COMP7.2 DATA MINING

Subject Code	Name of the Subject	Scheme of Instruction Hrs/Week			Scheme of Examination						
		L	T	P	Th Duration(Hrs)	Marks					
						Th	S	TW	P	O	Total
COMP7.2	Data Mining	3	1	2	3	100	25	--	--	--	**125**

Course Objectives:

1. This course will focus on imparting a complete introduction to data mining for students.
2. It will provide a sound understanding of the foundations including fundamental concepts and algorithms of data mining which intend to search through data for hidden relationships and patterns.

Course Outcomes:

The student after undergoing this course will be able to:

1. Understand the basic concept of preprocessing the data before using it in the mining algorithms and different types of data mining tasks.
2. Understand and apply the principles of various classification techniques
3. Understand and apply the principles of various association mining techniques.
4. Study and apply various clustering algorithms and anomaly detection techniques.

UNIT -1 (12 Hours)

Introduction –Challenges, Origin of Data Mining, Data Mining Tasks

Data -- *Types of Data*:Attributes and Measurement , Types of Data Sets
Data Quality:Measurement and Data Collection Issues, Issues Related to Applications
Data Pre-processing : Aggregation, Sampling, Dimensionality Reduction, Feature Subset Selection, Feature Creation, Discretization and Binarization, Variable Transformation.

UNIT - 2 (12 Hours)

Measures of Similarity and Dissimilarity

Similarity and Dissimilarity between Simple Attributes Dissimilarities between Data Objects Similarities between Data Objects Examples of Proximity Measures Issues in Proximity Calculation Selecting the Right Proximity Measures.

Summary Statistics

Frequencies and the Mode, Percentiles,

Measures of Location: Mean and Median

Measures of Spread: Range and Variance

Multivariate Summary Statistics, Other Ways to Summarize the Data

Data Cube and OLAP

Data Cube: A multidimensional data model.*Schemas for Multidimensional data model* : Star, Snowflakes and Fact Constellation schemas. *Dimensions*: The role of Concept Hierarchies. *Measures*: Categorization and Computation. *OLAP Operations.*

UNIT - 3 (12 Hours)

Classification

General Approach to Solving a Classification Problem

Decision Tree Induction: Working, Construction, Methods for Expressing Attribute Test Conditions, Measures for Selecting the Best Split, Algorithm and Characteristics for Decision Tree Induction.

Model Overfitting: Overfitting Due to Presence of Noise, Overfitting Due to Lack of Representative Samples, Overfitting and the Multiple Comparison Procedures, Estimation of Generalization Errors, Handling Overfitting in Decision Tree Induction.

Rule-Based Classifier: Concept, Rule-Ordering Schemes, Building a Rule-Based Classifier Direct Methods for Rule Extraction, Indirect Methods for Rule Extraction, Characteristics of Rule-Based Classifiers.

Nearest-Neighbor classifiers: Algorithm, Characteristics of Nearest-Neighbor Classifiers.

UNIT - 4 (12 Hours)

Association Analysis

Frequent Itemset Generation, The Apriori Principle, Frequent Itemset Generation in the Apriori Algorithm, Candidate Generation and Pruning, Support Counting, Computational Complexity, *Rule Generation*: Confidence-Based Pruning, Rule Generation in Apriori Algorithm, Maximal Frequent Itemsets, Closed Frequent Itemsets.

Alternative Methods for Generating Frequent Itemsets.

Cluster Analysis

K-means :The Basic K-means Algorithm, K-means: Additional Issues, Bisecting K-means K-means and Different Types of Clusters, Strengths and Weaknesses.

Agglomerating Hierarchical Clustering: Basic Agglomerative Hierarchical Clustering Algorithm, Key Issues in Hierarchical Clustering, Strengths and Weaknesses.

Anomaly Detection

Statistical Approaches :Detecting Outliers in a Univariate Normal Distribution, Outliers in a Multivariate Normal Distribution, A Mixture Model Approach for Anomaly Detection

Strengths and Weaknesses.

List of Experiments:

(At least 8 experiments should be conducted from the list of experiments. A certified journal reporting the experiments conducted should be submitted at the end of the term)

1. Implementation of data preprocessing techniques
2. Implementation of measures of similarity and dissimilarity.
3. Study of Data Cube and OLAP Operations.
4. Implementation of K-Nearest Neighbor Classifier.
5. Implementation of Classification by Decision Tree Induction.
6. Implementation of Apriori Algorithm.
7. Implementation of FP Tree.
8. Implementation of K-Means Clustering Algorithm.
9. Implementation of Bisecting K-Means Clustering Algorithm.
10. Implementation of Agglomerative Hierarchical Algorithm.

Recommended Readings:

TEXT BOOK

1. Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson Education, ISBN:81-317-1472-1

REFERENCE BOOK

1. Data Mining - Concepts and Techniques by Jiawei Han and Micheline Kamber, Elsevier, Second Edition, Original ISBN: 978-1-55860-901-3, Indian Reprint ISBN: 978-81-3120535-8

Date 17-07-19

Introduction

Data mining :-

A process that uses various techniques to discover patterns or knowledge from data using visualization and machine learning algorithm.

In the other words it refers to extracting or mining knowledge from large amount of data.

eg :- fraud detection
new medicine
marketing
stocks
revenue management etc.

EB (Exa bytes)
PB (Peta bytes)
TB (Tera bytes) ↑
GB (Giga bytes)
MB (Mega bytes)
KB (Kilo bytes)
Bytes
Bits

