Final Exam

COMP9318

By: Arth Sanskar Patel
z5228942

Q1

a. When there are n dimensions, the total total cuboids in the data cube are 2^n. Hence here as n = 3, the total cuboids will be $2^3 = 8$.

Considering Dimensions as (A, B, C), the Cuboids will be $(A, B, C), (A, B), (A, C), (B, C), (A), (B), (C), ()$. The Empty $()$ will be the ALL row in the data cuboid.

b. Considering that each dimension has p distinct values and there are n dimensions, the maximum tuples and minimum tuples are as follows:

$$MaximumTuples = p^n$$
$$MinimumTuples = p$$

Hence here $n = 3$ there fore assuming p distinct values,

$$MaximumTuples = p^3$$
$$MinimumTuples = p$$

c. The table will be as follows:

A	B	C	M
1	1	2	10
1	1	ALL	10
1	2	1	20
1	2	ALL	20
1	ALL	1	20
1	ALL	2	10
1	ALL	ALL	30
2	1	1	30
2	1	ALL	30
2	ALL	1	30
2	ALL	ALL	30
ALL	1	1	30
ALL	1	2	30
ALL	1	ALL	40
ALL	2	1	20
ALL	2	ALL	20

A	B	C	M
ALL	ALL	1	50
ALL	ALL	2	10
ALL	ALL	ALL	60

As there are 3 tuples, the single tuple optimisation will be applied 3 times.

Q.2>

ID	x_1	x_2
1	1.90	0.97
2	5.98	2.68
3	2.68	1.18
4	3.14	4.24
5	1.54	1.80
6	3.82	4.50
7	5.74	3.84
8	2.46	1.86
9	3.17	4.96
10	5.44	3.18

For the table above:

For point ID = 1. (1.90, 0.97).

$$D(1,2) = \sqrt{(1.9-5.98)^2 + (2.68-0.97)^2}$$

$D(1,2) = 4.423$

similarly.

$D(1,3) = 0.807$

$D(1,4) = 3.497$

$D(1,5) = 0.904$

$D(1,6) = 4.018$

$D(1,7) = 4.794$

$D(1,8) = 1.051$

$D(1,9) = 4.872$

$D(1,10) = 4.173$

For ID = 6.

$D(6,1) = 4.018$

$D(6,2) = 2.82$

$D(6,3) = 3.51$

$D(6,4) = 0.72$

$D(6,5) = 3.53$

$D(6,7) = 2.03$

$D(6,8) = 2.96$

$D(6,9) = 0.796$

$D(6,10) = 2.089$.

For ID =

$D(10,1) = 4.173$

$D(10,2) = 0.735$

$D(10,3) = 3.408$

$D(10,4) = 2.532$

$D(10,5) = 4.136$

$D(10,6) = 2.089$

$D(10,7) = 0.724$

$D(10,8) = 3.259$

$D(10,9) = 2.84$

Hence the first Cluster will be as follows.

(1) $\{1, 3, 5, 8\}$

$\{6, 4, 9\}$

$\{10, 2, 7\}$

b) The new centroids for the clusters will be.

$\{1, 2, 3, 5, 8\} \Rightarrow (2.145, 1.45)$

$\{6, 4, 9\} \Rightarrow (3.376, 4.56)$

$\{10, 7\} \Rightarrow (5.72, 3.23)$

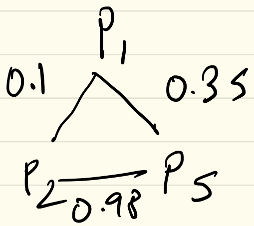
These values are calculated mean for all the clusters hence new centroids.

c) No, even when using manhattan distance, the way of choosing centroids will still be taking the mean of distances found in a cluster.

