Justin Yeo, yeoj0029@e.ntu.edu.sg

Question 1:

(a):

i. T
ii. F
iii. F
iv. F
v. T
vi. T
vii. F
viii. F
ix. F
x. F

(b):

i. $SMC = \frac{M00+M11}{M00+M01+M10+M11} = \frac{3+2}{10} = \frac{1}{2}$
ii. $Hamming\ Distance = \sum_{k=1}^{n} |p_k - q_k| = |1-0| + |0-1| + |0-2| + |0-0| + |1-1| + |0-0| + |0-2| + |3-3| + |1-0| + |1-1| = 7$
iii. $Jaccard = \frac{M11}{M01+M10+M11} = \frac{3}{3+2+3} = \frac{3}{8}$
iv. $Euclidean = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2} = \sqrt{(1-0)^2 + (0-1)^2 + \cdots + (1-0)^2 + (1-1)^2} = \sqrt{11} \approx 3.3166$
v. Supremun Distance: I had no idea what this was so I just put N/A
vi. $cosine = \frac{(p \cdot q)}{\|p\|\|q\|} = \frac{1}{10}$
vii. $Correlation = \frac{covariance(p,q)}{std(p) \times std(q)} = 2$

(c):

Advantage: It is easier to view the difference as the differences between colors are very obvious.

Disadvantage: It does not show absolute values of difference. Some people may be colorblind, thus they cannot see the differences.

(d):

Advantage: It is easy to obtain the sampling data using simple sampling.

Disadvantage: If there are different density groups between the data objects, the sampling data may be skewed towards the higher density data groups.

Question 2:

(a):

$$Gini(parent) = 1 - \left(\frac{3}{10}\right)^2 - \left(\frac{7}{10}\right)^2 = 0.42$$

$$Gini(Cuurent\ Phone = iPhone) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.44444,$$

$$Gini(Cuurent\ Phone = Samsung) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.4444,$$

$$Gini(Cuurent\ Phone = Sony) = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0,$$

$$Gini(Children) = \frac{3}{10}(0.4444) + \frac{3}{10}(0.4444) + \frac{4}{10}(0) = 0.26666,$$

$$Gini(Gain) = 0.42 - 0.26666 = 0.15334 \approx 0.153$$

$$Gini(Drive\ Car = Yes) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48,$$

$$Gini(Drive\ Car = Yes) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375,$$

$$Gini(Children) = \frac{5}{10}(0.48) + \frac{4}{10}(0.375) = 0.39,$$

$$Gini(Gain) = 0.42 - 0.39 = 0.03$$

(b):

Buy	No	No	No	No	Yes	Yes	No	Yes	No	No
	Age									
Sorted Values	18	24	24	28	30	38	40	40	50	50

Split Positions	16		21		24		26		29		34		39		40		45		50		55	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	0	3	1	2	2	1	2	1	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	4	3	4	3	4	3	5	2	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.300		0.400		0.417		0.419		0.375		0/400		0.420	

Therefore, best Gini split is age <= 29 vs age >29.

(c):

$$Gini(Age < 20) = 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 = 0$$

$$Gini(20 \le Age < 30) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

$$Gini(30 \le Age < 40) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

$$Gini(Age \ge 40) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Gini(Children) = \frac{4}{10} \times 0.375 = 0.15$$

(d):

For Q2(b),

$$Error = \frac{0 + (11 \times 0.5)}{10} = 0.55$$

For Q2(c),

Age < 20, Yes = 0, No = 1

20 <= Age < 30, Yes = 0, No = 3

30 <= Age < 40, Yes = 2, No = 0

40 <= Age, Yes = 1, No = 3

$$Error = \frac{1 + (4 \times 0.5)}{10} = 0.3$$

Therefore, the better splitting strategy is Q2(c).