5 marks

Data mining as a step in the process of knowledge discovery

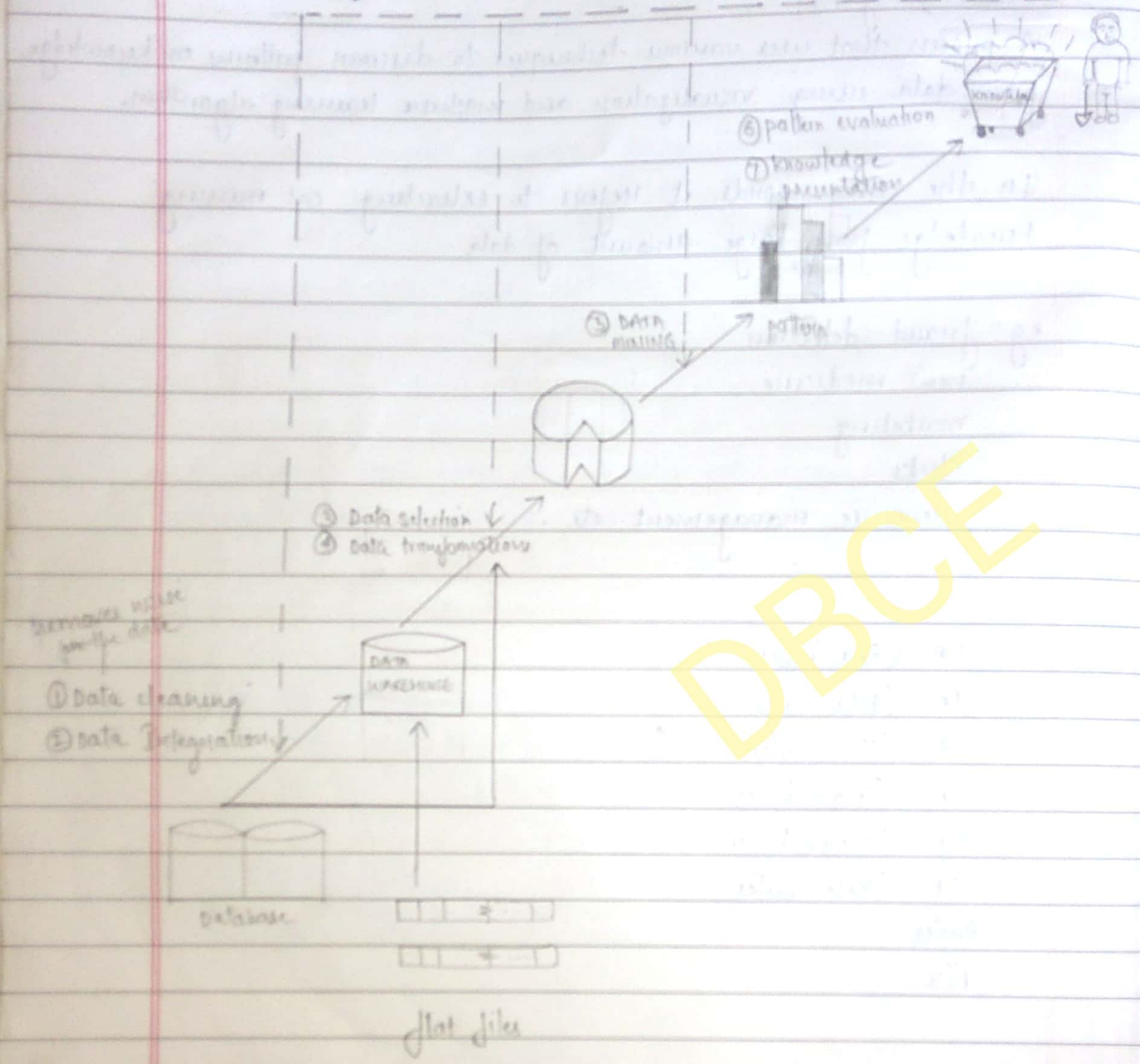


Figure: DM as a step in the process of knowledge discovery.

DBCE

classmate
Date 17/7/19
Page

1) Data cleaning
To remove noise and inconsistent data.

2) Data integration
Here multiple data sources may be combine.

3) Data selection
Here data relevant to the analysis task are retrieve from the database.

4) data transformation
Here data are transformed or consolidated in a form appropriate for mining by performing summary or aggregation operation.

5 Data mining
It is an essential process intelligent method are applied inorder to extract data patterns.

6 Pattern evaluation
To identify truely ~~intellio~~ interesting patterns representing knowledge based on some interestingness measures.

7) Knowledge presentation
Here visualization and knowledge representation techniques are used to present the mind knowledge to the user.

Date 18/07/19

What is not Data mining?

1 Searching for a phone number in phonebook or directory.
2 Searching for keywords on google.
3 Generating histogram of salaries for different age groups.
4 Generating multidimensional cube of relational table.
5 Issuing SQL query to a b. database and reading its reply.

What is Data Mining?

1 It is finding a people with similar hobbies.
2 ~~There are~~ chances of getting cancer are higher if you live near a powerline.
3 Grouping together similar documents returned by search engine according to their context.

Origins of data mining

- Draws ideas from machine learning / AI, pattern recognition statistics and database system for applications that have anormous data (large data)
- Traditional techniques may be unsuitable due to data i.e
  large scale
  complex
  high dimensional
  distributed
  heterogeneous
  unstructured
- A key components of the emerging field of data science & data driven discovery.

classmate
Date 18/07/19
Page

Q9
6 marks

Challenges in data mining

1 noisy and incomplete data
The real world data is heterogenous, incomplete and noisy due to errors of the instruments that measures the data or because of human errors. All these result in inaccurate, unreliable and incomplete data which makes data mining challanging.

2 distributed data/ ~~Data~~ Ownership and Distribution
Real world data is stored on different platforms. in distributed computing environment. So it is practically very difficult to bring all data to centralized data reposatory due to ~~orangal~~ Organizational and technical reasons. so data mining demands development of tools and algorithms that enables mining of distributed data.

~~Heterogeneous and~~
3 Complex Data
Realworld data is heterogenous and it could be multimedia data including images, audio, video, time series, natural language text and so on.
So new tools and methods would have to be developed to extract relevant information from such different kinds of data.

4. Performance
The performance of data mining system depends on efficiency of algorithm and techniques used if they are not up to the mark, it will affect the performance of data mining.

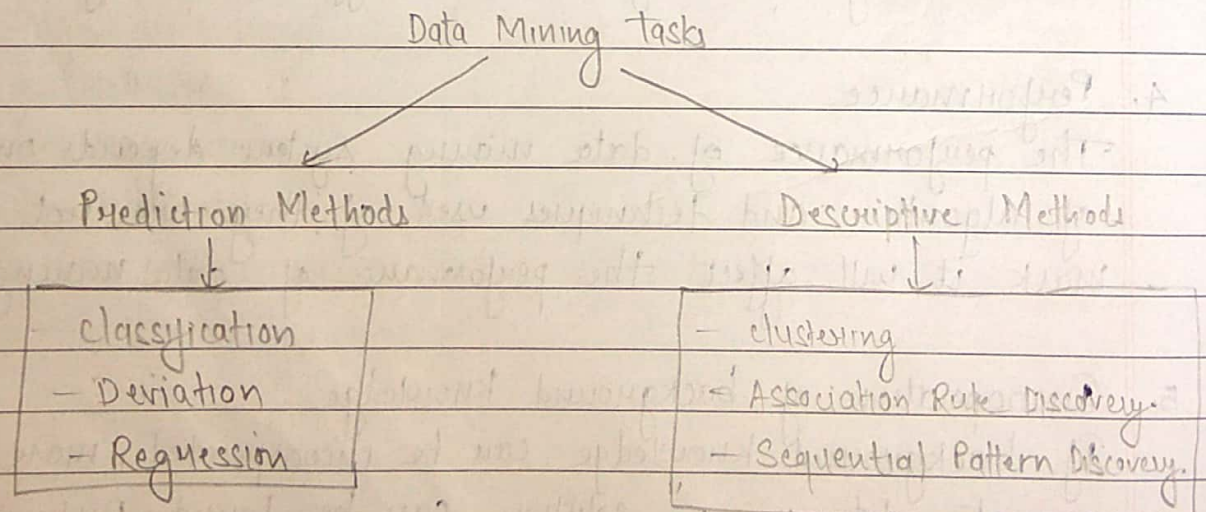
5. Incooperation of background knowledge
If background knowledge can be incooperated, more reliable & accurate data mining solutions can be found. Distriptive task can come up with more useful funding and predictive task can make more accurate prediction. but collecting an encorpatating

background knowledge is also a complex process.

