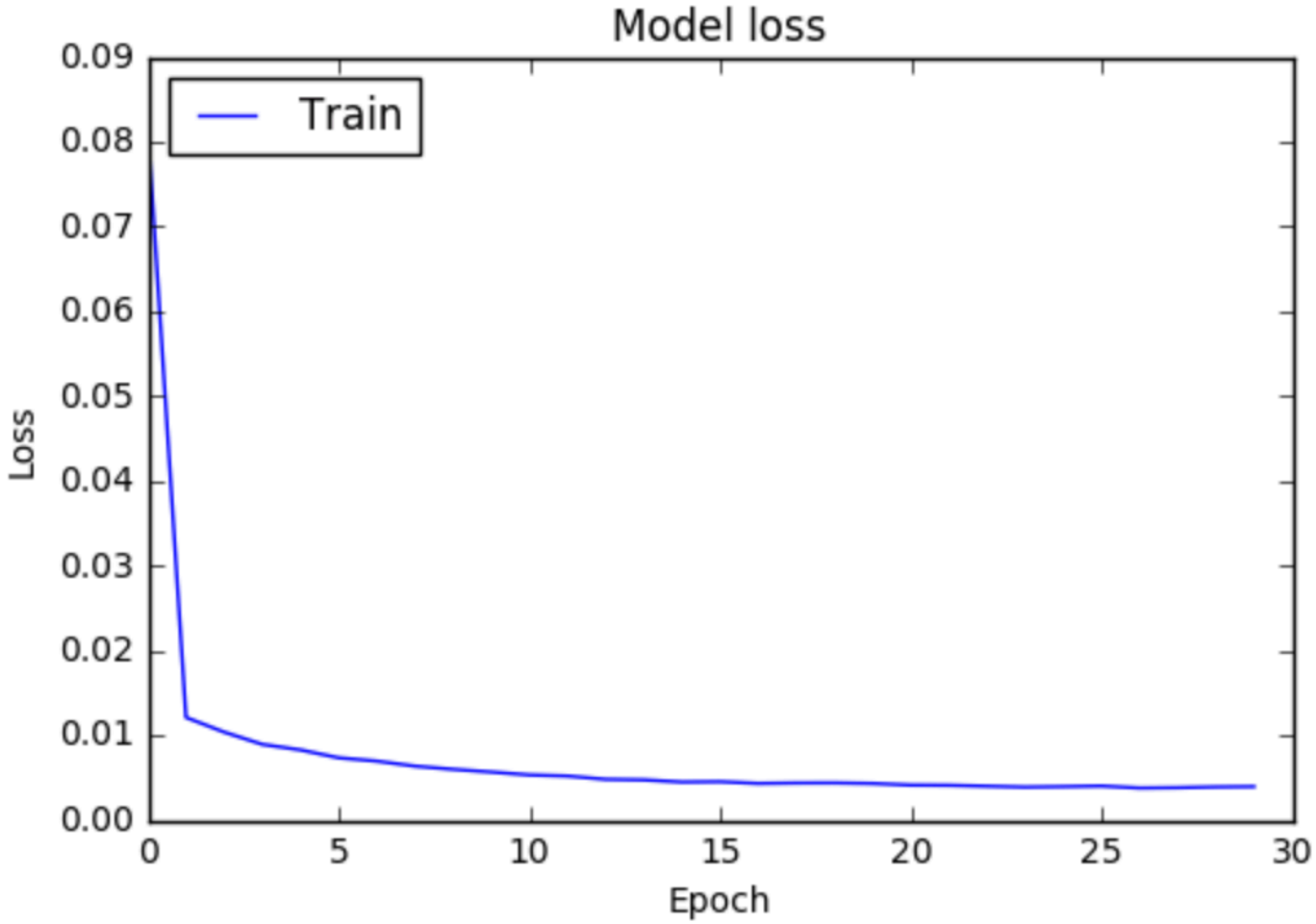
方法：K-mean

解釋每個群集特性



類神經網路

訓練過程
and
測試結果



套件介紹

- 讀取資料 numpy、pandas
- 機器學習 Scikit-learn
- 類神經網路 Kearas
- 資料視覺化 Matplotlib、Seaborn

一般流程(回歸、分類)

1. 讀取資料並把不相關的欄位去掉
2. 設定應變數(y)和自變數們(x)
3. 訓練集(test)和測試集(train)劃分
4. 挑選模型(回歸or分類)利用訓練集訓練
5. 拿測試集資料進行預測
6. 驗證(回歸：決定係數 or 分類：混淆矩陣)測試集資料預測結果

一般流程(分群)

1. 讀取資料並把不相關的欄位去掉
2. 設定多少個群集(clusters)
3. 資料訓練
4. 將分群結果畫出來解釋每個群集的特性

一般流程(類神經網路)

1. 讀取資料並把不相關的欄位去掉
2. 設定應變數(y)和自變數們(x)
3. 訓練集(test)和測試集(train)劃分
4. 設計類神經網路模型架構
5. 利用訓練集訓練模型
6. 拿測試集資料進行預測
7. 驗證(回歸：決定係數 or 分類：混淆矩陣)測試集資料預測結果