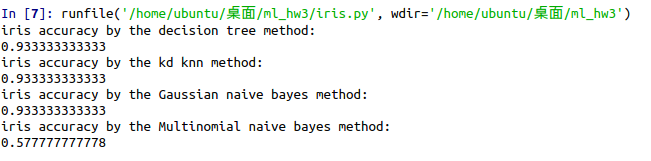
機器學習 作業三報告

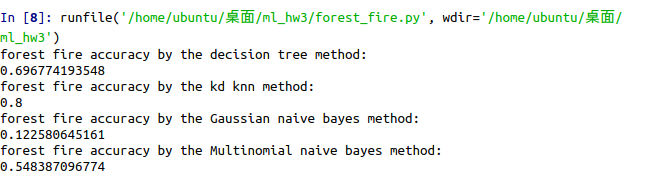
0416094 黃兆宇

一、執行結果 :

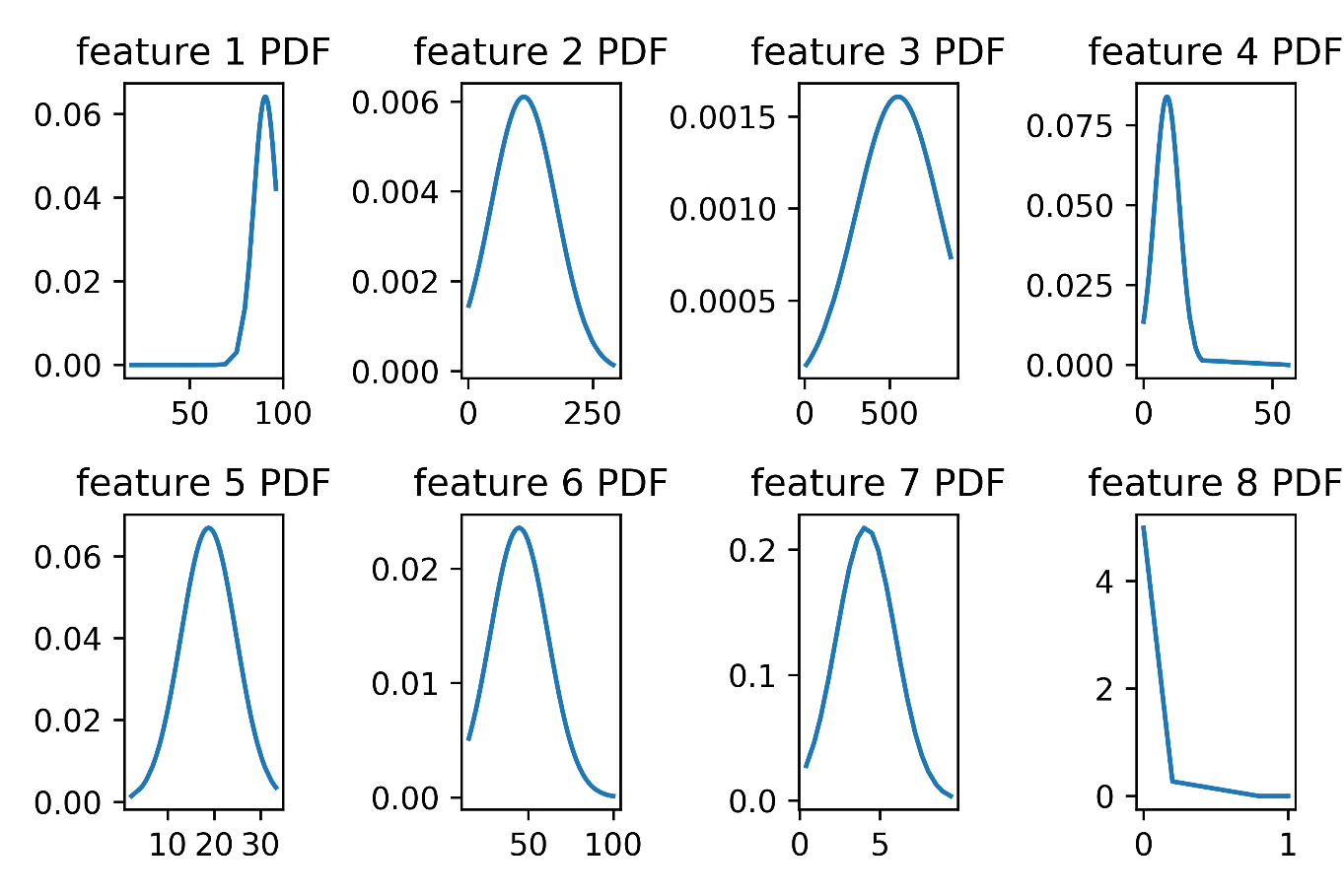
Iris.data :



