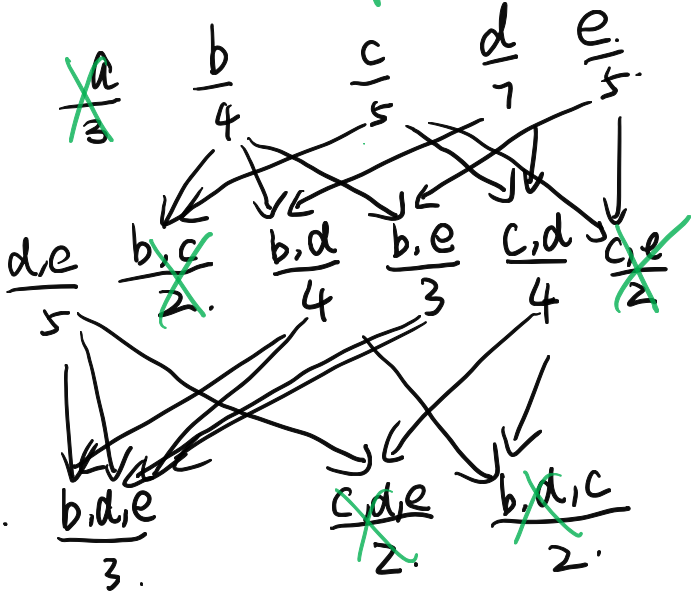
19 Q4.

a	b	c	d	e	f
1		1			
	1	1	1		
1	1		1	1	
1			1	1	
	1	1	1	1	
		1	1		1
	1		1	1	
		1	1	1	

i) maximal itemset.
直接超集都不频繁

ii) closed ~
~ 都没有与之相同的 sup.



a) i) Find. sup ≥ 3

maximal itemsets:

{b,d,e}, {c,d}

closed itemsets:

{c}, {d}, {d,e}, {b,d}, {c,d}, {b,d,e}.

~~{b,c,d,e}~~

ii) $b \to de$

support: $\frac{S(d \cup e \cup b)}{N} = \frac{3}{8}$

coffidence: $\frac{S(d \cup e \cup b)}{S(b)} = \frac{3}{4}$

lift: $\frac{C(b \to de)}{S(de)} = \frac{3/4}{5} = \frac{3}{20}$

b) i) step 1:

set 1: {c1 = 10 | 10}.

set 2: {c2 = 20 | 20}.

set 3: {c3 = 30 | 30, 40, 50, 60} ⇒ c3 = 45.

step 2:

set 1: {c1 = 10 | 10}

set 2: {c2 = 20 | 20, 30} ⇒ c2 = 25

set 3: {c3 = 45 | 40, 50, 60} ⇒ c3 = 50

step 3:

set 1: {c1 = 10 | 10}

set 2: {c2 = 25 | 20, 30}

set 3: {c3 = 50 | 40, 50, 60}

ii). $SSE = \sum\sum dist(c_i, x)^2$

$= 0 + 5^2 + 5^2 + 10^2 + 10^2$

$= 250$

$BSS/SSB = \sum |C_i| (c_i - \bar{c})^2$

$= 1 \times (10-35)^2 + 2 \times (25-35)^2 + 3 \times (50-35)^2$

$= 1500$

$TSS = \sum (x - \bar{c})^2$

$= 25^2 \times 2 + 15^2 \times 2 + 5^2 \times 2$

$= 625 \times 2 + 225 \times 2 + 50$

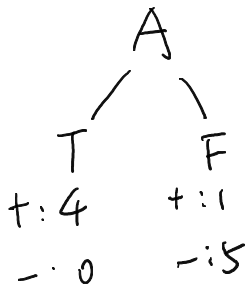
$= 1250 + 450 + 50$

$= 1750$

$= SSE/WSS + SSB/BSS$

18 Q4.

a)



i)

Gini : $1 - \sum P(i|t)$

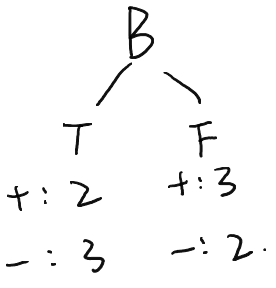
Gini index:

A, T $= 1 - 1 = 0$

A, F $= 1 - (\frac{1}{6})^2 - (\frac{5}{6})^2 = \frac{10}{36} = \frac{5}{18}$.

