r11921091 電機所 楊冠彥

Problem 1

Step 1

Parent (Class)

C0	11
C1	9

$$Gini = 1-\left(\frac{11}{20}\right)^2-\left(\frac{9}{20}\right)^2$$

$$= \frac{198}{400} = 0.495$$

Gender(1) : G(1)

	M	F
C0	7	4
C1	3	6

$$Gini(M) = 1-\left(\frac{7}{10}\right)^2-\left(\frac{3}{10}\right)^2 = \frac{42}{100}$$

$$Gini(F) = 1-\left(\frac{4}{10}\right)^2-\left(\frac{6}{10}\right)^2 = \frac{48}{100}$$

$$Gini_{G(1)} = \left(\frac{10}{20}\right)\left(\frac{42}{100}\right)+\left(\frac{10}{20}\right)\left(\frac{48}{100}\right) = 0.45$$

$$Gain_{G(1)} = \frac{198}{400} - 0.45 = \frac{9}{200} = 0.045$$

Car Type (1) : C(1)

	{Family, Sports}	{Luxury}
C0	8	3
C1	5	4

$$Gini\{Family, Sports\} = 1-\left(\frac{8}{13}\right)^2-\left(\frac{5}{13}\right)^2 = \frac{80}{169}$$

$$Gini\{Luxury\} = 1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2 = \frac{24}{49}$$

$$Gini_{CarType(1)} = \left(\frac{13}{20}\right)\left(\frac{80}{169}\right)+\left(\frac{7}{20}\right)\left(\frac{24}{49}\right)$$

$$= \frac{218}{455} \doteqdot 0.4791$$

$$Gain_{C(1)} = \frac{198}{400} - \frac{218}{455} = \frac{289}{18200} \doteqdot 0.0159$$

Car Type (2) : C(2)

	{Family, Luxury}	{Sports}
C0	10	1
C1	4	5

$$Gini\{Family, Luxury\} = 1-\left(\frac{10}{14}\right)^2-\left(\frac{4}{14}\right)^2 = \frac{20}{49}$$

$$Gini\{Sports\} = 1-\left(\frac{1}{6}\right)^2-\left(\frac{5}{6}\right)^2 = \frac{10}{36}$$

$$Gini_{CarType(2)} = \left(\frac{14}{20}\right)\left(\frac{20}{49}\right)+\left(\frac{6}{20}\right)\left(\frac{10}{36}\right) = \frac{31}{84} \doteqdot 0.369$$

$$Gain_{C(2)} = \frac{198}{400} - \frac{31}{84} \doteqdot 0.126$$

Car Type (3) : C(3)

	{Sports, Luxury}	{Family}
C0	4	7
C1	9	0

$$Gini\{Sports, Luxury\} = 1-\left(\frac{4}{13}\right)^2-\left(\frac{9}{13}\right)^2 = \frac{72}{169}$$

$$Gini\{Family\} = 1-\left(\frac{0}{7}\right)^2-\left(\frac{7}{7}\right)^2 = 0$$

$$Gini_{CarType(3)} = \left(\frac{13}{20}\right)\left(\frac{72}{169}\right)+\left(\frac{7}{20}\right)(0) = \frac{18}{65} \doteqdot 0.2769$$

$$Gain_{C(3)} = \frac{198}{400} - \frac{18}{65} = 0.2181$$

Let Shirt Size : SS

Small : Sm

Medium : Me

Large : La

Extra Large : ExLa

Shirt Size(2)

	{Me}	{Sm, La, ExLa}
C0	2	9
C1	5	4

$Gini(Me) = \frac{20}{49}$

$Gini(\{Sm, La, ExLa\}) = \frac{72}{169}$

$Gini_{SS(2)} = (\frac{7}{20})(\frac{20}{49}) + (\frac{13}{20})(\frac{72}{169}) = \frac{191}{455} \doteq 0.4198$

