Data Mining

Project2

P76071349

王鈺云

2018/11/20

● Introduction

此次作業希望我們學習分類器的原理，尤其是 Decision Tree Classifier，以及利用自己設計的 absolutely right’ rules 來產生資料，並實做 Decision Tree Classifier，在這過程中，觀察資料是否有依照自己定下的 rules 來分類。

● Dataset

此次作業要由自己生成資料，我選擇了”判斷是否為吸血鬼”的問題，吸血鬼的典型形象特徵為牙齒尖長，皮膚蒼白，眼睛發紅，擁有長生不老的身軀，白天時很想睡覺，晚上卻睡不著，一天睡眠時間極少，且非常害怕照射到陽光，通常生活在陰暗的環境下，另外，他們也懼怕十字架與大蒜的味道，食物來源通常是新鮮的人血。

這次設計了五種 features 來生成資料集，分別是以下五種 featrures 以及可能出現的數值代表意義：

1. 食物來源是否為人血
    a. 0：否
    b. 1：是
2. 可以忍受照射陽光幾分鐘
    a. 0：不能忍受
    b. 1：1 分鐘
    c. 2~
3. 是否害怕大蒜的味道
    a. 0：否
    b. 1：是
4. 皮膚顏色
    a. 0：蒼白
    b. 1：普通白
    c. 2：中等

d. 3：偏黑

e. 4：黝黑

5. 一天睡多久

a. 0：1 小時內

b. 1：1 小時

c. 2~

我設定了以下三個 absolutely right’ rules：

1. 吸血鬼只能忍受陽光照射 4 分鐘以下

2. 吸血鬼的皮膚顏色為蒼白或是普通白

3. 吸血鬼一天睡眠時間為 2 小時以下

以這三個 ‘absolutely right’ rules 來生成 positive 和 negative data，第一種資料集的資料總數為 30 筆，其中的 25 筆為 training data，有 12 筆 positive data，13 筆 negative data，剩下的 5 筆資料為 testing data。

	X[0]	X[1]	X[2]	X[3]	X[4]	
	human_ blood	sunshine	afraid_ garlic	skin	sleep	vampire
1	1	0	1	0	0	1
2	0	0	1	1	0	1
3	1	1	0	1	0	1
4	0	4	1	1	1	1
5	1	1	0	0	1	1
6	1	2	0	1	1	1
7	1	2	1	0	2	1

8	1	3	0	1	2	1
9	0	4	1	0	2	1
10	0	6	0	2	0	0
11	1	2	0	4	2	0
12	0	3	1	1	3	0
13	0	5	1	0	1	0
14	0	4	0	2	2	0
15	0	5	0	3	4	0
16	1	8	1	0	5	0
17	0	9	0	3	2	0
18	1	4	1	2	7	0
19	0	4	1	3	3	0
20	1	3	0	1	1	1
21	0	2	0	0	1	1
22	1	0	1	1	2	1
23	0	5	1	3	5	0
24	0	7	1	2	2	0
25	0	2	0	4	1	0
26	0	4	0	1	2	1
27	1	3	0	1	1	1
28	1	8	0	2	5	0
29	0	9	1	2	3	0
30	0	6	0	3	6	0

