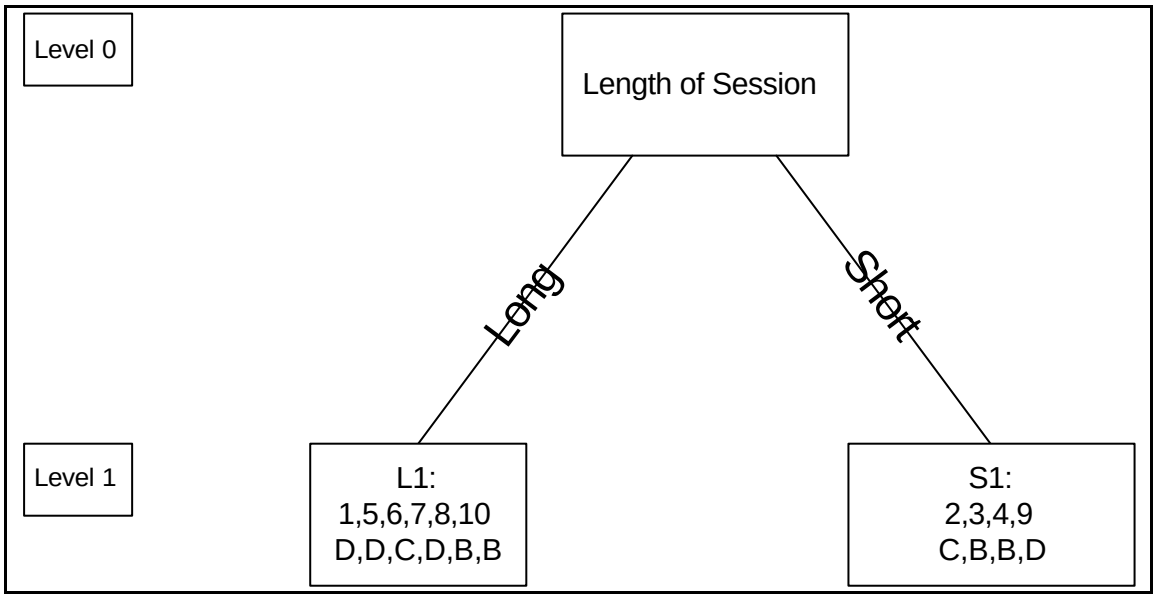
Supplementary Notes #9

COMP 578 Data Mining and Data Warehousing

Mid-term Quiz Suggested Solution

Q1. *(Total 15 marks)*

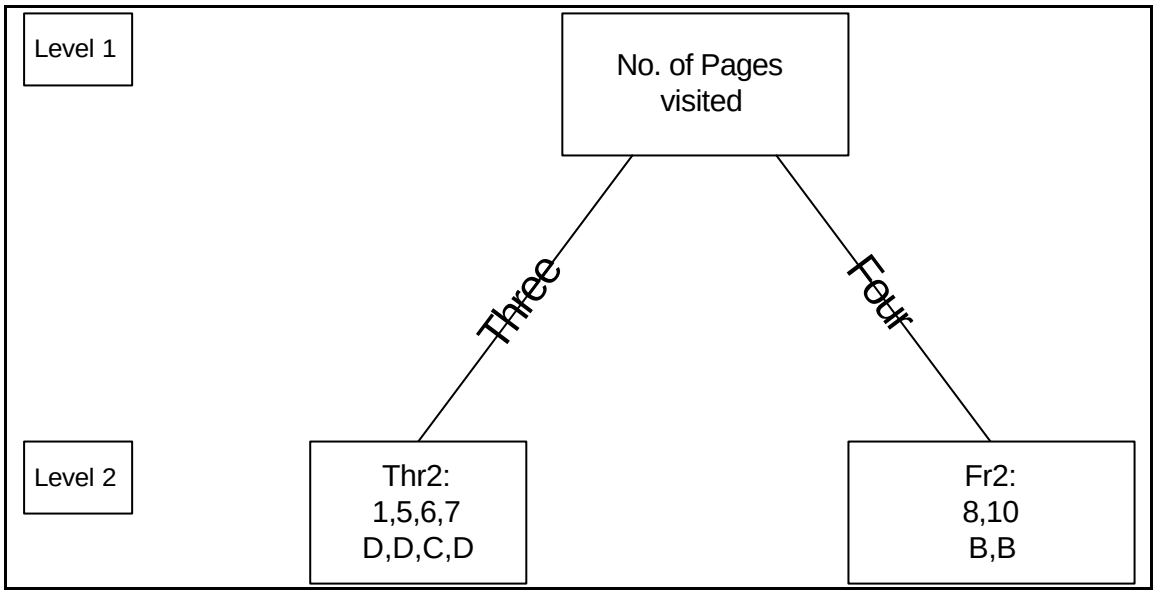
Given "Length of Session" as the root of the tree (let D is "Did not buy", C is "Bought computer book" and B is "Bought business book", session 0013 – 1, 0024 – 2, 0035 – 3, 0014 – 4, 0085 – 5, 0099 – 6, 0102 – 7, 1011 – 8, 1339 – 9, 2021 - 10):



