CS422: Data Mining A20402686

Homework 2

Que 1.1 – Tan, Chapter 3

Exercise 8 - Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11?

Ans –
(a) The data is symmetrically distributed, minimum of the 75% of the data between first and third quartiles, if the line which is representing median of the data is at the middle of box. For rest of the data, the length of the whiskers and outliers is also an indication because these features don't involve as many points, they may be misleading.
(b) Sepal width and length is relatively symmetrically distributed, petal length is skewed, and petal width is little bit skewed.

Exercise 9 - Compare sepal length, sepal width, petal length, and petal width, using Figure 3.12.

Ans –
For Setosa, sepal length > sepal width > petal length > petal width.
For Versicolour and Virginiica, sepal length > sepal width and petal length > petal width,
Even if sepal length > petal length, petal length > sepal width.

Exercise 10 - Comment on the use of a box plot to explore a data set with four attributes: age, weight, height, and income.

Ans –
A lot of information can be obtained by looking at the box plots for each and, also for particular attribute with various categories of a second attribute. For example, if we compare the box plots of age for different categories of ages, we would see that weight and income increases with age.

Que 1.2 – Tan, Chapter 4

Exercise 2 - Consider the training examples shown in Table 4.1 for a binary classification problem.

(a) Compute the Gini index for the overall collection of training examples.
Ans-
Gini = $1 - 2 \times 0.5^2 = 0.5$.

(b) Compute the Gini index for the Customer ID attribute.
Ans-
The gini for each Customer ID value is 0. Hence, the overall gini for Customer ID attribute is 0.

(c) Compute the Gini index for the Gender attribute.

Ans-

The gini for Male is $1 - 2 \times 0.5^2 = 0.5$. The gini for Female is also 0.5.
Therefore, gini for Gender is $0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$.

(d) Compute the Gini index for the Car Type attribute using multiway split.

Ans-

The gini for Family car is 0.375, Luxury car is 0.2188, and Sports car is 0. The overall gini is 0.1625.

(e) Compute the Gini index for the Shirt Size attribute using multiway split.

Ans-

The gini for Small shirt size is 0.48, Medium shirt size is 0.4898, Large shirt size is 0.5, and Extra Large shirt size is 0.5. The overall gini for Shirt Size attribute is 0.4914.

(f) Which attribute is better, Gender, Car Type, or Shirt Size?

Ans-

Car Type is better because it has the lowest gini among the three attributes.

(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

Ans-

The attribute has no predictive power because new customers are assigned to new Customer IDs.

Exercise 3 - Consider the training examples shown in Table 4.2 for a binary classification problem.

(a) What is the entropy of this collection of training examples with respect to the positive class?

Answer:

There are four positive examples and five negative examples.
Hence, $P(+) = 4/9$ and $P(-) = 5/9$. The entropy of the training examples is,
$-4/9 \log_2(4/9) - 5/9 \log_2(5/9) = 0.9911$.

(b) What are the information gains of a1 and a2 relative to these training examples?

Answer:

The entropy for a1 is
$4/9\ [-(3/4) \log_2 (3/4) - (1/4) \log_2 (1/4)] + 5/9\ [-(1/5) \log_2 (1/5) - (4/5) \log_2 (4/5)] = 0.7616$
Therefore, the information gain for a1 is $0.9911 - 0.7616 = 0.2294$.

The entropy for a2 is
$5/9\ [-(2/5) \log_2 (2/5) - (3/5) \log_2 (3/5)] + 4/9\ [-(2/4) \log_2 (2/4) - (2/4) \log_2 (2/4)] = 0.9839$
Therefore, the information gain for a2 is $0.9911 - 0.9839 = 0.0072$.

(c) For a3, which is a continuous attribute, compute the information gain for every possible split.