Q.3>

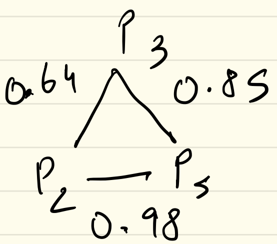
	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	0.1	0			
P_3	0.41	0.64	0		
P_4	0.55	0.47	0.44	0	
P_5	0.35	0.98	0.85	0.76	0

1) Highest is $P_2 P_5$, hence merge $P_2 P_5$.



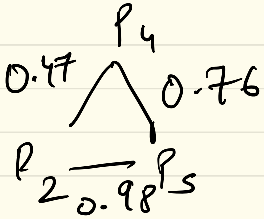
$$\frac{(0.98 + 0.1 + 0.35) \times 2}{(2+1) \times (3-1)}$$

$$= 0.477$$



$$= \frac{(0.98 + 0.64 + 0.85) \times 2}{(2+1) \times (3-1)}$$

$$= 0.823$$



$$\frac{(0.98 + 0.47 + 0.76) \times 2}{(2+1) \times (3-1)}$$

$$= 0.737.$$

Hence new matrix is

	P_1	P_2P_5	P_3	P_4
P_1	0			
P_2P_5	0.477	0		
P_3	0.41	0.823	0	
P_4	0.55	0.737	0.44	0

2) Here again highest is P_3, P_2P_5, hence merge P_2P_5 and P_3.

doing similar calculation from above.

$$\text{sim}(235, 1) = \frac{(0.98 + 0.64 + 0.85 + 0.1 + 0.35 + 0.41) \times 2}{(3+1) \times (4-1)}$$

$$sim(235, 1) = 0.555.$$

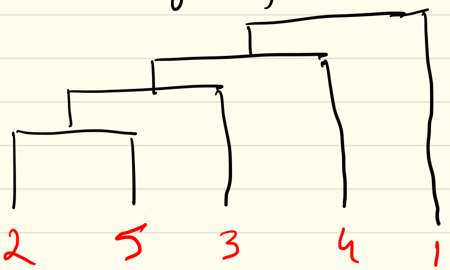
similarly.

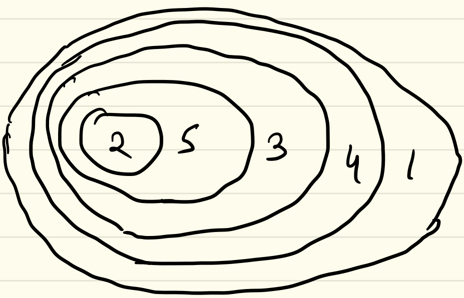
$$sim(235, 4) = 0.69.$$

Hence new matrix is

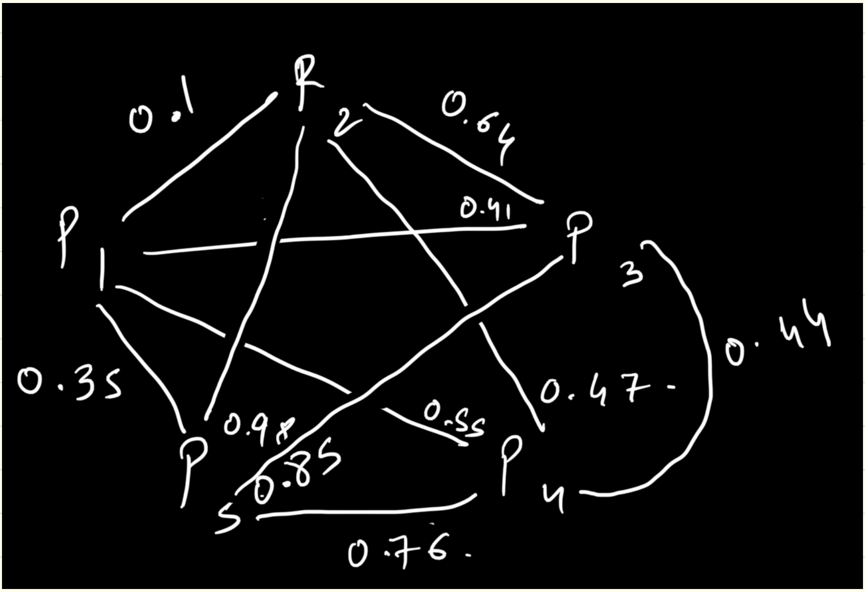
	P_1	$P_2 P_5 P_3$	P_4
P_1	0		
$P_2 P_5 P_3$	0.555	0	
P_4	0.55	0.69	0

