Project 2

Classification

N96094196 張維峻

工程科學系

摘要 [0]

File structure[1]

```
+ – main
| + --dataset_creation.py
| + --decisiontree.py
| + --naive_bayes.py
| + --svm.py
+ – data
| + – train_unlabel.csv
| + – train_label.csv
| + – val_unlabel.csv
| + – val_label.csv
+ – Project2_Report_N96094196.pdf
```

- main/dataset_creation.py: 產生資料集
- main/decisiontree.py: 決策樹分類器
- main/naive_bayes.py: 樸素貝葉斯分類器
- main/svm.py: 支持向量機分類器
- data/train_unlabel.csv:訓練資料的輸入。
- data/train_label.csv:訓練資料的標籤。
- data/val_unlabel.csv:測試資料的輸入。
- data/val_label.cdv:測試資料的標籤。
- Project2_Report_N96094196.pdf:報告。

Data^2

採用職業球團選秀的情境做分類。定義 Attitude(態度)、Technology(技術)、Potential(潛能)、Age(年紀)、Physical_Fitness(體能)、Psych_Quality(內心強度)、Label(選秀結果)

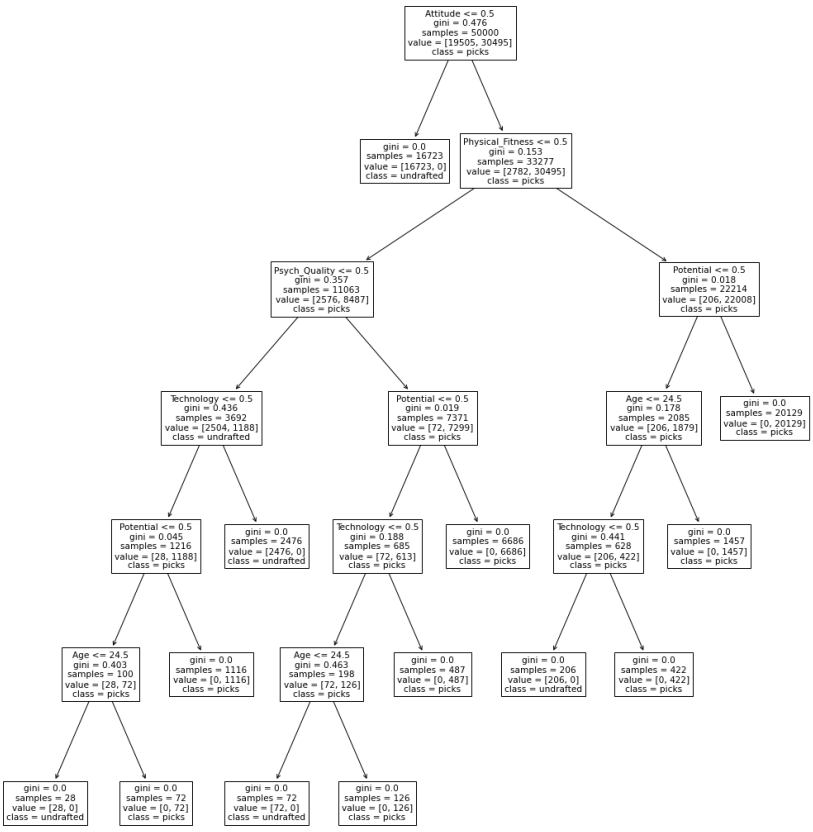
- Attitude:0~2 (差,普通,積極)
- Technology:0~2(差,平凡,教科書)
- Potential:0~10
- age:22~40
- Physical_Fitness:0~2(差,普通,好)
- Psych_Quality:0~2(玻璃,普通,鋼鐵)
- Label:-1、1(落選,選中)

Absolutely right 定義如下:

```
|--- Attitude <= 1
|   |--- class: -1
|--- Attitude >  1
|   |--- Technology < 1
|   |   |--- Potential < 5
|   |   |   |--- class: -1
|   |   |--- Potential >=  5
|   |   |   |--- Age <= 24
|   |   |   |   |--- class: 1
|   |   |   |--- Age >  24
|   |   |   |   |--- class: -1
|   |--- Technology >=  1
|   |   |--- Physical_Fitness < 1
|   |   |   |--- Psych_Quality <= 1
|   |   |   |   |--- class: -1
|   |   |   |--- Psych_Quality >  1
|   |   |   |   |--- class: 1
|   |   |--- Physical_Fitness >=  1
|   |   |   |--- class: 1
```

Decision tree[3]

