Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

EIGHTH SEMESTER B.TECH DEGREE EXAMINATION(S), OCTOBER 2019

Course Code: CS402

Course Name: DATA MINING AND WAREHOUSING

Max. Marks: 100 Duration: 3 Hours

PART A

Answer all questions, each carries 4 marks. Marks

1	How is data warehouse different from a database? How are they similar?	(4)
2	Compare star and snowflake schema dimension table.	(4)
3	Use the two methods below to normalize the following group of data: 100,200,300,500,900 i) min-max normalization by setting min=0 and max=1 ii) z-score normalization	(4)
4	Explain the attribute selection method in decision trees .	(4)
5	Distinguish between hold out method and cross validation method.	(4)
6	Explain prepruning and postpruning approaches in decision tree algorithm.	(4)
7	Differentiate between support and confidence.	(4)
8	How to compute the dissimilarity between objects described by binary variables?	(4)
9	Differentiate between Agglomerative and Divisive hierarchical clustering method.	(4)
10	Explain web content mining?	(4)

PART B

Answer any two full questions, each carries 9 marks.

11		The following data is given in increasing order for the attribute age: 13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,36,40,45,46,52,70.	
	a)	Use smoothing by bin boundaries to smooth these data, using bin depth of 3.	(3)
	b)	How might you determine outliers in the data?	(3)
	c)	What other methods are there for data smoothing?	(3)
12)		Explain the following procedures for attribute subset selection	
	a)	Stepwise forward selection	(3)
	b)	Stepwise backward elimination	(3)
	c)	A combination of forward selection and backward elimination	(3)

13 a) Suppose a datawarehouse consists of three measures customer, account and branch and two measures count (number of customers in the branch) and balance. Draw the schema diagram using snowflake schema. (4)

b) Real-world data tend to be incomplete, noisy, and inconsistent. What are the various approaches adopted to clean the data? (5)

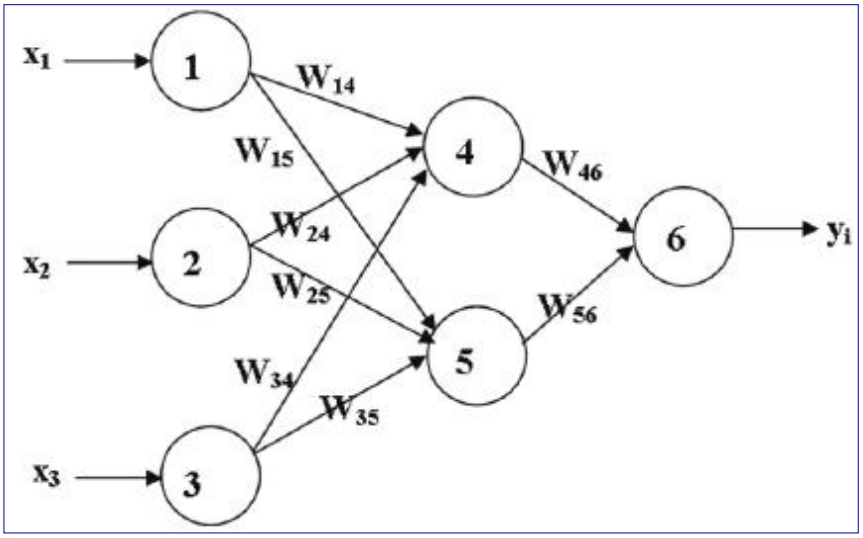
PART C

Answer any two full questions, each carries 9 marks.

14 Given the following data on a certain set of patients seen by a doctor, can the doctor conclude that a person having chills, fever, mild headache and without running nose has the flu?(Use Naive Bayes algorithm for prediction) (9)

chills	running nose	headache	fever	has flu
Y	N	mild	Y	N
Y	Y	no	N	Y
Y	N	strong	Y	Y
N	Y	mild	Y	Y
N	N	no	N	N
N	Y	strong	Y	Y
N	Y	strong	N	N
Y	Y	mild	Y	Y

15 The following figure shows a multilayer feed-forward neural network. Let the learning rate be 0.9. The initial weight and bias values of the network is given in the table below.The activation function used is the sigmoid function. (9)



x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2	0.1

Show weight and bias updation with the first training sample (1,0,1) with class label 1, using backpropagation algorithm

16 a) Explain classification by C4.5 algorithm. (6)

b) What is meant by Maximum Marginal Hyperplane (MMH)? (3)

PART D

Answer any two full questions, each carries 12 marks.

17 Consider the transaction database given below. Set minimum support count as 2 and minimum confidence threshold as 70%

Transaction ID	List of Item_Ids
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

a) Find the frequent itemset using Apriori Algorithm. (8)

b) Generate strong association rules . (4)

18 a) Explain DBSCAN algorithm . (8)

b) State the pros and cons of DBSCAN method. (4)

19 a) Explain clustering by k-medoid algorithm. (6)

b) Explain Apriori based frequent subgraph mining. (6)

****

I & II

MAY 2019

PART A

1. How is datamining related to business intelligence? (Mod I)

Datamining is the process of converting raw data into useful information. It is an important analytics process designed to explore data. Data mining is also refered to as data/knowledge discovery. It is about understanding the data you already have to make better business decisions.

Business Intelligence as a discipline is made up of several related activities, including data mining, online analytical processing, querying and reporting. It often relies on data mining to look for sol^{n}s to problems in the data already available. It can provide insights of things that you didnot know / it can confirm / deny a hypothesis you may have. Either way it allows you to make business decisions.

>37
>16 2. Differentiate b/w OLTP and OLAP. (Mod I)
>15

Basis for comparison	OLTP	OLAP
Basic	It is an online transactional s/m & manages database modification	It is an online data retrieving & data analysis s/m
Focus	Insert, Delete, Update info from the database	Extract data for analysing that helps in decision making
Data	OLTP and its transactions are the original source of data	Diff OLTPs database become the source of data for OLAP
Transaction	Short transactions	Long transactions
Time	Processing time of a transaction is comparatively less in OLTP	Processing time of a transaction is comparatively more in OLAP
Queries	Simple	Complex
Normalization	Tables in OLTP database are normalized (3NF)	Tables in OLAP database are normalized
Integrity	OLTP database must maintain data integrity constraint	OLAP database doesnot get frequently modified. Hence, data integrity is not affected.

1
2 3 Why do we need data transformation? What are the different ways of data transformation? (Mod II)

Data transformations (eg normalization) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements.

These techniques are not mutually exclusive, they may work together. For eg, data cleaning can involve transformations to correct wrong data such as by transforming all entries for a date field to a common format.

* Smoothing :- Remove noise from data
* Aggregation :- Summarization
* Generalization :- Concept hierarchy climbing
* Normalization :- Scaled to fall within a small, specified range
  - → min-max normalization
  - → z-score normalization
  - → normalization by decimal scaling.
* Attribute / feature construction.
  - → New attributes constructed from the given ones.

→ Min-max normalization to (new_min_A, new_max_A)

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Eg:- Let income range $12000 to $98000 normalized to [0.0, 1.0]

Then $73600 is mapped to $\frac{73600 - 12000}{98000 - 12000}(1 - 0) + 0$

$= 0.716$

→ z-score normalization (μ : mean, σ : standard deviation)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Eg:- Let $\mu = 54000$, $\sigma = 16000$. Then 73600 is mapped to

$$= \frac{73600 - 54000}{16000} = 1.225$$

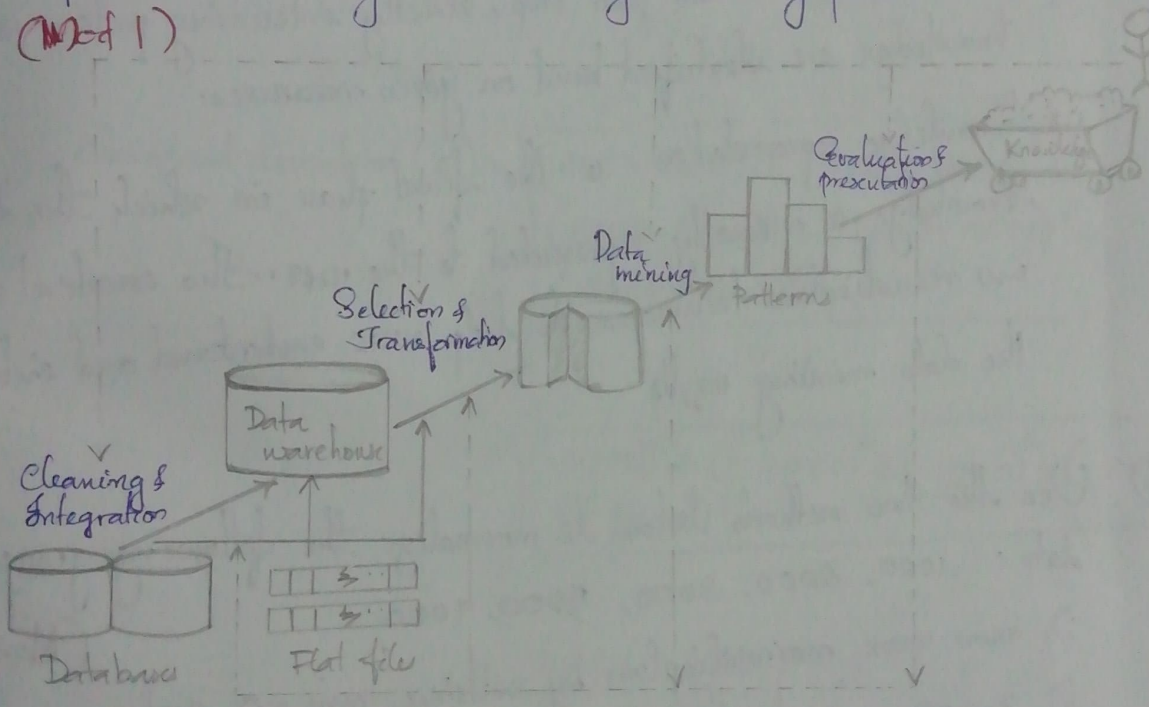
→ Normalization by decimal scaling.

$$v' = \frac{v}{10^j}$$ where j is the smallest integer such that $\max(|v'|) < 1$

j = no: of digits in max no:

PART B

11 a) Explain various stages in knowledge discovery process with neat diagram (Mod 1)



Knowledge Discovery in database (KDD) is the process of finding useful information and patterns in data. Datamining is the use of algorithms to extract the information and patterns derived by the KDD process

The iterative process consists of the following steps:

* Data cleaning: also known as data cleansing, it is a phase in

A

which noise data and irrelevant data are removed from the collection.

* Data integration :- at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
* Data selection :- At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
* Data transformation :- also known as data consolidation it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
* Data mining :- is the crucial step in which clever techniques are applied to extract patterns potentially useful.
* Pattern evaluation : in this step, strictly interesting patterns representing knowledge are identified based on given measures.
* Knowledge representation :- is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

b) Use the two methods below to normalize the following group of data : 1000, 2000, 3000, 5000, 9000. (Mod II).

i) min max normalization by setting min = 0 and max = 1

ii) z-score normalization.

