Machine Learning Program assignment #1

0416303 楊博凱

Implementation :

這次作業一開始，我先 import 一些之後會用到的 module 到程式中；再來的爬蟲我運用 request.get() 來將網頁中的資料抓下來，並經由兩次的 split() 與 numpy.array() 將資料改成 numpy.array 的格式。在將資料的 target 與 attributes 分開之後，將數字部分從 string 轉成 float。

在 for 迴圈中，由於我採用的是 10-fold cross validation，所以讓 for 迴圈執行 10 次，每次都利用 sklearn 的 KFold module 中的函式將資料 random 分成 training data 與 testing data，並建出含有 5 棵樹的 forest，設定 max depth 為 3。將每次 for 迴圈跑出的結果總和取平均後得到最後的 Confusion Matrix、Normalized Confusion Matrix 與 Accurancy Score。

Results :

Resubstitution 的準確度平均為 96.7%，而 K Fold Cross Validation 的準確度則是 100%

```
:w !python
**************** Resubstitution Result ****************
Confusion Matrix:
[[ 50.    0.    0. ]
 [  0.   48.4   1.6]
 [  0.    3.3  46.7]]
Normalize Confusion Matrix:
[[ 1.     0.     0.   ]
 [ 0.     0.968  0.032]
 [ 0.     0.066  0.934]]
Accuracy Score:
0.967333333333
***************** 10-Fold Result ****************
Confusion Matrix:
[[ 5.  0.  0.]
 [ 0.  5.  0.]
 [ 0.  0.  5.]]
Normalize Confusion Matrix:
[[ 1.  0.  0.]
 [ 0.  1.  0.]
 [ 0.  0.  1.]]
Accuracy Score:
1.0
```

Using library :

request、numpy、sklearn、os、pydotplus

```
import requests
import numpy as np
from sklearn import tree
import os
import pydotplus
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import KFold
from sklearn.metrics import accuracy_score
```

Environment & Language :

Language : python 2.7

Environment :

```
kai@kai-UX305FA:~$ sudo lsb_release -a
[sudo] password for kai:
No LSB modules are available.
Distributor ID: Ubuntu
Description:    Ubuntu 16.04.2 LTS
Release:        16.04
Codename:       xenial
```

Forest for the first for loop :