$Gain_{SS(2)} = \frac{198}{400} - \frac{191}{455} \doteq 0.0752$

Shirt Size(4)

	{ExLa}	{Sm, Me, La}
C0	4	7
C1	0	9

$Gini(ExLa) = 0$

$Gini(Sm, Me, La) = \frac{65}{128}$

$Gini_{SS(4)} = 0 + (\frac{16}{20})(\frac{65}{128}) = \frac{13}{32} \doteq 0.4063$

$Gain_{SS(4)} = \frac{198}{400} - \frac{13}{32} = 0.0888$

Shirt Size(1)

	{Sm}	{Me, La, ExLa}
C0	2	9
C1	3	6

$Gini(Sm) = \frac{12}{25}$

$Gini(Me, La, ExLa) = \frac{12}{25}$

$Gini_{SS(1)} = (\frac{5}{20})(\frac{12}{25}) + (\frac{15}{20})(\frac{12}{25}) = 0.48$

$Gain_{SS(1)} = 0.495 - 0.48 = 0.015$

Shirt Size(3)

	{La}	{Sm, Me, ExLa}
C0	3	8
C1	1	8

$Gini(La) = \frac{3}{8}$

$Gini(\{Sm, Me, ExLa\}) = 0.5$

$Gini_{SS(3)} = (\frac{4}{20})(\frac{3}{8}) + (\frac{16}{20})(0.5) = \frac{19}{40} = 0.475$

$Gain_{SS(3)} = 0.495 - 0.475 = 0.02$

$\Rightarrow$ Max Gain : $Gain_{C(3)}$

$\Rightarrow$

$\boxed{\text{Car type}}$

{Sports, Luxury} ← → Family → C0

Step 2.

