資料探勘HW3 科管所一 0753524 邱薇如

1. **以income作為label / 2-a. 資料前處理(dummy, normalize)**

|  |
| --- |
| data = pd.read\_csv('D://data.csv')  data = pd.get\_dummies(data,  columns= ['workclass','education', 'marital\_status','occupation',  'relationship','race','sex','native\_country'])  for i in ['age','fnlwgt','education\_num','capital\_gain', 'capital\_loss',  'hours\_per\_week']:  temp = [data[i].values]  data[i] = preprocessing.normalize(temp)[0]  data['income'] = data['income'].replace({' <=50K':0," >50K":1}) |
|  |

**2-b. 請使用Gradient Boosting進行分類 / 2-c. 請寫自行撰寫function進行k-fold cross-validation(不可使用套件)並計算Accuracy**

|  |
| --- |
| def K\_fold\_CV(k, data):  size = data.shape[0]//k  acc=[]  for i in range(k):  test\_set = data[i\*size:(i+1)\*size]  train\_set = pd.concat([data[0:i\*size],data[(i+1)\*size:]],ignore\_index=True)  X\_train = train\_set.drop(['income'],axis=1)  Y\_train = train\_set['income']  X\_test = test\_set.drop(['income'],axis=1)  Y\_test = test\_set['income']  GDBT = GradientBoostingClassifier()  GDBT.fit(X\_train, Y\_train)  acc.append(GDBT.score(X\_test,Y\_test))  print(acc)  return np.mean(acc) |
|  |

1. **請計算k=10的Accuracy，並上傳程式碼與報告**

|  |
| --- |
| print ('Mean accuracy of 10-fold CrossValidation：',K\_fold\_CV(10, data)) |
| Mean accuracy of 10-fold CrossValidation： 0.8659090909090909 |