Step 1: Considering the "Long" branch at Level 1:

Let U(L1) is an uncertainty associated with the L1 dataset:
U(L1)
$= -(3/6)\log_2(3/6) - (2/6)\log_2(2/6) - (1/6)\log_2(1/6)$
= 0.5 + 0.5283 + 0.4308
= 1.4591

In order to split the branch to Level 2, we need to calculate the information gain of "No. of Pages visited" and "Date of week":
Consider splitting "No. of Pages visited" *(3 marks)*



Let U(Thr2) is an uncertainty associated with the Thr2 dataset:
U(Thr2)

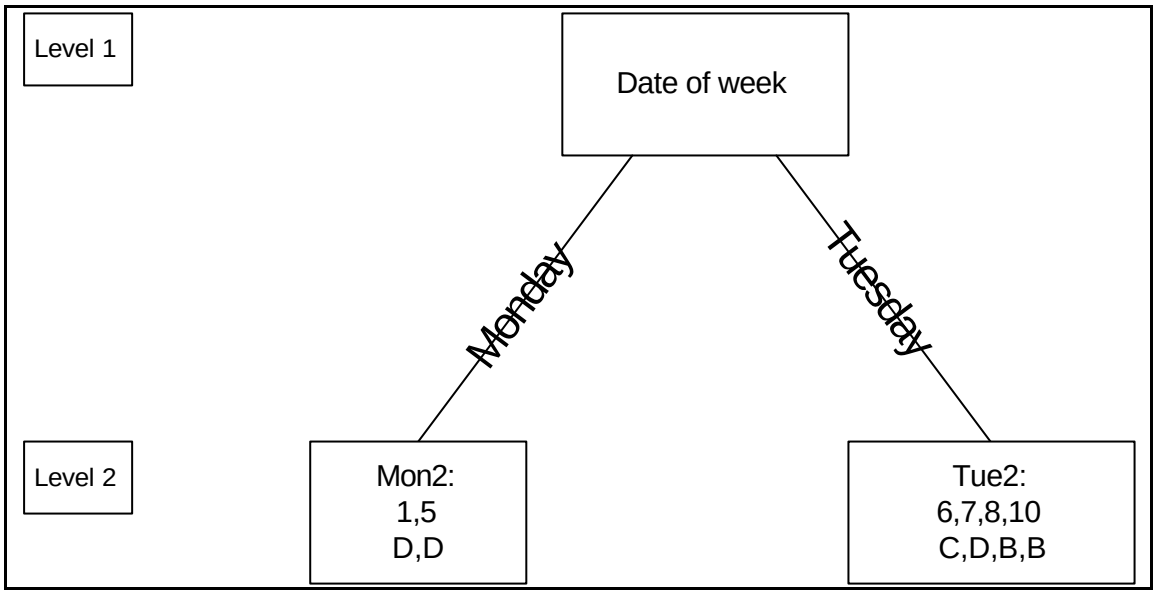
$= -(3/4) \log_2 (3/4) - (1/4) \log_2 (1/4)$
$= 0.3113 + 0.5$
$= 0.8113$

Let U(Fr2) is an uncertainty associated with the Fr2 dataset:
U(Fr2)
= ?