D) Min max normalization $v' = \frac{v - min}{max - min}(new\_max - new\_min) + new$

min = 1000 max = 9000.

new_min = 0 new_max = 1

$v = 1000 \Rightarrow v' = \frac{1000 - 1000}{9000 - 1000}(1 - 0) + 0 = 0$

$v = 2000 \Rightarrow v' = \frac{2000-1000}{9000-1000}(1-0)+0 = 0.125$

$v = 3000 \Rightarrow v' = \frac{3000-1000}{9000-1000}(1-0)+0 = 0.25$

$v = 5000 \Rightarrow v' = \frac{5000-1000}{9000-1000}(1-0)+0 = 0.5$

$v = 9000 \Rightarrow v' = \frac{9000-1000}{9000-1000}(1-0)+0 = 1$

Values	1000	2000	3000	5000	9000
Nomalized	0	0.125	0.25	0.5	1

ii) Z-score normalization

$$\text{Mean } (\mu) = \frac{1000+2000+3000+5000+9000}{5}$$

$$= 4000$$

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{(1000-4000)^2+(2000-4000)^2+(3000-4000)^2+(5000-4000)^2+(9000-4000)^2}{5-1}}$$

$$= \sqrt{\frac{(-3000)^2+(-2000)^2+(-1000)^2+(-1000)^2+(5000)^2}{2}}$$

$$= \frac{6324.5532}{2}$$

$$= 3162.2776$$

Z-score normalization $v' = \frac{v-\mu_A}{\sigma_A}$

$v = 1000 \Rightarrow v' = \frac{1000-4000}{3162.2776} = -0.9486$

$v = 2000 \Rightarrow v' = \frac{2000-4000}{3162.2776} = -0.6324$

6 $v = 3000 \Rightarrow v' = \frac{3000 - 4000}{3162.2776} = -0.3162$

$v = 5000 \Rightarrow v' = \frac{5000 - 4000}{3162.2776} = 0.3162$

$v = 9000 \Rightarrow v' = \frac{9000 - 4000}{3162.2776} = 1.5811$

Values	1000	2000	3000	5000	9000
Normalised	-0.9486	-0.6324	-0.3162	0.3162	1.5811

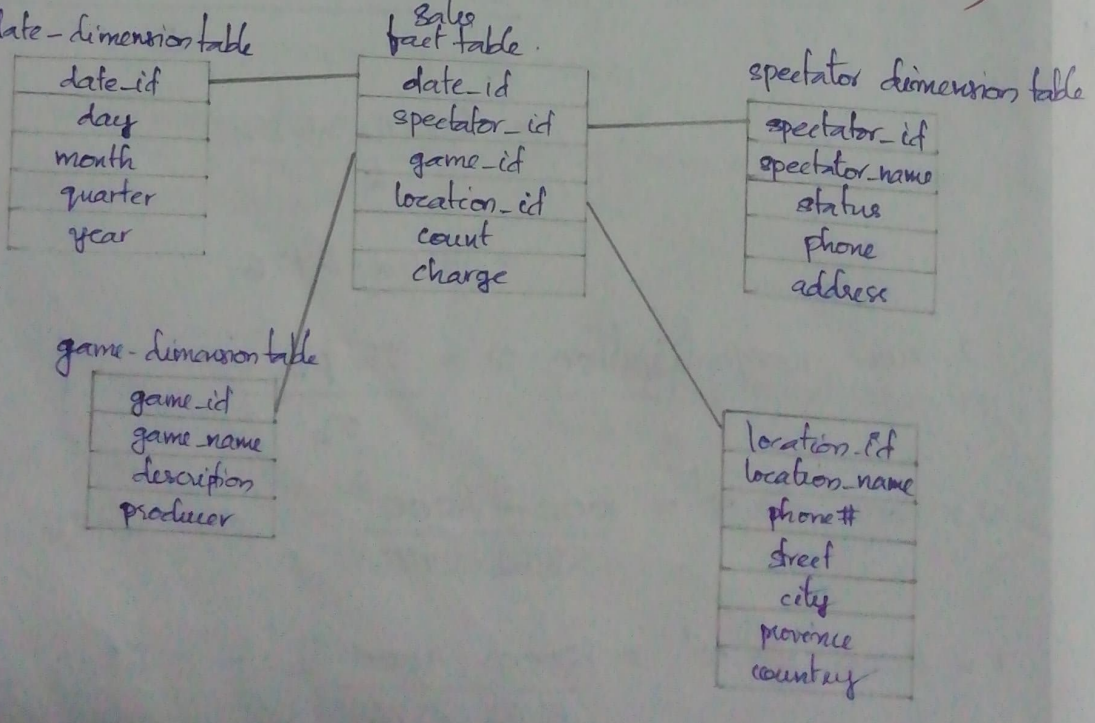
12 Suppose that a datawarehouse for University consists of four dimensions date, spectator, location and game and two measures count and charge, where charge is the fare that a spectator pays when watching a game on the given date. Spectator may be students, adults or seniors, with each category having its own charge rate.

a) Draw a star schema for the data warehouse

b) Starting with the basic cuboid [date, spectator, location, game] what specific OLAP operation should be performed inorder to list the total charge paid by student spectators at GM_PLACE in 2010.

(Mod II)

a)



b) → Roll up on date from date_id to year
→ Roll up on game from game_id to all
→ Roll up on location from location_id to location_name
→ Roll up on spectator from spectator_id to status.
→ Dice with status = "students", location name = "GM place" and year = 2004.

13 Summarize the various preprocessing activities involved in data mining. (Mod II)

There are several data preprocessing techniques:

*Data Cleaning can be applied to remove noise and correct inconsistencies in data

* Data Integration merged data from multiple sources into a coherent data ~~source~~ store such as a data warehouse

* Data Reduction can reduce any data size by, for instance aggregating, eliminating redundant features or clustering.

* Data Transformations (eg normalization) maybe applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements

These techniques are not ~~such~~ mutually exclusive; they may work together. For eg, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a ~~date~~ date field to common format.

Data processing techniques, when applied before mining can substantially improve the overall quality of the patterns mined

& and / or the time required for mining

OCTOBER 2019

PART A

1. How is datawarehouse different from a database? How are they similar? (Mod 1)

* Database is a collection of related data that represents some elements of the real world whereas Data warehouse is an information s/m that stores historical and commutative data from single or multiple sources.
* Database is designed to record data whereas the Data warehouse is designed to analyse data
* Database is application-oriented - collection of data whereas Data warehouse is the subject oriented collection of data.
* Database uses Online Transactional Processing (OLTP) whereas Data warehouse uses Online Analytical Processing (OLAP)
* Database tables and joins are complicated because they are normalized whereas Data Warehouse tables and joins are easy because they are denormalized.
* ER modelling techniques are used for designing Database. whereas Data modelling techniques are used for designing Data Warehouse.

2. Compute star and snowflake schema dimension table (Mod II)

STAR SCHEMA	SNOWFLAKE SCHEMA
* It consists of fact table with a single table for each dimension (dimension table)	* It is a variation of star schema, which have multiple level of dimension table.
* Highly denormalized.	* Normalized (in dimension table)
* Category wise single dimension table	* Dimension table further splits into additional table
* More data dependency and redundancy	* Less
* No need of complicated joins	* Require complicated joins

3. Use the two methods below to normalize the following group of data: 100, 200, 300, 500, 900

i) min max normalization by setting min = 0 and max = 1

ii) z score normalization (Mod II)

i) Min max normalization. $v' = \frac{v - min}{max - min}(new\_max - new\_min) + new\_min$

min = 100 max = 900

new_min = 900 new_max = 1

$v = 100 \Rightarrow v' = \frac{100 - 100}{900 - 100}(1 - 0) + 0 = 0$

$v = 200 \Rightarrow v' = \frac{200 - 100}{900 - 100}(1 - 0) + 0 = 0.125$

10

$v = 300 \Rightarrow v' = \frac{300-100}{900-100}(1-0)+0 = 0.25$

$v = 500 \Rightarrow v' = \frac{500-100}{900-100}(1-0)+0 = 0.5$

$v = 900 \Rightarrow v' = \frac{900-100}{900-100}(1-0)+0 = 1$

Value	100	200	300	500	900
Normalized	0	0.125	0.25	0.5	1

ii) Z-score normalization

Mean (μ) $= \frac{100+200+300+500+900}{5}$

$= 400$

Standard deviation (σ) $= \sqrt{\frac{(100-400)^2+(200-400)^2+(300-400)^2+(500-400)^2+(900-400)^2}{5-1}}$

$= \sqrt{\frac{(-300)^2+(-200)^2+(-100)^2+(100)^2+(500)^2}{2}}$

$= \frac{632.455532}{2}$

$= 316.227766$

Z-score normalization $= \frac{v-\mu}{\sigma_A}$

$v = 100 \Rightarrow v' = \frac{100-400}{316.2277} = -0.9486$

$v = 200 \Rightarrow v' = \frac{200 - 400}{316.2277} = -0.6324$

$v = 300 \Rightarrow v' = \frac{300 - 400}{316.2277} = -0.3162$

$v = 500 \Rightarrow v' = \frac{500 - 400}{316.2277} = 0.3162$

$v = 900 \Rightarrow v' = \frac{900 - 400}{316.2277} = 1.5811$

Values	100	200	300	500	900
Normalized	-0.9486	-0.6324	-0.3162	0.3162	1.5811

PART B

11 The following data is given in increasing order for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 3536, 40, 45, 46, 52, 70.

a) Use smoothing by bin boundaries to smooth these data, using bin depth of 3. Bin mean ⇒ Replace all by the bin value. Bin median ⇒ Replace 50% by median. Bin boundaries ⇒ Replace with min/max values according to closer.

b) How might you determine outliers in the data?

c) What other methods are there for data smoothing? (Mod II)

Partition N/D

a) Bin depth (D) = 3; No: of attributes (N) = 27; Each bin contains = $\frac{N}{D} = 9$ items.

	Smoothed value		Smoothed value
Bin 1 : 13, 15, 16	13, 13, 16	Bin 8 : 36, 40, 45	36, 36, 45
Bin 2 : 16, 19, 20	16, 16, 20	Bin 9 : 46, 52, 70	46, 46, 70
Bin 3 : 20, 21, 22	20, 20, 22		
Bin 4 : 22, 25, 25	22, 22, 25		
Bin 5 : 25, 25, 30	25, 25, 30		
Bin 6 : 33, 33, 35	33, 33, 35		
Bin 7 : 35, 35, 35	35, 35, 35		

12

b) Outliers in data can be identified in several ways

→ By dividing the data into equi-width histograms and identifying the outlying histogram

→ By clustering the data into groups. Any data that doesn't fall into a group can be taken as outliers.

→ In general, fit a model to the data. Any data points that deviate significantly (based on some threshold) from the model can be considered outliers.

c) Other methods that can be used for data smoothing include alternate forms of binning such as:

→ Smoothing by bin medians

→ Smoothing by bin boundaries

Alternatively, equal width bins can be used to implement any of the forms of binning, where the interval range of values in each bin is constant.

Methods other than binning include using regression techniques to smooth data by setting it to a function ~~such~~ such as through linear or multiple regression

Classification techniques can be used to implement concept hierarchies that can smooth the data by rolling-up lower level concepts to higher level concepts.

12. Explain the following procedures for attribute subset selection.
a) Stepwise forward selection
b) Stepwise backward elimination
c) A combination of forward selection and backward elimination

a) Stepwise forward selection.

This procedure start with an empty set of attributes as the minimal set. The most relevant attributes are chosen (having minimal p-value) and are added to the minimal set. In each iteration, one attribute is added to a reduced set.

b) Stepwise backward selection

Here all the attributes are considered in the initial set of attributes. In each iteration, one attribute is eliminated from the set of attributes whose p-value is higher than significance level.

c) A combination of forward selection and backward elimination.

The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently. This is the most common technique which is generally used for attribute selection.

