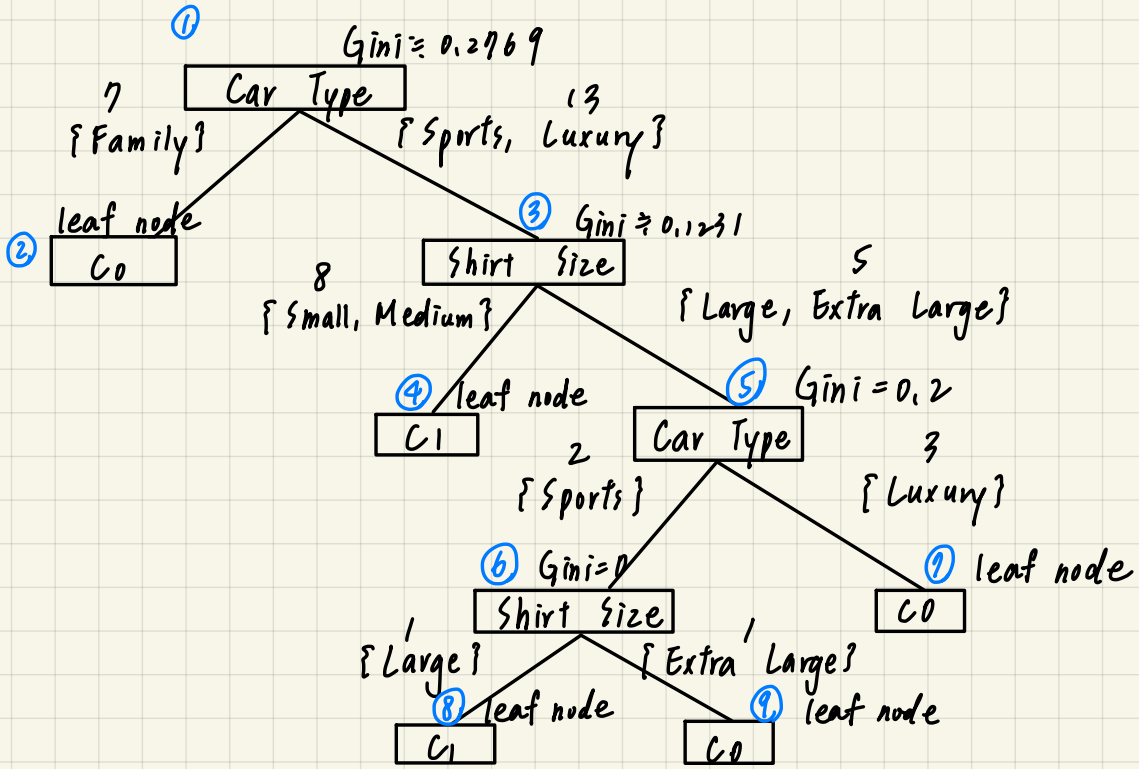
Data Science HW5

R11922066 林庭安

Problem 1.




```
======= step1 =======
Gender: {'M'} {'F'} -> gini = 0.45
Car Type: {'Family'} {'Sports','Luxury'} -> gini = 0.27692307692307694
Car Type: {'Sports'} {'Family','Luxury'} -> gini = 0.369047619047619
Car Type: {'Luxury'} {'Family','Sports'} -> gini = 0.47912087912087914
Shirt Size: {'Small'} {'Medium','Large','Extra Large'} -> gini = 0.48
Shirt Size: {'Medium'} {'Small','Large','Extra Large'} -> gini = 0.41978021978021973
Shirt Size: {'Large'} {'Small','Medium','Extra Large'} -> gini = 0.47500000000000003
Shirt Size: {'Extra Large'} {'Small','Medium','Large'} -> gini = 0.39375000000000004
Shirt Size: {'Small','Medium'} {'Large','Extra Large'} -> gini = 0.3541666666666667
Shirt Size: {'Small','Large'} {'Medium','Extra Large'} -> gini = 0.49494949494949503
Shirt Size: {'Small','Extra Large'} {'Medium','Large'} -> gini = 0.4727272727272728
---> min gini = 0.27692307692307694

======= step2 =======
Gender: {'M'} {'F'} -> gini = 0.0
Shirt Size: {'Small'} {'Medium','Large','Extra Large'} -> gini = 0.0
Shirt Size: {'Medium'} {'Small','Large','Extra Large'} -> gini = 0.0
Shirt Size: {'Large'} {'Small','Medium','Extra Large'} -> gini = 0.0
Shirt Size: {'Extra Large'} {'Small','Medium','Large'} -> gini = 0.0
Shirt Size: {'Small','Medium'} {'Large','Extra Large'} -> gini = 0.0
Shirt Size: {'Small','Large'} {'Medium','Extra Large'} -> gini = 0.0
Shirt Size: {'Small','Extra Large'} {'Medium','Large'} -> gini = 0.0
---> min gini = 0.0
leaf node: C0

======= step3 =======
Gender: {'M'} {'F'} -> gini = 0.4153846153846154
Car Type: {'Sports'} {'Luxury'} -> gini = 0.3919413919413919
Shirt Size: {'Small'} {'Medium','Large','Extra Large'} -> gini = 0.36923076923076925
Shirt Size: {'Medium'} {'Small','Large','Extra Large'} -> gini = 0.3076923076923077