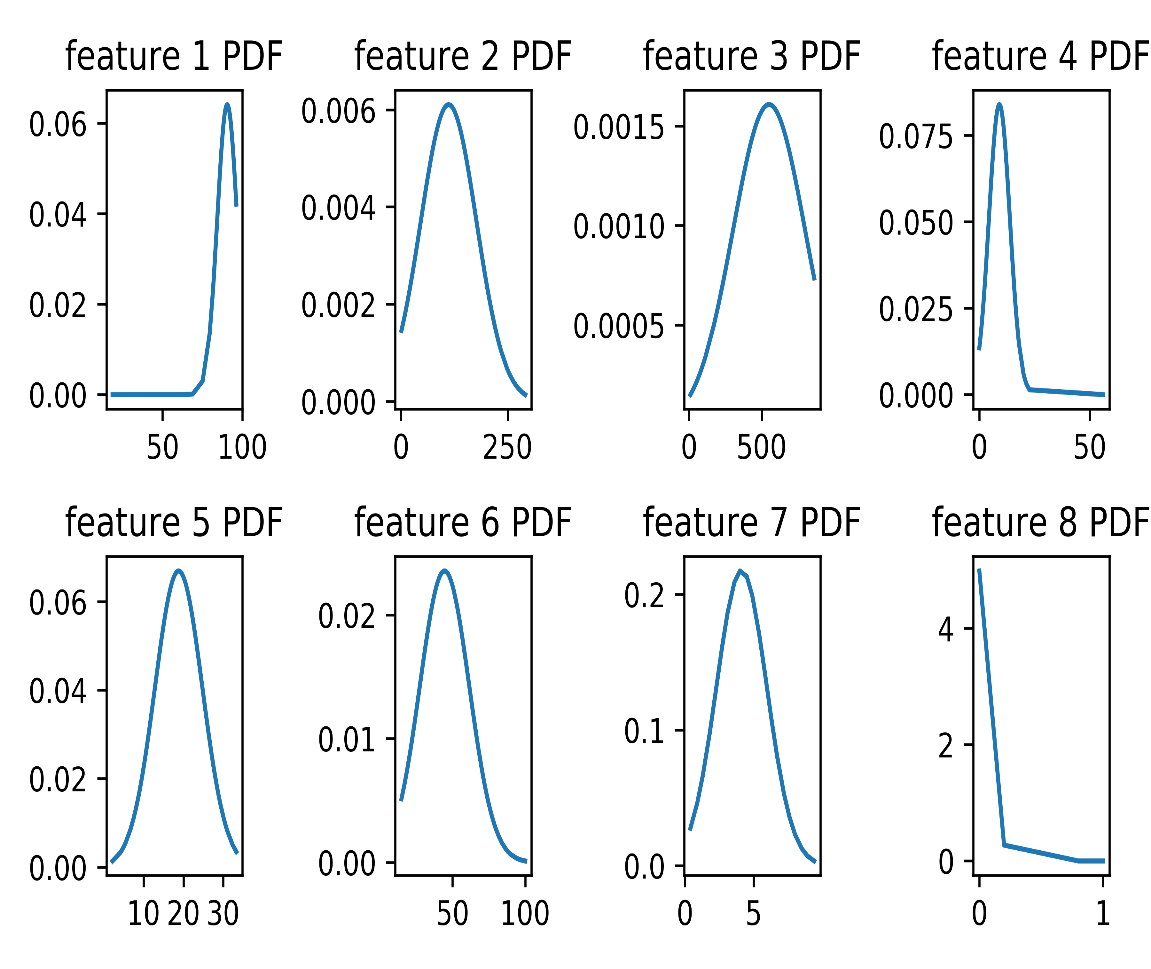
forest\_fire.csv :