6. Data visualization
It is a main process of data mining that displays output in presentable manner to the user.
The information extracted should convey exact meaning of what it actually intense to convey. but many times it is really very difficult.

7. Data privacy and security
Data mining normally leads to serious issues interms of data security privacy and governance
for eg when a retailer analysis the purchase details, it reveals information about buying habbits and preferances of customers without their permissions.

Data mining tasks
In data mining tasks there are 2 methods
1) Prediction Methods
2) Discriptive Methods



* Predictive task

- The objective of these task is to predict the value of particular attribute based on the values of other attributes
- The attribute to be predicted is known as target or dependent variable while attribute used for making prediction is known as explanatory or independent variables.
- For example ① The type of a flower ② market basket analysis ③ Document clustering ④ Credit card fraud detection

* Descriptive task

The objective is to derive patterns (correlations, trends, clusters and anomalies) that summarized the underlying relationships in data

- Discriptive data mining task are often exploratory in nature. and frequently require post processing technique to validate and explain the results.

* Data mining and knowledge discovery.

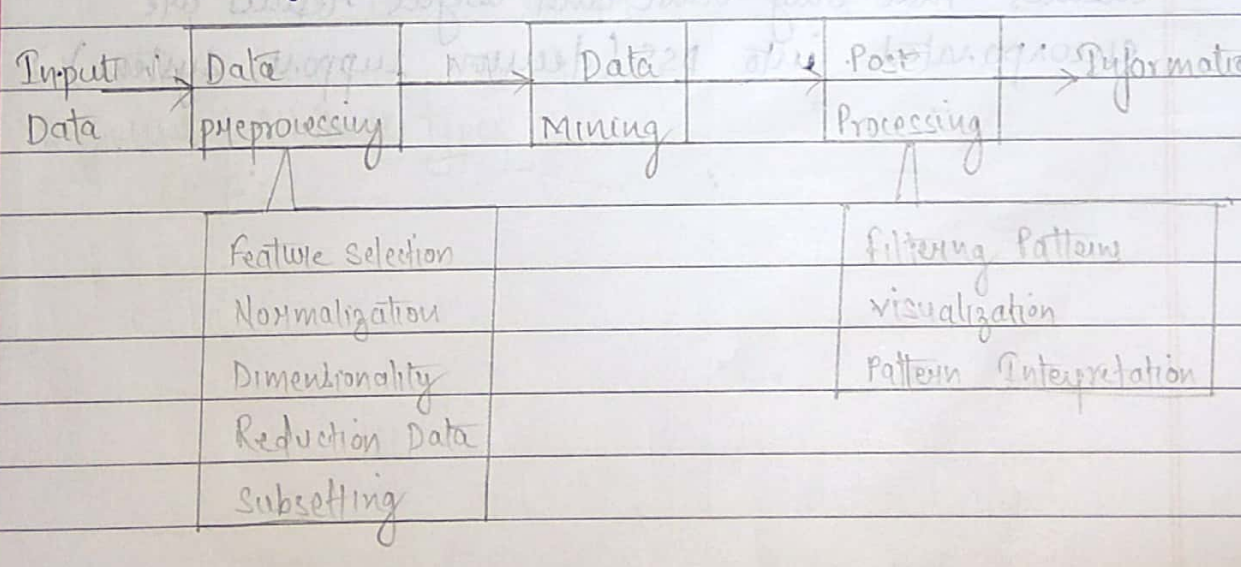


fig:- Process of KDD (Knowledge Discovery in Databases)

DBCE

Date 19/07/19

Data mining is an integral part of KDD which deals with converting raw data into information as shown in the figure above.
The input data can be stored in variety of formats like flat files, spreadsheets, or relational tables and may reside in centralized data repository or be distributed across multiple sides.
The purpose of preprocessing is to transform raw input data into appropriate format for subsequent analysis.

Steps involved in data processing include fusing data from multiple sources, cleaning data to remove noise & duplicate observations, selecting records and features that are relevant to data mining task at hand.
Closing the loop is the phraze used to refer to the process of integrating data mining results into decision support system
Such integration requires a post processing steps that ensures that only valid and useful results are encorparated into DSS (decision support system).

Date 19/07/19

Chapter 2 : Data

- Collection of data objects is called as datasets.
- Othernames for data objects are record, point, vector, pattern, event case, sample, observation or entity.
- Data objects are describe by a number of attributes as shown below that captures the basic characteristics of an objects.
- Othernames for attributes are variables characteristics, fields, features or dimensions.

Attributes

Data objects

Studid	Name	class	Payed fees
100	Arun	FE	Yes
101	Bill	TE	No
102	Caron	SE	No
103	Derick	FE	Yes
104	Elston	BE	Yes

8 marks

- Different Attribute types

(P.T.O)

Q3 & Q7

Categories of Data

Qualitative: (Attribute / count)

- ✓ Non numerical
- ✓ Mathematical operations are MEANINGLESS

(colours, political party, religions, gender, car color, car body type, etc.)