13) a) Suppose a datawarehouse consists of three measures customer, account and branch and two measures count (no: of customers in the branch) and balance. Draw the schema diagram using snowflake schema.

b) Real world ~~entity~~ data tend to be incomplete, noisy and inconsistent. What are the various approaches adopted to clean the data?

b) Data cleaning tasks

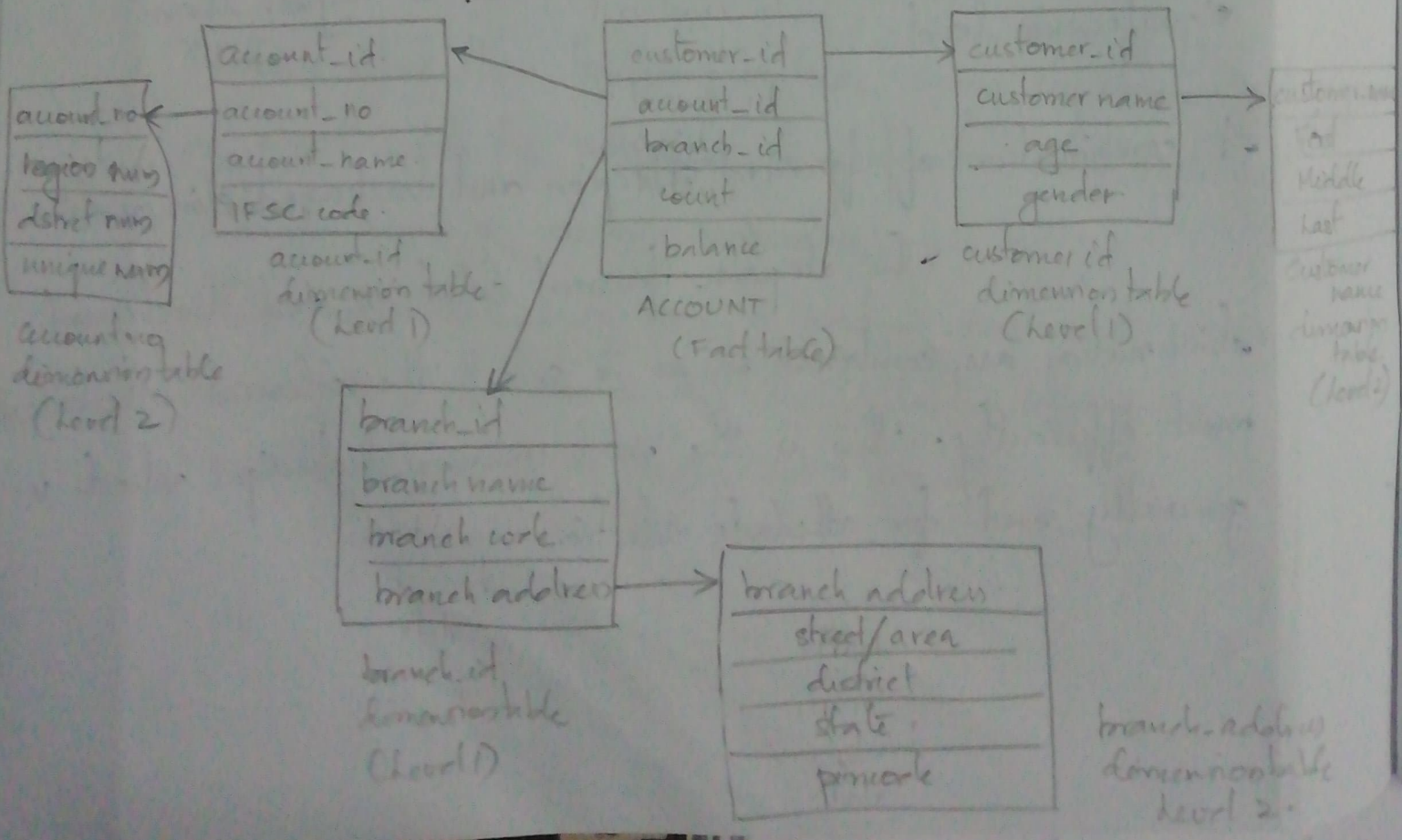
→ Fill in missing values

→ Identify outliers and smooth out noisy data

→ Correct inconsistent data

→ Resolve redundancy caused by data integration

a)



III & IV

MAY 2019

PART A

4 An airport security screening station wants to determine if passengers are criminals or not. To do this, the faces of passengers are scanned and kept in a database. Is this a classification or prediction task? Justify. (Mod III)

Classification can be defined as the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.

The face of the passenger is scanned and its basic pattern (distance b/w eyes, size and shape of mouth, head, etc) is identified. This pattern is compared to entries in a database to see if it matches any patterns that are associated with known offenders.

5 Where do we use linear regression? Explain linear regression (Mod IV)

A straight line regression analysis involves a response variable y and a single predictor variable x

ie $y = b + wx$

b and w are regression coefficients, specifying y intercept and slope of the line respectively.

16

→ The regression coefficients can be thought of as weights.

$$y = w_0 + w_1 x$$

→ These coefficients can be solved for ~~any~~ by the method of least square.

→ Let D be the training set consisting of values of predictor variable x and their associated response variable y. The training set contains |D| datapoints of the form $(x_1, y_1), (x_2, y_2) \ldots (x_{|D|}, y_{|D|})$.

The regression coefficients can be estimated by.

$$w_1 = \frac{\sum_{i=1}^{|D|} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|} (x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

$\bar{x}$ = Mean value of $x_1, x_2, \ldots, x_{|D|}$

$\bar{y}$ = Mean value of $y_1, y_2, \ldots, y_{|D|}$

→ Multiple linear regression is an extension of straight line regression which involves more than one predictor variable.

→ It allows response variable y to be modelled as a linear fnc of n predictor variables or attributes $A_1, A_2, \ldots, A_n$ describing a tuple X

Example of multiple linear regression model based on two predictor attributes are

$$y = w_0 + w_1 x_1 + w_2 x_2$$

x_1 and x_2 → Value of attributes A_1 & A_2

The method of least squares can be extended to calculate w_0, w_1 & w_2.

→ The major uses for regression analysis are
- a) Determining the strength of predictors
- b) Forecasting an effect
- c) Trend forecasting.

6. What is the significance of tree pruning in decision tree algorithms (Mod III)

→ Main approach to avoid overfitting is pruning

→ Technique that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.

→ Reduces complexity of the final classifier and hence improves predictive accuracy by the reduction of overfitting.

→ The pruning phase might remove redundant comparisons or remove subtrees to achieve better performance.

→ When decision trees are built, many of the branches may reflect noise or outliers in the training data.
Tree pruning address methods address this problem of overfitting the data.

→ Tree pruning methods attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.

18

→ Decision trees suffer from repetition and replication, making them overwhelming to interpret

→ Repetition occurs when an attribute is repeatedly ~~tested~~, making them ~~over~~ tested along a given branch of the tree

→ In replication, duplicate subtrees exist within the tree

→ These situations can impede the accuracy and comprehensibility of a decision tree

PART C

14 Based on the following data determine the gender of a person having height 6 ft, weight 130 lbs and foot size 8 in (Naive Bayes algorithm)

Person	Height (feet)	Weight (lbs)	foot size (inches)
male	6.00	180	10
male	6.00	180	10
male	5.50	170	8
male	6.00	170	10
female	5.00	130	8
female	5.50	150	6
female	5.00	130	6
female	6.00	150	8

(Mod III)

Person	Height (feet)			Weight (lbs)				foot size (inches)		
	6.00	5.50	5.00	180	170	150	130	10	8	6
Male	3	1	0	2	2	0	0	3	1	0
Female	1	1	2	0	0	2	2	0	2	2

Conditional Probability table

Person	Height (feet)			Weight (lbs)				foot size (inches)		
	6.00	5.50	5.00	180	170	150	130	10	8	6
Male	3/4	1/4	0	2/4	2/4	0	0	3/4	1/4	0
Female	1/4	1/4	2/4	0	0	2/4	2/4	0	2/4	2/4

$$P(\text{Male}) = \frac{\text{No: of male}}{\text{Total persons}} = \frac{4}{8} = \frac{1}{2}$$

$$P(\text{Female}) = \frac{\text{No: of female}}{\text{Total persons}} = \frac{4}{8} = \frac{1}{2}$$

Probability if the conditions satisfied in male

$$q_1 = P(6\text{ft}/\text{Male}) \times P(130\text{lbs}/\text{Male}) \times P(8\text{ in}/\text{Male}) \times P(\text{Male})$$

$$= \frac{3}{4} \times 0 \times \frac{1}{4} \times \frac{1}{2}$$

$$= 0$$

$$q_2 = P(6\text{ft}/\text{Female}) \times P(130\text{lbs}/\text{Female}) \times P(8\text{ in}/\text{Female}) \times P(\text{Female})$$

$$= \frac{1}{4} \times \frac{2}{4} \times \frac{2}{4} \times \frac{1}{2}$$

$$= \frac{1}{32}$$

$$= 0.03125$$

$$\max(q_1, q_2) = \max(0, 0.03125)$$

$$= 0.03125 = q_2$$

So the gender of the person having height 6ft, weight 130lbs and foot size 8inch is Female

$$P(A|C) = \frac{P(C|A)\,P(A)}{P(C)}$$

$P(A|C)$ → Posterior probability – which represents the degree to which we believe a given model accurately describes the situation given the available data and all of our prior information.

$P(C|A)$ → Likelihood – which describes how well the model predicts the data

$P(A)$ → Prior probability – which describes the degree to which we believe the model accurately describes reality based on all of our prior information.

$P(C)$ → Normalizing constant – The constant that makes the posterior density integrate to one.

15 The "Restaurant A" sells burger with optional flavours: Pepper, Ginger and Chilly. Every day this week you have tried a burger (A to F) and kept a record of which you liked. Using Hamming distance, show how the 3NN classifier with majority voting would classify.

{pepper = false; ginger = true; chilly = true}.

	Pepper	Ginger	Chilly	Liked
A	true	true	true	False
B	true	false	false	True
C	false	true	true	False
D	false	true	false	True
E	true	false	false	True

(M-1 N)

The easiest local distance function known as the overlap fnc, returns 0 if the two values are equal and 1 otherwise

$$dist_A(x.A, q.A) = def \begin{cases} 0 & \text{if } x.A = q.A \\ 1 & \text{otherwise} \end{cases}$$

Hamming Distance:

Let Qn be the {pepper = false, ginger = true, chilly = true}.

Dist (A, Qn) = 1 + 0 + 0 = 1

Dist (B, Qn) = 1 + 1 + 1 = 3

Dist (C, Qn) = 0 + 0 + 0 = 0

Dist (D, Qn) = 0 + 0 + 1 = 1

Dist (E, Qn) = 1 + 1 + 1 = 3

true & true ⇒ 0
true & false ⇒ 1

0, 1, 1, 3, 3
C, A, D, B, E

Arrange the distances in Ascending order and the least 3 distances are taken. So A, C, D are taken

While considering them A — False E — False D — True

So the majority voting is taken and 2 votes for false

and 1 rule for true ~~So considering one~~

Hence No: of false > No of true

So concluding that class liked = False for the condition {pepper = false; ginger = true; chilly = true}

---

16 a) How CA·5 differs from ID3 algorithm? (Mod III)

b) How does backpropagation algorithm works? (Mod IV)

a)

	ID3	CA·5
Splitting criteria	Information Gain	Gain ratio
Attribute type	Handles only categorical value	Handles only categorical & numerical values
Missing values	Donot handle missing values	Handle missing value
Pruning strategy	No pruning is done	Error based pruning is used
Outlier detection	Susceptible to outliers	Susceptible to outliers

b) Backpropagation learns by iteratively processing a data set of training tuples, comparing the network's prediction for each each tuple with the actual known target value.

The target value may be known class label of the training tuple (for classification prob) or continuous value (for numeric prediction)

For each training tuple, the weights are modified so as to minimize the mean squared error b/w the n/w's prediction and the actual target value.