Car Type(21) : C21

	Sports	Luxury
C0	1	3
C1	5	4

$Gini(Sports) = \frac{5}{18}$

$Gini(Luxury) = \frac{24}{49}$

$Gini_{C21} = (\frac{6}{13})(\frac{5}{18}) + (\frac{7}{13})(\frac{24}{49})$

$= \frac{107}{273} \doteq 0.3919$

$Gain_{C21} = \frac{72}{169} - \frac{107}{273}$

$\doteq 0.0341$

Gender(2) : G2

	M	F
C0	2	2
C1	3	6

$Gini(M) = \frac{12}{25}$

$Gini(F) = \frac{3}{8}$

$Gini_{G2} = (\frac{5}{20})(\frac{12}{25}) + (\frac{15}{20})(\frac{3}{8})$

$= \frac{321}{800} \doteq 0.40125$

$Gain_{G2} = \frac{72}{169} - 0.40125$

$\doteq 0.0248$

Shirt Size(21) : SS21

	{Sm}	{Me, La, ExLa}
C0	0	4
C1	3	6

$Gini(Sm) = 0$

$Gini(\{Me, La, ExLa\}) = 0.48$

$Gini_{SS21} = 0 + (\frac{10}{20})(0.48) = 0.24$

$Gain_{SS21} = \frac{72}{169} - 0.24$

$= 0.186$

Parent(Class)2

C0	4
C1	9

$Gini = \frac{72}{169}$

$\doteq 0.426$

Shirt Size(22) : SS22

	{Me}	{Sm, La, ExLa}
C0	2	2
C1	5	4

$Gini(Me) = \frac{20}{49}$

$Gini(\{Sm, La, ExLa\}) = \frac{4}{9}$

$Gini_{SS22} = (\frac{7}{13})(\frac{20}{49}) + (\frac{6}{13})(\frac{4}{9})$

$= \frac{116}{273} = 0.4249$

$Gain_{SS22} = \frac{72}{169} - \frac{116}{273} \doteqdot 0.0011$

Shirt Size(23) : SS23

	{La}	{Sm, Me, ExLa}
C0	2	2
C1	1	8

$Gini(La) = \frac{4}{9}$

$Gini(\{Sm, Me, ExLa\}) = 0.36$

$Gini_{SS23} = (\frac{3}{13})(\frac{4}{9}) + (\frac{10}{13})(0.36)$

$= \frac{74}{195} \doteqdot 0.3795$

$Gain_{SS23} = \frac{72}{169} - \frac{74}{195} \doteqdot 0.0465$

Shirt Size(24) : SS24

	{ExLa}	{Sm, Me, La}
C0	2	2
C1	0	9

$Gini(ExLa) = 0$

$Gini(Sm, Me, La) = \frac{36}{121}$

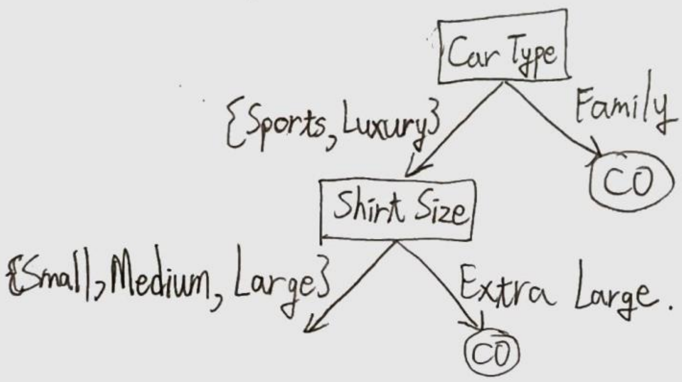
$Gini_{SS24} = 0 + (\frac{11}{13})(\frac{36}{121})$

$= \frac{36}{143} \doteqdot 0.2517$

$Gain_{SS24} = \frac{72}{169} - \frac{36}{143} \doteqdot 0.1743$

$\Rightarrow$ Max Gain : $Gain_{SS24} = 0.1743$

$\Rightarrow$



Step3.