Average
$= (4/6)(0.8113) + (2/6)(0)$
$= 0.5409$

Information gain
$= 1.4591 - 0.5409$
$= 0.9182$

Consider splitting "Date of week" *(3 marks)*



Let U(Mon2) is an uncertainty associated with the Mon2 dataset:
U(Mon2)
= ?

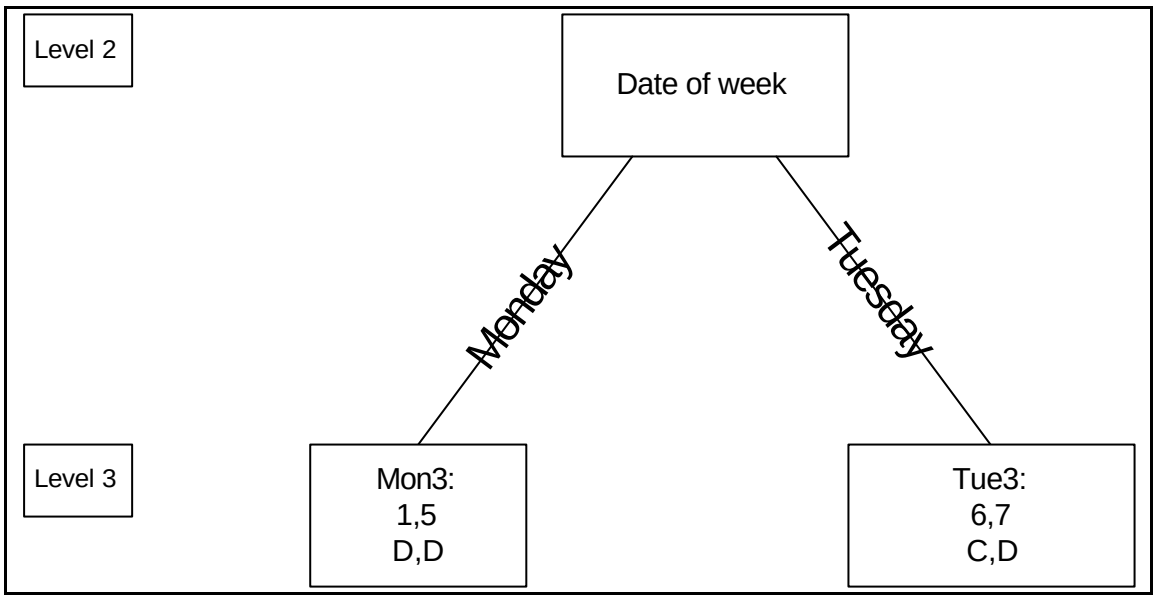
Let U(Tue2) is an uncertainty associated with the Tue2 dataset:
U(Tue2)
$= -(2/4) \log_2 (2/4) - (1/4) \log_2 (1/4) - (1/4) \log_2 (1/4)$
= 0.5 + 0.5 + 0.5
= 1.5

Average
= (2/6)(0) + (4/6)(1.5)
= 1

Information gain
= 1.4591 – 1
= 0.4591

Comparing the information gain of the above two attributes, “No. of Pages visited” is selected.

In order to split the branch to Level 3, now we consider splitting “Date of week” at the node “Thr2”: *(3 marks)*



Let U(Mon3) is an uncertainty associated with the Mon3 dataset:
U(Mon3)
= ?

Let U(Tue3) is an uncertainty associated with the Tue3 dataset:
U(Tue3)
$= -(1/2) \log_2 (1/2) - (1/2) \log_2 (1/2)$
= 0.5 + 0.5
= 1