(e):

If Drive Car = Yes:

- $Sample\ Mean = \frac{40+28+38+50+50}{5} = 41.2$
- $Sample\ Variance = (40-41.2)^2 + \cdots + (50-41.2)^2 = 341$

If Drive Car = No:

- $Sample\ Mean = \frac{24+18+30+24}{4} = 24$
- $Sample\ Variance = (24-24)^2 + \cdots + (24-24)^2 = 72$

$P(Drives\ Car = Yes|Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)$

$$= \frac{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No|Drives\ Car = Yes\ \times P(Drives\ Car = Yes)}{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)}$$

$$= \frac{P(Age = 40|Drives = Yes) \times P(Current\ Phone = Sony|Drives = Yes) \times P(Buys = No|Drives = Yes) \times P(Drives = Yes)}{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)}$$

$$P(Age = 40|Drives = Yes) = \frac{1}{\sqrt{2 \times \pi \times 341}} e^{-\frac{(40-41.2)^2}{2x341}} = 0.0215$$

$$P(Current\ Phone = Sony|Drives = Yes) = \frac{P(Current\ Phone = Sony \cap Drives = Yes)}{P(Drives = Yes)} = \frac{1}{5}$$

$$P(Buy = No|Drives = Yes) = \frac{P(Buy = No \cap Drives = Yes)}{P(Drives = Yes)} = \frac{3}{5}$$

$$P(Drives = Yes) = \frac{5}{9}$$

$$\begin{aligned} &P(Age = 40|Drives = Yes) \times P(Current\ Phone = Sony|Drives = Yes) \\ &\qquad \times P(Buys = No|Drives = Yes) \times P(Drives = Yes) = 0.0215 \times \frac{1}{5} \times \frac{3}{5} \times \frac{5}{9} \\ &\qquad = 0.00143 \end{aligned}$$

$P(Drives\ Car = No|Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)$

$$= \frac{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No|Drives\ Car = Yes\ \times P(Drives\ Car = No)}{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)}$$

$$= \frac{P(Age = 40|Drives = No) \times P(Current\ Phone = Sony|Drives = No) \times P(Buys = No|Drives = No) \times P(Drives = No)}{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)}$$

$$P(Age = 40|Drives = No) = \frac{1}{\sqrt{2 \times \pi \times 72}} e^{-\frac{(40-24)^2}{2x72}} = 0.0079$$

$$P(Current\ Phone = Sony|Drives = No) = \frac{P(Current\ Phone = Sony \cap Drives = No)}{P(Drives = No)} = \frac{2}{4}$$

$$P(Buy = No|Drives = No) = \frac{P(Buy = No \cap Drives = No)}{P(Drives = No)} = \frac{3}{4}$$

$$P(Drives = No) = \frac{4}{9}$$

$$P(Age = 40|Drives = No) \times P(Current\ Phone = Sony|Drives = No)$$
$$\times P(Buys = No|Drives = No) \times P(Drives = No) = 0.0079 \times \frac{2}{4} \times \frac{3}{4} \times \frac{4}{9}$$
$$= 0.00131$$

Therefore, Customer 009 does not drive a car as the probability is lower (0.00131 < 0.00143).

(f):

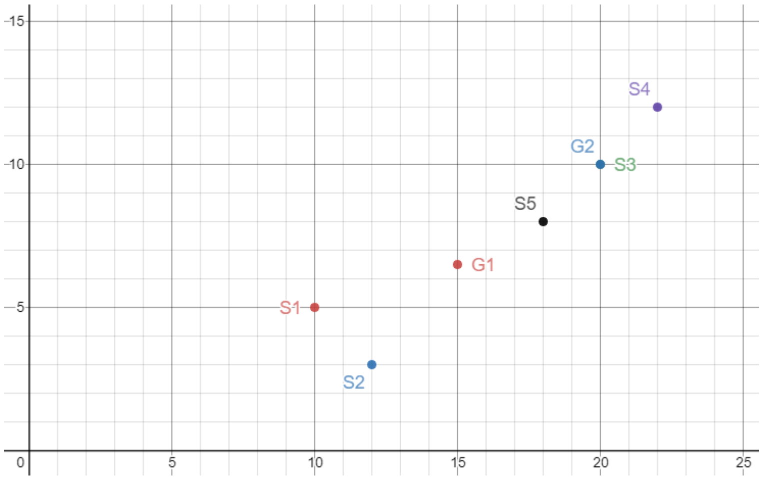
$$P(Current\ Phone = iPhone|Buy = Yes) = \frac{2}{3}$$
$$P(Current\ Phone = Samsung|Buy = Yes) = \frac{1}{3}$$
$$P(Current\ Phone = Sony|Buy = Yes) = 0$$
$$P(Current\ Phone = iPhone|Buy = No) = \frac{1}{7}$$
$$P(Current\ Phone = Samsung|Buy = No) = \frac{2}{7}$$
$$P(Current\ Phone = Sony|Buy = No) = \frac{4}{7}$$

$$P(Drive\ Car = Yes|Buy = Yes) = \frac{2}{3}$$
$$P(Drive\ Car = No|Buy = Yes) = \frac{1}{3}$$
$$P(Drive\ Car = Yes|Buy = No) = \frac{3}{7}$$
$$P(Drive\ Car = No|Buy = No) = \frac{3}{7}$$

For age, it depends on the age of the customer.