Quantitative:

- ✓ Natural numericals
- ✓ Mathematical operations are MEANINGFULL

(height, weight, wages, temperature, time, age, volume)

Discrete

- ✓ countable
- ✓ finite
- ✓ only certain values are possible

(ex. grades, A, B, C, D)

Continuous

Countable / variable / Measurable / Quantitable

- ✓ Not countable
- ✓ Infinite possible values usually a measurements

(temperature, pressure, length, time)

NOMINAL (=, ≠)

- ✓ Name only
- ✓ labels
- ✓ Not Ordered
- ✓ eg. gender, religion

ORDINAL (<, ≤, >, ≥)

- ✓ Meaningful difference betn units
- ✓ Size (small, medium, large)
- ✓ list of Rich people

INTERVAL (+, −)

- ✓ Ordered
- ✓ Meaningful or equal difference betn units
- ✓ temp in °C or F

RATIO (×, ÷)

- ✓ Ordered
- ✓ meaningful or equal difference betn units
- ✓ Temp in °K, dollar

Binary

Non Binary

classmate Date 19/07/19 Page

Date 23/07/19

Categorical [Qualitative]
- * Nominal :- Distinctness
- * Ordinal :- Distinctness and order

Numeric [Quantitative]
- * Interval :- Distinctness ~~and~~, order and addition
- * Ratio :- Distinctness ~~and~~ order ~~and~~, addition and Ratio

Data Sets

Characteristics of data sets

1) Dimensionality
2) Sparsity
3) Resolution

①

Stud id	class		Stud id	class		Stud id	class
1	FE						
2	FE						
3	FE						
41	BE						
51	BE						

②

Empid	Salary	Address
1	10000	ABC
2	10000	
3	0	XYZ
4	5000	PQR

Q5

Datasets

General characteristics of Data sets

1) Dimensionality

- It is the number of attributes that the objects in the dataset possess.
- Data with a small number of dimensions tends to be qualitatively different than moderate or high dimensional data
- Difficulties associated with analysing high dimensional data are sometimes reffered as curse of dimensionality (COD)

2) Sparsity

- Dataset with assymmetric features most attribute of an objects have of zero, fewer than 1% of the entries are nonzero.
- It is an advantage because usually only nonzero values need to be stored and manipulated. which results in saving computation time and storage.
- Such data are called as sparse data.

3) Resolution

It is possible to obtain data at different levels of resolution. and properties of data are different at different resolution.

Q5 types of datasets

The datasets are grouped into 3 type

1) Record data
2) Graph base data
3) Ordered data

DRCE

classmate
Date 23/7/19
Page

1) Record data

- It is a collection of records, each of which consists of a fix set of data fields or attributes.
- They are categorised as shown below

Eid	Name	Phone no.	Paid sem fees	Class
100	Arun	9876543210	Yes	FE
101	Bala	9090123456	Yes	SE
102	Cauj	9010654321	Yes	SE
103	Derick	9923454545	Yes	TE

(a) Record Data

Tid	Elements
1	Bread, soda, Milk
2	Beer, Soda, Milk
3	Beer, Bread, Soda

(b) Transaction data

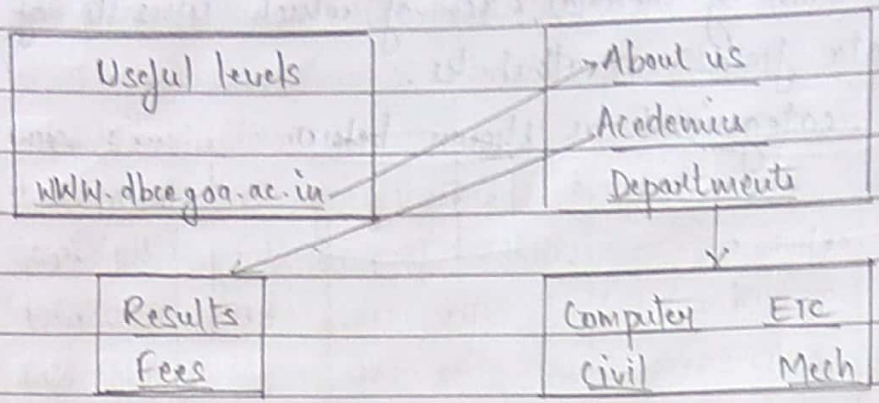
Project of 'x' load	Projectn of 'y' load	Distance	load	thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.21	22	1.1
13.54	7.23	17.34	23	1.2

(c) data matrix

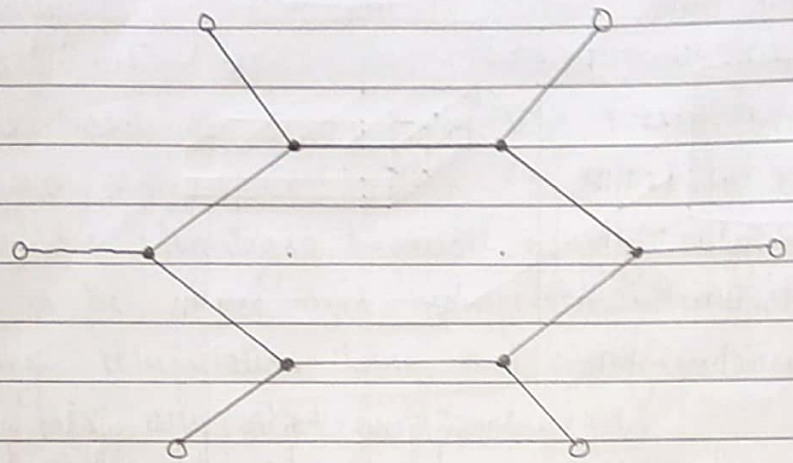
	Team	Coach	Play	Ball	lost	win
Document 1	3	0	5	0	0	2
Document 2	1	7	0	2	0	3
Document 3	1	1	0	0	2	0

(d) Document-term matrix

② Graph Based Data



(a) linked web pages



(b) Benzene Module

3) Ordered Data

a) Sequential transaction data (Temporal data)

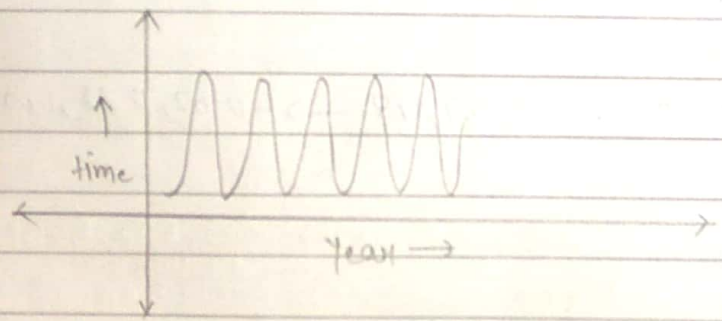
Time	Customer	Item Purchased
t_1	C_1	A, B
t_2	C_2	A, C
t_3	C_3	C, D
t_4	C_4	A, D

Customer	Time & Item Purchased
C_1	(t_1: A, B) (t_3: C, D)
C_2	(t_4: A, D)
C_3	(t_2: A, C)