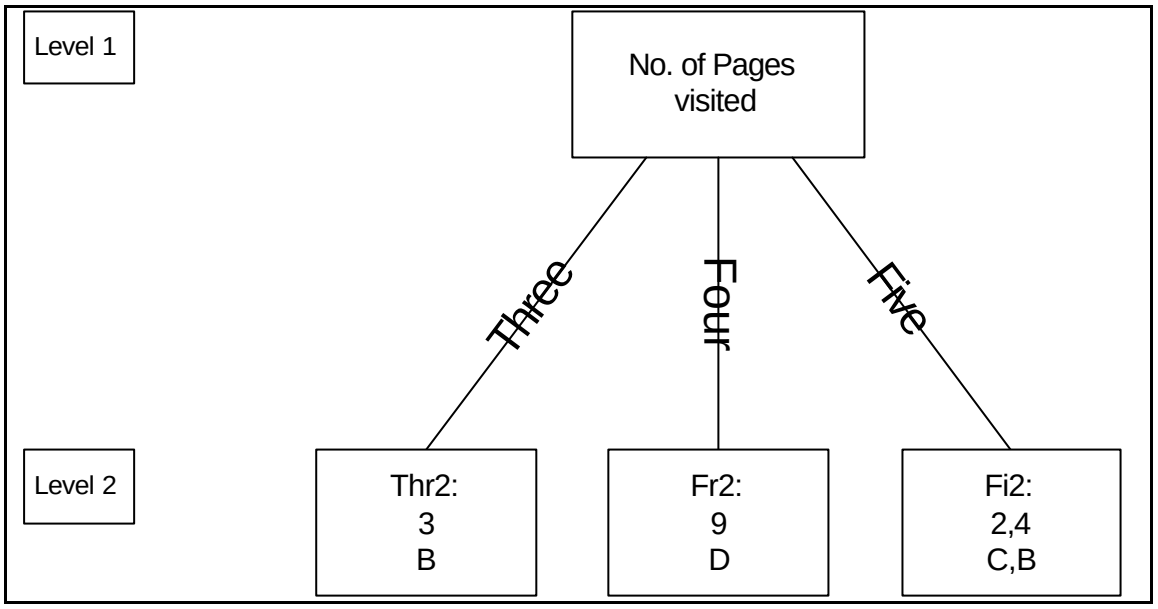
Average
= (2/4)(0) + (2/4)(1)
= 0.5

Information gain
= 0.8113 – 0.5
= 0.3113

Therefore, we select "Date of week" to split at the node "Thr2". Since, all attributes were considered in this branch, we couldn't split the node Tue3.

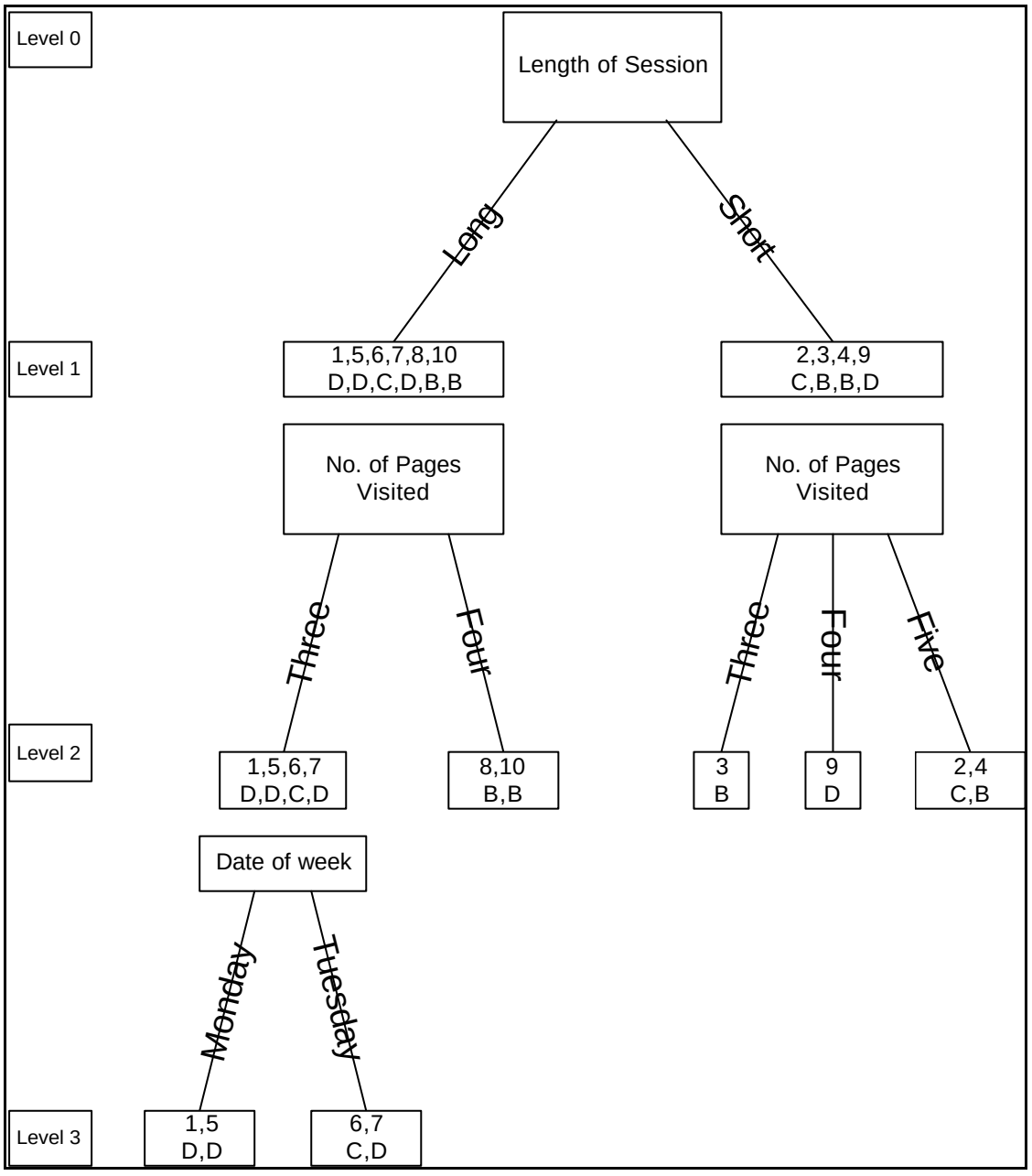
Step 2: Considering the "Short" branch at Level 1:

Given that the node is split by “No. of Pages visited” attribute,



At the node Fi2, 2 & 4 belong to the same value “Monday” of attribute “Date of week”, so we cannot split it. *(2 marks)*

The final ID3 decision tree is as follows: *(2 marks)*



According to the above decision tree, the classification accuracy of the testing set is 0.2 (20%). *(2 marks)*

Q2. *(Total 15 marks)*

a)

Data normalization: *(1 mark)*

Age = (X – 26)/(66 - 26), Income = (X – 66)/(92 – 66), and
Loan size = (X – 125)/(199 – 125)

Record No.	Age	Income	Loan Size
1	0.75	1	0.473
2	0.525	0.846	0.189
3	0.15	0.923	1
4	0	0.231	0.716
5	1	0	0

Iteration 1: *(3 marks)*

Cluster 1 Center (0.75, 1, 0.473)		**Cluster 2 Center (0.525, 0.846, 0.189)**	
Record No.	**Euclidean Distance**	**Record No.**	**Euclidean Distance**
3	*0.802*	3	0.897
4	1.101	4	*0.965*
5	1.134	5	*0.989*

The discovered clusters are:
Cluster 1 – {1, 3}, Cluster 2 – {2, 4, 5}

Iteration 2: *(3 marks)*

New center of cluster 1
= {(0.75+0.15)/2, (1+0.923)/2, (0473+1)/2}
= (0.45, 0.962, 0.737)

New center of cluster 2
= {(0.525+0+1)/3, (0.846+0.231+0)/3, (0.189+0.716+0)/3}
= (0.508, 0.359, 0.302)

Cluster 1 Center (0.45, 0.962, 0.737)		**Cluster 2 Center (0.508, 0.359, 0.302)**	
Record No.	**Euclidean Distance**	**Record No.**	**Euclidean Distance**
1	*0.401*	1	0.706
2	0.565	2	*0.5*
3	*0.401*	3	0.966
4	0.859	4	*0.668*
5	1.331	5	*0.68*

The discovered clusters are:
Cluster 1 – {1, 3}, Cluster 2 – {2, 4, 5}

There are no changes in the grouping -> Stop

Discussion: *(2 marks)*
Need to mention that any evidence to support your friend' s belief based on your discovered results.

b) Condorset algorithm *(4 marks)*

After transforming the quantitative variables into qualitative variables:

Record No.	Age	Income	Marital Status	Account Balance	Loan Size
1	Old	High	M	High	Medium
2	Middle	High	M	Low	Small
3	Young	High	S	Low	Large
4	Young	Low	S	High	Large
5	Old	Low	M	Low	Small

Step 1 -
Record 1 is assigned to cluster 1.

