NAME – MOHIT AKHOURI

ROLL NO – 19UCC023

DATA MINING ASSIGNMENT – 1

Q 1) ----------

Ans 1)

Proximity measures mean the measure the similarity or dissimilarity between two data objects.

Similarity Measures : It is the numerical measure of how much two data objects are alike. It falls in the range of [0,1]. Similarity is higher when two objects are alike. The measurement of similarity for different types of attributes are as follows:

Attribute type	Similarity (s)
Nominal	1 (when x=y) , 0 (when x≠y)
Ordinal	1-d , d=dissimilarity
Interval / Ratio	-d , d=dissimilarity

Dissimilarity Measures : It is the numerical measure of how much two data objects are not alike. The lowest value is 0 and highest value may vary. It is lower when two objects are similar.

Attribute type	Dissimilarity (d)
Nominal	0 (when x=y) , 1 (when x≠y)
Ordinal	| x-y |/ (n-1) , n=number of values
Interval / Ratio	| x-y |

Given :

X = (0,1,0,0,1,1,0,0,1,1)

Y = (1,0,0,0,1,1,0,0,0,1)

(a) Cosine Similarity = The cosine similarity is defined as :

$$\cos(X,Y) = \frac{<x,y>}{||x||.||y||}$$

<x,y> = 0*1+1*0+0*0+0*0+1*1+1*1+0*0+0*0+1*0+1*1 = 3

||x|| = sqrt (0^2 + 1^2 + 0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2 + 1^2)
= sqrt (5) = 2.236

||y|| = sqrt (1^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 1^2)
= sqrt (4) = 2

Hence , cos(X,Y) = 3/(2.236*2) = **0.6708**

(b) Jacard Similarity is defined as :

$$J = \frac{Number\ of\ matching\ presences\ (f11)}{Number\ of\ attributes\ not\ involved\ in\ 00\ matches\ (f10+f01+f11)}$$

f_{11} = Number of values where x=1 and y=1 = **3**

f_{10} = Number of values where x=1 and y=0 = **2**

f_{01} = Number of values where x=0 and y=1 = **1**

J = (3) / (2*1) = 3/2 = **1.5**

(c) Pearson's Correlation Coefficient is as follows :

X	**Y**	**X.Y**	**X^2**	**Y^2**
0	1	0	0	1
1	0	0	1	0
0	0	0	0	0
0	0	0	0	0
1	1	1	1	1
1	1	1	1	1
0	0	0	0	0
0	0	0	0	0
1	0	0	1	0
1	1	1	1	1
$\sum X =$ **5**	$\sum Y =$ **4**	$\sum X.Y =$ **3**	$\sum X^2 =$ **5**	$\sum Y^2 =$ **4**

The Pearson's Correlation Coefficient formula is :

$$r = \frac{N*\sum xy - (\sum x)(\sum y)}{sqrt\ ([\ N*\sum x^2 - (\sum x)^2\ ]\ .\ [\ N*\sum y^2 - (\sum y)^2\ ])}$$

N = 10

r = [(10*3) – (5*4)] / [sqrt ([10*5 - 25] . [10*4 - 16])]

= 10 / sqrt (25 * 24) = 10 / 24.49 = **0.4083**

Q 2) ----------

Ans 2)

(a) The Gini Index for the overall collection of training examples is :

1 – 2*0.5*0.5 = **0.5**

(b) The Gini Index of the Customer ID Attribute is as follows :

Gini value for each instance of the customer ID from 1 to 20 is 0. Now the Overall Gini is given as : (Number of instances in child) / (Number of instances in parent) * Gini (Child).

Hence , Gini Index for the Customer ID Attribute is : **0**

(c) The Gini index for the Gender Attribute is as follows :

There are two possible values of the Gender Attribute which are Male and Female. Now , first calculate the Gini index for the "Male" and "Female" and then calculate the overall weighted Gini Index.

Gini Index for male = 1 – 2*0.5*0.5 = 0.5

Gini Index for female = 1 – 2*0.5*0.5 = 0.5

Number of Male instances = 10

Number of Female instances = 10

Number of instances in the parent class = 20

Overall Gini index for the Gender = (10/20) * 0.5 + (10/20) * 0.5

= 0.5*0.5 + 0.5*0.5

= **0.5**

(d) Gini index for the car-type attribute using multiway split is as follows :

There are 3 possible values of car type which are :

Possible Values	Gini Index	Number of instances
Family	0.375	4
Sports	0	8
Luxury	0.2188	8

Now the overall Gini index is as follows : (4/20)*0.375 + (8/20)*0 + (8/20)*0.2188 = **0.1625**

(e) Gini index for the shirt-size attribute using multiway split is as follows :

There are 4 possible values of shirt size which are :

Possible Values	Gini Index	Number of instances
Small	0.48	5
Medium	0.4898	7
Large	0.5	4
Extra-Large	0.5	4

Now the overall Gini index is as follows : (5/20)*0.48 + (7/20)*0.4898 + (4/20)*0.5 + (4/20)*0.5 = **0.4914**

(f) **Car Type** is better attribute since it has the **lowest Gini index** among all , which is 0.1625

(g) When new customers are added to the dataset , the customer ID keeps on increasing , which means that there is no fixed set of values for customer ID that could help in prediction. In other words , customer ID has **no predictive power** due to increasing dataset.