These modifications are made in the "backwards" direction

(ie from output layer) through each hidden layer down to the first hidden layer (-hence the name propagation).

Although it is not guaranteed, in general the weights will eventually converge and the learning process stops.

Algorithm

Input :- D, a data set con the training tuples and their associated target values
l, the learning rate
network, a multilayer feed-forward n/w

Output :- a trained neural network

Method:

1. Initialize all weights and bias in n/w
2. while terminating condition not satisfied {
3. for each training tuple X in D {
4. // Propagate the inputs forward
5. for each input layer unit j {
6. $O_j = I_j$; // o/p of an i/p unit is its actual i/p value
7. for each hidden or o/p layer unit j {
8. $I_j = \sum_i w_{ij} O_i + \theta_j$; // compute the net i/p of unit j wrt previous layer, i
9. $O_j = \frac{1}{1+e^{-I_j}}$; // compute the o/p of each unit j
10. // Backpropagate the errors
11. for each unit j in the o/p layer
12. $Err_j = O_j(1-O_j)(T_j - O_j)$.
// compute the error

13. for each unit j in the hidden layers, from the last to the first hidden layer

14. $Err_j = O_j(1 - O_j) \sum_k Err_k w_{jk}$; // compute the error wrt the next higher layer, k

15. for each w_{ij} in network {

16. $\Delta w_{ij} = (l) Err_j O_i$; // weight increment

17. $w_{ij} = w_{ij} + \Delta w_{ij}$;} // weight update

18. for each bias θ_j in network {

19. $\Delta \theta_j = (l) Err_j$; // bias increment

20. $\theta_j = \theta_j + \Delta \theta_j$;} // bias update

21. }}

Terminating condition: Training stops when

→ All Δw_{ij} in the previous epoch are so small as to be below some specified threshold or

→ The percentage of tuples misclassified in the previous epoch is below some threshold or

→ A prespecified number of epochs has expired.

OCTOBER 2019

PART A

2q

A. Explain the attribute selection method in decision trees (Mod III)

→ Entropy :- Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

where $S \rightarrow$ current state

$P_i \rightarrow$ probability of an event i of state S or percentage of class i in a node of state S.

→ Information Gain : Information Gain is a Statistical property that measures how well a given attribute separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute that returns the highest information gain and the smallest entropy.

$$\text{Information Gain} = \text{Entropy (before)} - \sum_{j=1}^{K} \text{Entropy (j, after)}$$

where "before" is the dataset before the split, K is the no. of subsets generated by the split, and (j, after) is subset j after the split.

→ Gini Index – Gini Index is a cost function used to evaluate splits in the dataset. It is calculated by subtracting the sum of the squared probabilities of each class from one.

$$\text{Gini} = 1 - \sum_{i=1}^{c} (P_i)^2$$

Gini index works with the categorical target variable "Success" or "Failure". It performs only Binary splits.

→ Gain Ratio : Gain Ratio overcomes the problem with information gain by taking into account the number of branches that would result before making the split. It corrects information gain by taking the intrinsic information of a split into account.

$$\text{Gain ratio} = \frac{\text{Information Gain}}{\text{Split Info}} = \frac{\text{Entropy(before)} - \sum_{j=1}^{K} \text{Entropy}(j, \text{after})}{\sum_{j=1}^{K} w_j \log_2 w_j}$$

→ Reduction in Variance - Reduction in Variance is an algorithm used for continuous target variables (regression problems). This algorithm uses the standard formula of variance to choose the split. The split with lower variance is selected as the criteria to split the population:

$$\text{Variance} = \frac{\sum (X - \bar{X})^2}{n}$$

→ Chi Square - It is one of the oldest tree classification methods. It finds out the statistical significance b/w the differences between sub-nodes and parent node. We measure it by the sum of squares of standardized diff b/w observed and expected frequencies of the target variable.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$ where

χ^2 = Chi square obtained

$\sum$ = the sum of

O = observed score

E = expected score.

26 5 Distinguish b/w hold out method and cross validation method (Mod II)

Cross validation is usually preferred method because it gives your model the opportunity to train on multiple train-test splits. This gives you a better indication of how well your model will perform on unseen data. Hold out on other hand, is dependent on just on train-test split. That makes the holdout method more dependent on how the data is split into train and test sets.

The hold out method is good to use when you have a very large dataset, you're on a time crunch, or you are starting to build an initial model in your data science project. Keep in mind that because cross-validation uses multiple-train-test splits, it takes more computational power and time to run than using the holdout method.

6 Explain prepruning and post pruning approaches in decision tree algorithm. (Mod III)

Tree pruning is performed inorder to remove anomalies in the training due to noise or outliers. The pruned trees are smaller and less complex.

Types of pruning

Post pruning – The tree is pruned by halting its construction early

Pre pruning – This approach removes a subtree from a fully grown tree

The cost complexity is measured by following two parameters

* No: of leaves in the tree
* Error rate of the tree.

Post Pruning

→ First build the tree

→ Then prune it

* Fully grown tree shows all attribute interactions

→ Problem: Some subtrees might be due to chance effects

→ Two pruning operations

* Subtree replacement

* Subtree raising

→ Possible strategies

* Error estimation

* Significance testing

* MDL principle.

Pre Pruning

→ Based on statistical significance test.

* Stop growing the tree when there is no statistically significant association b/w any attribute and the class at a particular node.

* Use all available data for training and apply statistical test to estimate whether expanding / pruning a node is to produce an improvement beyond the training set.

→ Most popular test — Chi squared test

→ ID3 used chi square test in addition to information gain

* Only statistically significant attributes were allowed to be selected by information gain procedure.

→ Early stopping :- Pre pruning may stop the growth process prematurely.

PART C

28

14 Given the following data on a certain set of patients seen by a doctor, can the doctor conclude that a person having chills, fever, mild headache and without running nose has the flu? (Use Naive Bayes algorithm for prediction)

Chills	Running nose	Headache	Fever.	Has flu
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

(Mod III)

Has flu	Chills		Running nose		Headache			Fever	
	Yes	No	Yes	No	Mild	Strong	No	Yes	No
Yes	3	2	4	1	2	2	1	4	1
No	1	2	1	2	1	1	1	1	2

Conditional Probability Table ↓

Has flu	Chills		Running Nose		Headache			Fever	
	Yes	No	Yes	No	Mild	Strong	No	Yes	No
Yes	3/5	2/5	4/5	1/5	2/5	2/5	1/5	4/5	1/5
No	1/3	2/3	1/3	2/3	1/3	1/3	1/3	1/3	2/3

$P(\text{Has flue} = \text{'yes'}) = \frac{\text{No: of Has flue = "yes"}}{\text{Total no:}} = \frac{5}{8}$

$P(\text{Has flue} = \text{"no"}) = \frac{\text{No: of Has flue = "no"}}{\text{Total no:}} = \frac{3}{8}$

Probability of having chills, fever, mild headache & without running nose has flu

$q_1 = P(\text{Chills} = \text{yes} / \text{Has flu} = \text{yes}) \times P(\text{Fever} = \text{yes} / \text{Has flu} = \text{'yes'}) \times P(\text{Mild headache} / \text{Has flu} = \text{yes}) \times P(\text{Running nose} = \text{no} / \text{Has flu} = \text{yes}) \times P(\text{Has flu} = \text{'yes'})$

$= \frac{3}{5} \times \frac{4}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{5}{8}$

$= \frac{12}{125}$

$= 0.096$

$q_2 = P(\text{Chills} = \text{yes} / \text{Has flu} = \text{'no'}) \times P(\text{Fever} = \text{yes} / \text{Has flu} = \text{'no'}) \times P(\text{Mild headache} / \text{Has flu} = \text{'no'}) \times P(\text{Running nose} = \text{no} / \text{Has flu} = \text{'no'}) \times P(\text{Has flu} = \text{'no'})$

$= \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{3}{8}$

$= \frac{1}{54}$

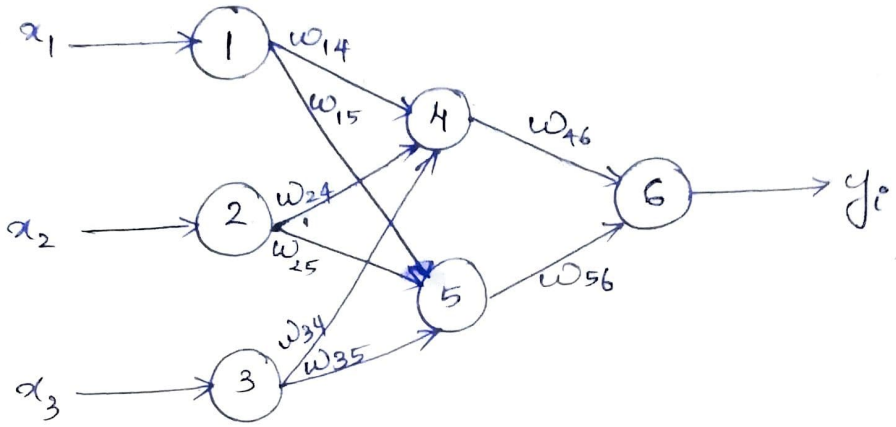
$= 0.0185$

$\max(q_1, q_2) = \max(0.096, 0.0185) =$

$= 0.096 = q_1$

So the person has flue = Yes

30 15 The following figure shows a multilayer feed-forward neural n/w. Let the learning rate be 0.9. The initial weights and bias values of the network is given in the table below. The activation function used is the sigmoid function.



x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5
1	0	1	0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2

Show weight and bias updation with the first training sample (1, 0, 1) with class label 1, using backpropagation algorithm (Mod IV)

Net Input and Output Calculations

Unit j	Net Input I_j	Output, O_j
4	$x_1w_{14} + x_2w_{24} + x_3w_{34} + \theta_4$ $= (1 \times 0.2) + (0 \times 0.4) + (1 \times -0.5) + (-0.4)$ $= 0.2 + 0 - 0.5 - 0.4$ $= -0.7$	$\frac{1}{1+e^{+0.7}}$ $= 0.332$
5	$x_1w_{15} + x_2w_{25} + x_3w_{35} + \theta_5$ $= (1 \times -0.3) + (0 \times 0.1) + (1 \times 0.2) + (0.2)$ $= -0.3 + 0 + 0.2 + 0.2$ $= 0.1$	$\frac{1}{1+e^{-0.1}}$ $= 0.525$

6	$w_{46} \times O_4 + w_{56} \times O_5 + \theta_6$ $= (-0.3)(0.332) + (0.2)(0.525) + 0.1$ $= -0.105$	$\frac{1}{1+e^{-0.105}}$ $= 0.474$

Calculation of the error at each node

Unit j	Err_j
6	$(0.474)(1-0.474)(1-0.474) = 0.1311$
5	$(0.525)(1-0.525)(0.1311)(-0.2) = -0.0065$
4	$(0.332)(1-0.332)(0.1311)(-0.3) = -0.0087$

Calculation for Weight and Bias Updating.

Weight / Bias	New Value	
w_{46}	$-0.3 + (0.9)(0.1311)(0.332)$	$= -0.261$
w_{56}	$-0.2 + (0.9)(0.1311)(0.525)$	$= -0.138$
w_{14}	$0.2 + (0.9)(-0.0087)(1)$	$= 0.192$
w_{15}	$-0.3 + (0.9)(-0.0065)(1)$	$= -0.306$
w_{24}	$0.4 + (0.9)(-0.0087)(0)$	$= 0.4$
w_{25}	$0.1 + (0.9)(-0.0065)(0)$	$= 0.1$
w_{34}	$-0.5 + (0.9)(-0.0087)(1)$	$= -0.508$
w_{35}	$-0.2 + (0.9)(-0.0065)(1)$	$= 0.194$
θ_6	$0.1 + (0.9)(0.1311)$	$= 0.218$
θ_5	$0.2 + (0.9)(-0.0065)$	$= 0.194$
θ_4	$-0.4 + (0.9)(0.0087)$	$= -0.408$

32

16 a) Explain classification by C4.5 algorithm (Mod III)

b) What is meant by Maximum Margenal Hyperplane (MMH) (Mod IV)

a) C4.5 algorithm is used to generate a decision tree which was developed by Ross Quinlan.