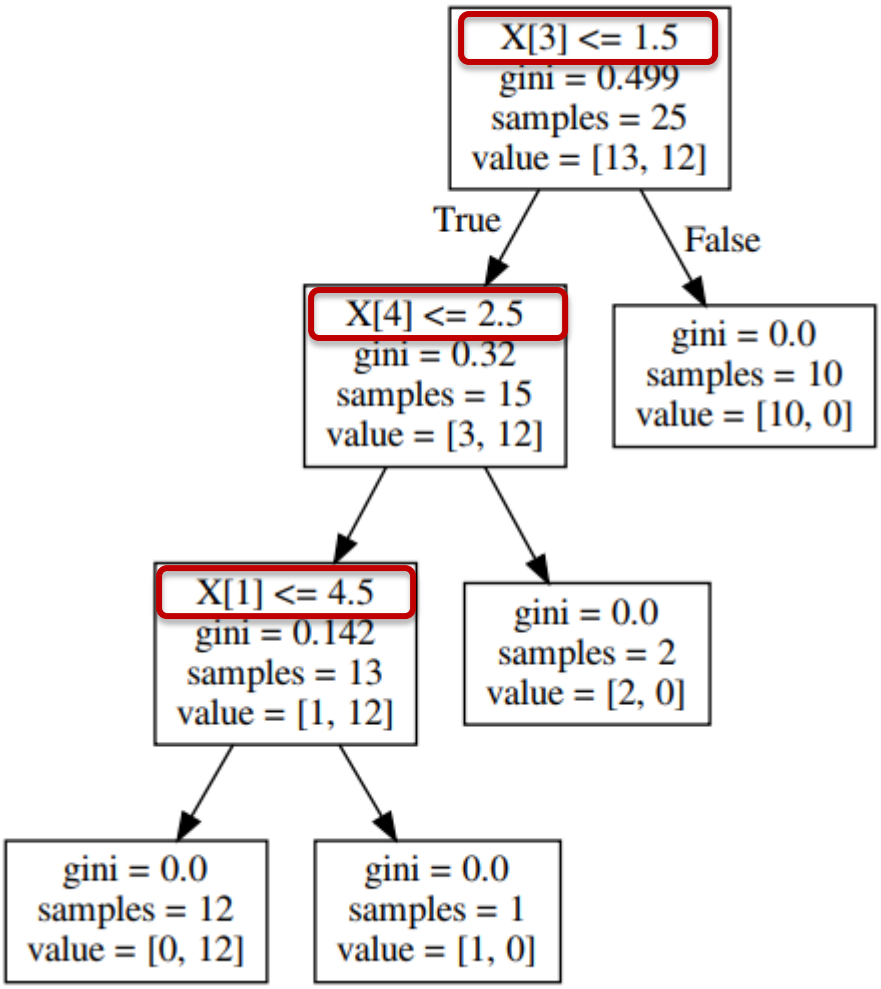
第二種資料集的資料總數為 34 筆，其中的 30 筆為 training data，有 12 筆 positive data，13 筆 negative data，5 筆 noise data，noise data 生成規則為符合上述三條 absolutely right’ rules，但卻不是吸血鬼，剩下的 4 筆資料為 testing data，其中最後 2 筆資料用來測試 noise 是否能預測正確。

	X[0]	X[1]	X[2]	X[3]	X[4]	
	human_ blood	sunshine	afraid_ garlic	skin	sleep	vampire
1	1	0	1	0	0	1
2	0	0	1	1	0	1
3	1	1	0	1	0	1
4	0	4	1	1	1	1
5	1	1	0	0	1	1
6	1	2	0	1	1	1
7	1	2	1	0	2	1
8	1	3	0	1	2	1
9	0	4	1	0	2	1
10	0	6	0	2	0	0
11	1	2	0	4	2	0
12	0	3	1	1	3	0
13	0	5	1	0	1	0
14	0	4	0	2	2	0
15	0	5	0	3	4	0
16	1	8	1	0	5	0
17	0	9	0	3	2	0

18	1	4	1	2	7	0
19	0	4	1	3	3	0
20	1	3	0	1	1	1
21	0	2	0	0	1	1
22	1	0	1	1	2	1
23	0	5	1	3	5	0
24	0	7	1	2	2	0
25	0	2	0	4	1	0
26	1	4	1	1	1	0
27	0	2	0	0	2	0
28	1	3	1	0	0	0
29	1	0	0	1	1	0
30	0	2	1	0	0	0
31	1	3	0	1	1	1
32	1	8	0	2	5	0
33	1	1	0	1	2	0
34	0	4	0	0	0	0