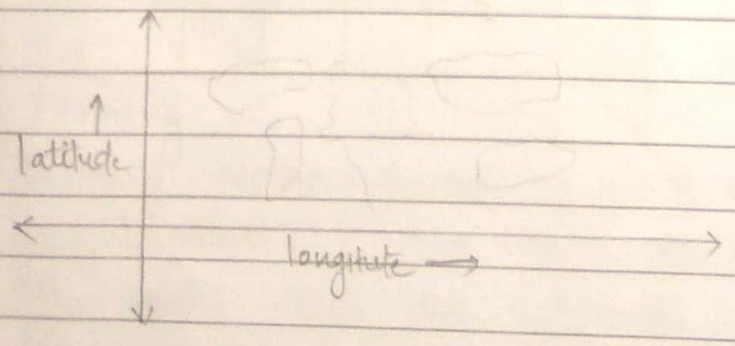
b) Genomic sequence data

GGTTCGCCTTCACTTACG
CGCTCGTTCAGGTTCGCA
GAGGATTCAGGTTCGCC

c) Temperature time series



d) Spatial temperature date



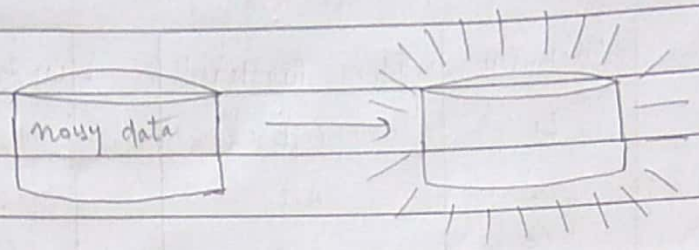
Q) Write in short - what is data cleaning
data integration, data reduction
data transformation

classmate
Date 24/07/19
Page

Data preprocessing

Why preprocess the data?

1) Data cleaning



2) Data Integration

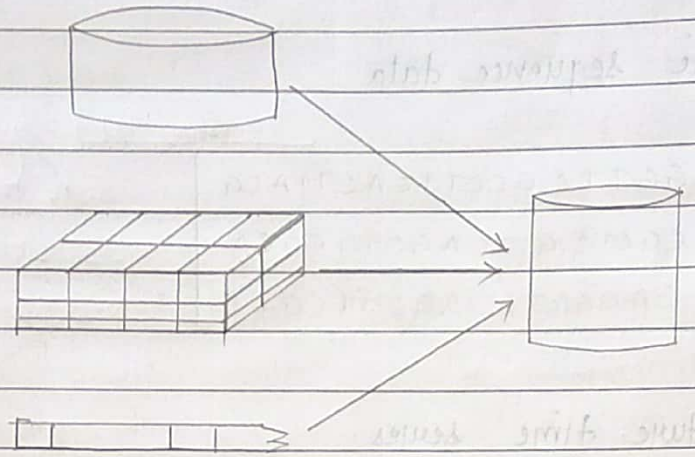


figure : Forms of Data preprocessing

3) Data transformation : $-2, 32, 100, 59, 48 \rightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

4) Data reduction

Attributes $-2, 32, 100, 59, 48 \rightarrow -0.02, 0.32, 1, 0.59, 0.48$

Transactions	A_1	A_2	A_3		A_{126}
T_1					
T_2					
...					
T_{2000}					

→

Transactions	A_1	A_2	...	A_{115}
T_1				
T_2				
...				
T_{400}				

classmate
Date 24/07/19
Page

1) Noisy (containing errors or outliers values that deviate from the expected)

2) Inconsistent (containing discrepancies in department codes used to categorize items)

3) Incomplete : (lacking attribute values or certain attributes of interest, or containing only aggregate data)

I Data cleaning

- Real world data tend to be incomplete, noisy and inconsistent
- Data cleaning routings attempts to fill in missing values, smooth out noise while identifying outliers & correct inconsistencies in the data.
- Basic methods for data cleaning are categorised based on
  1) Noisy data
  2) Missing values
  3) Inconsistent data

[1 Noisy Data]

- Noise is an random error or variance in a measured variable.
- Given a numeric attribute such as price how can we smooth out data to remove noise is explain using different ~~st~~ data smoothing technique.

a) Binning

Binning methods smooth a sorted data value by consulting its neighbourhood i.e the values around it.
The sorted values are distributed into a number of buckets or bins. where each bin performs local smoothing as shown in the example below.

DBCE

eg 1 Perform smoothing on binning for the sorted data for price (in dollars) : 4, 8, 15, 21, 21, 24, 25, 28, 34 with bins of depth 3.

→ Step 1

Partition into (equidepth) bins

Bin 1 : 4, 8, 15

Bin 2 : 21, 21, 24

Bin 3 : 25, 28, 34

Step 2

Smoothing by BIN means

Bin 1 : 9, 9, 9 $4+8+15 = 27/3 = 9$

Bin 2 : 22, 22, 22 $21+21+24 = 66/3 = 22$

Bin 3 : 29, 29, 29 $25+28+34 = 87/3 = 29$

Step 3

Smoothing by BIN Boundaries

Closest value

Bin 1 : 4, 4, 15

Bin 2 : 21, 21, 24

Bin 3 : 25, 25, 34

Step 4

Smoothing By BIN Median

centre value

Bin 1 : 8, 8, 8

Bin 2 : 21, 21, 21

Bin 3 : 28, 28, 28

b) Clustering

Outliers may be detected by clustering, where similar values are organized into groups or cluster. Values that fall outside of the set of clusters may be considered as outliers.

c) Combined computer and human Inspection

Outliers may be identified through a combination of computer and human inspection.

- Outlier patterns may be informative or garbage. Patterns whose surprize content is above a threshold are output to a list.
- A human can than sort through the patterns in the list to identify the actual garbage values. This is much faster than having to manually search through the entire database.
- The Garbage pattern than can be excluded from used in subsequent data mining

d) Regression

Data can be smoothen by fitting the data to a function. such as with regression.

- Linear regression involves finding the best line to fit 2 variables so that one variable can be used to predict the other.
- Multiple linear regression is an extension of linear regression where more than 2 variables are involved and data are fit to a multidimensional surface.
- Regression is defined as predicting a value of a even continuous valued variable based on the values of other variables.

2 Missing values

To fill in the missing values different methods are used they are as follows:

a) Ignore the tuple
b) fill in the missing values manually
c) use global constants to fill in missing values. ∞ or unknown
d) use attribute mean
e) use attribute mean for all samples belonging to the same class
f) use Most Probable value to fill the missing value.

3. Inconsistent Data

There may be inconsistency in the data recorded for some transactions which ~~may~~ can be correct manually using external references.

- This may be coupled with routines design to help correct the inconsistent use of codes.
- knowledge engineering tools may also be used to detect violation of known data constraints.
- There may also be inconsistency due to data integration where a given attribute can have different name in different databases.

01/08/19

II Data Integration

It is the merging of data from multiple data stores in other words it combines from other sources into a coherant data store as in data where housing.