Shirt Size: {'Large'} {'Small','Medium','Extra Large'} -> gini = 0.34871794871794864
Shirt Size: {'Extra Large'} {'Small','Medium','Large'} -> gini = 0.25174825174825166
Shirt Size: {'Small','Medium'} {'Large','Extra Large'} -> gini = 0.12307692307692303
Shirt Size: {'Small','Large'} {'Medium','Extra Large'} -> gini = 0.4249084249084249
Shirt Size: {'Small','Extra Large'} {'Medium','Large'} -> gini = 0.4153846153846154
---> min gini = 0.12307692307692303
```

```
======= step4 =======
Gender: {'M'} {'F'} -> gini = 0.0
Car Type: {'Sports'} {'Luxury'} -> gini = 0.0
Shirt Size: {'Small'} {'Medium'} -> gini = 0.0
leaf node: C1

======= step5 =======
Gender: {'M'} {'F'} -> gini = 0.26666666666666666
Car Type: {'Sports'} {'Luxury'} -> gini = 0.2
Shirt Size: {'Large'} {'Extra Large'} -> gini = 0.26666666666666666
---> min gini = 0.2

======= step6 =======
Gender: {'M'} {'F'} -> gini = 0.5
Shirt Size: {'Large'} {'Extra Large'} -> gini = 0.0
---> min gini = 0.0

======= step7 =======
Gender: {'M'} {'F'} -> gini = 0.0
Shirt Size: {'Large'} {'Extra Large'} -> gini = 0.0
---> min gini = 0.0
leaf node: C0

======= step8 =======
leaf node: C1

======= step9 =======
leaf node: C0
```

Problem 2.

$P(M \mid C0) = \frac{7}{11}$ $\quad P(M \mid C1) = \frac{3}{9}$

$P(\text{Sports} \mid C0) = \frac{1}{11}$ $\quad P(\text{Sports} \mid C1) = \frac{5}{9}$

$P(\text{Medium} \mid C0) = \frac{2}{11}$ $\quad P(\text{Medium} \mid C1) = \frac{5}{9}$

$P(C0) = \frac{11}{20}$ $\quad P(C1) = \frac{9}{20}$

$$P(C0 \mid M, \text{Sports}, \text{Medium}) = \frac{P(M \mid C0)\,P(\text{Sports} \mid C0)\,P(\text{Medium} \mid C0)\,P(C0)}{P(M, \text{Sports}, \text{Medium})}$$

$$= \frac{\frac{7}{11} \times \frac{1}{11} \times \frac{2}{11} \times \frac{11}{20}}{P(M, \text{Sports}, \text{Medium})} = \frac{\frac{7}{1210}}{P(M, \text{Sports}, \text{Medium})}$$

