

機器學習作業 Lab01 報告

資訊工程學系 黃右萱(0416323)

題敘：

使用 Iris 資料庫來訓練 Classification Model，並以 K-fold Cross Validation、Resubstitution Validation 來進行驗證。

做法：

建造 Classification Decision Model，其中：

1. 使用預設 Gini Index 做為特徵篩選指標
2. 使用 Pre-Pruning，限定樹深不超過 3
3. 使用 5 個 Decision Tree 構成 Forest，以 voting 做為選擇依據

Library:

1. Scikit-Learn 建構選擇樹、挖掘資料庫資料
2. Numpy 做矩陣運算
3. Math 做數學運算

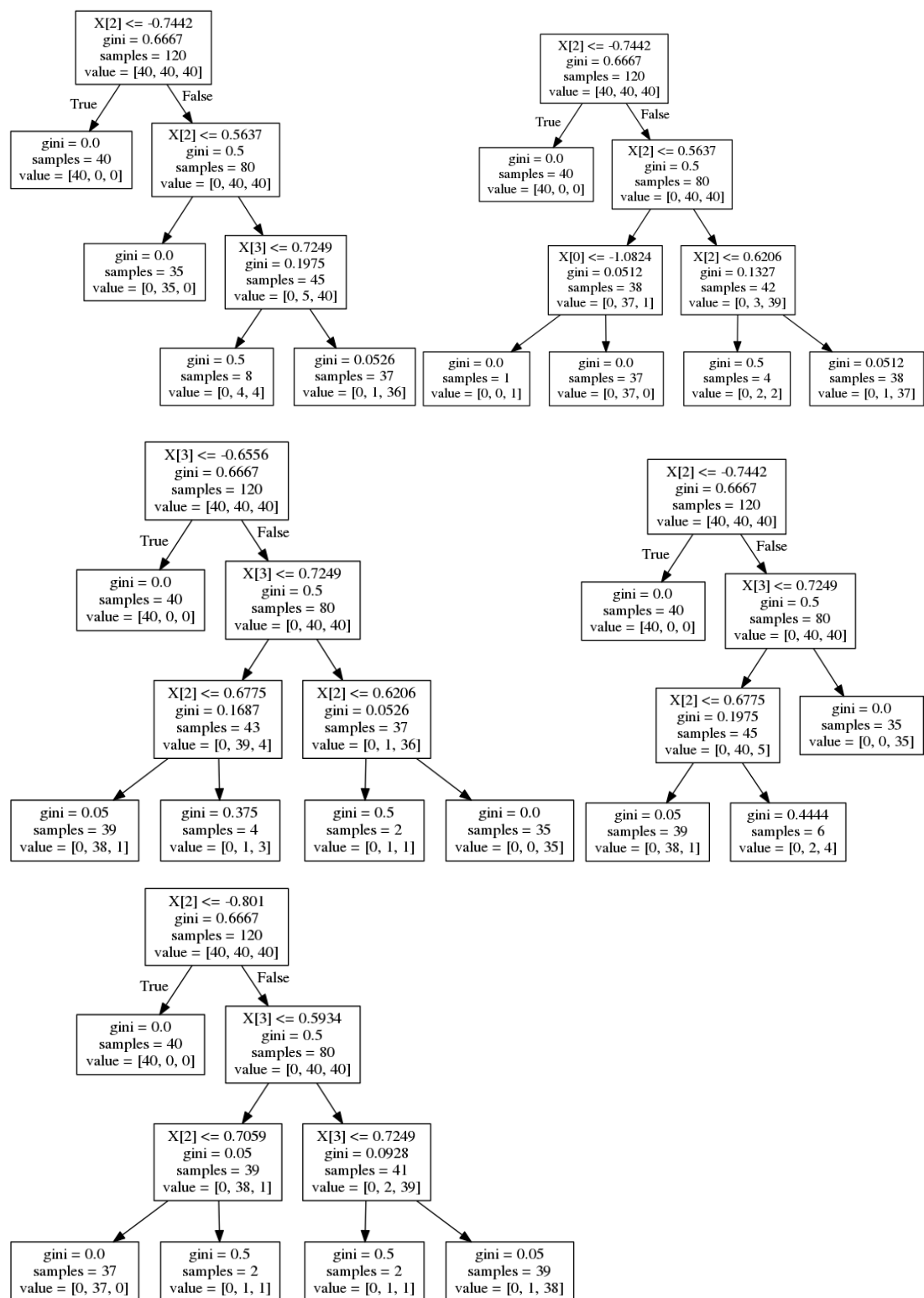
結果：

如圖，採用 Resubstitution Validation 準確率約為 97.3%；而採用 K-fold Validation 準確率約為 96%，訓練結果良好，沒有明顯的 overfit 產生，又從 Confusion Matrix 可見，主要錯誤應是來自 1, 2 兩項的誤判，而第 0 項鮮少誤判。

```
Resubstitution valid:  0.973333333333
Resubstitution Confusion Matrix:
[[50  0  0]
 [ 0 47  3]
 [ 0  1 49]]

k fold valid:  0.96
k fold confusion matrix:
[[50  0  0]
 [ 0 47  3]
 [ 0  3 47]]
```

全資料訓練模型圖：



備註：

1. 語言：Python 2.7.6
2. 環境：Ubuntu 14.04.5
3. K-fold: 採用 10-fold