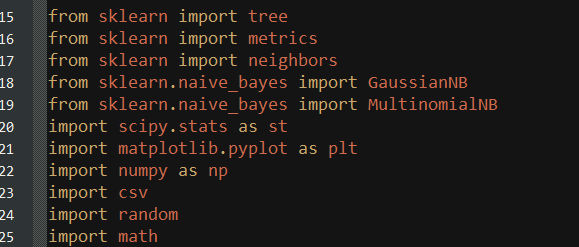
Iris.data 之PDF



forest\_fire.csv之 PDF



二、使用的library :



功能 :

Sklearn tree : decision tree

Sklearn metrics : accuracy

Sklearn neighbors : kd tree

Sklearn.naive\_bayes GaussianNB : Gaussian naive bayes

Sklearn.naive\_bayes MultinomialNB : multinomial naive bayes

Scipy.states matplotlib.pyplot : 畫出pdf

Numpy : 使用numpy array、mean、std

Csv : 讀取csv檔

Random: 將資料進行random shuffle 的預處理

Math: log 運算

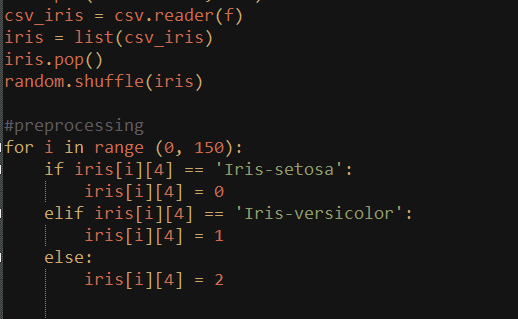
三、使用語言 :

Python 3.6.2

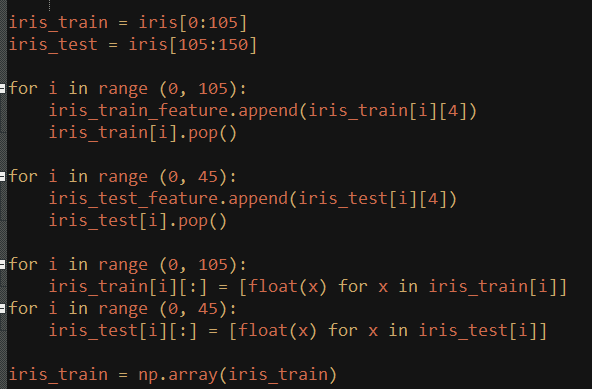
四、程式說明 :

1.預處理 :

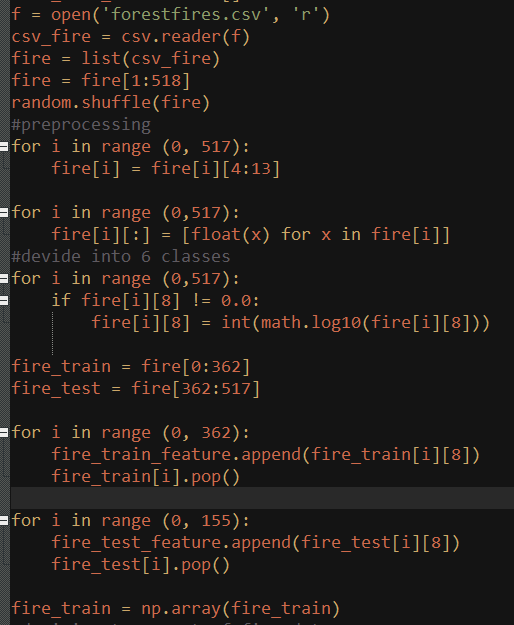
Iris.py : 讀入iris.data 的資料，為了配合package的輸入需求，將iris-setosa 改為0、iris-versicolor改為1、iris-verginica改為2