Car Type(31) : C31

	Sports	Luxury
C0	0	2
C1	5	4

$Gini(Sports) = 0$

$Gini(Luxury) = \frac{4}{9}$

$Gini_{C31} = 0 + (\frac{6}{11})(\frac{4}{9})$

$= \frac{8}{33} \doteqdot 0.2424$

$Gain_{C31} = \frac{36}{121} - \frac{8}{33} = 0.0551$

Gender(3) : G3

	M	F
C0	0	2
C1	3	6

$Gini(M) = 0$

$Gini(F) = \frac{3}{8}$

$Gini_{G3} = 0 + (\frac{8}{11})(\frac{3}{8})$

$= \frac{3}{11} \doteqdot 0.2727$

$Gain_{G3} = \frac{36}{121} - \frac{3}{11} \doteqdot 0.0248$

Shirt Size(31) : SS31

	{Sm}	{Me, La}
C0	0	2
C1	3	6

$Gini(Sm) = 0$

$Gini(\{Me, La\}) = \frac{3}{8}$

$Gini_{SS31} = 0 + (\frac{8}{11})(\frac{3}{8})$

$= \frac{3}{11} \doteqdot 0.2727$

$Gain_{SS31} = \frac{36}{121} - \frac{3}{11} \doteqdot 0.0248$

Parent(Class)3

C0	2
C1	9

$Gini = \frac{36}{121}$

$\doteqdot 0.2975$

Shirt Size(32) : SS32

	{Me}	{Sm, La}
C0	0	2
C1	5	4

$Gini(Me) = 0$

$Gini(\{Sm, La\}) = \frac{4}{9}$

$Gini_{SS32} = 0 + (\frac{6}{11})(\frac{4}{9})$

$= \frac{8}{33} \doteqdot 0.2424$

$Gain_{SS32} = 0.0551$

Shirt Size(33) : SS33

	{La}	{Sm, Me}
C0	2	0
C1	1	8

$Gini(La) = \frac{4}{9}$

$Gini(\{Sm, Me\}) = 0$

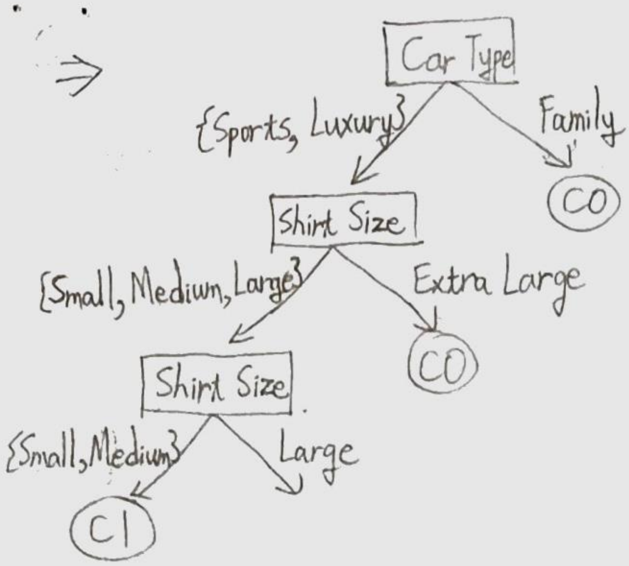
$Gini_{SS33} = (\frac{4}{9})(\frac{3}{11}) + 0$

$= \frac{4}{33} \doteqdot 0.1212$

$Gain_{SS33} = \frac{36}{121} - \frac{4}{33} \doteqdot 0.1763$

$\Rightarrow$ Max Gain : $Gain_{SS33} \doteqdot 0.1763$

⇒



Step 4.