∴ Gini of A:

$\frac{6}{10} \times \frac{5}{18} = \frac{1}{6}$



B, T $= 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = \frac{12}{25}$

B, F $= \frac{12}{25}$

∴ Gini of B:

$\frac{12}{25}$

choose A.

ii)

A:

	+	−
T	4	0
F	1	5

B:

	+	−
T	2	3
F	3	2

iii) Gain in Gini index

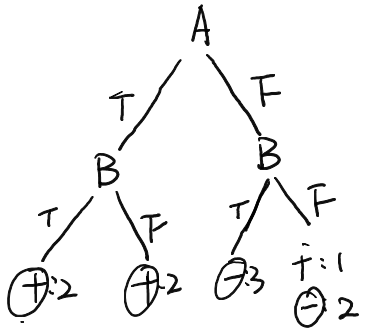
Overall Gini before splitting:

$1 - (0.5)^2 - (0.5)^2 = 0.5$

$G_A = 0.5 - \frac{1}{6} = \frac{1}{3}$

$G_B = 0.5 - \frac{12}{25} = \frac{1}{50}$

b>



c>

label / predict	+	−
+	4 tp	0 fp
−	1 fn	5 tn

acc = 0.9.

precision = 1 = $\frac{\text{true positive}}{(tp + fp)}$

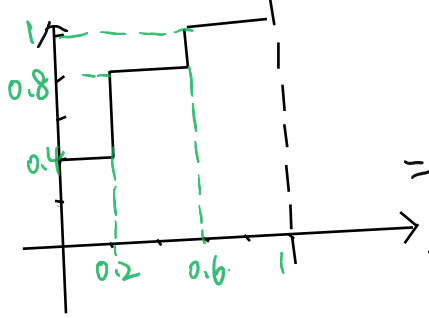
recall = 0.8 = $\frac{tp}{(tp+fn)}$

$F_1 = \frac{2 \times 1 \times 0.8}{1+0.8}$

$= \frac{1.6}{1.8} = \frac{8}{9}$

d>

	−	−	+	−	−	+	+	−	+	+	
ID	x10	9	8	7	6	5	4	3	2	1	
P	0.35	0.4	0.45								
TP	5	5	5	4	4	4	3	2	2	1	0
FP	5	4	3	3	2	1	1	1	0	0	0
TN		1	2	2	3	4	4	4	5	5	5
FN				1	1	1	2	3	3	4	5
TPR	1	1	1	0.8	0.8	0.8	0.6	0.4	0.4	0.2	0
FPR	1	0.8	0.6	0.6	0.4	0.2	0.2	0.2	0	0	0



Area under ROC

$= 1 - 0.6 \times 0.2 - 0.4 \times 0.2$

$= 0.8$

17 Q4

step 1:

(a) C_1 : 1100

S_1 : {0, 200, 300, 900, 1100}.

$c_1 \rightarrow 500$.

C_2 : 1600

S_2 : {1600}

Step 2: 2100, 1050

C_1 : 500 $\quad \frac{1400}{4} = 35$

S_1 : {0, 200, 300, 900}

$C_1 \rightarrow 350$

C_2 : 1600

S_2 : {1100, 1600} 850.

$C_2 \rightarrow 1350$

Step 3

C_1 : 350

S_1 : {0, 200, 300}.

$C_1 : \frac{500}{3}$

C_2 : 1350

S_2 : {900, 1100, 1600}

C_2 : 1200

Step 4.

Same as step 3

(b) ✓

(c)

	P1	P2	P3	P4	P5	P6
P1	0	200	300	900	1100	1600
P2	200	0	100	700	9	14
P3	3	1	0	6	8	13
P4	9	7	6	0	2	7
P5	11	9	8	2	0	5
P6	16	14	13	7	5	0