→ It is an extension of Quinlan's ID3 algorithm.

→ C4.5 generates decision trees which can be used for classification & therefore C4.5 is often referred to as statistical classifier.

→ It is better than ID3 algorithm because it deals with both continuous and discrete attributes and also with the missing values and pruning trees after construction

→ C5.0 is the commercial successor of C4.5 because it is a lot faster, more memory efficient and used for building smaller decision tree C4.5 perform by default a tree pruning process

→ This leads to the formation of smaller trees, more simple rules and produces more intuitive interpretations.

C4.5 follows 3 steps in tree growth

i) For splitting of categorical ~~values~~ attributes, C4.5 follows the similar approach to ID3 algo. Continuous attribute always generate binary splits.

ii) Selecting attribute with highest gain ratio

iii) These steps are applied to new tree branches and growth of the tree is stopped after checking of stop criterion. Information gain bias the attribute with more no: of values. Thus, C4.5 use gain ration which is a less biased selection criterion.

Advantages of C4.5

i) Easy to implement

ii) Builds models that can be easily interpreted

iii) It can deal with noise and deal with missing value attributes

Disadvantages of C4.5

i) A small variation in data can lead to diff decision trees when using C4.5

ii) For a small training set, C4.5 doesnot work well.

Maximum margin - the max distance of $d_+ + d_-$ where d_+ is the shortest dist of +ve eg from hyperplane (similarly d_- for -ve eg)

b) There are infinite number of separating lines that could be drawn and we want to find the best one which have the minimum classification error on previously unseen tuples

If our data were 3-D (ie three attributes), we would want to find the best separating plane

Generalizing to n dimensions, we want to find the best hyperplane

Hyperplane refers to the decision boundary that we are seeking, regardless of the number of input attributes.

An SVM approaches this problem by searching for the maximum marginal hyperplane.

In fig both hyperplanes can classify all the given data tuples. However we expect the hyperplane with the larger margin to be more accurate at classifying future data tuples than the hyperplane with the smaller margin.

This is why (during the learning or training phase) the SVM searches for the hyperplane with the largest margin, ie, the maximum marginal hyperplane (MMH). The associated margin gives the largest separation b/w classes.

The shortest distance from a hyperplane to one side of its margin is equal to the shortest distance from the hyperplane to the other side of its margin, where the "sides" of the margin are parallel to the hyperplane When dealing with MMH, this distance is in fact the shortest dist from MMH to closest training tuple of either class

V & VI

MAY 2019

PART A

7. What are two measures used for rule interestingness? (Mod V)

The strength of a rule is measured by its support and confidence.

Support (X) is the no: of times X in the transaction divided by total no: of transaction

Confidence of the association rule $X \rightarrow Y$ is defined as the ratio of X and Y together to the support of X.

$$\text{Support}(X) = \frac{\text{No: of items X appears}}{\text{Total no: of transaction}} = P(X)$$

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support of X \& Y together}}{\text{Support of X}} = \frac{P(X \cap Y)}{P(X)}$$

$$= P(Y|X)$$

Support is a useful measure because if it is low, the rule may just occur due to chance. Furthermore in a business environment a rule covering too few cases (or transactions) may not be useful because it doesnot make business sense to act on such a rule (not profitable).

For eg., In the database, the itemset {milk, bread, butter} has a support of $\frac{1}{5} = 0.2$ since it occurs in 20% of all transactions (1 out of 5 transactions).

Confidence thus determines the predictability of the rule. If the confidence of a rule is too low, one cannot reliably infer or predict Y from X. A rule with low predictability is of limited use.

For eg., the rule {butter, bread} $\Rightarrow$ {milk} has confidence of $\frac{0.2}{0.2} = 1.0$ in the database, which means that for 100% of transactions containing butter and bread the rule is correct.

36 Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8). Compute the Manhattan distance b/w the two objects

Manhattan distance = $|22-20| + |1-0| + |42-36| + |10-8|$

(City block distance) = $|2| + |1| + |6| + |2|$

= 11

$$\text{distance} = \sum_{i=0}^{n-1} |(x[i] - y[i])|$$

Manhattan distance – The distance b/w two points measured along axes at right angles. In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$

9 How can density based clustering varies from other methods? (Mod IV)

→ Density based method finds ~~clusters~~ clusters of arbitrary shape. It grows the clusters with as many points as possible till some threshold is met. The e-neighbourhood of a point is used to find dense regions in the database.

→ Density based algorithms typically regard clusters as dense regions of objects in the data space that are separated by regions of low density. The main idea of density-based approach is to find regions of high density and low density, with high-density regions being separated from low-density regions. These approaches can make it easy to discover arbitrary clusters.

→ A common way to divide the high dimensional space into density-based grid units. Units containing relatively high densities are the cluster centres and the boundaries b/w clusters fall in the regions of low-density units

→ Diff type of density based methods → DBSCAN, OPTICS, DENCLUE

10 Differentiate web content mining and web structure mining? (Mod VI)

Web Content mining

→ Web Content mining is the application of extracting useful information from the content of the web documents

→ Web Content consists of several types of data - text, image, audio, video etc

→ Content data is the group of facts that a web page is designed.

→ It can provide effective and interesting patterns about user needs

→ Text documents are related to text mining, machine learning and natural language processing

→ This mining is also known as text mining

→ This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.

Web Structure mining

→ Web Structure mining is the application of discovering structure information from the web

→ The structure of the web graph consists of web pages as nodes, and hyperlinks as edge connecting related pages.

→ Structure mining basically shows the structured summary of a particular website

→ It defines relationship b/w web pages linked by information or direct link connection

→ To determine the connection b/w two commercial websites, Web Structure mining can be very useful.

→ This type of datamining can be performed ↔ document level (Intra page) / hyperlink level (Inter page)

PART D

36 17. Consider the transaction database given below. Set minimum support count as 2 and minimum confidence threshold as 70%

Transaction ID	List of Items_Ids
T100	I_1, I_2, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	I_1, I_2, I_4
T500	I_1, I_3
T600	I_2, I_3
T700	I_1, I_3
T800	I_1, I_2, I_3, I_5
T900	I_1, I_2, I_3

a) Find the frequent itemset using FP Growth Algorithm (Mod 5)

b) Generate strong association rules (Mod 5)

a) Minimum support count = 2

Total no: of transactions = 9.

Step: Construct list of itemsets table with support count

Itemset	Support count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2.

~~Table 2~~

Table 2

Step II: Write the itemset in descending order of support count

Items	Sup. count
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

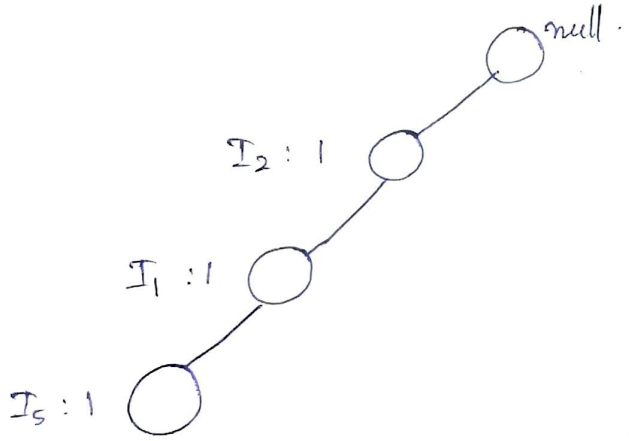
Table 3

Step III: Start drawing the FP tree with root node as null.

Step IV: Now consider the first transaction in table (T_{100}) and arrange in descending order as of sup count.

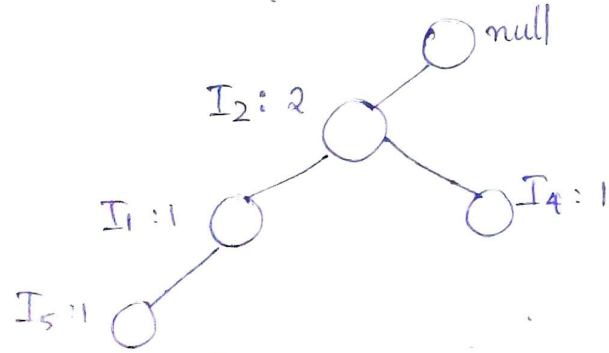
So $T_{100} \Rightarrow I_2(7), I_1(6), I_5(2)$.

Start from left node.



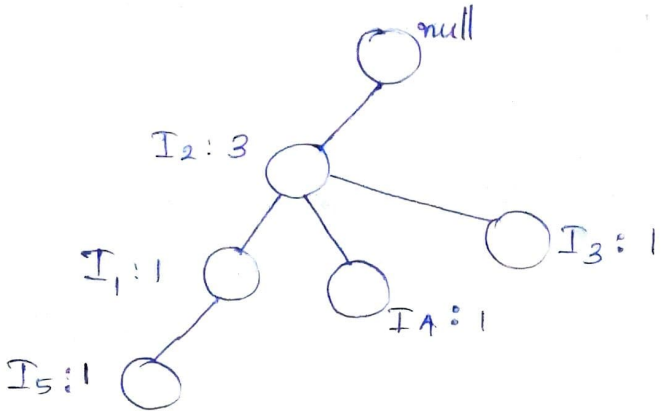
Next transaction I_{200} arrange it $\Rightarrow I_2, I_4$

Now check. I_2 is there from the root node if there increment the count of I_2 then check whether I_4 is there after I_2 if not then add another branch from I_2 and write I_4 with count 1

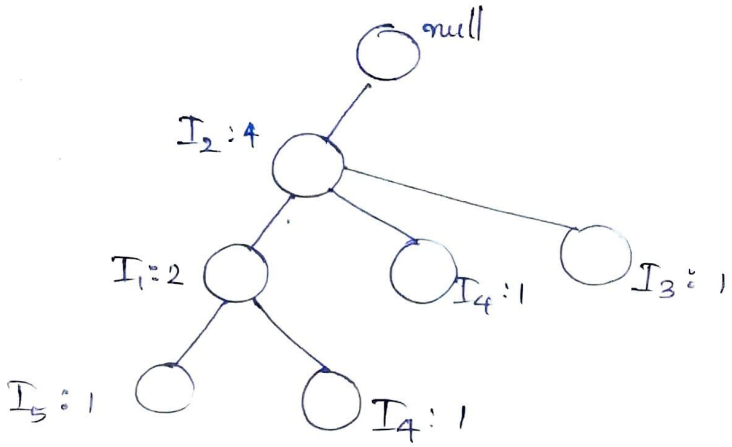


A0

Next transaction T_{300} ⇒ Arranging ⇒ I_2 I_3.