Parent (Class) 4

C0	2
C1	1

$Gini = \frac{4}{9} \fallingdotseq 0.4444$

Gender(4) : G4

	M	F
C0	0	2
C1	1	0

$Gini(M) = 0$

$Gini(F) = 0$

$Gini_{G4} = 0$

$Gain_{G4} = 0.4444$

Car Type(41) : C41

	Sports	Luxury
C0	0	2
C1	1	0

$Gini(Sports) = 0$

$Gini(Luxury) = 0$

$Gini_{C41} = 0$

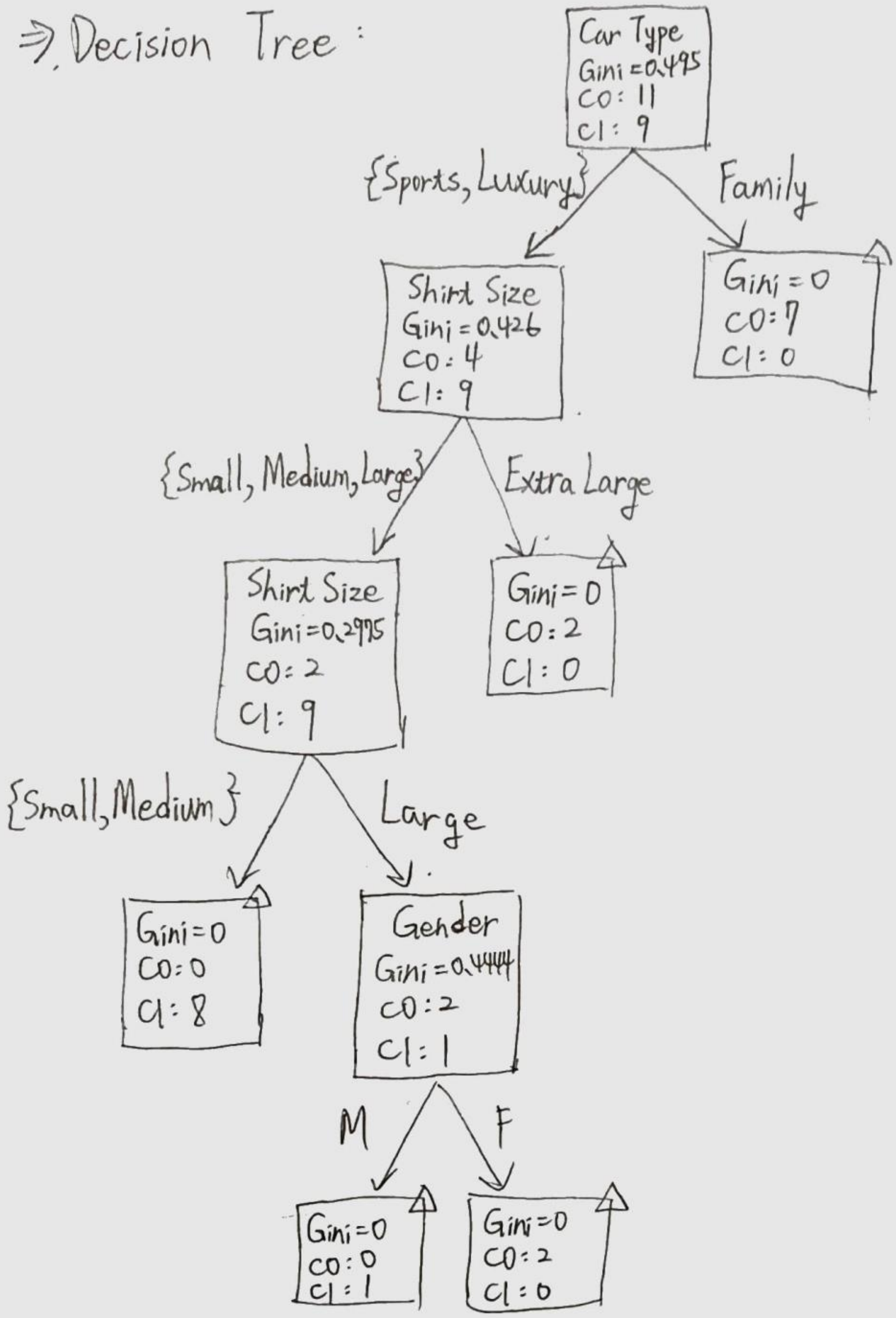
$Gain_{C41} = 0.4444$

⇒ Max Gain: $Gain_{G4} = 0.4444$ and $Gain_{C41} = 0.4444$

∴ Just pick one.

I choose Gender.

⇒ Decision Tree :



△: leaf node
→ can not split.
∵ Gini = 0
, and the classes have been classified.

Problem 2.

A: Attributes

$$P(A|C0) = \frac{7}{11} \times \frac{1}{11} \times \frac{2}{11} = \frac{14}{1331}$$

$$P(A|C1) = \frac{3}{9} \times \frac{5}{9} \times \frac{5}{9} = \frac{25}{243}$$

$$P(A|C0)P(C0) = \frac{14}{1331} \times \frac{11}{20} = 0.0058$$

$$P(A|C1)P(C1) = \frac{25}{243} \times \frac{9}{20} = 0.0463$$

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

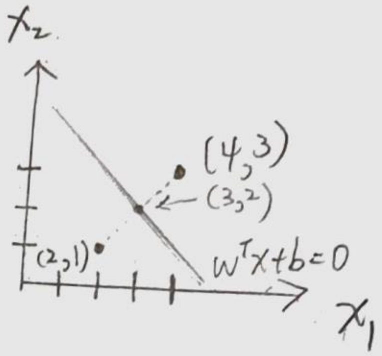
$\Rightarrow C1$

#

Problem 3.

distance	$y=1$ (4,3)	(4,8)	(7,2)
(-1,-2)	7.07	11.18	8.94
(-1,3)	5	7.07	8.06
$y=-1$ (2,-1)	4.47	9.22	5.83
(2,1)	2.83	7.28	5.10

$\Rightarrow$



Objectives: maximize $\frac{2}{\|w\|^2}$ subject to $y_i(w^T x_i - b) - 1 \geq 0 \quad \forall x_i$

