ECT_HW1

2020

第一大題

第一大題

用Weka軟體對mushrooms.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的錯誤率為多少？有多少Test dataset instances被分類到有毒的(poisonous)但實際上屬於可食用的(edible)?請利用Confusion matrix解釋。(25%)

第一大題(a)-解答

```
=== Summary ===

Correctly Classified Instances        7984               98.2767 %
Incorrectly Classified Instances       140                1.7233 %
Kappa statistic                          0.9655
Mean absolute error                      0.0222
Root mean squared error                  0.1209
Relative absolute error                  4.439  %
Root relative squared error             24.2041 %
Total Number of Instances             8124
```

```
=== Confusion Matrix ===

    a    b   <-- classified as
 3822   94 |    a = p
   46 4162 |    b = e
```

- 錯誤率=1.7233%
- 被分類為有毒卻但實際為可食用的：46個

第一大題(b)-題目

(b)在Output predictions結果中 “+”代表的意義為何，請截圖並解釋。(10%)

第一大題(b)-解答

5274	2:e	1:p	+	0.802
5275	1:p	1:p		1
5276	1:p	1:p		0.996
5277	1:p	1:p		1
5278	1:p	1:p		0.952
5279	1:p	2:e	+	1
5280	1:p	1:p		1
5281	2:e	2:e		1
5282	1:p	2:e	+	0.998
5283	1:p	1:p		1
5284	2:e	2:e		1
5285	2:e	2:e		1
5286	2:e	2:e		1

- +代表分類錯誤
- 1：Actual class, 2：Predicted class

第一大題(c)-題目

(c)請使用Visualize Classifier Errors, 解釋此圖與Confusion matrix之間的關係。(10%)

第一大題(c)-解答


```
=== Confusion Matrix ===

    a    b   <-- classified as
3820   96 |    a = p
  54 4154 |    b = e
```

- 左圖Visualize Classifier Errors，右圖為Confusion Matrix
- 兩者以不同方式表達相同概念

第二大題

第二大題

用python對mushrooms.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)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。(15%)

第二大題(a)-解答

請同學自由作答

第二大題(b)-題目

(b)請問mushrooms資料集中共有多少instance?是否包含空欄位(null)?(10%)

第二大題(b)-解答

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8124 entries, 0 to 8123
Data columns (total 12 columns):
type                      8124 non-null object
cap_shape                 8124 non-null object
cap_surface               8124 non-null object
cap_color                 8124 non-null object
odor                      8124 non-null object
stalk_shape               8124 non-null object
stalk_color_above_ring    8124 non-null object
stalk_color_below_ring    8124 non-null object
ring_number               8124 non-null object
ring_type                 8124 non-null object
population                8124 non-null object
habitat                   8124 non-null object
dtypes: object(12)
memory usage: 761.8+ KB
```

- 共有8124筆資料
- 無空值

第二大題(c)-題目

(c)請問欄位“stalk_color_above_ring”有幾種不同的value?(5%)

第二大題(c)-解答

	type	cap_shape	cap_surface	cap_color	odor	stalk_shape	stalk_color_above_ring
count	8124	8124	8124	8124	8124	8124	8124
unique	2	6	4	10	9	2	9
top	e	x	y	n	n	t	w
freq	4208	3656	3244	2284	3528	4608	4464

- 9種

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

```
     e       p
[[3296   912]  e
 [ 493  3423]] p
```

第二大題(e)-題目

(e)請利用metrics.classification_report()列出模型的準確率,並與Weka的結果比較何者較高?(10%)

第二大題(e)-解答

	precision	recall	f1-score	support
0	0.87	0.78	0.82	4208
1	0.79	0.87	0.83	3916
accuracy			0.83	8124
macro avg	0.83	0.83	0.83	8124
weighted avg	0.83	0.83	0.83	8124

- 準確率83%
- 較Weka低