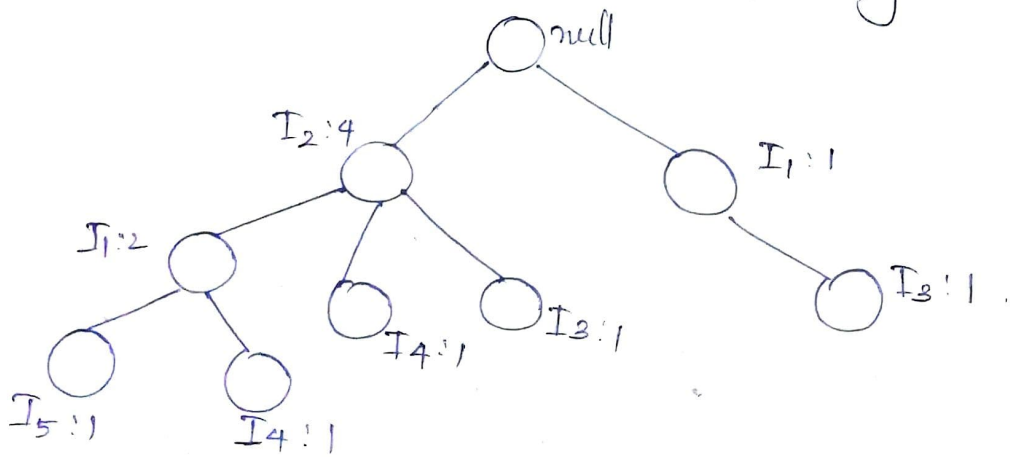


T_{400} ⇒ Arranging ⇒ I_2 I_1 I_4

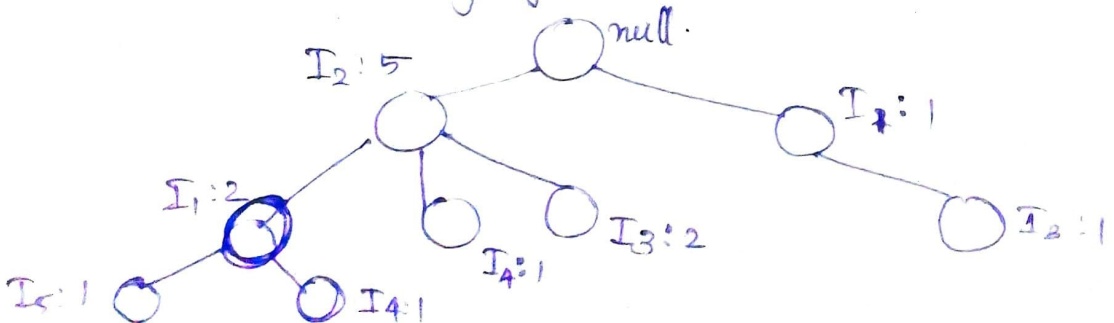


T_{500} ⇒ Arranging ⇒ I_1 I_3.

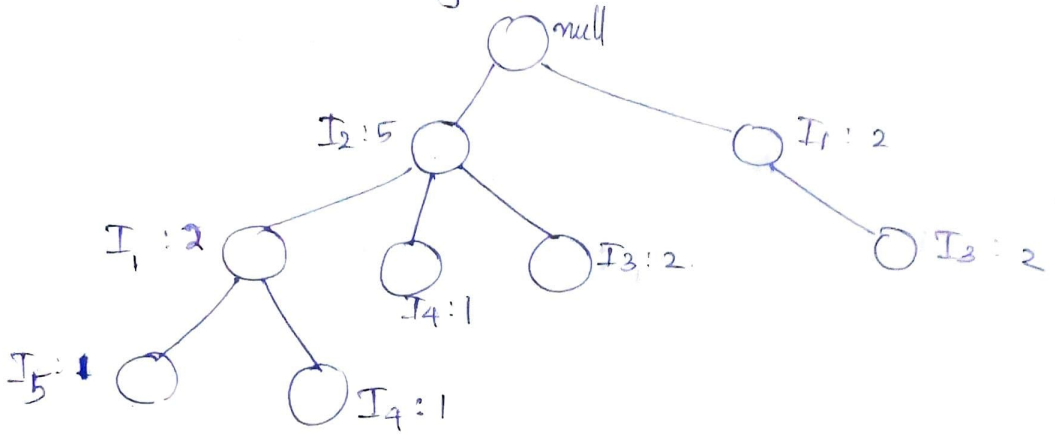
There is no I_1 from null so generate another branch.



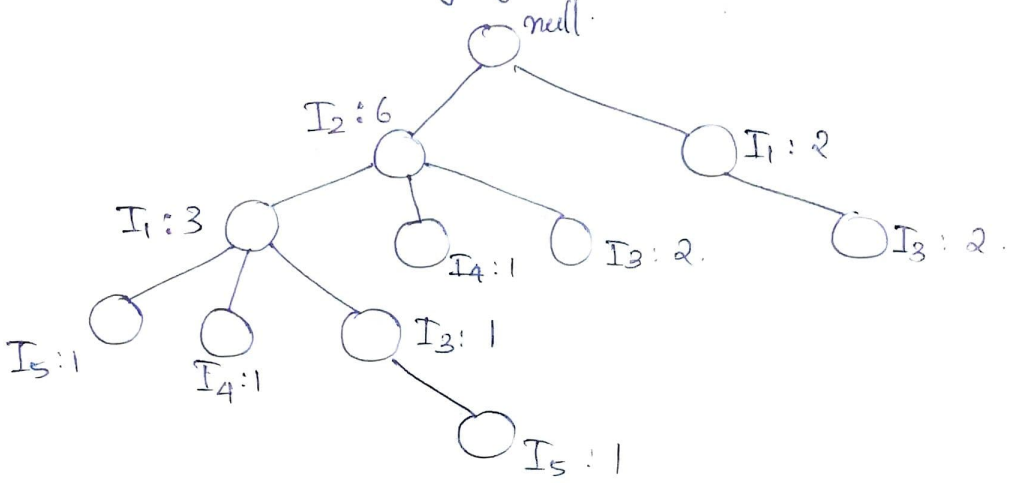
T_{600} ⇒ Arranging ⇒ I_2, I_3.



$T_{700} \Rightarrow$ Arranging $\Rightarrow I_1, I_3$

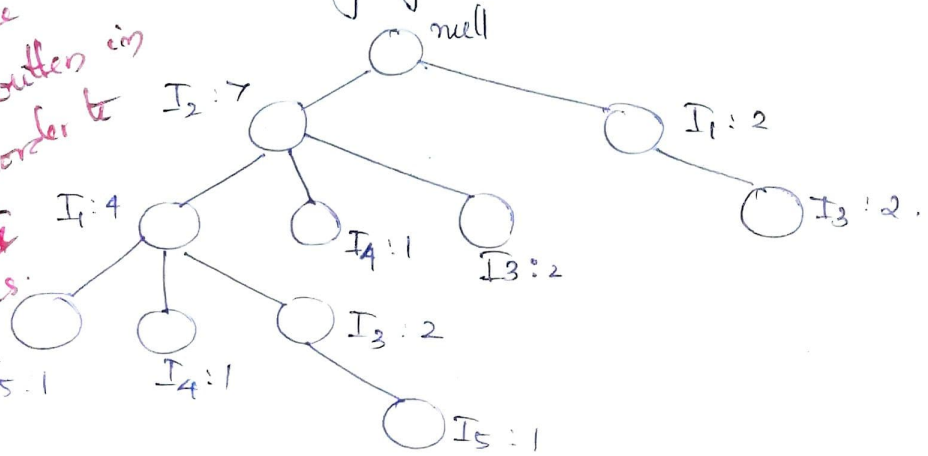


$T_{800} \Rightarrow$ Arranging $\Rightarrow I_2, I_1, I_3, I_5$.



$T_{900} \Rightarrow$ Arranging $\Rightarrow I_2, I_1, I_3$.

One FP tree is created and the counts can be re-written in the same tree. In order to understand them, I used separate trees.



Now check the total count of transactions in the tree and compare it with the table. If the counts are same then the tree is correct.

A2

Item	Conditional Pattern Base	Conditional FP tree
I_5	$\{\{I_2, I_1 : 1\}, \{I_2, I_1, I_3 : 1\}\}$	$\{I_2 : 2, I_1 : 2\}$
I_4	$\{\{I_2, I_1 : 1\}, \{I_2 : 1\}\}$	$\{I_2 : 2\}$
I_3	$\{\{I_2, I_1 : 2\}, \{I_2 : 2\}, \{I_1 : 2\}\}$	$\{I_2 : 4, I_1 : 2\}\{I_1 : 2\}$
I_1	$\{\{I_2 : 4\}\}$	$\{I_2 : 4\}$

Frequent patterns generated:

$\{I_2, I_5 : 2\}, \{I_1, I_5 : 2\}, \{I_2, I_1, I_5 : 2\}$.

$\{I_2, I_4 : 2\}$.

$\{I_2, I_3 : 4\}\{I_1, I_3 : 4\}\{I_2, I_1, I_3 : 2\}$

$\{I_2, I_1 : 4\}$.

Consider the database (Table 3) check for least sup count first.
So here I_5 / I_4 can be taken. Taking I_5.

Check FP tree

How we traverse from root node to I_5 is checked and then writing each path and writing the respective sup count of I_5 in that path in the respective paths. Then are ~~for~~ under conditional patterns base

Now in the column conditional FP tree write the count for each item. In item $I_5 \Rightarrow I_2 = 2$, $I_1 = 2$, $I_3 = 1$

But we don't write I_3 since its not satisfying min count = 2

So neglecting I_3.

Now in freq patterns generated, joining the each sets in conditional FP tree with item.

So joining. I_2, I_5 with min count $\Rightarrow \{I_2, I_5 : 2\}$

" I_1, I_5 " " " $\Rightarrow \{I_1, I_5 : 2\}$.

" I_2, I_1, I_5 " " " $\Rightarrow \{I_2, I_1, I_5 : 2\}$.

Min count checked from conditional FP tree.

Similarly generate for I_4, I_3 and I_1 too.
Generate paths wherever the items are located.
For I_3 check for both RHS and LHS
while writing the conditional FP tree write the paths in separate brackets when obtained from different LHS and RHS branch.

For $I_3 \Rightarrow \{I_2, I_3 : 2\}$ first but again from RHS branch we get again $\{I_1, I_3 : 2\}$ so adding the count and combining so $\{I_1, I_4 : 4\}$

For $\{I_2, I_1, I_3 : 2\}$ since the least count = 2.

Thus frequent patterns are generated.

$\{I_2, I_5 : 2\}$ $\{I_1, I_5 : 2\}$, $\{I_2, I_1, I_5 : 2\}$

$\{I_2, I_4 : 2\}$

$\{I_2, I_3 : 4\}$, $\{I_1, I_3 : 4\}$, $\{I_2, I_1, I_3 : 4\}$

$\{I_2, I_1 : 4\}$.

Strong Association rules from itemsets (where strong association rules satisfy both minimum support and minimum confidence)
Association rule can be generated as follows

→ For each frequent itemset l, generate all nonempty subsets of l

→ For every nonempty subset of l, output the rule

"$s \Rightarrow (l - s)$" if $\frac{\text{support\_count}(l)}{\text{support\_count}(s)} \geq \text{min\_conf}$,

where min_conf is the minimum confidence threshold.

$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{\text{support count}(A \cup B)}{\text{support count}(A)}$$

3. Frequent itemsets, taking I_1, I_2, I_5

Take the non empty subsets of I_1, I_2, I_5. ϕ set is not taken

	Rule $s \Rightarrow (l-s)$
I_1	$\longrightarrow$ $I_2 \cap I_5$
I_2	$\longrightarrow$ $I_1 \cap I_5$
I_5	$\longrightarrow$ $I_2 \cap I_5$
$I_1 \cap I_2$	$\longrightarrow$ I_5
$I_1 \cap I_5$	$\longrightarrow$ I_2
$I_2 \cap I_5$	$\longrightarrow$ I_1
$I_1 \cap I_2 \cap I_5$	$\longrightarrow$ ϕ Not considering since there is no RHS.

6 rules generated.

Now we have to check $s \rightarrow l-s \geq$ ~~min conf~~.

Support count of $\{I_1, I_2, I_5 : 2\}$ is 2.