Class	<=1	>1	<=3	>3	<=4	>4	<=5	>5	<=6	>6	<=7	>7	<=8	>8
-	0	5	1	4	1	4	3	2	3	2	4	1	5	0
+	1	3	1	3	2	2	2	2	3	1	4	0	4	0
Total	1	8	2	7	3	6	5	4	6	3	8	1	9	0

For 1.0 –

For class <=1.0,
Entropy = -(1/1) * log (1/1) - (0/1) * log (0/1) = 0
For class >1.0,
Entropy = -(3/8) * log (3/8) - (5/8) * log (5/8) = 0.95444
Information gain = (1.0) = 0.991 - (1/9) *0 - (8/9) *0.9544 = 0.1427

For 3.0 –

For class <=3.0,
Entropy = = -(1/2) * log (1/2) - (1/2) * log (1/2) = 1
For class >3.0,
Entropy = -(3/7) * log (3/7) - (4/7) * log (4/7) = 0 .9852
Information gain = 0.991 - (2/9) * 1 - (7/9) * 0.9852 = 0.0026

For 4.0 –

For class <=4.0,
Entropy = = -(2/3) * log (2/3) - (1/3) * log2(1/3) = 0.9183
For class >4.0,
Entropy = -(2/6) * log (2/6) - (2/6) * log (2/6) = 0.9183
Information gain = 0.991 - (3/9) * 0.9183 - (6/9) * 0.9183 = 0.0728

For 5.0 –

For class <=5.0,
Entropy = = -(2/5) * log (2/5) - (3/5) * log (3/5) = 0.971
For class >5.0,
Entropy = = -(2/4) * log (2/4) - (2/4) * log (2/4) = 1
Information gain = 0.991 - (5/9) * 0.971 - (4/9) * 1 = 0.0072

For 6.0 –
For class <=6.0,
Entropy = -(3/6) * log (3/6) - (3/6) * log (3/6) = 1
For class >6.0,
Entropy = -(1/3) * log (1/3) - (2/3) * log (2/3) = 0.9183
Information gain = 0.991 - (6/9) * 1 - (3/9) * 0.39 = 0.0183

For 7.0 –

For class <=7.0,
Entropy = = -(4/8) * log (4/8) - (4/8) * log (4/8) = 1
For class >7.0,
Entropy = -(0/1) * log (0/1) - (1/1) * log (1/1) = 0
Information gain = 0.991 - (8/9) * 1 - (1/9) * 0 = 0.1021

For 8.0 –

For class <=8.0,
Entropy = -(4/9) * log (4/9) - (5/9) * log (5/9) = 0.9911
For class >8.0,
Entropy = -(0/0) * log (0/0) - (0/0) * log (0/0) = 0
Information gain = 0.991 - (9/9) * 1 - (0/9) * 0= 0

(d) What is the best split (among *a*1, *a*2, and *a*3) according to the information gain?
Ans - According to information gain, a1 produces the best split.

(e) What is the best split (between *a*1 and *a*2) according to the classification error rate?
Ans –
For a1,
Classification error = 1 – max(p(i|t))
Error(T) = 1- max(3/4,1/4) = 1 – ¾ = 0.25
Error(F) = 1 – max(1/5,4/5) = 1 – 4/5 = 0.2

Total classification error = (4/9) * 0.25 + (5/9) * 0.2 = 0.2222

For a2,
Classification error = 1 – max(p(i|t))
Error(T) = 1- max(2/5,3/5) = 1 – 3/5 = 0.4
Error(F) = 1 – max(2/4,2/4) = 1 – 2/4 = 0.5

Total classification error = (5/9) * 0.4 + (4/9) * 0.5 = 0.4444
Therefore, according to error rate, a1 produces the best split.

f) What is the best split (between a1 and a2) according to the Gini index?
Ans –
For attribute a1, the gini index is
$4/9\ [1 – (3/4)^2 – (1/4)^2\ ] + 5/9\ [\ 1- (1/5)^2 – (4/5)^2\ ] = 0.3444$

For attribute a2, the gini index is

$5/9\ [1 – (2/5)^2 – (3/5)^2] + 4/9\ [1- (2/4)^2 – (2/4)^2] = 0.4889$

The gini index for a1 is smaller, hence it produces the better split.

Exercise 5 - Consider the following data set for a binary class problem.

(a) Calculate the information gain when splitting on *A* and *B*. Which attribute would the decision tree induction algorithm choose?

Ans –

The contingency tables after splitting on attributes A,

	A =T	A=F
+	4	0
-	3	3

The contingency tables after splitting on attributes B,

	B =T	B=F
+	3	1
-	1	5

Overall entropy before splitting,
$\text{Entropy}_{\text{before}} = − 0.4 \log 0.4 − 0.6 \log 0.6 = 0.9710$

Information gain at A = Entropy before - (7/10) * 0.9852 - (3/10) * 0 = 0.2813
Information gain at B = Entropy before - (4/10) * 0.8113 - (6/10) * 0.6500 = 0.2565

Therefore, attribute A will be chosen to split the node.