$\Rightarrow$ Obviously, max Margin $= \frac{2}{\|w\|^2} = \sqrt{(4-2)^2+(3-1)^2} = 2\sqrt{2} \Rightarrow \|w\|^2 = \frac{1}{\sqrt{2}}$

In (3,2),

$$w^T x + b = 0$$
$$w^T x = -b$$

boundary line: $x_2 = x_1 - 1$

If we guess $w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$,

then $w^T x = -b$

$$\Rightarrow \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = -b = 5$$

$$\Rightarrow \quad b = -5$$

Margin $= \frac{2}{\sqrt{2}} < 2\sqrt{2}$, $\therefore$ not the answer.

Let boundary line: $cx_1 + cx_2 - 5c = 0$

$\therefore w = \begin{bmatrix} c \\ c \end{bmatrix}$, $b = -5c$

$$\frac{2}{\|w\|} = \frac{2}{\sqrt{2}c} = 2\sqrt{2}$$

$$\Rightarrow \frac{1}{\sqrt{2}c} = \sqrt{2} \Rightarrow c = \frac{1}{2}$$

$\therefore w = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$, $b = -2.5$

$\therefore$ Support vector : $[2, 1]$, $[4, 3]$

hyperplane : $y = [0.5 \ \ 0.5]X - 2.5$

Objectives : Margin $= 2\sqrt{2}$

Constraints :

$$f(\bar{x}_i) = \begin{cases} 1 & \text{if } [0.5 \ \ 0.5]X - 2.5 \geq 1 \\ -1 & \text{if } [0.5 \ \ 0.5]X - 2.5 \leq -1 \end{cases}$$

Appendix.

