科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/6/4

Data Analytics – HW3-2

Predict the score of restaurants is high or low

1. 資料欄位 (data from UCI Machine Learning Repository)

train.csv (為 Google 旅遊評價，共 4909 筆 entries)

- User **[type: object]**

  用戶 ID

  Ex. “User 7”

- Category 1 ~ Category 24 **[type: float64、object]**

  皆是對某個地點的評分，評分範圍為 0 ~ 5 分，其中只有 Category 11 有出現 object type

  Ex. “3.63”

- Unnamed: 25 **[type: float64]**

  最莫名其妙的一個欄位，只有兩個值不是 NaN，在 Data Set 的網頁中並沒有講到這一個欄位是關於甚麼的，而模型中也不會用到這個欄位，因此不討論

test.csv (為 Google 旅遊評價，共 547 筆 entries)

- User **[type: object]**

  用戶 ID

  Ex. “User 4915”

- Category 1 ~ Category 24 **[type: float64]**

  皆是對某個地點的評分，評分範圍為 0 ~ 5 分

  Ex. “2.77”

- Unnamed: 25 **[type: float64]**

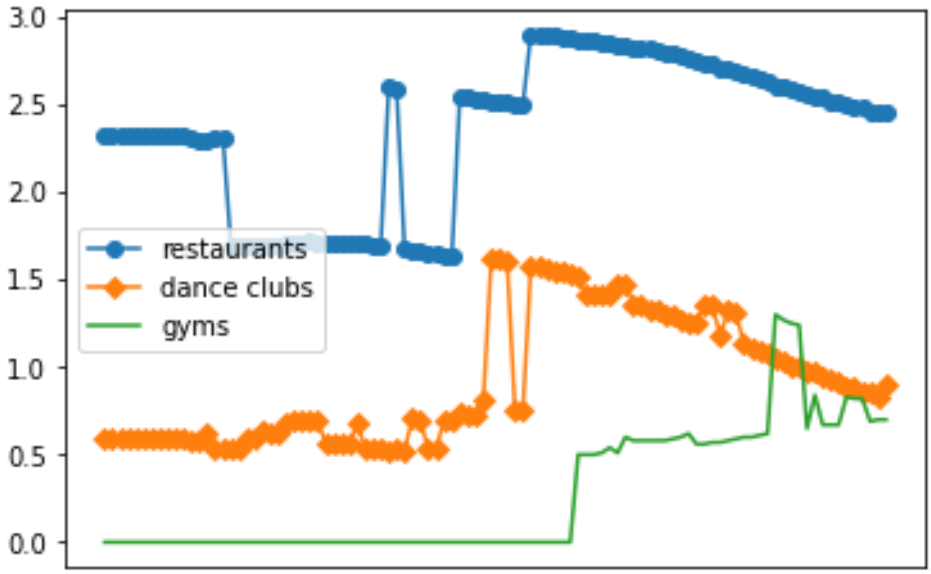
科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/6/4

最莫名其妙的一個欄位，皆為 NaN，在 Data Set 的網頁中並沒有講到這一個欄位是關於甚麼的，而模型中也不會用到這個欄位，因此不討論

2. 資料前處理

- **預覽資料並進行資料視覺化**
  - 先利用 head()、info()、describe() 來預覽資料，得知以下訊息:
    1. 資料欄位的型別
    2. 是否有缺失值
    3. 哪些欄位有可能當作特徵值
  - 將資料欄位重新命名
    - 由於資料欄位名稱僅以 Category 1 ~ 24 來區別，無法很直觀看出某個欄位的數值意義，因此改用所評分的地點來作為欄位名稱。
  - 接著利用 matplotlib 來將資料視覺化，此時可以更加深入了解資料欄位間的關聯性，並決定要選擇哪些欄位作為特徵值；從下方折線圖中可以發現三條線的走勢基本上相去不大 (忽略掉少數的 outliers)，因此可以作為特徵值餵給模型來找出彼此間的關聯性。

科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/6/4



- **生成驗證測試的答案**
  - **目標:**
    - 以餐廳評價高低作為驗證測試的答案 (**分類問題**)
  - **rs pred:** 以 2.5 為分界點，若餐廳評分大於 2.5，即為高 (1)；若小於等於 2.5，即為低 (0)。
- **加入特徵值**
  - 將舞蹈俱樂部 (dance clubs) 和健身房 (gyms) 作為特徵值。

3. 模型訓練與測試結果

- 三個模型先用 5-fold cross validation 對訓練資料進行測試及驗證，再將全部訓練資料丟進模型進行訓練，分別套用到以下 3 個模型:

1. Logistic Regression
2. NN – Multilayer Perceptron Classifier
3. K Nearest Neighbor Classifier

科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/6/4

- 其中 K Nearest Neighbor Classifier 在 5-fold cross validation 和丟入全部訓練資料後的表現都是最好的，準確度高的同時確保不會 overfitting。
- 套用至 Testing Set 時，也如預期是 K Nearest Neighbor Classifier 表現最好。

4. 結論

- 而之所以 K Nearest Neighbor Classifier 在這次的預測中表現較另外兩種亮眼，可能取決於:
  - 特徵值間的配合對預測結果有正相關
- K Nearest Neighbor Classifier 在進行分類問題時，需要注意 K 值的選擇:
  - K 值太大可能會讓準確度下降
  - K 值太小可能造成 overfitting
- 如何改良 Classifiers?
  - 試用並比較不同的 classifiers，選擇表現最好的
  - 選用不同的特徵值

    剛開始有嘗試將游泳池 (swimming pools) 作為其中一個特徵值，但在 5-fold cross validation 中的準確度只有 76%左右，但改用舞蹈俱樂部 (dance clubs) 後，準確度就能提升到 83%。

科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/6/4

- 嘗試改變超參數的值

  K Nearest Neighbor Classifier 在 n_neighbors 小於 3 的時候，會有 overfitting 的問題產生，而改用 default 的 5，準確度上升且誤差值變小。