(b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
Ans –
The overall gini before splitting is:
$\text{Gini}_{\text{before}} = 1− 0.42 − 0.62 = 0.48$

The gain in gini after splitting on A is:
Gini of A(T) = $1− (4/7)^2 – (3/7)^2 = 0.4898$

Gini of A(F) = $1− (3/3)^2 – (0/3)^2 = 0$

Gain at A = $\text{Gini}_{\text{before}}$ − 7/10 Gini of A(T) − 3/10 Gini of A(F) = 0.1371

The gain in gini after splitting on B is:
Gini of B(T) = $1− (1/4)^2 – (3/4)^2 = 0.3750$

Gini of B(F) = $1− (1/6)^2 – (5/6)^2 = 0.2778$

Gain at B = $\text{Gini}_{\text{before}}$ − 4/10Gini of B(T) − 6/10Gini of B(F) = 0.1633

Therefore, attribute B will be chosen to split the node

(c) Figure 4.13 shows that entropy and the Gini index are both monotonously increasing on the range [0, 0.5] and they are both monotonously decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain

Ans –
Yes, even if these measures have similar range and somewhat similar behavior , their respective gains which are scaled differences of the measures, don't behave in the same way, looking at the results we got in above question a and b.

Que. 1.3 Tan, Chapter 5.

Exercise 20- Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "−." Half of the data set is used for training while the remaining half is used for testing.

(a) Suppose there are an equal number of positive and negative records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?
Ans –
If there are total 100 records, then 50 will be positive and 50 will be negative. As decision tree classifier predicts every test record to be positive, remaining 50 % will be of negative.
Hence answer is 50%.

(b) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2.
Ans-
If there are total 100 records, then 50 will be positive and 50 will be negative. Error rate of positive class will be 20 %(10 records) and 30%(40 records) of negative class. Hence 40+10 = 50 which is 50% of 100.
Hence answer is 50%.

(c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?
Ans –
If there are total 100 records, then 67 will be positive and 33 will be negative.
Hence answer is 33%.

(d) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3.

Ans –

If there are total 100 records, then 67 will be positive and 33 will be negative.
33 % error rate in positive (22.2) and 67% error rate in negative (which is 22.2)
Hence answer is 44.4%.

Question 1.4, Zaki, Chapter 8

Exercise 6 - Consider Figure 8.10. It shows a simple taxonomy on some food items. Each leaf is a simple item and an internal node represents a higher-level category or item. Each item (single or high-level) has a unique integer label noted under it. Consider the database composed of the simple items shown in Table 8.5 Answer the following questions:

(a) What is the size of the itemset search space if one restricts oneself to only itemsets composed of simple items?

Ans –

Size of itemset search = 2^{11}

(b) Let X = {x1, x2, ..., xk} be a frequent itemset. Let us replace some xi ∈ X with its parent in the taxonomy (provided it exists), to obtain X′, then the support of the new itemset X′ is:

i. more than support of X
ii. less than support of X
iii. not equal to support of X
iv. more than or equal to support of X
v. less than or equal to support of X