- This sources may include multiple databases, datacubes, flatfiles etc.
- The issues during data integration are

① Redundancy

An attribute may be redudant if it can be derived

from another table.
for eg. Annual income can be derived from monthly income such things are detected by correlation analysis for eg. given two attributes such analysis can measure how strongly one attribute implies the other based on available data

- The correlation betn attribute A and B can be measured by

$$r_{A,B} = \frac{\sum (A-\bar{A})(B-\bar{B})}{(n-1)\,\sigma_A \cdot \sigma_B}$$

mean of A is $\bar{A} = \frac{\sum A}{n}$

$\sigma_A = \sqrt{\sum (A-\bar{A})^2}$

where n is number of tuples
$\bar{A}$ & $\bar{B}$ means or mean values of A & B
σ_A & σ_B are standard deviation of A & B.

If $r_{A,B} > 0$ than A & B are positively correlated. meaning the values of A increase as the values of B increase.

If $r_{A,B} = 0$ than A & B are independent and there is no correlation between them.

If $r_{A,B} < 0$ than A & B are negetivity correlated where values of one attribute increase as the values of the other attribute decrease.

(11) Schema integration can be tricky.
How can equivalent real world entities from multiple datasource be matched up is dealt here. This is refferred to as entity identification problem.
for eg. how can be data analyslist or the computer be sure that cust_id in one database & cust_number in another database refer to the same entity. Such metadata (data about the data) can be used to help avoid errores in schema

integration.

③ Detection and resolution of data value conflicts
For the same real world entity, attributes values from different sources may differ.
This may be due to differences in representation, scaling or encoding.

Handling all the above issues of data integration can help improve the accuracy and speed of data mining process.

III ② Data Transformation
Here data are transform or consolidated into forms appropriate for mining.
Data transformation can involve the following.

i) Smoothing
Which works to remove the noise from data such techniques include binning, clustering & regression.

ii) Aggregation
Here summary or aggregation operations are applied to the data.
for eg. Daily sales may be aggregated to compute monthly and annual total amount/sales.
This step is typically used in constructing for datacubes for analysis of data.

iii) Generalization of data
Here low level or primitive (raw) data are replaced by Higher level concepts through the use of concept hierarchy.

DBCE

for eg. Attribute, streets can be generalized to higher level concept like city or country.

iv) Attribute construction or feature construction
Here new attributes are constructed and added from the given set of attributes to help data mining.

v) Normalization
Here attributes data are scaled so that they fall within a small specified range.
There are 3 types of normalization
A) Min-Max normalization
B) Z-score normalization
C) Normalization by decimal scaling

A) Min-Max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new_max_A - new_min_A) + new_min_A$$

eg.1 Suppose that minimum and maximum values for the attribute income are Rs 12,000 & Rs. 98,000. We would like to map income to the range of [0.0, 1.0]. by min-max normalizat^n, a value of Rs 73,600 for income is transformed to ?

Rs 12,000	Rs 73,600	Rs 98,000
0	?	1

→ $min_A = 12,000$
$max_A = 98,000$
$v = 73,600$
$v' = ?$

$$v' = \frac{73600 - 12000}{98000 - 12000}(1.0 - 0) + 0 = \frac{61600}{86000} \times 1 = 0.716$$

Date 02/09/19

Z-score normalization

$$\text{mean} = \bar{x} = \mu = \sum_{i=1}^{N} x_i / n$$

$$\text{standard deviation}, \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

$$\text{Variance}, v = \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Q) Use following method to normalize following group of data: 200, 300, 400, 600, 1000 (a) use Z-score normalization

→

$$\text{formula for Z-score} = v' = \cancel{\frac{v - \min_A}{\max_A - \min_A}} \quad \frac{V - \bar{A}}{\sigma_A}$$

Step 1: calculate mean for the given data

$$\bar{A} = \frac{200 + 300 + 400 + 600 + 1000}{5}$$

$$\bar{A} = 500$$

Step 2: Calculate standard deviation

$$\sigma = \sqrt{\frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}{5}}$$

$$= \sqrt{\frac{(-300)^2 + (-200)^2 + (-100)^2 + (100)^2 + (500)^2}{5}}$$

$$= \sqrt{\frac{90000 + 40000 + 10000 + 10000 + 250000}{5}}$$

$$= 282.8427$$

Date 8/8/19

Step 3: For $v = 200$

$$v' = \frac{200 - 500}{282.8}$$

$$= -1.06$$

For $v = 300$

$$v' = \frac{300 - 500}{282.8}$$

$$= -0.70$$

For $v = 400$

$$v' = \frac{400 - 500}{282.8}$$

$$= -0.35$$

For $v = 600$

$$v' = \frac{600 - 500}{282.8}$$

$$= 0.35$$

For $v = 1000$

$$= \frac{1000 - 500}{282.8}$$

$$= 1.76$$

V	200	300	400	600	1000
V'	-1.06	-0.70	-0.35	0.35	1.76

② Use min-max normalization method to normlize following group of data.

data : 200, 300, 400, 600, 1000

assume (min as 0 and max as 1)

→

$$v' = \frac{200 - 200}{1000 - 200}(1-0) + 0 = 0$$

$$v' = \frac{300 - 200}{1000 - 200}(1-0) + 0 = 0.125$$

$$v' = \frac{400 - 200}{1000 - 200}(1-0) + 0 = 0.25$$

$$v' = \frac{600 - 200}{1000 - 200}(1-0) + 0 = 0.5$$

$$v' = \frac{1000 - 200}{1000 - 200}(1-0) + 0 = 1$$

v	200	300	400	500	1000
v'	0	0.125	0.25	0.5	1

v.v.v.v Imp

Suppose that the data for analysis include attribute frequency of the stopword in document the values are given in increasing order. 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 30, 33, 33, 35, 35, 35, 36, 40, 45, 46, 52, 70

Apply following methods.

a) use data smoothing by Means, Boundaries, median with depth 3.

b) Use mean max normalization to transform value 40 into range from [0.0 to 1.0]

c) use z-score normalization to normalize value 70.