到cluster中其它点的平均dist

min(到其它cluster中平均dist)

	a_i	b_i	s_i
1	2.5	12	
2	1.5	10	0.85
3	2	9	
4	4.5	22/3	
5	3.5	28/3	
6	6	43/3	

$s_i = (b_i - a_i)/\max(a_i, b_i)$

$\frac{\frac{28}{3} - 3.5}{\frac{28}{3}} = \frac{9\frac{1}{3} - 3.5}{\frac{28}{3}} = \frac{4.5\frac{1}{3}}{\frac{28}{3}} = \frac{4.5\times 3 + 1}{28} = \frac{29}{56}$

d

① dist(2,3) = 100

② dist(4,5) = 200

③ dist {{2,3}, 1) = 300 ✓ → {2,3,1}

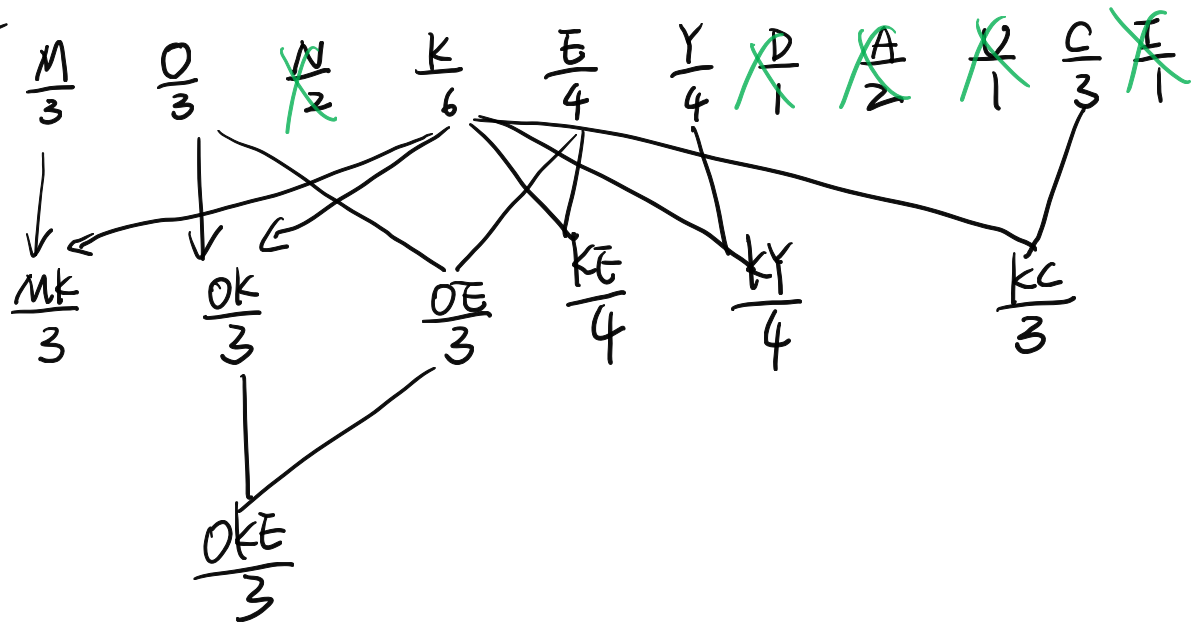
dist({4,5}, 6) = 700 → {4,5,6}

dist({2,3},{4,5}) = 1100 X

16.

Q5

(a)



(b) $con = \frac{sup\{E,C\}}{sup\{E\}} = \frac{1}{4}$.

$lift = \frac{C\{E \to C\}}{s(C)} = \frac{1}{12}$

weakly related.

(c). $h(\{E,C\}) = \frac{sup\{E,C\}}{max(s(E), s(C))} = \frac{1}{max(4,3)} = \frac{1}{4}$.

不会.

$X \subset Y$

$sup(X) \geq sup(Y)$

$max(s(X_1), s(X_2)) \leq max(s(Y_1), s(Y_2), \cdots)$

$\therefore h(X) \geq h(Y)$

$\therefore$ it's anti-monotone.

$$\zeta(\{A,B\}) = \min[C(A\to B), C(B-A)]$$

$$= \min\left[\frac{sup(A,B)}{sup(A)}, \frac{sup(A,B)}{sup(B)}\right]$$

$$= \frac{sup(A,B)}{\max(sup(A), sup(B))}$$

$$C(\{A,B,C\}) = \min[C(A\to BC), C(B\to AC), C(C\to AB)]$$

$$= \frac{sup(A,B,C)}{\max\{sup(A), sup(B), sup(C)\}}$$

$\{A,B\} \subset \{A,B,C\}$

$\therefore sup(A,B) \geq sup(A,B,C)$

$\max\{sup(A), sup(B)\} \leq \max\{sup(A), sup(B), sup(C)\}$

$\therefore \zeta(A,B) \geq \zeta(A,B,C)$

$\therefore$ anti-monotone

15 Ex. 7 on Page 37

~~a)~~