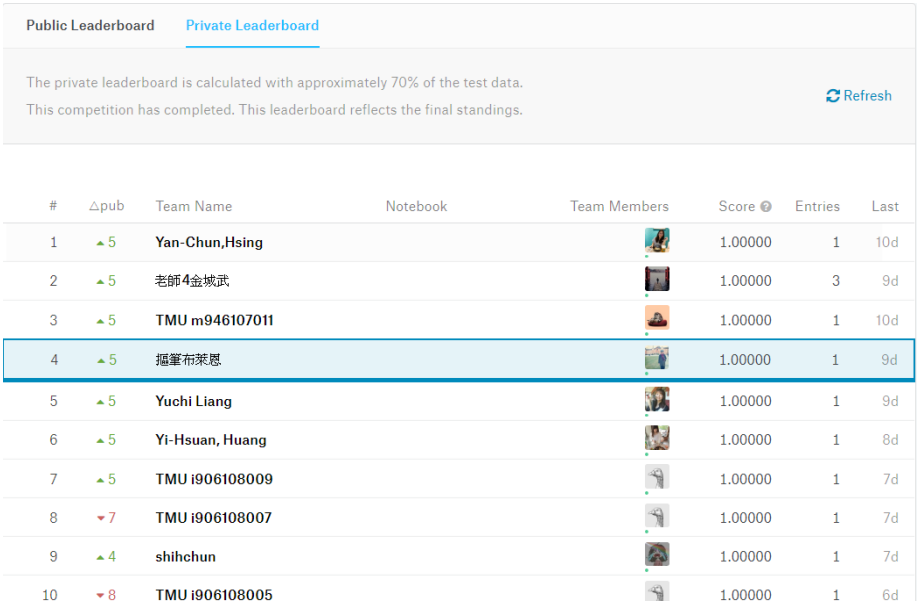
Kaggle Warm Up 課堂比賽

-- 使用鳶尾花資料(iris dataset)為模型訓練對象

created by 黃彥鈞 (Weber, YC Huang) 2019-10-28

---

1. 比賽結果：



Public Leaderboard | Private Leaderboard

The private leaderboard is calculated with approximately 70% of the test data.
This competition has completed. This leaderboard reflects the final standings.

Refresh

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲5	Yan-Chun,Hsing			1.00000	1	10d
2	▲5	老師4金城武			1.00000	3	9d
3	▲5	TMU m946107011			1.00000	1	10d
4	▲5	摳筆布萊恩			1.00000	1	9d
5	▲5	Yuchi Liang			1.00000	1	9d
6	▲5	Yi-Hsuan, Huang			1.00000	1	8d
7	▲5	TMU i906108009			1.00000	1	7d
8	▼7	TMU i906108007			1.00000	1	7d
9	▲4	shihchun			1.00000	1	7d
10	▼8	TMU i906108005			1.00000	1	6d

2. 模型策略：

- 主要應用簡單貝葉演算法，其中原因為此方法簡單直觀，適合應用例如鳶尾花這類特徵值清楚明確的訓練資料。不同於課堂上的練習，我使用 **GaussianNB** 演算法來訓練資料。三種 **Naive Bayes** 演算法差異主要為：
    - **Bernoulli Naive Bayes**：博努力簡單貝葉演算法主要用於處理二元**特徵資料(Binary feature)**，例如：0, 1。
    - **Multinomial Naive Bayes**：多項式簡單貝葉演算法主要用於**離散資料 (Discrete data)**，例如，電影評分。這些離散資料會有特定頻率計數。
    - **Gaussian Naive Bayes**：高斯簡單貝葉演算法主要用於**連續型資料 (Continuous data)**。
- 由於iris 訓練資料型態為**連續型資料**，如 sepal width, petal width, sepal length, petal length 中的特徵值。因此我選擇高斯簡單貝葉演算法來訓練模型。過程直接匯入訓練與測試資料，改變訓練集標籤為數值資料後，依照訓練資料 x_train, y_train 訓練出模型，並直接應用於測試資料求出測試集的結果標籤，沒有做 training_validation。
- 比賽期結果分數為：0.975，有私底下嘗試使用其餘算法如回歸(LR)與支援向量機(SVM)，做 training_validation，不過結果均無高斯簡單貝葉遠算法來的好，因此維持原上傳結果。

3. 參考資源：

What is the difference between the the Gaussian, Bernoulli, Multinomial and the regular Naive Bayes algorithms?