Ans –

iv. more than or equal to support of X

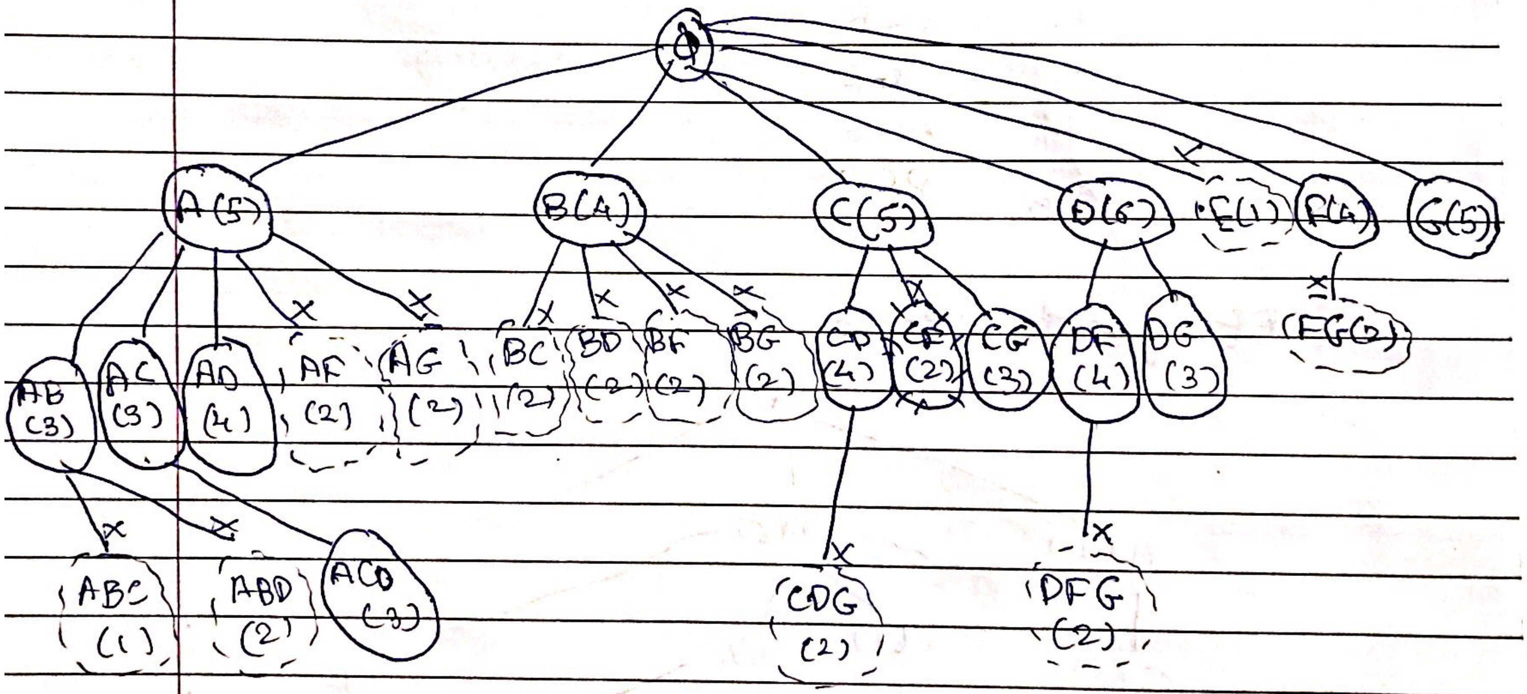
Q. 1-4 Zaki, chapter 8

Q. 1) a)

ans: -

minsup = 3

tid	itemset
t1	ABCD
t2	ACDF
t3	ACDEG
t4	ABDF
t5	BCG
t6	DFG
t7	ABG
t8	CDFG



Frequent itemset

Sup	Itemsets
6	D
5	A, C, G
4	B, F, AD, CD, DF
3	AC, CG, DG, ACD, AB

Q. 1) (b) Answer:

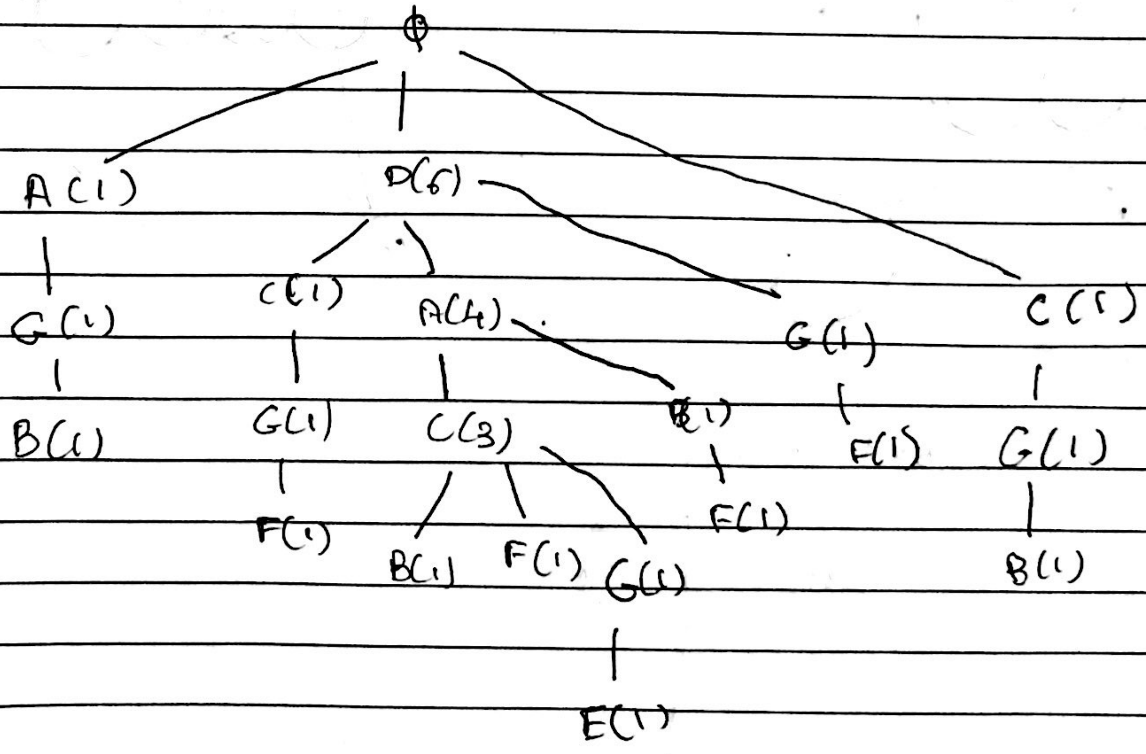
From given datasets,

D(6) A(5) C(5) G(5) B(4) F(4) E(1)

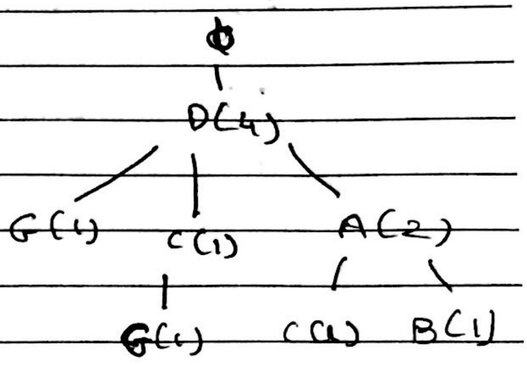
∴

t_{id}	itemsets
t_1	DACB
t_2	DACF
t_3	DACGE
t_4	DABF
t_5	CGB
t_6	DGF
t_7	AGB
t_8	DCGF

FP growth tree :-

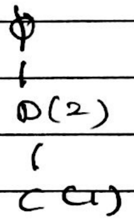


R_F(projection on F)



D GF	cnt (F) = 1