Step 2 -
Addition of record 2:
Overall score for cluster 1 = -1
- ⇨ Record 2 is assigned to a new cluster (cluster 2).

Step 3 –
Addition of record 3:
Overall score for cluster 1 = -3
Overall score for cluster 2 = -1
- ⇨ Record 3 is assigned to a new cluster (cluster 3).

Step 4 –
Addition of record 4:
Overall score for cluster 1 = -3
Overall score for cluster 2 = -5
Overall score for cluster 3 = 1
- ⇨ Record 4 is assigned to cluster 3.

Step 5 –
Addition of record 5:
Overall score for cluster 1 = -1
Overall score for cluster 2 = 1
Overall score for cluster 3 = -6

⇨ Record 5 is assigned to cluster 2

Discussion: *(2 marks)*
- The discovered clusters are:
Cluster 1: {1}, Cluster 2: {2, 5}, Cluster 3: {3, 4}
- Any interesting patterns found in the discovered clusters?

Q3. *(Total 15 marks)*

a) Apriori algorithm *(10 marks)*

Let A – Orange, B – Coke, C – Apple, D – Diapers, E – Pepsi, and F – Lemon, min. support – 25% (0.25), min. confidence – 40% (0.4).

Itemset	**Count**	**Support**
A	6	0.75
B	5	0.625
C	6	0.75
D	5	0.625
E	3	0.375
F	4	0.5

Itemset	**Count**	**Support**
AB	4	0.5
AC	5	0.625
AD	3	0.375
AE	2	0.25
AF	3	0.375
BC	5	0.625
BD	4	0.5
BE	*1*	*0.125 (deleted)*
BF	*1*	*0.125 (deleted)*
CD	4	0.5
CE	*1*	*0.125 (deleted)*
CF	2	0.25
DE	*1*	*0.125 (deleted)*
DF	2	0.25
EF	2	0.25

Itemset	**Count**	**Support**

ABC	4	0.5
ABD	3	0.375
ACD	3	0.375
ACF	2	0.25
ADF	*1*	*0.125 (deleted)*
AEF	*1*	*0.125 (deleted)*
BCD	4	0.5
BCF	*1*	*0.125 (deleted)*
CDF	*1*	*0.125 (deleted)*

All 3-itemset discovered:
{ABC, ABD, ACD, ACF, BCD}

All 3-item interesting association rules in the data set:

Rule	**Confidence**	**Rule**	**Confidence**
A -> BC	0.67	AC -> D	0.6
B -> AC	0.8	AD -> C	1
C -> AB	0.67	CD -> A	0.75
AB -> C	1	*A -> CF*	*0.33 (deleted)*
AC -> B	0.8	*C -> AF*	*0.33 (deleted)*
BC -> A	0.8	F -> AC	0.5
A -> BD	0.5	AC -> F	0.4
B -> AD	0.6	AF -> C	0.67
D -> AB	0.6	CF -> A	1
AB -> D	0.75	B -> CD	0.8
AD -> B	1	C -> BD	0.67
BD -> A	0.75	D -> BC	0.8
A -> CD	0.5	BC -> D	0.8
C -> AD	0.5	BD -> C	1
D -> AC	0.6	CD -> B	1

b) Lift ratio >= 1.75 *(2 marks)*

Rule	**Lift Ratio**	**Rule**	**Lift Ratio**
A -> BC	1.072	AC -> D	0.96
B -> AC	1.28	AD -> C	1.33
C -> AB	1.34	CD -> A	1
AB -> C	1.33	F -> AC	0.8
AC -> B	1.28	AC -> F	0.8
BC -> A	1.067	AF -> C	0.893
A -> BD	1	CF -> A	1.33
B -> AD	1.6	B -> CD	1.6
D -> AB	1.2	C -> BD	1.34
AB -> D	1.2	D -> BC	1.28
AD -> B	1.6	BC -> D	1.28
BD -> A	1	BD -> C	1.33