tags: `Data Science`

系級：電機所碩一　姓名：楊冠彥　學號：R11921091

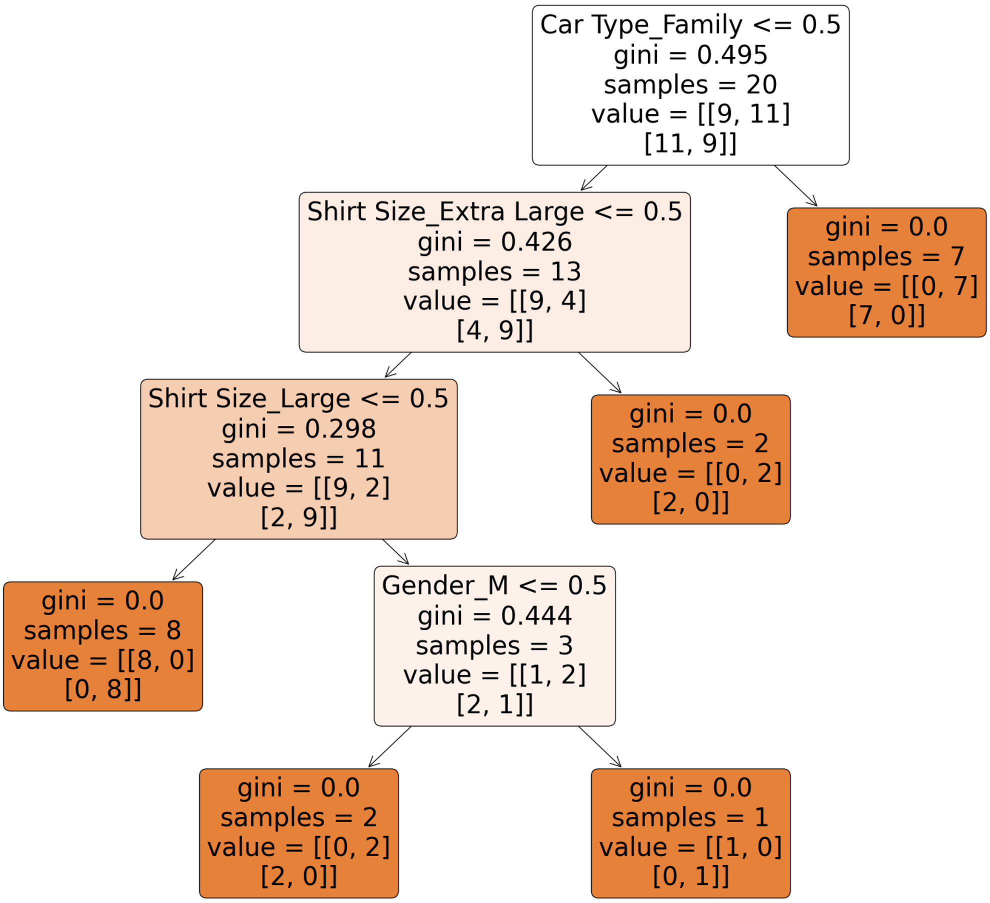
針對第一及第三題，我也另外寫了程式來驗證我的結果是否正確。

首先，我將Table寫進csv檔，資料讀入程式後如下所示：

	Customer ID	Gender_F	Gender_M	Car Type_Family	Car Type_Luxury	Car Type_Sports	Shirt Size_Extra Large	Shirt Size_Large	Shirt Size_Medium	Shirt Size_Small	Class_C0	Class_C1
0	1	0	1	1	0	0	0	0	0	1	1	0
1	2	0	1	1	0	0	0	0	1	0	1	0
2	3	0	1	1	0	0	1	0	0	0	1	0
3	4	0	1	0	0	1	1	0	0	0	1	0
4	5	0	1	1	0	0	0	1	0	0	1	0
5	6	0	1	1	0	0	1	0	0	0	1	0
6	7	0	1	0	1	0	1	0	0	0	1	0
7	8	1	0	1	0	0	0	0	0	1	1	0
8	9	1	0	1	0	0	0	0	1	0	1	0
9	10	1	0	0	1	0	0	1	0	0	1	0
10	11	1	0	0	1	0	0	1	0	0	1	0
11	12	0	1	0	0	1	0	0	1	0	0	1
12	13	0	1	0	0	1	0	1	0	0	0	1
13	14	0	1	0	0	1	0	0	1	0	0	1
14	15	1	0	0	0	1	0	0	0	1	0	1
15	16	1	0	0	1	0	0	0	0	1	0	1
16	17	1	0	0	1	0	0	0	0	1	0	1
17	18	1	0	0	0	1	0	0	1	0	0	1
18	19	1	0	0	1	0	0	0	1	0	0	1
19	20	1	0	0	1	0	0	0	1	0	0	1

Problem 1.

