ECT_HW1

2019

第一大題

第一大題

用 Weka 軟體對 diabetes.arff 利用 Naïve Bayes 進行Supervised learning, 選擇 “Use training set” , 設定 Attribute: class 為 Output , 在過程中對重要步驟截圖加以說明, 並回答以下問題:

第一大題(a)-題目

(a) 解釋Classifier Output，Test data 的錯誤率是多少 ?有多少百分比的 Test dataset instances 被分類到 tested_negative class 但實際上屬於 tested_positive class ?請利用Confusion matrix 解釋。(15%)

第一大題(a)-解答

```
=== Summary ===

Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4674
Mean absolute error                      0.2811
Root mean squared error                  0.4133
Relative absolute error                 61.8486 %
Root relative squared error             86.7082 %
Total Number of Instances              768

=== Confusion Matrix ===

   a   b   <-- classified as
 421  79 |   a = tested_negative
 103 165 |   b = tested_positive
```

- 錯誤率=23.6979%
- 被分類到Tested_negative，但實際為Tested_positive：103個
- 103/768 = 0.1341

第一大題(b)-題目

(b) 在 Output predictions 的結果中，欄位 error 出現 “+” 代表意思為何 ?請截圖並解釋之。(10%)

第一大題(b)-解答

318	1:tested_negative	2:tested_positive	+	0.775
319	1:tested_negative	1:tested_negative		0.508
320	1:tested_negative	1:tested_negative		0.729
321	2:tested_positive	2:tested_positive		1
322	2:tested_positive	2:tested_positive		0.661
323	1:tested_negative	1:tested_negative		0.975
324	1:tested_negative	2:tested_positive	+	0.716
325	2:tested_positive	1:tested_negative	+	0.936
326	2:tested_positive	2:tested_positive		0.872
327	1:tested_negative	1:tested_negative		0.954
328	1:tested_negative	1:tested_negative		0.958
329	1:tested_negative	1:tested_negative		0.896
330	1:tested_negative	2:tested_positive	+	0.615