Here highest is $P_2 P_5 P_3$ and P_4. Hence merging those to get final dendogram:





5) 1) The graph is below.



2)

Q4

a. Transformed Table will be as follows:

ID	a_1	a_2	a_3'	Class
1	T	T	L	Y
2	T	T	H	Y
3	T	F	M	N
4	F	F	M	Y
5	F	T	H	N
6	F	T	M	N
7	F	F	H	N
9	T	F	H	Y
9	F	T	M	N

b. The first splitting condition will be as follows:

Distribution for a_1 and Class		a_1	
		T	F
Class	Yes	3	1
	No	1	4
Total		**4**	**5**

Distribution for a_2 and Class		a_2	
		T	F
Class	Yes	2	2
	No	3	2
Total		**5**	**4**

Distribution for a_3 and Class		a_1		
		L	M	H
Class	Yes	1	1	2
	No	0	3	2
Total		**1**	**4**	**4**

Attribute	Gini Index
a_1	0.344
a_2	0.489
a_3	0.389

Hence here the lowest Gini Index is for both a_1 so we can choose a_1 as the first split for our Tree.

c. See Below

Q4) c) a_1

	Y	N	P(Y)	P(N)
T	3	1	3/4	1/5
F	1	4	1/4	4/5
	4	5		

a_2.

	Y	N	P(Y)	P(N)
T	2	3	2/4	3/5
F	2	2	2/4	2/
	4	5		

$P(yes) = 4/9$

$P(No) = 5/9$

a_3

	Y	N	P(Y)	P(N)
L	1	0	1/4	0
M	1	3	1/4	3/5
H	2	2	2/4	2/4
	4	5		

$10 = \{T, F, X\}$

$$P(yes|10) = \frac{P(T|yes) \times P(F|yes) \times P(X|yes) \times P(yes)}{P(10)}$$

$$P(No|10) = \frac{P(T|No) \times P(F|No) \times P(X|\cancel{No})\ P(No)}{P(10)}.$$

$$P(Yes|X) \propto 3/4 \times 2/4 \times \frac{1+0.5}{4+3\times 0.5} \times \frac{4}{9} \simeq 0.045$$

$$P(No|X) \propto 1/5 \times 2/5 \times \frac{1+0.5}{5+3\times 0.5} \times 5/9 \simeq 0.0102$$

Now as

$$P(Yes|10) + P(No|10) = 1$$

$$\therefore\ P(Yes|10) = \frac{0.045}{0.045+0.0102}$$

$$= 0.815$$

$$P(No|10) = 0.185$$

Since $P(Yes \mid 10) > P(No \mid 10)$.

hence class for $10 = \{T, F, X\}$

is Yes or Y

Q6

a.

First lets prove that s’ is also a frequent items

Let s be a frequent itemset. Let min sup be the minimum support. Let D be the task-relevant data, a set of database transactions. Let |D| be the number of transactions in D. Since s is a frequent itemset support count(s) = min sup × |D|.
Let $s^{'}$ be any nonempty subset of s. Then any transaction containing itemset s will also contain itemset $s^{'}$. Therefore, support count(s’) ≥ support count(s) = min sup × |D|. Thus, $s^{'}$ is also a frequent itemset.