Now check $\frac{\text{Support\_count}(l)}{\text{Support\_count}(s)} \geq \text{min\_conf}$.

	Rule	Confidence
1)	$I_1 \rightarrow I_2 \cap I_5$	$2/6 = 33.3\%$
2)	$I_2 \rightarrow I_1 \cap I_5$	$2/7 = 30\%$
3)	$I_5 \rightarrow I_1 \cap I_2$	$2/2 = 100\%$
4)	$I_1 \cap I_2 \rightarrow I_5$	$2/4 = 50\%$
5)	$I_1 \cap I_5 \rightarrow I_2$	$2/2 = 100\%$
6)	$I_2 \cap I_5 \rightarrow I_1$	$2/2 = 100\%$

Min_confidence = 70%. So rules greater than 70% are.

$I_1 \cap I_5 \longrightarrow I_2$

$I_2 \cap I_5 \longrightarrow I_1$

$I_5 \longrightarrow I_1 \cap I_2$

18 a) Explain BIRCH Clustering Method (Mod VI)

b) What are the advantages of BIRCH compared to other methods (Mod VI)

a) BIRCH – Balanced Iterative Reducing and Clustering using Hierarchies.

→ It is a scalable clustering method
→ Designed for very large data sets.
→ Only one scan of data is necessary.
→ It is based on the notation of CF (Clustering feature) a CF tree
→ CF tree is a height balanced tree that stores the clustering features for a hierarchical clustering
→ Cluster of data points is represented by a triple of numbers (N, LS, SS) where

N – no: of items in the sub cluster (zeroth moment of cluster)
LS – Linear sum of the points $\sum_{i=1}^{n} x_i$ (1st moment of cluster)
SS – Sum of the squared of the points $\sum_{i=1}^{n} x_i^2$ (2nd moment of cluster)

→ A CF tree structure is given as below:

2 parameters
* Branching Factor – max no: of children per non leaf node
* Threshold – max diameter of subclusters stored at leaf nodes of the tree

* Each non-leaf node has at most B entries
* Each leaf node has at most L CF entries which satisfies threshold T, a maximum diameter of radius
* P (page size in bytes) is the maximum size of a node
* Compact: each leaf node is a subcluster, not a data point.

3 points (2,5)(3,2)(4,3) in a cluster C_1.

$CF = \langle 3, ((2+3+4, 5+2+3), (2^2+3^2+4^2, 5^2+2^2+3^2)\rangle = \langle 3, (9,10), (29, 38)\rangle$

CF Tree structure

Phase I: Scans database to build initial CF tree
Phase II: Applies a clustering algo to leaf nodes of CF tree,
* removes sparse clusters
* groups dense clusters to larger ones

Basic algorithm

→ Phase I : Load data into m/y

Scan DB and load data into m/y by building a CF tree.
If m/y is exhausted rebuild the tree from the leaf node.

→ Phase II : Condense data (optional)

Resize the dataset by building a smaller CF tree
Remove more outliers
Condensing is optional

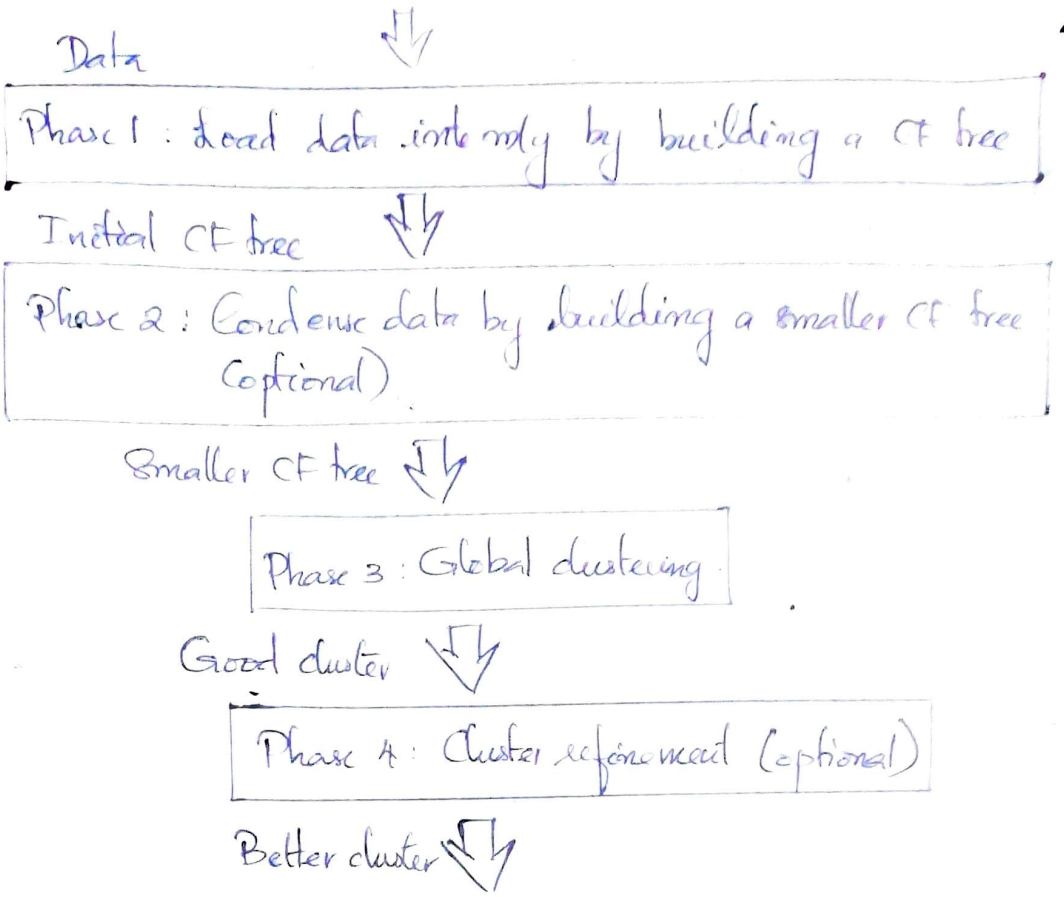
→ Phase III : Global clustering

Use existing clustering algorithms (eg KMEANS, HC)
on CF subsets

→ Phase IV : Cluster refining (optional)

Refining is optional
Fixes the problem with CF trees where same related data points may be assigned to diff. leaf entries.

CF tree is dynamically built as obj are inserted
→ Obj is inserted into closest leaf entry → If diameter of the subcluster stored in leaf node after insertion is larger than threshold value, then leaf & possibly other nodes split
→ Size of CF tree is determined threshold
→ If size of tree is larger reduce threshold
→ Similar to insertion & node split in B+ tree complexity. O(n) → no. of obj to be clustered



b) Advantages of BIRCH method

- ⇒ finds a good clustering with a single scan and improve the quality with a few additional scans
- → It is local in that each clustering decision is made without scanning all data points and currently existing clusters
- → It ~~makes~~ exploits the observation that data space is not usually uniformly occupied and not every data point is equally important.
- → It makes full use of available mdy to derive the finest possible sub-clusters while minimizing I/O costs
- → It is an incremental method that doesn't require the whole dataset in advance.

Disadvantages :- Handles only numeric data

19. a) Explain K-means partition algorithm. What is the dr
A8 of K means? (Mod V)

b) Term frequency matrix given in the table shows frequenc
terms per document. Calculate the TF-IDF value for the
term T_A in document 3.

Document/ Term	T_1	T_2	T_3	T_4	T_5	T_6
D_1	5	9	4	0	5	6
D_2	0	8	5	3	10	8
D_3	3	5	6	6	5	0
D_4	4	6	7	8	4	4

(Mod VI)

a) K-means algorithm takes the input parameter k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the interc similarity is low.

Cluster similarity is measured in regard to the mean of the objects in a cluster which can be viewed as the cluster's centroid or centre of gravity.

Algorithm

Input :- K - The no: of clusters
D - Dataset containing n objects

Output :-
A set of arbitrary clusters

Method

i) Arbitrarily choose k objects from D as the initial cluster centres

ii) Repeat

iii) (Re) Assign each object to the cluster to which the object is similar, based on the mean value of the objects in the cluster

iv) Update the cluster means; calculate the mean value of the objects for each cluster.

v) Until no change

First it randomly select k of the objects, each of which initially represents a cluster mean or center.
For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance b/w the object and cluster mean.
It then computes the mean for each cluster.
This process iterates until the criterion function.
Typically square error criterion is used

$$E = \sum_{i=1}^{k} \sum_{P \in C_i} |P - m_i|^2$$

P - represents an element in cluster C_i
m_i - mean of cluster C_i

Computational complexity

$$O(nkt)$$

n - no: of objects
k - no: of clusters
t - no: of iterations

Disadvantages of K means

→ Cannot be applied for data containing categorical attributes

→ Users need to specify the no of clusters in advance

→ Not suitable for discovering clusters with non convex shapes or clusters of very different size

→ It is sensitive to noise and outlier datapoints

19 b) Term Frequency (TF) – Measures how frequently a term occurs in a document

$$TF(d,t) = \begin{cases} 0 & , \text{if } freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & , \text{otherwise} \end{cases}$$

$t \rightarrow$ term

$d \rightarrow$ document given

Inverse Document Frequency (IDF) – Measures how important a ~~lesson~~ term is. It represents the scaling factor or importance of t.

$$IDF(t) = \log \frac{1 + |d|}{|d_t|}$$

where d is the document collection

d_t is the set of documents containing the term t

$$TF\_IDF(d,t) = TF(d,t) \times IDF(t)$$

$$TF(d_3, t_4) = 1 + \log(1 + \log(freq(d_3, t_4)))$$

$$= 1 + \log(1 + \log(6))$$

$$= 1.249968$$

$$IDF(t_4) = \log \frac{1 + |d|}{|d_t|}$$

$d \rightarrow$ total no: of documents = 4
D_1, D_2, D_3, D_4

$d_t \rightarrow$ set of documents containing t_4
$\{d_2, d_3, d_4\}$
so $d_t = 3$

$$= \log \frac{1+4}{3}$$

$$= \log \frac{5}{3}$$

$$= 0.22184$$

$$TF\_IDF(d_3, t_4) = TF(d_3, t_4) \times IDF(t_4)$$

$$= 1.249 \times 0.221$$

$$= \underline{0.2760}$$

OCTOBER 2019

PART A

7. Differentiate b/w support and confidence (Mod V)

May 2019 PART A Qn 7

8. How to compute the dissimilarity b/w objects described by binary variables? (Mod V)

→ The distance between objects may be calculated based on a contingency table.

→ A binary attribute is symmetric if both of its states are equally ~~variable~~ valuable

→ In that case, using the simple matching coefficient can assess dissimilarity b/w 2 objects.

$$d(x_i, x_j) = \frac{r + s}{q + r + s + t}$$

where,

q = no: of attributes that equal 1 for both obj

t = no: of attributes that equal 0 for both obj

s and r = no: of attributes that are unequal for both obj

→ A binary attribute is asymmetric, ~~because~~ if states are not equally important (usually +ve outcome is considered more important)

→ In this case, the denominator ignores the unimportant negative matches (t). This is called Jaccard coefficient

$$d(x_i, x_j) = \frac{r + s}{q + r + s}$$

9. Differentiate b/w Agglomerative and Divisive hierarchical clustering method. (Mod VI)

Agglomerative Hierarchical Clustering

This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a cluster or until certain termination conditions are ~~required~~ satisfied. Most hierarchical clustering methods belong to this category. They differ only in their definition of inter-cluster similarity.

Divisive Hierarchical Clustering

This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the distance b/w the two closest clusters is above a certain threshold distance.