D ABF	cnt(F) = 1
DACF	cnt (F) = 1
DCGF	cnt (F) = 1

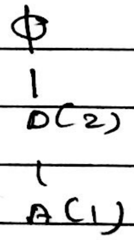
$(R_F)_G$: FG(2)



DG	cnt(G) = 1
D CG	cnt (G) = 1

output FGD (2)

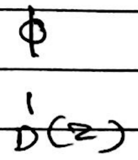
$(R_F)_C$: FC(2)



DC	cnt (c) = 1
DAC	cnt (c) = 1

Output = FCD(2)

$(R_F)_A$: FA(2)



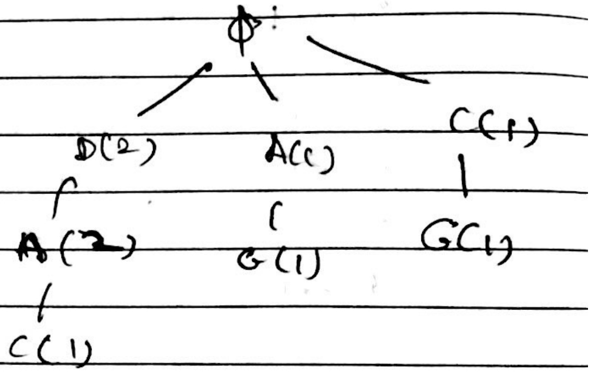
DA , cnt(A) = 2

output = FAD(2)

$(R_F)_D$: FD(4)

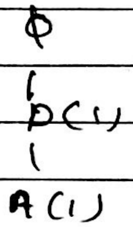
R_B (Projection on B)

CGB	cnt (B) = 1
DAB	cnt (B) = 1
DACB	cnt (B) = 1
AGB	cnt (B) = 1



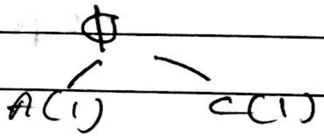
$(R_B)_C$: BC (2)

DAC	cnt (C) = 1
C	cnt (C) = 1



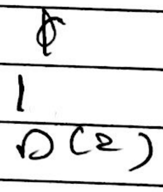
$(R_B)_G$: BG (2)