→

Step 1: Partition into (equidepth) bins

Bin 1 : 13, 15, 16

Bin 2 : 16, 19, 20

Bin 3 : 20, 21, 22

Bin 4 : 22, 25, 25

Bin 5 : 30, 33, 33

Bin 6 : 35, 35, 35

Bin 7 : 36, 40, 45

Bin 8 : 46, 52, 70

Step 2 Smoothing by BIN means

Bin 1 : 15, 15, 15

Bin 2 : 18, 18, 18

Bin 3 : 21, 21, 21

Bin 4 : 24, 24, 24

Bin 5 : 32, 32, 32

Bin 6 : 35, 35, 35

Bin 7 : 40, 40, 40

Bin 8 : 56, 56, 56

DBCE

Step 3: Smoothing by Bin Boundaries

Bin 1 : 13, 16, 16

Bin 2 : 16, 20, 20

Bin 3 : 20, 20, 22

Bin 4 : 22, 25, 25

Bin 5 : 30, 33, 33

Bin 6 : 35, 35, 35

Bin 7 : 36, 36, 45

Bin 8 : 46, 46, 70

Step 4 Smoothing by Bin median

Bin 1 : 15, 15, 15

Bin 2 : 19, 19, 19

Bin 3 : 21, 21, 21

Bin 4 : 25, 25, 25

Bin 5 : 33, 33, 33

Bin 6 : 35, 35, 35

Bin 7 : 40, 40, 40

Bin 8 : 52, 52, 52

b)

v	13	40	70
v'	0	0.47	1

$$v' = \frac{40-13}{70-13}(1-0)+0 = 0.47$$

c) Step 1: calculate mean

$\bar{x} = \cancel{28.29}$ $\frac{724}{24} = 30.16667 = 30.17$

Step 2: Calculate standard deviation

$$\sigma = \sqrt{\frac{(13-30.17)^2 + (15-30.17)^2 + (16-30.17)^2 + (16-30.17)^2 + (19-30.17)^2 + (20-30.17)^2 + (20-30.17)^2 + (21-30.17)^2 + (22-30.17)^2 + (22-30.17)^2 + (25-30.17)^2 + (25-30.17)^2 + (30-30.17)^2 + (33-30.17)^2 + (33-30.17)^2 + (35-30.17)^2 + (35-30.17)^2 + (35-30.17)^2 + (36-30.17)^2 + (40-30.17)^2 + (45-30.17)^2 + (46-30.17)^2 + (52-30.17)^2 + (70-30.17)^2}{24}}$$

$$\sigma = \sqrt{\frac{294.81 + 230.13 + 200.79 + 200.79 + 124.77 + 103.43 + 103.43 + 84.09 + 66.75 + 66.75 + 26.73 + 26.73 + 0.03 + 8.01 + 8.01 + 23.33 + 23.33 + 23.33 + 33.99 + 96.63 + 219.93 + 250.59 + 476.55 + 1586.43}{24}}$$

$$\sigma = \sqrt{\frac{4295.23}{24}} = 13.3778$$

step3: For $v = 70$

$$v' = \frac{70 - 30.17}{13.3778}$$

$$= 2.98$$

forms of data processing — U.V.V. Imp

classmate
Date 9/08/19
Page

(3) Normalization by decimal scaling

Here normalization is done by moving the decimal pts of values of attribute A. The number of decimal pts moved depends on the maximum absolute value of A.

A value of v of A is normalized to v' by computing

$v' = \frac{v}{10^j}$ where j is the smallest intergor such that

$Max(|v'|) < 1$.

Problem

(1) Suppose the recorded value of A is 35 than use normalization by decimal scaling to find v'.

→ $v = \frac{35}{10^2} = \frac{35}{100} = 0.35$ $\quad |35| = 35 \quad j=2$

(2) Suppose the recorded value of A range from −986 to 917. find the absolute value of all values and normalize by decimal scaling.

$|-986| \quad |-985| \ldots\ldots\ldots\ldots |910| \quad |917|$

⇓

986 so $j=3$

$v = \frac{986}{10^3} = \frac{986}{1000} = 0.986$

classmate
Date 9/8/'19
Page

IV Data Reduction

They are the technique that can be apply to obtain a reduce representation of dataset that is smaller in volumes, yet closely maintain the integrity of original data and thus makes mining process more efficient by producing the same analytical results.

Strategies for data reduction are

1. Data cube aggregation
2. Dimension Reduction → heuristic methods: Forward selection, Backward selection; → Decision Tree
3. Data compression → lossy technique, lossless
4. Numerosity Reduction → Regression & linear Models, Histogram, clustering, Sampling
5. Discritization & concept Hierarchy generation.

Data Cube Aggregation

1. It is any process in which information is gathered & expressed in summary form for statistical analysis.

OLAP (online analytical processing) is a simple type of data aggregation in which marketer uses online reporting mechanism to process information.

eg. We collect the data of all electronic sales per quarter for the year 1997 to 1999. but you are interested in annual sale rather than total per quarter thus data can be aggregated so that the resulting data summarize the total sales per year instead of per quarter. this results in dataset which is smaller in volume without loss of information necessary for analysis tasks. as shown in the figure below.

Year = 1999

Year = 1998

Year = 1997

Quarter	Sales
Q1	Rs 24000
Q2	Rs 408000
Q3	Rs 350000
Q4	Rs 586000

Year	Sales
1997	Rs. 15,68,000
1998	Rs. 25,56,000
1999	Rs 35,94,000

Data cube is a structure is used to represent data along some measure of interest it can be 2 dimensional 3 dimensional and higher dimensional.

- OLAP cube is a multi dimensional database that is optimize for data warehouse.
- Data cube provides fast access to pre computed summarized data there by benefitting online analytical processing as well as data mining.
- The cube created at lower level of abstraction is called as base cuboid which corresponds to an individual entity of interest such as sales
- A cube for the highest level of abstraction is apex cuboid which gives the total that is in the above eg. the total sales for all 3 years, for all item types or for all branches.
- Data cubes created for varing levels of abstract^n are called as a cuboids.

classmate
Date 13/08/19
Page

2. Dimensionality Reduction

- It reduces the dataset size by removing such attributes or dimensions from data.
- Typically methods of attribute/features subset selection are applied whose main goal is to find minimum set of attribute such that the resulting probability distribution of data classes is as close as possible to original distribution obtained using all attributes.
- There are 3 standard approaches of feature selection

A embedded approach
Feature selection occurs naturally as a part of data mining algorithm. Specifically during the operation of DM algorithm the algorithm itself decide which attribute to use and which to ignore.

B Filter approach (approches which are taken should not be good)
Features are selected before data mining algorithm is run. using some approach that is independent of DM tasks

C Wrapper approach (best attribute to be taken)
This methods use the target DM algorithm as a blackbox to find the best subset of attributes in a way similar to that of ideal algorithm but typically without enumerating all possible subset.

The best & worst attribute are typically determined using test of stastitical significance like above approches. which assume that attributes are independente of one another.