(h) The ID3 algorithm is used to generate a decision tree from the dataset. The steps involved in ID3 algorithm is : Begin with the original set S as root node. On each iteration , iteration is through all unused attribute and algorithm calculates the entropy of the attribute. The attribute with the smallest entropy is selected and then set S is split according to the selected attribute into subsets of the data. The recursion on subsets is then followed.

(i) The C4.5 algorithm is also used to generate a decision Tree. The steps involved in the C4.5 algorithm is as follows : First notice the base , calculate the normalized information gain ratio by splitting between X . Find the Attribute with the highest normalized information gain and then split the dataset and create a decision node. Repeat the algorithm on the subsets obtained.

Q 3) ----------

Ans 3)

(a) If we follow the optimistic approach : **e'(t) = e(t)** , then the generalization error rate of the tree is as follows : 3/10 = **0.3**

(b) If we follow the pessimistic approach : **e'(t) = e(t) + N * 0.5** (N = no. of leaf nodes) , so the calculation of **e'(t)** is as follows :

e(t) = 3/10
N = 4
Penalty factor = 0.5
Number of instances in the training set = 10
Therefore the generalization error rate using pessimistic approach is :
(3/10) + (4*0.5)/10 = **0.5**

(c) For Calculation of generalization error rate using validation set , the steps which are followed are : Split the training set into two components (one which is used for model building , and the second one is used for computing generalization error. For calculation of generalization error rate , we follow the reduced error pruning approach and the answer is : 4/5 = **0.8**

Q 4) ----------

Ans 4)

The **Minimum Description Length Principle** (MDL Principle) states that if there are two models , we should choose the one that minimizes the cost to describe a classification. The total description length of a tree is defined as :

Each internal node of the tree is encoded using the ID of the splitting attribute. If there are m attributes , cost of encoding each attribute is $\log_2 m$ bits and each leaf is encoded using the ID of class associated with. If there are k classes , the cost of encoding a class = $\log_2 k$ bits.

Cost(tree, data) = Cost(tree) + Cost(data | tree)

Cost (tree) = Cost of encoding all nodes in the tree

Cost (data | tree) = encoded using the classification errors the tree commits on the training set. Each error is encoded by $\log_2 n$ bits , where n = total number of training instances.

Given :

Number of attributes = **25** , Cost of each internal node in tree = $\log_2 25$ = **5**

Number of classes = **3** , Cost of each leaf in tree = $\log_2 3$ = **2**

Cost of each misclassification error = $\log_2 n$

Number of errors in the left decision tree = 9

Number of errors in the right decision tree = 6

Number of internal nodes in left decision tree = 2

Number of leaves in the left decision tree = 3

Number of internal nodes in right decision tree = 4

Number of leaves in the right decision tree = 5

Total description length of left decision tree = $2*5 + 3*2 + 9*\log_2 n$ = **$16+ 9*\log_2 n$**

Total description length of right decision tree = $4*5 + 5*2 + 6*\log_2 n$ = **$30+ 6*\log_2 n$**

Now , according to the MDL Principle : Let us consider 2 cases , when **n < 25** (Let's say n=24) and when **n > 25** (Let's say n=26).

Value of n	Total Description Length of left tree	Total description Length of right tree
24	57.26	57.50
26	58.30	58.20

Hence , we can say that :

When n<25 , Left Decision tree is better than right decision tree.

When n>25 , Right Decision tree is better than left decision tree.

Q 5) ----------

Ans 5)

a) **NO , rules are not mutually exclusive.**
   The mutually exclusive rules means : If rules are independent of each other and every record is covered by atmost one rule. In other words, we can say that set is mutually exclusive if no two rules are triggered by the same record.

b) **YES , the rule set is exhaustive**.
   Exhaustive rule set means each record is covered by atleast one rule.

c) **YES , the ordering is needed for this set of rules.**
   Ordering is needed since we saw that rules are not mutually exclusive means that a test instance may trigger more than one rule, but if rules are ordered, then only the highest ranked rule will be triggered.

d) **NO , we do not need a default class for the rule set.**
   Default class is not needed since we saw that rules are exhaustive , means every instance of the test set will trigger atleast one rule , so no need of default class (default class is assigned to that test instance for which none of the rules fired) .