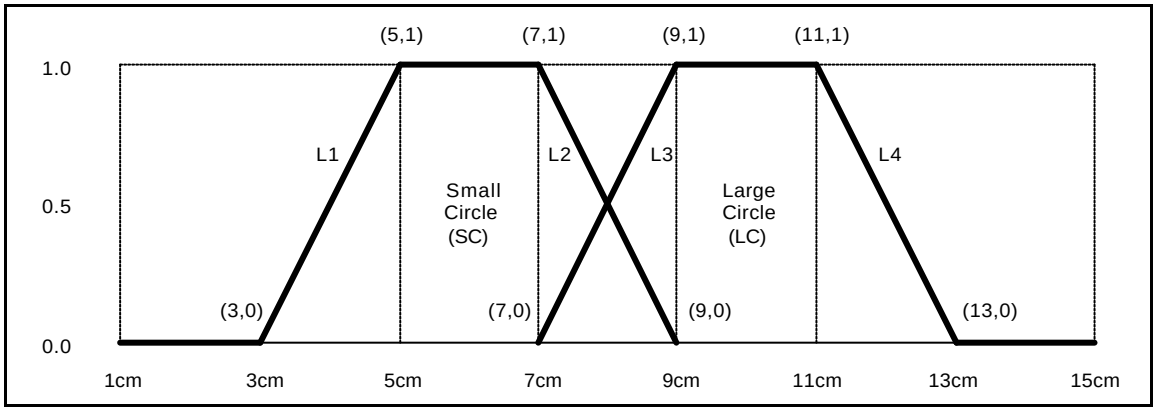
A -> CD	1	CD -> B	1.6
C -> AD	1.33		
D -> AC	0.96		

All rules discovered in a) are not interesting.

c) Discussion *(3 marks)*

For example, you may suggest increasing the min. confidence to reduce the number of interesting rules discovered in a).

Q4. Fuzzy membership function for a circle *(Total 5 marks)*



(x-axis is radius, and y-axis is degree of membership)

First, calculate the equation of L1, L2, L3 and L4:

L1: passing through two points (3,0) and (5,1),
$y = mx + c$
$0 = 3m + c$ ----- (1)
$1 = 5m + c$ ----- (2)
⇨ $y = (x - 3)/2$

L2: passing through two points (7,1) and (9,0),
$y = mx + c$
$1 = 7m + c$ ----- (1)
$0 = 9m + c$ ----- (2)
⇨ $y = (-x + 9)/2$

L3: passing through two points (7,0) and (9,1),

y = mx + c
0 = 7m + c ----- (1)
1 = 9m + c ----- (2)
⇨ y = (x – 7)/2

L4: passing through two points (11,1) and (13,0),
y = mx + c
1 = 11m + c ----- (1)
0 = 13m + c ----- (2)
⇨ y = (-x + 13)/2

Then, we can represent the fuzzy set SC and LC as follows:

$$SC = \left\{ (X, m_{SC}(X) \mid X \in [0cm, 15cm], m_{SC}(X) = \begin{bmatrix} 0 & SC(X) \le 3cm \\ (SC(X) - 3)/2 & 3cm \le SC(X) \le 5cm \\ 1 & 5cm \le SC(X) \le 7cm \\ (-SC(X) + 9)/2 & 7cm \le SC(X) \le 9cm \\ 0 & SC(X) \ge 9cm \end{bmatrix} \right\}$$

$$LC = \left\{ (X, m_{LC}(X) \mid X \in [0cm, 15cm], m_{LC}(X) = \begin{bmatrix} 0 & LC(X) \le 7cm \\ (LC(X) - 7)/2 & 7cm \le LC(X) \le 9cm \\ 1 & 9cm \le LC(X) \le 11cm \\ (-LC(X) + 13)/2 & 11cm \le LC(X) \le 13cm \\ 0 & LC(X) \ge 13cm \end{bmatrix} \right\}$$