接著就能用程式處理算出gini index(two way split)及畫出decision tree。



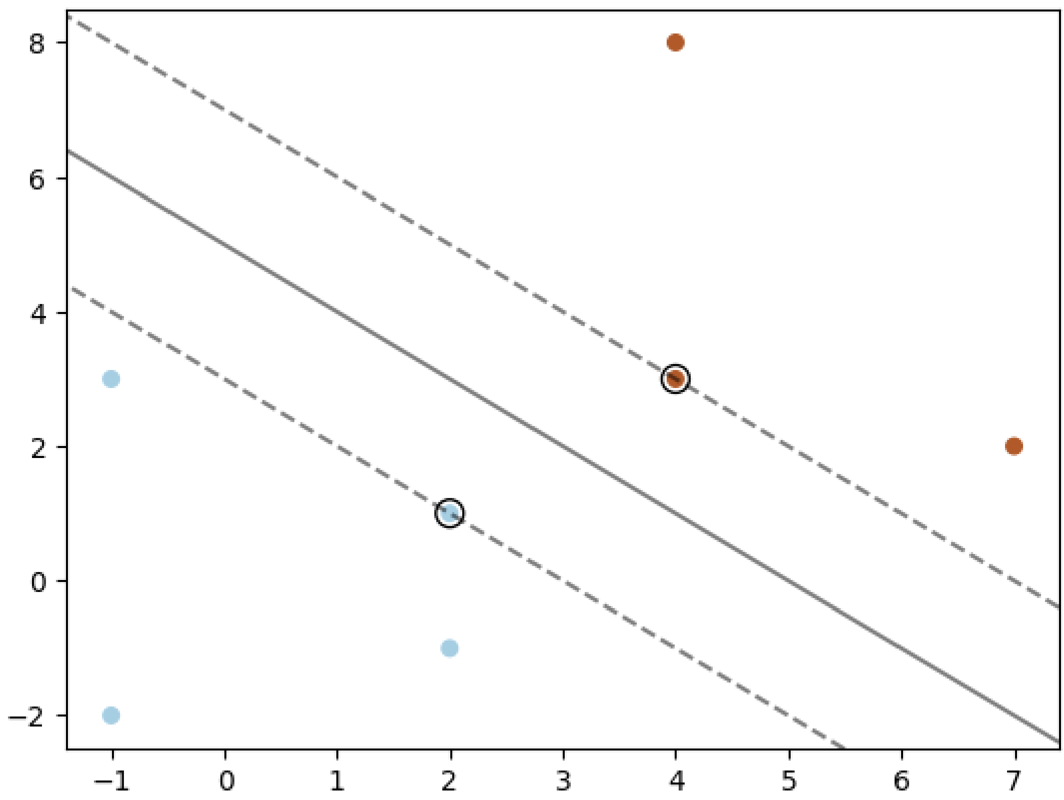
Source Code

```
import pandas as pd
import numpy as np
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from matplotlib import pyplot as plt
from sklearn import tree
df = pd.read_csv("hw5_data.csv",sep='\s*,\s*')
df_dum = pd.get_dummies(df)
feature_col = df_dum.columns[1:-2]
X = df_dum[feature_col]
label = pd.get_dummies(df['Class'])
print(label)
X_dum = pd.get_dummies(X)
clf = DecisionTreeClassifier()
model = clf.fit(X,label)
# plot decision tree
fig = plt.figure(figsize=(25,20))
_ = tree.plot_tree(clf,
                   feature_names=feature_col,
                   class_names=['0','1'],
                   filled=True,
                   rounded=True)
fig.savefig("decistion_tree.png")
```

Problem 3.

以下為SVM部分，程式跑出之結果：

```
w:  [[0.5 0.5]] b:  [-2.5]
```



Source Code

```
from sklearn.svm import SVC
import matplotlib.pyplot as plt
import numpy as np
svm = SVC(kernel='linear', probability=True)
# samples
X_train = np.array([[4, 3], [4, 8], [7,2], [-1, -2], [-1, 3], [2, -1], [2, 1]])
y = [1, 1, 1, -1, -1, -1, -1]
svm.fit(X_train,y)
print(svm.coef_)
print(svm.intercept_)

plt.scatter(X_train[:, 0], X_train[:, 1], c=y, s=30, cmap=plt.cm.Paired)
# plot the decision function
ax = plt.gca()
xlim = ax.get_xlim()
ylim = ax.get_ylim()

# create grid to evaluate model
xx = np.linspace(xlim[0], xlim[1], 30)
yy = np.linspace(ylim[0], ylim[1], 30)
YY, XX = np.meshgrid(yy, xx)
xy = np.vstack([XX.ravel(), YY.ravel()]).T
Z = svm.decision_function(xy).reshape(XX.shape)

```

```python
# plot decision boundary and margins
ax.contour(XX, YY, Z, colors='k', levels=[-1, 0, 1], alpha=0.5,
           linestyles=['--', '-', '--'])
# plot support vectors
ax.scatter(svm.support_vectors_[:, 0], svm.support_vectors_[:, 1], s=100,
           linewidth=1, facecolors='none', edgecolors='k')
plt.show()
```