AG	cnt (G) = 1
CG	cnt (G) = 1



$(R_B)_A$: BA (3)

A	cnt(A) = 1
DA	cnt(A) = 2

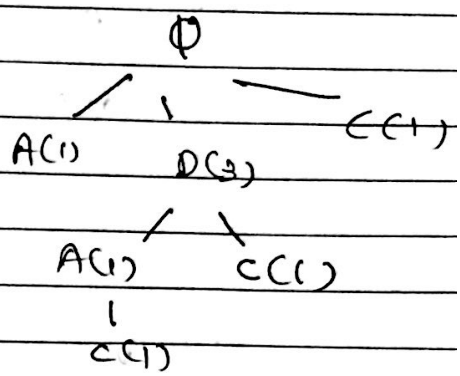


∴ output : BAD (2)

$(R_B)_D$: output = BD(2)

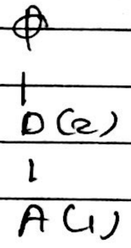
R_G (Projection on G)

CG	cnt (G) = 1
DG	cnt (G) = 1
DACG	cnt (G) = 1
DCG	cnt (G) = 1
AG	cnt (G) = 1



$(R_G)_C$ output = GC(3)

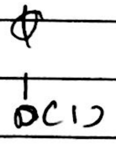
DAC	cnt (C) = 1
DC	cnt (C) = 1
C	cnt (C) = 1



Output = GCD(2)

$(R_G)_A$ - output = GA(2)

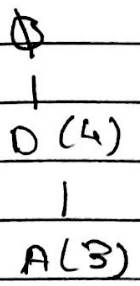
DA	cnt (A) = 1
A	cnt (A) = 1



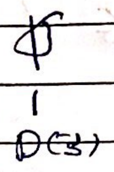
$(R_G)_D$: output = GD(3)

- R_C (projection on C)

C	cnt(C) = 1
DAC	cnt(C) = 3
DC	cnt(C) = 1



$(R_C)_A$: CA(3)

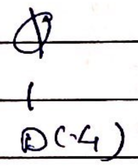


DA cnt(A) = 3

output = CAD(3)

$(R_C)_D$ = CD(4)

R_A (Projection on A)



A cnt(A) = 1

DA cnt(A) = 4

$(R_A)_D$ = AD(4)

R_D (Projection on D)

D cnt (D) = 6

Frequent itemsets :-

Sup	Itemsets
6	D
5	A, C, G
4	B, F, FD, AD, CD
3	BA, GC, GD, CA, CAD
2	FG, FGD, FC, FCD, FA, BC, FAD, BG, BAD, BD ~~GCD~~, GA, GCD

Q.4 <u>Ans:-</u>

Support Of ABE = 2

All subsets of ABE with support are:~

A(4), B(5), E(4), AB(3), AE(2), BE(4)

Rules:~

<u>For A(4)</u>

$\{A\} \rightarrow \{BE\}$ $conf(A \rightarrow BE) = \dfrac{SUP(ABE)}{SUP(A)}$

$= 2/4 = \underline{\underline{0.5}}$

<u>For B(5)</u>

$\{B\} \rightarrow \{AE\}$

$Conf(B \rightarrow AE) = \dfrac{SUP(ABE)}{SUP(B)} = \dfrac{2}{5} = \underline{\underline{0.4}}$

<u>For E(4)</u>

$\{E\} \rightarrow \{A, B\}$

$conf(E \rightarrow AB) = \dfrac{SUP(ABE)}{SUP(E)} = \dfrac{2}{4} = \underline{\underline{0.5}}$

For <u>AB(3)</u>

$\{AB\} \rightarrow \{E\}$

$conf(AB \rightarrow E) = \dfrac{Sup(ABE)}{Sup(AB)} = \dfrac{2}{3} = 0.66$

For AE(2)

$\{AE\} \rightarrow \{B\}$

$$\text{conf}(AE \rightarrow B) = \frac{\text{sup}(AEB)}{\text{sup}(AE)} = \frac{2}{2} = \underline{\underline{1}}$$

For BE(4)

$\{BE\} \rightarrow \{A\}$

$$\text{conf}(BE \rightarrow A) = \frac{\text{sup}(AEB)}{\text{sup}(BE)} = \frac{2}{4} = \underline{\underline{0.5}}$$