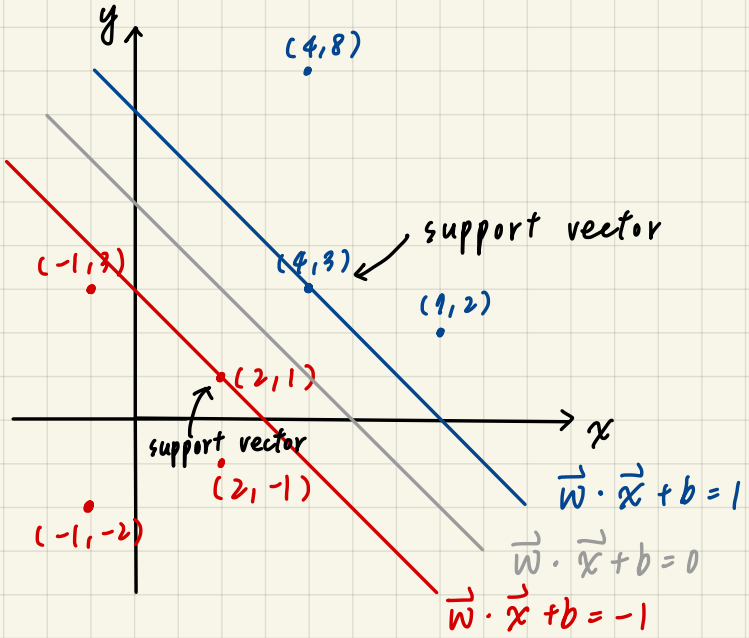
$$P(C1 \mid M, \text{Sports}, \text{Medium}) = \frac{P(M \mid C1)\,P(\text{Sports} \mid C1)\,P(\text{Medium} \mid C1)\,P(C1)}{P(M, \text{Sports}, \text{Medium})}$$

$$= \frac{\frac{3}{9} \times \frac{5}{9} \times \frac{5}{9} \times \frac{9}{20}}{P(M, \text{Sports}, \text{Medium})} = \frac{\frac{5}{108}}{P(M, \text{Sports}, \text{Medium})}$$

$\because P(C1 \mid M, \text{Sports}, \text{Medium}) > P(C0 \mid M, \text{Sports}, \text{Medium})$

$\therefore (M, \text{Sports}, \text{Medium}) \Rightarrow C1$

Problem 3.



Objective: Max margin $= \frac{2}{\|\vec{w}\|^2}$

i.e. min $\frac{\|\vec{w}\|^2}{2}$

Constraints: $f(\vec{x_{n'}}) = \begin{cases} 1 & \text{if } \vec{w}\cdot\vec{x}+b \geq 1 \\ -1 & \text{if } \vec{w}\cdot\vec{x}+b \leq -1 \end{cases}$

i.e. $y_{n'}(\vec{w}\cdot\vec{x_{n'}}+b) \geq 1$

$\Rightarrow y_{n'}(\vec{w}\cdot\vec{x_{n'}}+b)-1 \geq 0, \forall x_{n'}$

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{n'=1}^{N} \alpha_{n'}[y_{n'}(w^T x_{n'}-b)-1]$$

if $y_{n'}(w^T x_{n'}+b)-1=0$, $\alpha_{n'} \geq 0$

if $y_{n'}(w^T x_{n'}+b)-1>0$, $\alpha_{n'}=0$

$$\frac{\partial L}{\partial w} = w - \sum_{n'=1}^{N} \alpha_{n'} y_{n'} x_{n'} = 0$$

$$\frac{\partial L}{\partial b} = \sum_{n'=1}^{N} \alpha_{n'} y_{n'} = 0$$

$$w - \left(-\alpha_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 4 \\ 3 \end{bmatrix}\right) = 0, \quad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -2\alpha_1 + 4\alpha_2 \\ -\alpha_1 + 3\alpha_2 \end{bmatrix}$$

$$-\alpha_1 + \alpha_2 = 0, \quad \alpha_1 = \alpha_2 = \alpha, \quad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 2\alpha \\ 2\alpha \end{bmatrix}$$

On the boundary (2,1) and (4,3)

$-1((2\alpha,2\alpha)\cdot(2,1)+b) = -6\alpha - b = 1$

$1((2\alpha,2\alpha)\cdot(4,3)+b) = 14\alpha + b = 1$

$\Rightarrow 8\alpha = 2, \alpha = \alpha_1 = \alpha_2 = \frac{1}{4}, b = -2.5$

$$w = \begin{bmatrix} 2\alpha \\ 2\alpha \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \quad b = -2.5$$

hyperplane : $y = w^T x + b = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 2.5$