訓練後,決策數結果如下




```
|--- Attitude <= 0.50
|   |--- class: -1
|--- Attitude >  0.50
|   |--- Physical_Fitness <= 0.50
|   |   |--- Psych_Quality <= 0.50
|   |   |   |--- Technology <= 0.50
|   |   |   |   |--- Potential <= 0.50
|   |   |   |   |   |--- Age <= 24.50
|   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |--- Age >  24.50
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- Potential >  0.50
|   |   |   |   |   |--- class: 1
|   |   |   |--- Technology >  0.50
|   |   |   |   |--- class: -1
|   |   |--- Psych_Quality >  0.50
|   |   |   |--- Potential <= 0.50
|   |   |   |   |--- Technology <= 0.50
|   |   |   |   |   |--- Age <= 24.50
|   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |--- Age >  24.50
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- Technology >  0.50
|   |   |   |   |   |--- class: 1
|   |   |   |--- Potential >  0.50
|   |   |   |   |--- class: 1
|   |--- Physical_Fitness >  0.50
|   |   |--- Potential <= 0.50
|   |   |   |--- Age <= 24.50
|   |   |   |   |--- Technology <= 0.50
|   |   |   |   |   |--- class: -1
|   |   |   |   |--- Technology >  0.50
|   |   |   |   |   |--- class: 1
|   |   |   |--- Age >  24.50
|   |   |   |   |--- class: 1
|   |   |--- Potential >  0.50
|   |   |   |--- class: 1
```

Accuracy Score: 100.0 %

Comparison:

經比較 absolutely right 與學習出來的 Tree，發現只有 Root 的規則一致，再往後的分支就會與原本設定的 absolutely right 分支順序有些差異，甚至多出幾個 absolutely right 沒出現的判斷分支。

精準度的部份，發現 Decision Tree,雖然分支有所不同,但可能因為數據的複雜度較低,仍然可以到 100%的精度。

Naive_Bayes[4]

使用了 Naive_Bayes 對照，採用 scikit-learn 所提供之

Naive_Bayes 進行訓練資料，訓練出來的精準度如下:

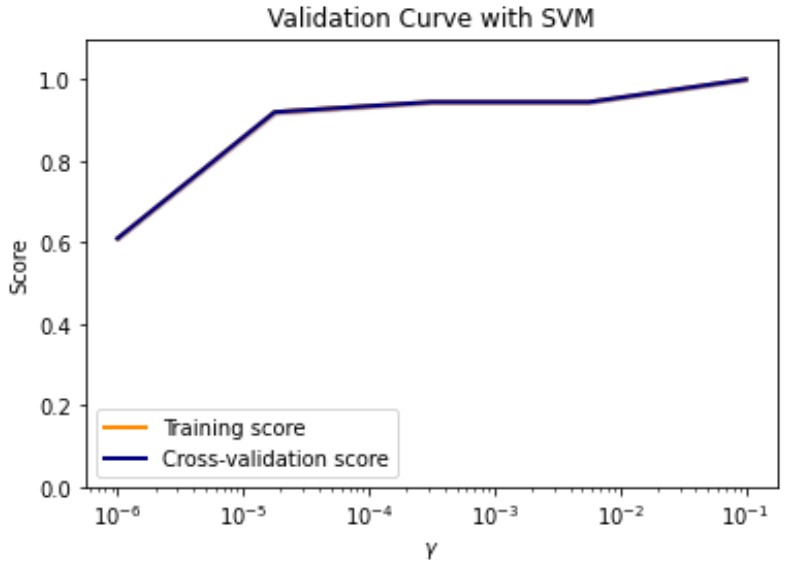
clf = CategoricalNB( )	Accuracy Score: 95.0 %

SVM[5]

使用了 SVM 來作對照，採用 scikit-learn 所提供之 SVC 進行訓練

資料，訓練出來的精準度如下:

kernel='rbf' ,C=1, gamma=0.01	Accuracy Score: 97.6 %
使用 Validation Curve 調整 SVM 參數	
kernel='rbf' ,C=100, gamma=0.1	Accuracy Score: 100.0 %



Summary[6]

資料集是由一連串的 if-else 所產生,每一次都只進行單一屬性判斷，這種型式和 Decision Tree 較為相近，因此 Decision Tree 較容易訓練出好結果。]

Naive_Bayes,由於模型的特性,迅速,有不錯的準度,常用來作為評估一個資料集,第一個使用得模型。

SVM 會把原本資料投射到高維度空間，再進行分類，在其中就會有一些不同情況，使得雖然在訓練時精準度較難提升，但由於資料集的規則太簡單,精準度仍可達到100%,若是遇到複雜度較高的資料集,訓練精度應會下降。

當想訓練一個未知類別的資料進行分類，應該要嘗試各種不同的模型綜合考量後，再做出模型選擇的決定。在這次實驗中 Decision Tree 有好的結果，而且其分支出來的判斷又能比較貼近人類的理解，是一種還不錯的模型選擇考量。但 SVM 經過調參之後也可以得到 100%的準度。因此殊途同歸,沒有最好的模型,只有最適合的。