Data Mining (Homework 3)

Due date : 2018/11/05

1. 資料集如附檔，label是income (<=50K或>50K兩種)

2. 作業要求：

a.資料前處理 (可參考HW1)

b.請使用Gradient Boosting進行分類

c.請寫自行撰寫function進行k-fold cross-validation計算Accuracy，不可使用

套件

b-1. input(k, data)，將data切成k份，其中1份當測試集，剩餘k-1份當訓

練集建立模型

b-2. 輪流將k份的每份資料都當 測試集，其餘當訓練集建立模型，因此會

進行k次，k次都計算出Accuracy

b-3. 將k次的Accuracy平均即為output

<參考範例>

def K\_fold\_CV(k, data):

#設定subset size 即data長度/k

#設定Accuracy初始值

for i in range(k):

#設定testing set與training set的資料起始點與結束點

#例如資料有100筆，testing set在本次iteration取第1到10筆，

則training set為第11到100筆；下次testing set為11~20，training set為21~100 & 1~10

#利用training set建立模型，testing set計算出Accuracy累加

return Accuracy/k

3. 請計算k=10的Accuracy，並上傳程式碼與報告