Homework no -1
Course ID : CSCE 5380

Name : Sai Harsha Reddy
ID : 11589323

Data Mining

1b) Data mining is the process of discovering patterns or relationships in large data sets, with the goals of extracting meaningful information and insights from the data.

Data Mining is a complex process that combines elements from several fields, including databases, statistics and machine learning. Data mining is more than just the simple application of technology from databases, statistics and machine learning. It involves a holistic approach to data analysis that integrates these fields and leverages their strengths to extract valuable insights from large and complex datasets.

2) a) Let's assume that the data set contains the recordings of patients weights height age, temperature, blood pressure, glucose level and heart rate for every visit they come.

Classification :

Row : Records all patients

Column : Various measures like temperature, blood pressure, blood glucose and heart rate.

Clustering

Row: Patient ID

Column: Various measures like temperature, blood pressure, blood glucose and heart rate.

Association rule mining

Row: Patient ID

column: various measure like temperature, blood pressure and heart rate.

Anomaly detection

Row: Patient ID

column: various measures like temperature, blood pressure, blood glucose and heart rate.

4)

a) Continous: Can be any value within the given range without any restriction

Discrete: Data can have only a specific value and it may be the case the number of values is infinite but each will be unique and nogery

a) Continous: quantitative, interval

For average use of the internet is Continous and it's a quantity

and this fund mutch in a ~~xxxxxxx~~ time of interval.

b) Discrete, qualitive, ordinal

The mark will bee considered a specific one. so its descrete then it will bee a quality of the peformance of the study.

c) Discrete as it will bee distinct and qualitative ordinal as needs to be assigned in order - for the credit card number.

d) Discrete, qualitative, ordinal

It implied a salary of all employees in an organisation.

5)

a) The limitation of the approach to compute similarity if wee replace the null values in sales commission by 0 is that we can't distinguish between sales and non sales person occupation.

If user wants to calculate the similarity between all the users based on their rates commision and replace mule

both 0 we have a problem null. stands for not applicable and hence means that the user does not need to fill in that value of column for that table. If we replace null with 0 will be unable to distinguish sales varad. non sales peoples as their sales peoples with 0 sales commission. Hence we won't have any method to distinguish between sales and non sales person occupation. This is the short coming of replacing null with 0 in the sales commission attribute

6)
a) Distance b/w given objects (22, 1, 42, 10) & (20, 0, 36, 8)

$$= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2+(p_2-p_1)^2}$$

$$= \sqrt{(20-22)^2+(0-1)^2+(36-42)^2+(8-10)^2}$$

$$= \sqrt{4+1+36+4}$$

$$= \sqrt{45}$$

$$= 6.7082$$

7)
a) let consider the following dataset we want to decide whether the Patients have lung cancer or not.

where

class P = s & lung cancer = "yes"

class N = s lung cancer = "No"

Step 1 : Compute the gini index of the overall collection of training set.

formula $Gini(D) = 1 - \sum_{i=1}^{m} P_i^2$

$$gini(D) = 1 - (5/10)^2 - (5/10)^2$$

$$= 0.5$$

Step 2:

select the best split among tabocco smoki Random exposure and chronic cough) For the root list all the split you consid together their corresponding values.

formula:

$$Gini_A(D) = \frac{|D_1|}{|D|}$$

$$= Gini(D_1) + \frac{|D_2|}{|D|}$$

$$Gini(D_2)$$

Let's now consider Tobacco smoking: {yes, No}
It is a Binary attribute

Gini
Tobacco Smoking $\in$ {yes} (D) = (5/10)

$$\left(1 - (4/5)^2 - (1/5)^2\right) + (5/10)\left(1 - (1/5)^2 - (4/5\right.$$

$$= 0.32 \Rightarrow$$ Gini Tabacco Smoking

let Now consider Random exposes {yes, no}
It is binary attribute

Gini random exposure $\in \{yes\}^{(D)}$ $= (2/10)$.

$(1-(2/2)^2-0) + (8/10)(1-(3/8)^2-(5/8)^2))$.

$= 0.375 \Rightarrow$ Gini random exposure $\in \{No\}^{(D)}$.

Gini chronic cough $\in \{yes\}^{(D)} = (7/10)$.

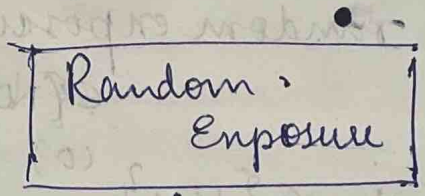
$= \left(1-(4/7)^2-(3/7)^2\right) + (3/10)(1-(1/3)^2-(2/3)^2)$

$= 0.476 =$ Gini chronic cough $\in \{No\}^{(D)}$

Attribute	Split	Gini Index	reduction in impurity Δ Gini = Gini (D) - Gini $(D)_A$
Tobacco	Binary	0.32	0.5 - 0.32 = 0.18
Random Exposure	Binary	0.375	0.5 - 0.375 = 0.125
Chronic cough	Binary	0.476	0.5 - 0.476 = 0.024.

The attribute that maximize the reduction in impurity or Equivalently has the minimum gini index Gini split (n) is selected as the splitting attribute

Tobacco smoking is selected with minimum gini index and highest reduction in impurity.



Yes

No

Tobacco smoking	chronic cough	lung cancer
Yes	Yes	Yes
No	No	Yes

Tobacco Smoking	Chronic cough	lung Cancer
Yes	Yes	Yes
Yes	Yes	Yes
Yes	Yes	Yes
Yes	No	No
No	Yes	No
No	Yes	No
No	Yes	No
No	No	No

Step 1 : Gini In $= 1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2$

$= 0.468$

Step 2 : Let now Tobacco smoking

$= \{Yes, No\}$

It is a binary attribute

gini Tobacco smoking $\in \{yes\}^{(0)}$

$= (4/8)\left(1-(3/4)^2-(1/4)^2\right) + (4/8)\left(1-0-(4/4)^2\right)$

$= 0.1875 + 0$

$= 0.187$

gini Tobacco smoking $\in \{No\}^{(0)}$

Let now consider chronic cough {yes, no}

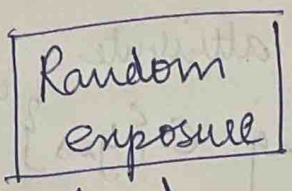
It is a binary attribute

gini chronic cough $\in \{yes\}^{(0)}$

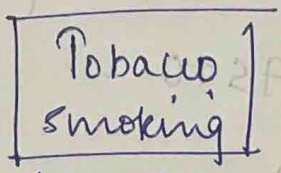
$= (6/8)\cdot\left(1-(3/6)^2-(3/6)^2\right) + (2/8)\left(1-0-(2/2)^2\right)$

$= 0.375$ = gini chronic cough $\in \{no\}^{(0)}$

Attribute	split	gini index	reduction in purity
Tobacco smoking	binary	0.187	0.281
Chronic cough	binary	0.375	0.093



lung = yes



yes

No.

Chronic cough	lung cancer
Yes	Yes
Yes	Yes
Yes	Yes
No	No

Chronic cough	lung cancer
Yes	No
Yes	No
Yes	No
No	No

8)

a) We can modify the basic decision tree algorithm by integrating the count of each tuple into the calculation of the attribute selection measure, taking the count into consideration to determine the most common class among the tuples.