Question 3:

(a):



$$G1\ Centroid: \left(\frac{10+20}{2}, \frac{3+10}{2}\right) = (15,6.5)$$

$$G2\ Centroid: \left(\frac{22+18}{2}, \frac{12+8}{2}\right) = (20,10)$$

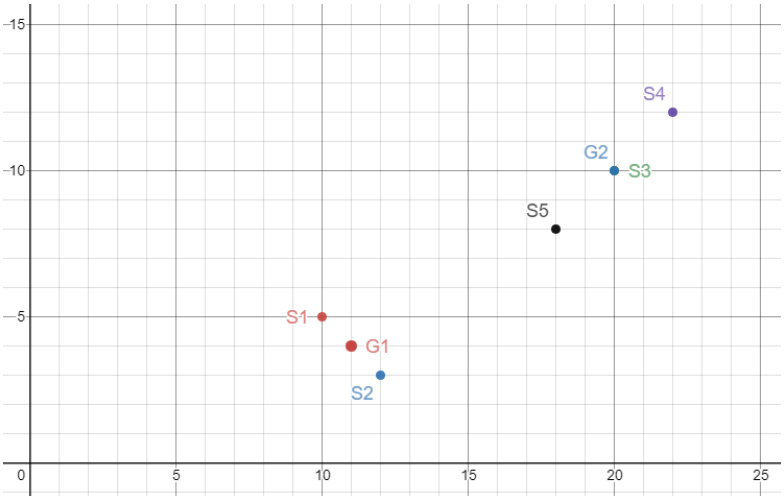
Next Phase,

G1: S1, S2

G2: S3, S4, S5

$$G1\ Centroid: \left(\frac{10+12}{2}, \frac{5+3}{2}\right) = (11,4)$$

$$G2\ Centroid: \left(\frac{22+18}{2}, \frac{12+8}{2}\right) = (20,10)$$



No more changes.

G1: S1, S2

G2: S3, S4, S5

(b):