More complex, more efficient and more accurate than that of agglomerative hierarchical clustering.

10 Explain Web Content mining? (Mod VI)

May 2019 PART A Qn 10.

The two diff approaches to web content mining

i) Agent Based Approach

Usually involves intelligent sys that can act autonomously or semi autonomously

It usually relies on autonomous agents (or bots) that can parse through the web, identifying websites that are relevant and then collect information from them.

These agents can be of 3 types → Intelligent Search Agents
→ Information Filtering "
→ Personalized Web "

ii) Data-based Approach

Used to organize semi-structured data present on the internet into more structured data sources

Once this done, standard data querying techniques can be used to harness the data.

The approach can have 2 subtypes → Multilevel database

PART D

14. Consider the transaction database given below. Let minimum support count as 2 and minimum confidence threshold as 70%

Transaction ID	List of Item Ids
T100	I_1, I_2, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	I_1, I_2, I_4
T500	I_1, I_3
T600	I_2, I_3
T700	I_1, I_3
T800	I_1, I_2, I_3, I_5
T900	I_1, I_2, I_3

a) Find frequent itemset using Apriori Algorithm (Mod V)
b) Generate strong association rules. (Mod V)

a) Apriori employs an iterative approach known as a level wise search where k itemsets are used to explore (k+1) itemsets.

Apriori property – All non empty subsets of a frequent itemset must also be frequent.

A two-step process is followed.

1. The join step: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself. The set of candidates is denoted by C_k.
2. The prune step: C_k is a superset of L_{k-1}, i.e., its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. To reduce the size of C_k the Apriori property is used.

Minimum support count = 2

Step 1: Scan Database D for count of each candidate.

C_1

Itemset	Sup count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

I_6.

Creating 1-frequent item

Compare candidate Support count with minimum Support count →

L_1

Itemset	Sup count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

Creating 2-frequent itemset, C_2.. created from L_1.

Generating C_2 from L_1 →

I_1 I_2
I_1 I_3
I_1 I_4
I_1 I_5
I_2 I_3
I_2 I_4
I_2 I_5
I_3 I_4
I_3 I_5
I_4 I_5

Scan D for each candidate's count. →

C_2

I_1 I_2	4
I_1 I_3	4
I_1 I_4	1
I_1 I_5	2
I_2 I_3	4
I_2 I_4	2
I_2 I_5	2
I_3 I_4	0
I_3 I_5	1
I_4 I_5	0

Compare candidate Support count with minimum Sup. count →

L_2

I_1 I_2	4
I_1 I_3	4
I_1 I_5	2
I_2 I_3	4
I_2 I_4	2
I_2 I_5	2

Creating 3-frequent itemset C_3 from L_2.

Generating C_3 from L_2 →

I_1 I_2 I_3
I_1 I_2 I_5
I_1 I_2 I_4
I_1 I_3 I_5
I_2 I_3 I_4
I_2 I_3 I_5

Scan D for count of each candidate →

C_3

I_1 I_2 I_3	2
I_1 I_2 I_5	2
I_1 I_2 I_4	1
I_1 I_3 I_5	1
I_2 I_3 I_4	0
I_2 I_3 I_5	1

Compare candidate Support count with minimum Sup. count →

L_3

I_1 I_2 I_3	2
I_1 I_2 I_5	2

56

Generating C_4 with 4-frequent itemset

$I_1\ I_2\ I_3\ I_5 \longrightarrow I_1\ I_2\ I_3\ I_5 \longrightarrow L_4 = \phi$

L_4 is empty so k is $L_{k-1} = L_3$.

So freq itemsets are $\{I_1\ I_2\ I_3\}$ & $\{I_1\ I_2\ I_5\}$.

So if $I_1\ I_2\ I_3$ is a frequent itemset then all its non-empty subsets must also be frequent, while checking we can see, I_1 ; I_2 ; I_3 ; $I_1 \cap I_2$; $I_1 \cap I_3$; $I_2 \cap I_3$; $I_1 \cap I_2 \cap I_3$ is also frequent. Ully $I_1\ I_2\ I_5$.

b) Generate strong association rules from itemsets (where strong association rules satisfy both min support and min confidence)

Association rules can be generated as follows

→ For each frequent itemset l, generate all nonempty subsets of l.

→ For every nonempty subset s of l, output the rule "$s \Rightarrow (l-s)$" if $\frac{\text{support\_count}(l)}{\text{support\_count}(s)} \geq \text{min\_conf}$, where min_conf is the minimum confidence threshold.

$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}$$

Taking any one from $\{I_1\ I_2\ I_3\}$ & $\{I_1\ I_2\ I_5\}$.

Taking $I_1\ I_2\ I_5$

Generating all non empty subsets from $I_1\ I_2\ I_5$

I_1 , I_2 , I_5 , $I_1 \cap I_2$, $I_1 \cap I_5$, $I_2 \cap I_5$, $I_1\ I_2\ I_5$

Support count of $l = \{I_1\ I_2\ I_5\}$ is 2 (from L_3)

Rule	s	l − s	Confidence
1	$I_1 \rightarrow$	$I_2 \cap I_5$	2/6 = 33.3%.
2	$I_2 \rightarrow$	$I_1 \cap I_5$	2/7 = 30%.
3	$I_5 \rightarrow$	$I_1 \cap I_2$	2/2 = 100 %
4	$I_1 \cap I_2 \rightarrow$	I_5	2/4 = 50 %
5	$I_1 \cap I_5 \rightarrow$	I_2	2/2 = 100 %
6	$I_2 \cap I_5 \rightarrow$	I_1	2/2 = 100 %
	$I_1 \cap I_2 \cap I_5$.	ϕ	So not considering since it empty in Rhs

Min. conf = 70% (given in ques).

Here Rule 3, 5, 6 has confidence > 70%

So strong association rules are

$I_5 \rightarrow I_1 \cap I_2$

$I_1 \cap I_5 \rightarrow I_2$

$I_2 \cap I_5 \rightarrow I_1$.

18 a) Explain DBSCAN algorithm (Mod VI)

b) State the pros and cons of DBSCAN method (Mod VI)

a) A density-based clustering method based on connected regions with sufficiently high density — DBSCAN.

DBSCAN is a density based clustering algorithm. The algorithm grows regions with sufficiently high density into clusters, and discovers clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal set of density-connected points

The basic ideas of density based clustering involve a number of new definitions → The neighborhood ~~within~~ within a radius ε of a given object is called ε-neighborhood of the object.

→ If the ε neighborhood of an object contains atleast a minimum number, MinPts, of objects, then the object is called a core object.

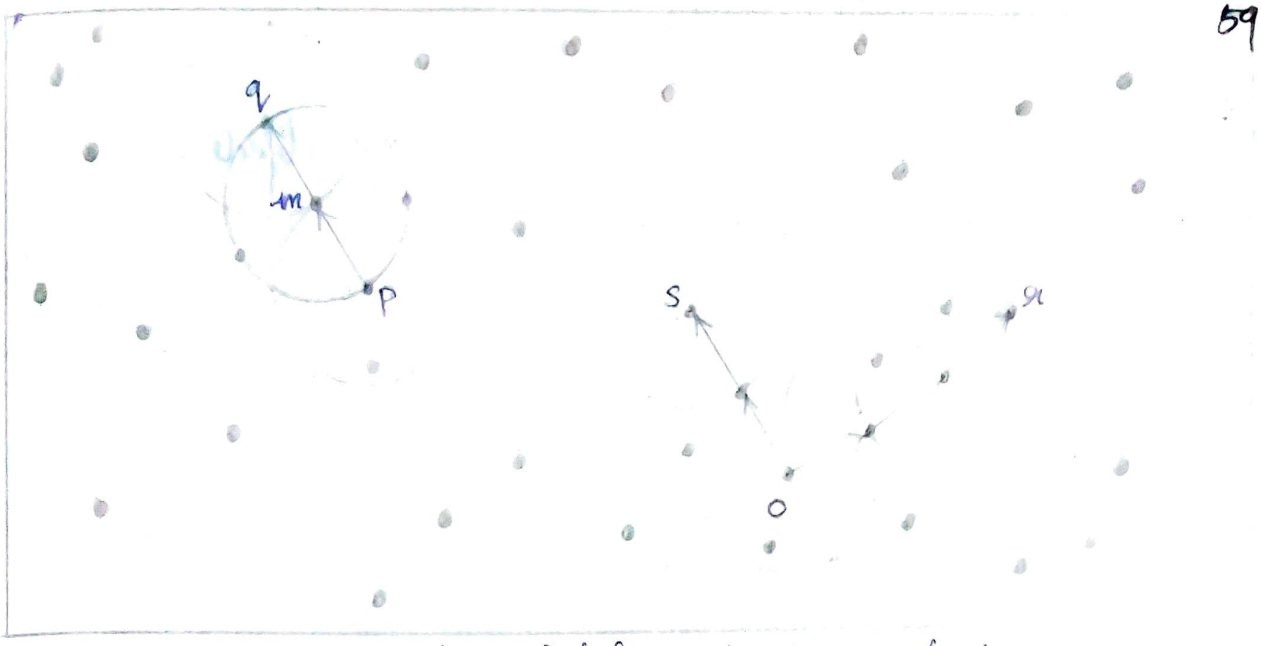
→ Given a set of objects, D, we say that an object 'p' is directly-density reachable from object 'q' if p is within the ε-neighborhood of q, and 'q' is a core object.

→ An object p is density reachable ~~from~~ from ~~from~~ object q wrt ε and MinPts in a set of objects D, if there is a chain of objects $p_1, p_2, \ldots, p_n$ where $p_1 = q$ and $p_n = p$ such that p_{i+1} is directly density reachable from p_i wrt ε and MinPts, for $1 \leq i \leq n$, $p_i \in D$.

→ An object 'p' is density connected to object q wrt ε and MinPts in a set of objects, D, if there is an object $o \in D$ such that both p and q are density reachable from o wrt ε and MinPts.

Density reachable is the transitive closure of direct density reachability and this relationship is asymmetric.
Only core objects are mutually density reachable.
Density connectivity, however, is a symmetric relation.

Consider the fig for a given ε represented by the radius of the circles, and let MinPts = 3.



Density reachability & Density connectivity

→ Of the labeled points M, P, O, and R are core objects since each is in an ε-neighborhood containing atleast 3 points

→ M is directly reachable from P and Q is directly density-reachable from M.

→ Based on the previous observation Q is (indirectly) density-reachable from P. However P is not density reachable from Q. Similarly R and S are density reachable from O

→ O, R and S are all density connected.

A density cluster is a set of density-connected objects that is maximal with respect to density reachability. Every object not contained in any cluster is considered to be noise.

How does DBSCAN find clusters ?

DBSCAN checks the neighborhood of each point in the database. If the ε-neighborhood of a point p contains more than MinPts, a new cluster with p as core obj is created. DBSCAN then iteratively collects directly density-reachable objects from these core objects, which may involve the merge of a few density-reachable clusters. The process terminates when no new point can be added to any cluster.

60

Algorithm :

I/p : D : a data set containing n objects

ϵ : the radius parameter

MinPts : the neighborhood density threshold

O/p : A set of density based clusters.

Method

1) Mark all objects as unvisited;
2) do
3) randomly select an unvisited object p
4) marke p as visited;
5) if the ϵ-neighborhood of p has atleast MinPts objects
6) create a new cluster C and add p to C
7) Let N be the set of objects in the ϵ-neighborhood of p
8) for each point p' in N
9) if p' is unvisited
10) marke p' as visited
11) if the ϵ neighborhood of p' has atleast MinPts points
12) add those points to N;
13) if p' is not yet a member of any cluster,
add p' to C
14) end for
15) output C
16) else marke p as noise
17) until no objects is unvisited.