將資料分成train set 和 test set ，在將target和feature分開，在將其轉成np array，並轉成float的型態



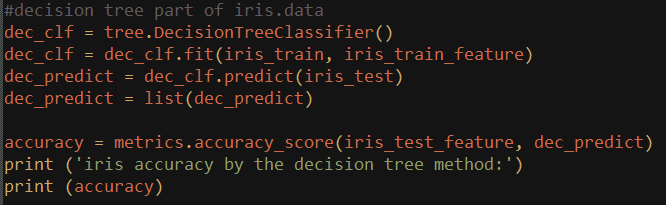
forest\_fire.py: 讀入檔案後，將前面四個feature去除去除X Y的原因主要因為位置和燃燒面積應受其他因素影響，如濕度、溫度、風速等等，某一地區的燃燒面積只需分析這些因素即可，不用分析位置，而月份和星期因為是循環性的資料，無法轉換成數字區間，不力分析，故也將其去除。最後data分成train set和test set。



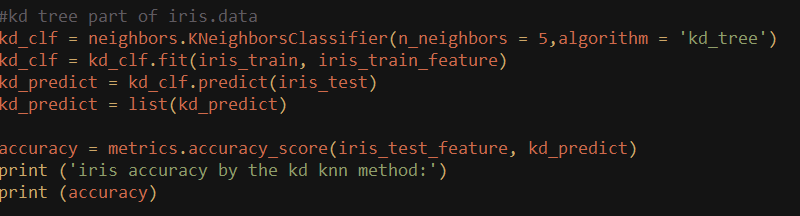
2.module 估算 :

Module估算在iris和forest fire兩筆資料的處理流程上大致相同，只需傳入不同的train set和test set，故以下僅以iris說明其流程。

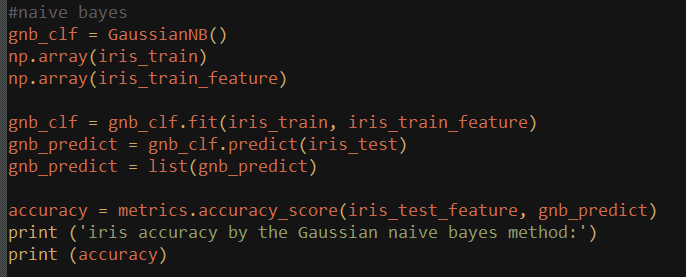
將資料傳入decision tree，並得出accuracy



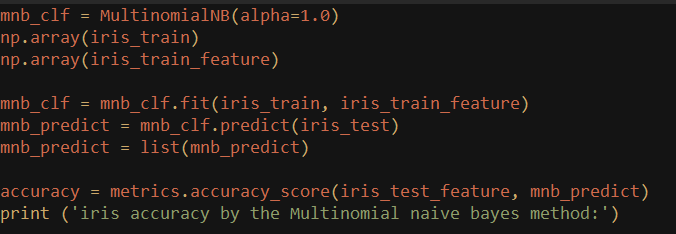
設定n = 5，algorithm = ‘kd\_tree’，用n = 5的knn去估計accuracy



用gaussian naive bayes去計算accuracy



用multinomial naive bayes，並用laplace smooth的方式去估算accuracy



五、結果分析:

Iris.data:

在使用四種方式分析後(decision tree、kd tree、gussianNB、multinomialNB)，發現除了multinomialNB的資料以外其餘的分析皆高達0.9以上，multinomialNB會比較不準的原因是因為由PDF可知，iris的結果分布接近常態(鐘形)分佈，用multinomial可能會失準，且比起decision tree和kd tree，guassianNB預估出的準確度波動幅度稍大，可能和選出的trainning set是否符合guassian分布有關，且多次試驗後，kd tree得結果會優於decision tree。

結論: kd tree > decision tree > guassian NB > multinomialNB

Forest fire:

此資料的分布較iris分布來的不均勻，即使分區間處理後，仍有大量為0的資料，非常態分布，故用guassian naive bayes的準確率較低。而資料也不適合用multinomial naive bayes，naive bayes較適合用來預估會不斷重複出現的資料次數(如文本分析)。比較多次decision tree和kd tree的結果後，kd tree的結果(0.7左右)，高於decision tree(0.6左右)，故以kd tree 結果最為優異。

結論: kd tree >= decision tree > guassian NB、multinomial NB