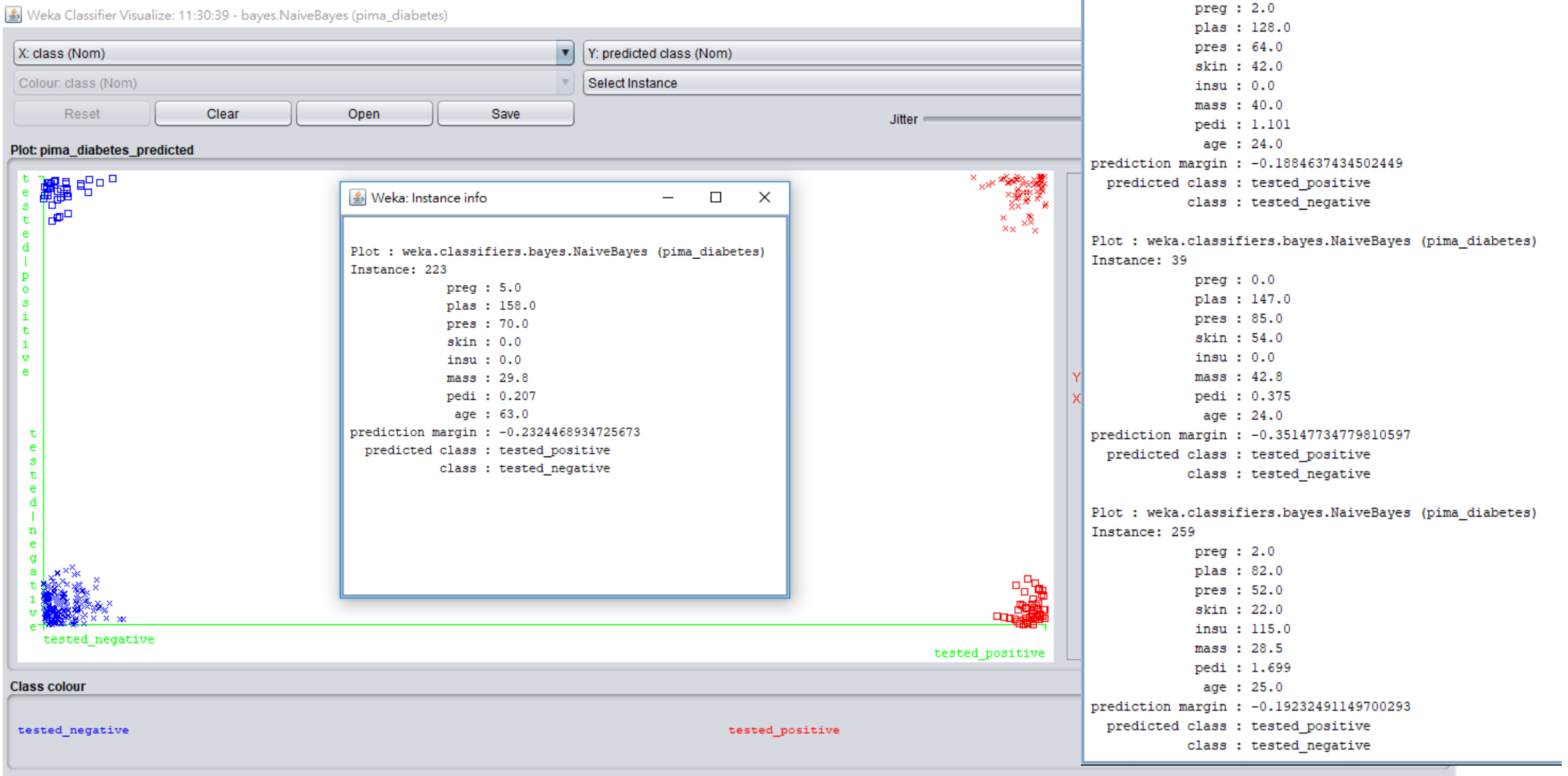
- 欄位Error出現“+”代表預測錯誤

第一大題(c)-題目

(c) 請利用 Visualize Classifier Errors，找出預測錯誤的資料點3個，並寫出各是第幾筆資料，請截圖操作步驟並解釋。(15%)

第一大題(c)-解答

為預測錯誤資料



第一大題(d)-題目

(d) 請使用 Visualize Classifier Errors，解釋產生的圖以及此圖與 Confusion matrix 之間的關係。 (10%)

第一大題(d)-解答


```
=== Confusion Matrix ===

   a   b   <-- classified as
 421  79 |   a = tested_negative
 103 165 |   b = tested_positive
```

- 左圖Visualize Classifier Errors，右圖為Confusion Matrix
  - 左圖的左下角為右圖的左上角，左圖的左上角為右圖的右上角
  - 兩者以不同方式表達相同概念

第二大題

第二大題

用python對 diabetes.csv 進行Supervised learning中的Naïve Bayes分析 ,並回答以下問題:

第二大題(a)-題目

(a)請問diabetes各屬性

('preg','plas','pres','skin','insu','mass','pedi','age')的平均值各為何? (5%)

第二大題(a)-解答

	preg	plas	pres	skin	insu	mass	pedi	age
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

preg=3.85, plas=120.89, pres=69.11, skin=20.54,
insu=79.80, mass=31.99, pedi= 0.47, age=33.24

第二大題(b)-題目

(b)在過程中對所有重要程式步驟進行截圖並加以說明，越詳盡越好。(10%)

第二大題(b)-解答

- **請同學自由作答**

第二大題(c)-題目

(c)請利用metrics.classification_report( )呈現出最後precision. recall. f1-score值，並截圖加以說明。(10%)

第二大題(c)-解答

	precision	recall	f1-score	support
0	0.80	0.84	0.82	500
1	0.68	0.62	0.64	268
micro avg	0.76	0.76	0.76	768
macro avg	0.74	0.73	0.73	768
weighted avg	0.76	0.76	0.76	768

- tested_negative

→presision=0.8, recall=0.84, f1-score=0.82

- tested_positive

→presision=0.68, recall=0.62, f1-score=0.64

第二大題(d)-題目

(d)請利用metrics.confusion_matrix ( )呈現出混淆矩陣，並截圖加以說明。(10%)

第二大題(d)-解答

```
[[421  79]
 [103 165]]
```

- 被分類到Tested_negative，但實際為Tested_positive：103個
- 被分類到Tested_positive ，但實際為Tested_positive：165個
- 被分類到Tested_negative，但實際為Tested_negative ：421個
- 被分類到Tested_positive ，但實際為Tested_negative ：79個

第二大題(e)-題目

(e)請問當'preg'=2, 'plas'=1, 'pres'=0, 'skin'=0, 'insu'=2, 'mass'=1, 'pedi=2', 'age'=20時，最終的output class為何。 (5%)

第二大題(e)-解答

```
predicted= model.predict([[2,1,0,0,2,1,2,20]])
print ("Predicted Value:", predicted)
```

```
Predicted Value: [1]
```

- 1 → tested_positive

第二大題(f)-題目

(f)請比較weka和python分析之結果，並加以說明。(10%)

第二大題(f)-解答

weka

```
=== Confusion Matrix ===

   a   b   <-- classified as
 421  79 |   a = tested_negative
 103 165 |   b = tested_positive
```

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.842	0.384	0.803	0.842	0.822	0.469	0.825	0.902	tested_negative
0.616	0.158	0.676	0.616	0.645	0.469	0.825	0.684	tested_positive
0.763	0.305	0.759	0.763	0.760	0.469	0.825	0.826	

2者結果一致!

python

	precision	recall	f1-score	support
0	0.80	0.84	0.82	500
1	0.68	0.62	0.64	268
micro avg	0.76	0.76	0.76	768
macro avg	0.74	0.73	0.73	768
weighted avg	0.76	0.76	0.76	768

```
[[421  79]
 [103 165]]
```