	S1	S2	S3	S4	S5
S1	0	0.0282	0.1118	0.1389	0.0854
S2		0	0.1063	0.1345	0.0781
S3			0	0.0282	0.0282
S4				0	0.0565
S5					0

Join S3 & S4

	S1	S2	S3&S4	S5
S1	0	0.0282	0.1253	0.0854
S2		0	0.1204	0.0781
S3&S4			0	0.0423
S5				0

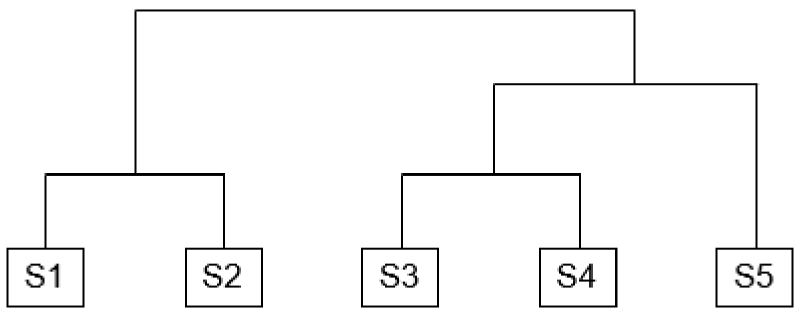
Join S1 & S2

	S1&S2	S3&S4	S5
S1&S2	0	0.1240	0.0817
S3&S4		0	0.0423
S5			0

Join S3&S4 & S5

	S1&S2	S3&S4&S5
S1&S2	0	0.1028
S3&S4		0

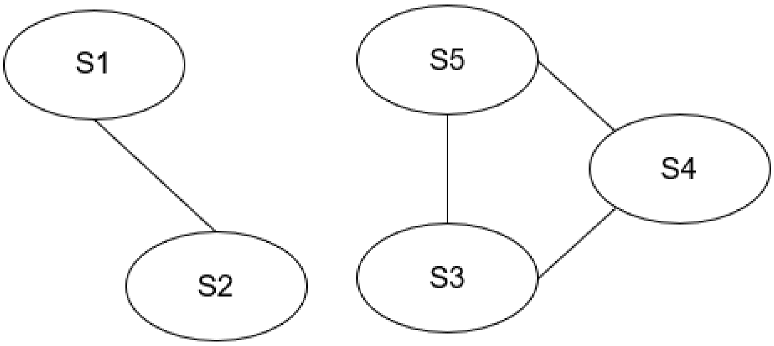
Join S1&S2 & S3&S4&S5



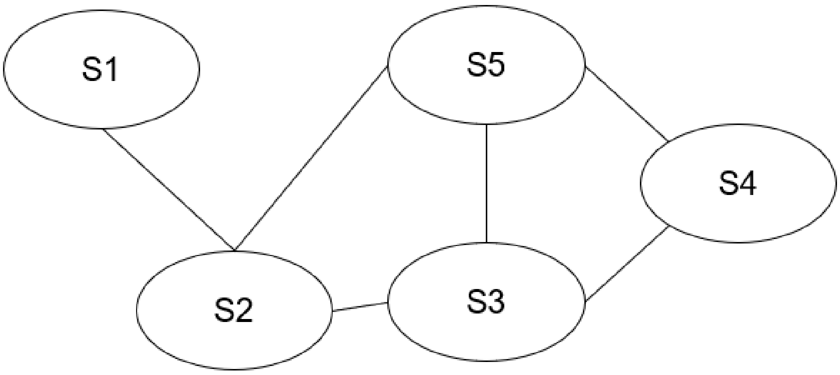
(c):

	Neighbor 1	Neighbor 2
S1	S2	S5
S2	S1	S5
S3	S4	S5
S4	S3	S5
S5	S3	S4

K = 2, T = 1



	Neighbor 1	Neighbor 2	Neighbor 3
S1	S2	S5	S3
S2	S1	S5	S3
S3	S4	S5	S2
S4	S3	S5	S2
S5	S3	S4	S2



Question 4:

(a):

(i):

A	5
B	5
C	5
D	2
E	4

→

AB	4
AC	3
AE	3
BC	4
BE	3
CE	2

→

ABC	3

Min sup = 50% > 3.5

Therefore, itemset = a (closed), b (closed), c (closed), e (max), ab (max), bc (max)

(ii):

a → b confidence = 4/5 = 0.8

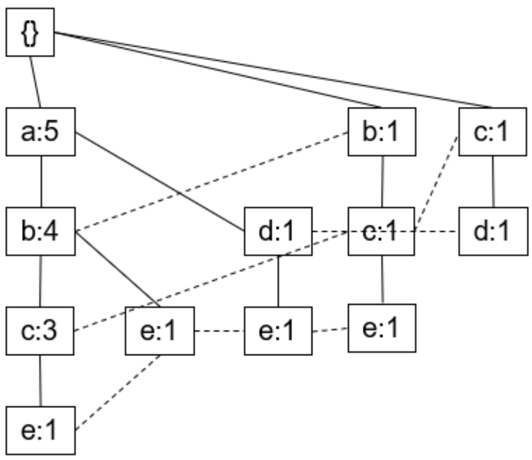
b → a confidence = 4/5 = 0.8

b → c confidence = 4/5 = 0.8

c → b confidence = 4/5 = 0.8

(iii):

a	5
b	5
c	5
e	4
d	2



(b):

(i):

Contextual Anomaly:

- An individual data instance is anomalous within a context.
- Requires a notion of context.

Collective Anomaly:

- Collection of related data instances is anomalous.
- Requires a relationship among data instances.

(ii):

Advantage:

- Utilize existing statistical modeling techniques to model various types of distributions.

Disadvantage:

- With high dimension, difficult to estimate distributions.
- Parametric assumptions often do not hold for real datasets.

Amendments to answer key:

1(b):

Both **simple matching and jaccard coefficient** do not apply here as the vectors are not binary.