Now,

Let D be the task-relevant data, a set of database transactions. Let |D| be the number of transactions in D. By definition,
support(s) = $\dfrac{support - count(s)}{|D|}$
Let $s^{'}$ be any nonempty subset of s. By definition, support(s’) = $\dfrac{support - count(s')}{|D|}$
From above we know that support(s’) ≥ support(s). This proves that the support of any nonempty subset s’ of itemset s must be as great as the support of s.

b. See below.

Q.6

b) To prove $conf(S \to A/S) \geq Conf(S' \to A|S')$

Lets assume the above statement to be not true the we can prove by contradiction.

lets assume.

$$conf(S \to A|S) < conf(S' \to A|S')$$

$$\therefore P((A-S)|S) < P((A-S')|S')$$

$$\therefore \frac{sup((A-S) \cup S)}{sup(S)} < \frac{sup((A-S') \cup S')}{sup(S')}$$

$$\therefore sup(S) > sup(S') \quad — (1)$$

Now from part (a) we know that

$$sup(S') > sup(S)$$

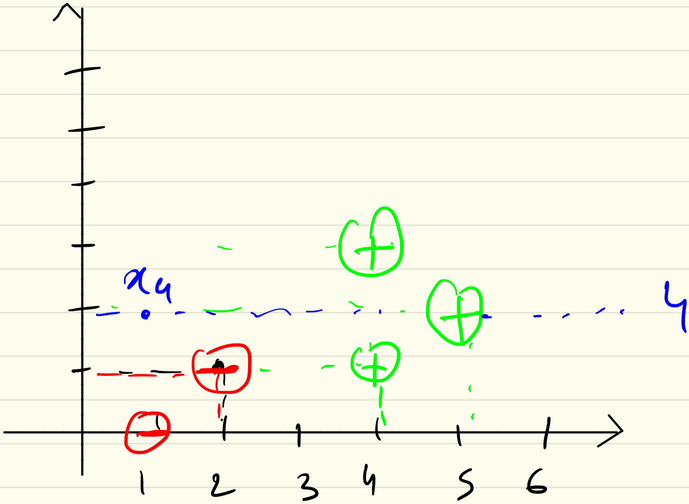
Hence (1) is a contradiction and our assumption is wrong.

Therefore.

$$conf(S \to A \setminus S) \geq conf(S' \to A \setminus S').$$

Q5)

a)

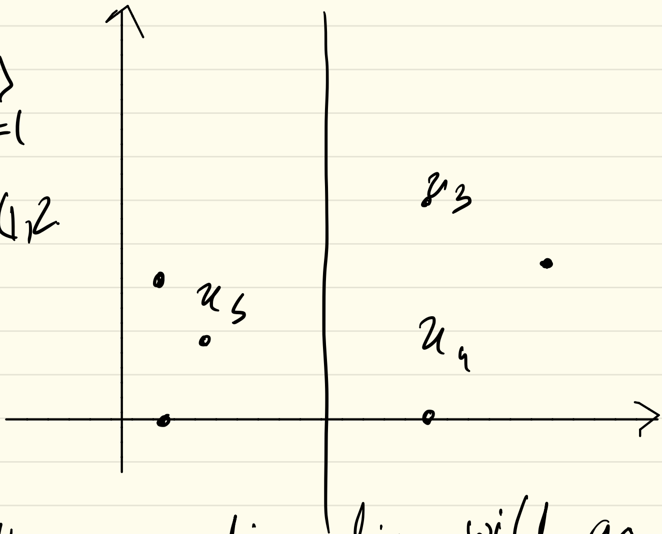


From the graph, $k < 4$.

b)

$k=1$

$x_6(1,2$



The seperation line will go though $x=3$

c) The closest point to separation line is x_5 (2, 1).

∴ The range for h in order not change the separation line is

$$\underline{\underline{h \leq 2}}$$

d) Critical training samples are the closest to the separation line. Here, $\underline{\underline{X_5, X_1, X_3}}$ are closest.