● Experiment - Decision Tree Classifier

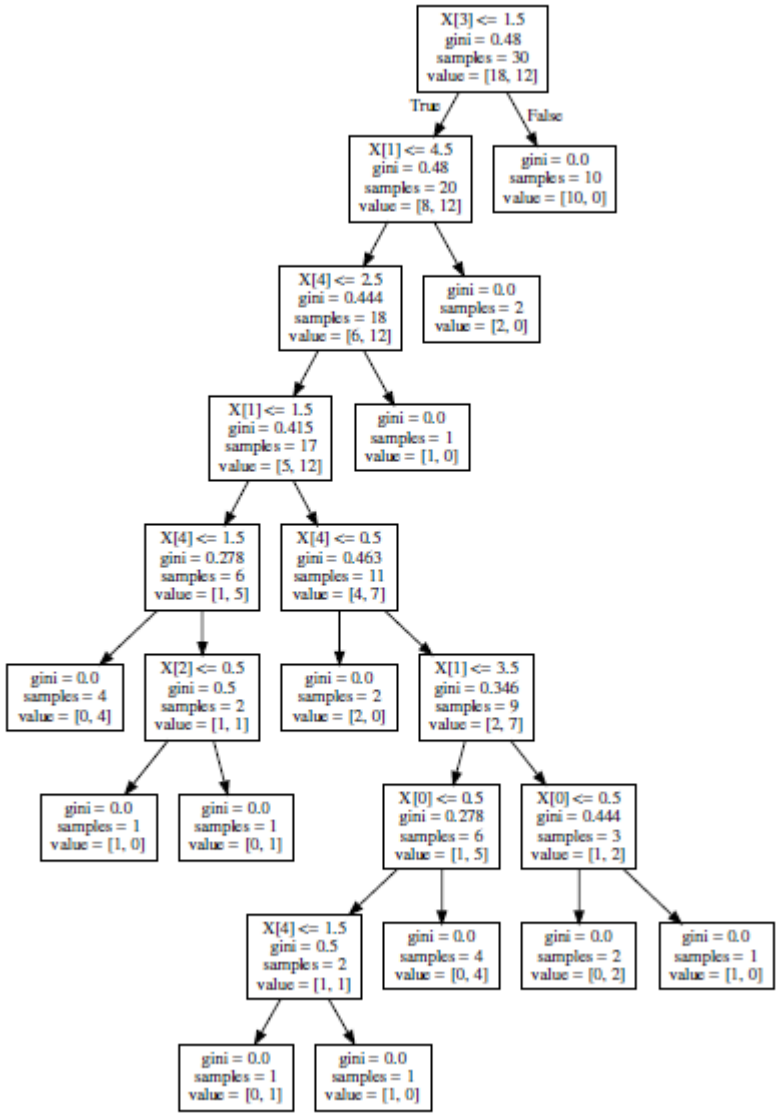
以第一種資料集跑出的 Decision Tree 為下圖：



可以看出，第一層先以 X[3] = skin 這個 feature 來判斷，皮膚顏色是 0 或 1 的被分到左邊，其餘分到右邊，此時右邊的就判斷為非吸血鬼。接下來看下一層，X[4] = sleep 為 0, 1, 2 的資料，即一天睡不到 2 小時的人被分到左邊，而 2 小時以上的就分到右邊且判斷為非吸血鬼。最後，X[1]為 0, 1, 2, 3, 4,的資料，即只能忍受照射太陽光 4 分鐘以下的被分到左邊，且判斷為吸血鬼，共有 12 筆資料，其餘 4 分鐘以上被分到右邊，為非吸血鬼。

此種分類方式與我設定的' absolutely right' rules 一樣，用三個 features 即可判斷出是否為吸血鬼。

以第二種資料集跑出的 Decision Tree 為下圖：



由於第二種資料集在 training 時加入 noise 資料，所以需要用到所有的 features 才能將所有的資料正確分類。

Testing 正確率為下圖所示：

```
test_y: ['1' '0' '0' '0']
y_pred: ['1' '0' '0' '0']
DecisionTree Accuracy Sore :  1.0
```

用 4 筆 testing data 測試，前兩筆符合‘ absolutely right’ rules 的 data 皆預測正確，剩下兩筆為 noise 的資料，即符合‘ absolutely right’ rules 但卻不是吸血鬼的案例，例如第 34 筆資料，只能忍受 4 分鐘的太陽光照射，皮膚蒼白，一天睡不到一個小時的人類，Classifier 可正確預測出他為人類，因為有加入 noise 的資料 training，model 學習到更多資訊，不會因為它符合那三條 rules 而判斷他為吸血鬼，

● Experiment – Random Forest Classifier

以第二種資料集跑 Random Forest Classifier，並以 4 筆 testing data 來看分類器預測的正確率，雖然 training 資料數不多，但訓練出的分類器可以正確預測，結果如下圖：

```
test_y: ['1' '0' '0' '0']
y_pred: ['1' '0' '0' '0']
RandomForest Accuracy Sore :  1.0
```

● Conclusion

用自己設定的‘ absolutely right’ rules 來產生資料，以”吸血鬼一定只能忍受 4 分鐘以下的陽光照射、皮膚顏色蒼白或普通白與一天睡眠時間不到兩個小時”來產生 positive 和 negative data，並使用部分資料來訓練 Decision Tree Classifier，以 Decision Tree 的圖可看出，訓練出的模型判斷方式與我設定的 rules 一樣。

除了 Decision Tree Classifier 以外，我還嘗試了 Random Forest Classifier，用有 noise 的資料集訓練模型，用有 noise 的資料去 test，也可以正確預測分類。

● Reference

https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c

https://www.jianshu.com/p/78594737b4b4

https://kknews.cc/zh-tw/other/2ae4q5e.html