Supremun Distance:

$$\max(|p_1 - q_1|, |p_2 - q_2|, \ldots, |p_{n-1} - q_{n-1}||p_n - q_n|) = 2$$

Cosine Similarity:

$$\frac{(p \cdot q)}{\|p\|\|q\|} = \frac{11}{\sqrt{13}\sqrt{20}} \approx 0.6822$$

Correlation:

$$\frac{covariance(p,q)}{std(p) \times std(q)} = \frac{4}{9 \times 0.9487 \times 1.0541} \approx 0.4444$$

2(d):

Question states to perform binarization into binary attributes in 2(b), hence for Q2(b):

$$\frac{3 + (2 \times 0.5)}{10} = 0.4$$

2(e):

The calculation is done using population formula and certain values are wrong, should have used **sample formula.**

If Drive Car = Yes:

- $Sample\ Mean = \frac{40+28+38+50+50}{5} = 41.2$
- $\boldsymbol{Sample\ Variance = (40 - 41.2)^2 + \cdots + (50 - 41.2)^2 = 85.2}$

If Drive Car = No:

- $Sample\ Mean = \frac{24+18+30+24}{4} = 24$
- $\boldsymbol{Sample\ Variance = (24 - 24)^2 + \cdots + (24 - 24)^2 = 24}$

$$P(Drives\ Car = Yes|Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)$$

$$= \frac{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No|Drives\ Car = Yes\ \times P(Drives\ Car = Yes)}{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)}$$

$$= \frac{P(Age = 40|Drives = Yes) \times P(Current\ Phone = Sony|Drives = Yes) \times P(Buys = No|Drives = Yes) \times P(Drives = Yes)}{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)}$$

$$\boldsymbol{P(Age = 40|Drives = Yes) = \frac{1}{\sqrt{2 \times \pi \times 85.2}} e^{-\frac{(40-41.2)^2}{2x85.2}} = 0.04359}$$

$$P(Current\ Phone = Sony|Drives = Yes) = \frac{P(Current\ Phone = Sony \cap Drives = Yes)}{P(Drives = Yes)} = \frac{1}{5}$$

$$P(Buy = No|Drives = Yes) = \frac{P(Buy = No \cap Drives = Yes)}{P(Drives = Yes)} = \frac{3}{5}$$

$$P(Drives = Yes) = \frac{5}{9}$$

$$\begin{aligned}\boldsymbol{P(Age = 40|Drives = Yes) \times P(Current\ Phone = Sony|Drives = Yes)} \\ \boldsymbol{\times P(Buys = No|Drives = Yes) \times P(Drives = Yes)} \\ \boldsymbol{= 0.04359 \times \frac{1}{5} \times \frac{3}{5} \times \frac{5}{9} = 0.002906}\end{aligned}$$

$$P(Drives\ Car = No|Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)$$

$$= \frac{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No|Drives\ Car = Yes\ \times P(Drives\ Car = No)}{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)}$$

$$= \frac{P(Age = 40|Drives = No) \times P(Current\ Phone = Sony|Drives = No) \times P(Buys = No|Drives = No) \times P(Drives = No)}{P(Age = 40\ and\ Current\ Phone = Sony\ and\ Buy = No)}$$

$$\boldsymbol{P(Age = 40|Drives = No) = \frac{1}{\sqrt{2 \times \pi \times 24}} e^{-\frac{(40-24)^2}{2x24}} = 16.86715}$$

$$P(Current\ Phone = Sony|Drives = No) = \frac{P(Current\ Phone = Sony \cap Drives = No)}{P(Drives = No)} = \frac{2}{4}$$

$$P(Buy = No|Drives = No) = \frac{P(Buy = No \cap Drives = No)}{P(Drives = No)} = \frac{3}{4}$$

$$P(Drives = No) = \frac{4}{9}$$

$$\begin{aligned}\boldsymbol{P(Age = 40|Drives = No) \times P(Current\ Phone = Sony|Drives = No)} \\ \boldsymbol{\times P(Buys = No|Drives = No) \times P(Drives = No)} \\ \boldsymbol{= 16.86715 \times \frac{2}{4} \times \frac{3}{4} \times \frac{4}{9} = 2.8112}\end{aligned}$$

Therefore, Customer 009 does not drive a car as the probability is lower (**0.4359 < 2.8112**).

3(a):

Iteration 0:

$$G1\ Centroid: \left(\frac{10+12+20}{3}, \frac{5+3+10}{3}\right) = (14, 6)$$

Iteration 1:

$$G2\ Centroid: \left(\frac{20+22+18}{3}, \frac{10+12+8}{3}\right) = (20, 10)$$

Question 4(a)(iii):

Removed d as it is not frequent.

a	5
b	5
c	5
e	4