Data Science Hw 5

工科海洋四 B07505015 梁瑞翔

Problem 1

Gini index for Gender:

	M	F
C0	8	4
C1	2	6

1-(8/10)^2-(2/10)^2=0.32

1-(4/10)^2-(6/10)^2=0.48

Gini index=0.32*0.5+0.48*0.5=0.40

Gini index for Car Type:

	F,S	L
C0	11	1
C1	2	6

Gini index=0.253

	F,L	S
C0	8	4
C1	6	2

Gini index=0.466

	S,L	F
C0	5	7
C1	8	0

Gini index=0.306

Gini index for Shirt Size:

	S	NS
C0	3	9
C1	2	6

Gini index=0.48

	M	NM
C0	3	9
C1	4	4

Gini index=0.451

	L	NL
C0	3	9

C1	1	7

Gini index=0.467

	E	NE
C0	3	9
C1	1	7

Gini index=0.467

Node_left:

Gender

	F	M
C0	1	0
C1	5	1

Gini index=0.231

Shirt size:

	L	NL
C0	1	0
C1	1	5

Gini index=0.142

Node_right:

Gender

	F	M
C0	3	8
C1	1	1

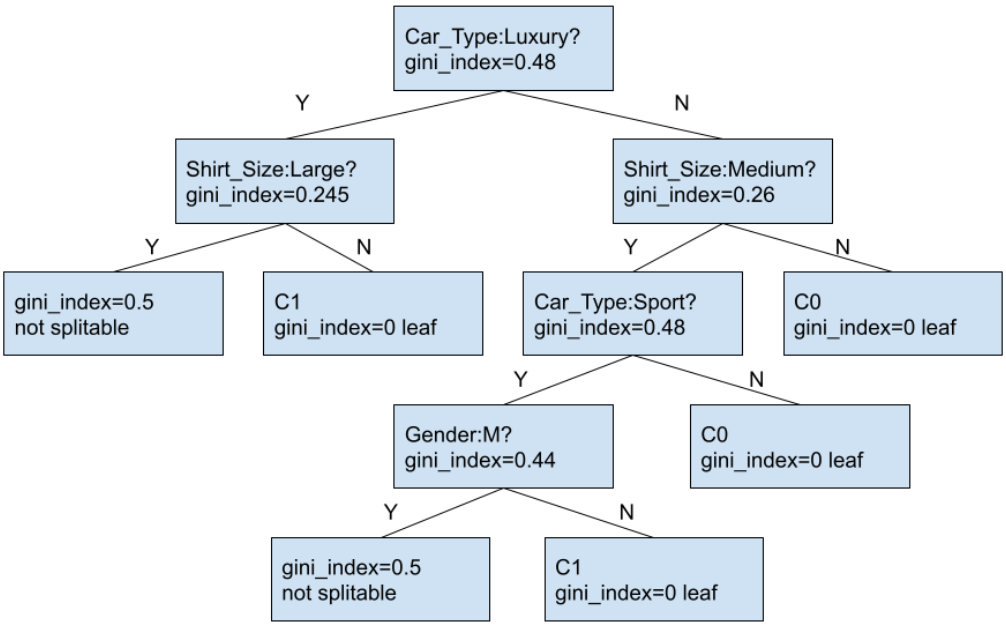
Gini index=0.252

Shirt size:

	M	NM
C0	3	8
C1	2	0

Gini index=0.185

The final decision I got:



Problem 2

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Family	Medium	C0
4	M	Sports	Large	C0
5	M	Family	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Family	Small	C0
8	F	Sports	Small	C0
9	F	Family	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C0
12	M	Family	Extra Large	C0
13	M	Sports	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Sports	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

(Gender=M, Car Type=Sports, Shirt Size=Medium)

Class: C0, C1

A: Attributes

P(A|C0)=8/12*4/12*3/12=0.05556

P(A|C1)= 2/8*2/8*4/8=0.03125

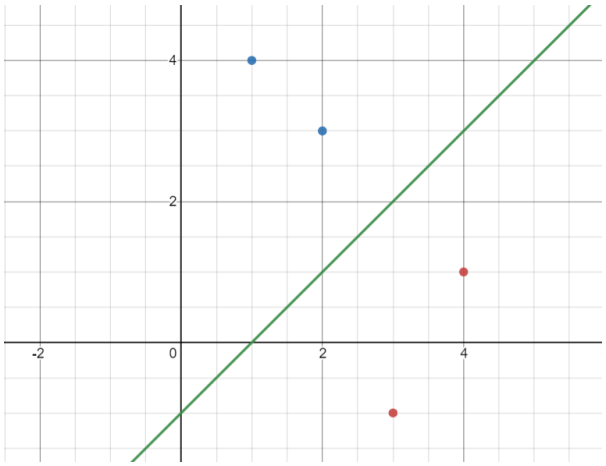
P(A|C0) P(C0)=0.05556*12/20=0.0333

P(A|C1) P(C1)=0.03125*8/20=0.0125

P(A|C0) P(C0)> P(A|C1) P(C1)
=> C0

Problem 3

x_i = (4,1), (7,2), (3,-1), (2,3), (1,4), (4.7), (-1,5)
$w^T x_i$-b>=1 for x_i = (4,1), (7,2), (3,-1)
$w^T x_i$-b<=-1 for x_i = (2,3), (1,4), (4.7), (-1,5)
$y_i(w^T x_i-b)>=1$
support vectors: (2,3), (4,1)



By inspection, the width between the support vectors is 2*sqrt(2)
Generalize the equation: cx1−cx2−c=0
2/||w||=2*sqrt2
2/(sqrt2*c)=2*sqrt2
b=-0.5
So w=[0.5,-0.5]^T, b=0.5
Hyperplane $y=w^Tx+b$
where w=[0.5,-0.5]^T, b=0.5

$$w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

In the case, $\alpha_5 = \frac{1}{6}, \alpha_6 = -\alpha_5 = -\frac{1}{6}, others: 0$

Compute with sklearn:

```python
import numpy as np
from sklearn.svm import SVC

X = np.array([[4,1],[7,2],[3,-1],[2,3],[1,4],[4,7],[-1,5]])
y = np.array([1,1,1,-1,-1,-1,-1])

clf = SVC(C = 1, kernel = 'linear')
clf.fit(X, y)

print('w = ',clf.coef_)
print('b = ',clf.intercept_)
print('Indices of support vectors = ', clf.support_)
print('Support vectors = ', clf.support_vectors_)
print('Number of support vectors for each class = ', clf.n_support_)
print('Coefficients of the support vector in the decision function = ', np.abs(clf.dual_coef_))
```

```
w =  [[ 0.5 -0.5]]
b =  [-0.5]
Indices of support vectors =  [3 0]
Support vectors =  [[2. 3.]
 [4. 1.]]
Number of support vectors for each class =  [1 1]
Coefficients of the support vector in the decision function =  [[0.25 0.25]]
```