Basic heuristic methods of attribute subset selection are used which include techniques like

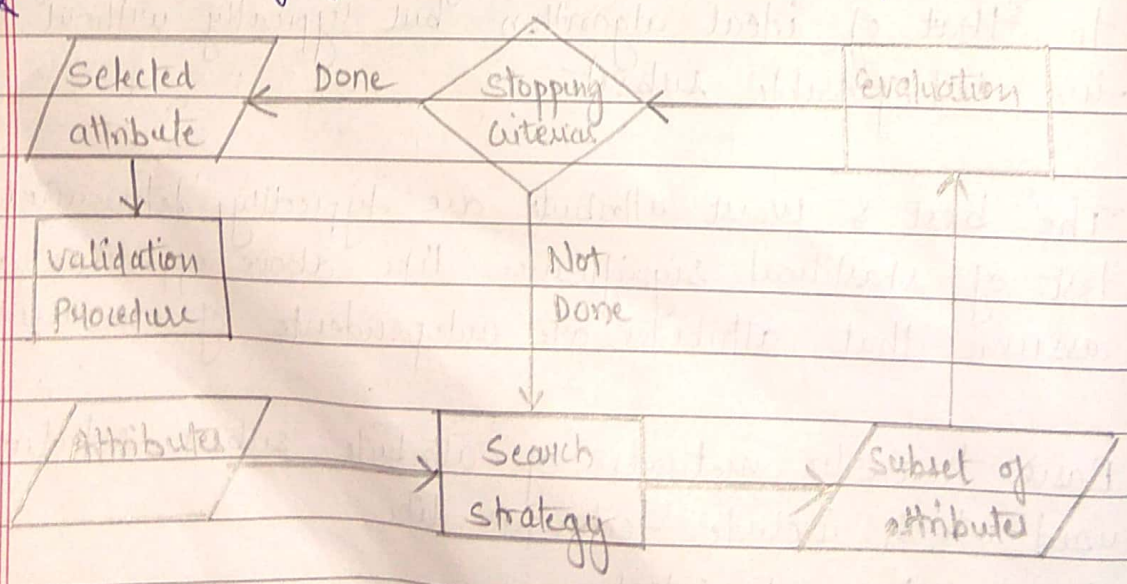
a) Stepwise forward selection

a Stepwise forward selection
The procedure starts with an empty set of attributes. The best of the original attribute is determined & added to the step. At each step best of the remaining original attribute is added to the set.

b Stepwise backward elimination
The procedure starts with full set of attributes. At each set, it removes the worst attribute remaining in the set.

Decision tree algorithms are also used for dimension production. Decision tree induction construct a flowchart like structure where each internal (non leaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test. and each external (leaf) node denotes a class prediction. At each node, the algorithm chooses the best attribute to partition the data into individual classes. All attributes that do not appear in the tree are assume ~~as~~ to be irrelevent.

14/08/19 Flowchart of feature subset selection process



Date 14/08/19

- The architecture for feature subset selection is shown above.
- The feature selection process consists of 4 parts
① measure for evaluating a subset. ② A search strategy that controls the generation of new subset of features.
③ Stopping criterian
④ validation procedure

- Feature subset selection is a search over all possible subsets of features
- Different types of search strategies can be used, but they should be computationally in expensive and should find optimal or near optimal sets of features.
- Evaluation step is used to judge how the current subset of features compares to others that have been considered. this requires an evaluation measure that attempts to determine the goodness of subset of attribute with respect to particular DM tasks such as classification or clustering.
- Bcoz the no. of subsets can be enormous & it is impractical to examine them all some set of stopping criteria is used. which involve one or more conditions like no. of iteration, whether the value of subset evaluation measure is optimal or exceeds certain threshold whether a subset which is obtained is of certain size whether simultaneous size and evaluation criteria have been achieved and whether any improvement can be achieved by option available to research strategy.
- Finally once the features or subsets of features has been selected the result of the target DM algorithm on the selected subset should be validated.
- A straight forward evaluation approache is to run the algorithm with full set of features & compare the full results to results obtained using subset of features.

- The subset of features will produce results that are better than or almost good as those produced when using all features
- Another validation approach is to used a number of different feature selection algorithm to obtain subsets of features & than compare the results of running the DM algorithm on each subset.

③ Data compression
- Here data encoding or transformation are applied to obtain a reduced or compress representation of original data.
- If original data can be reconstructed from compress data. without any loss of information the data compression technique use is lossless.
- If we can reconstruct only and approximation of original data than this technique is called lossy.
- The effective & popular methods of lossy data compression are wavelength transforms and PCA (Principle component analysis)

22-08-'19

④ Numerosity Reduction
Such techniques were applied to reduce the data volume by choosing alternative smaller forms of data representation.
- This techniques may be parametric or non parametric
- For parametric methods a model is used to estimate the data so that only data parameters need to be stored instead of actual data.
- eg. log linear models which estimate discrete multidimensional probability distributh.
- Non parametric methods are used for storing reduced representation of data which include ~~a hist~~

a) histogram

b) clustering
c) Sampling

a) Histograms
- Histograms used binning to approximate data distributions & are popular form of data reduction
- A histogram for an attribute 'A' partitions the data distribution of 'A' into disjoint subset or buckets.
- The buckets are displayed on horizontal axis while the height (or area) of a bucket reflects the average frequency of values represented by the bucket.
- If each bucket represents only a single attribute value or frequency pair the buckets are called singleton buckets.
- The several partitioning rules to determine the buckets and partition the attribute values include equidepth (equiheight), equiwidth, v-optimal, and maxdiff.
- v-optimal and maxdiff histogram tend to be most accurate & practical.

b) clustering
- clustering techniques consider data tuples as objects
- They partition object into groups or clusters so that objects within clusters are similar to one another and dis similar to objects in other cluster.
- The quality of the cluster may be represented by its diameter, the maximum distance betn any two objects in the cluster
- Centroid distance is an alternative measure of cluster quality and is define as the average distance of each cluster object from the cluster centroid

Date 22/8/19

Write short note on sampling?

c) Sampling

Sampling can be used as data reduction since it allows a large dataset to be representated by a much smaller random sample. all the subset of the data

- Suppose we have a large dataset 'D' and contains 'N' tuples than possible samples for 'D' are

1) Simple random sample without replacement. (SRSWOR) of size 'n':

- This is created by drawing 'n' of the 'N' tuples from 'D' ($n < N$) where the probability of drawing any tuple in 'D' is $\frac{1}{N}$ i.e all tuples are equally likely.

2) Simple Random sample with replacement of size n

- This is similar to SRSWOR that each time a tuple is drawn from 'D' it is recorded and then replace i.e after the a tuple is drawn it is placed back in D. so that it may be drawn again.

3) Cluster sample

If the tuples in 'D' are group into 'M' mutually disjoint clusters than simple random sample of 'm' clusters can be obtained where $m < M$

- For eg. tuples in a database are usually retrieved a page at a time so that each page is consider as a cluster.

4) Stratified Sample

If 'D' is divided into mutually disjoint parts called strata a stratified sample of 'D' is generated by obtaining a simple random sample at each stratum.

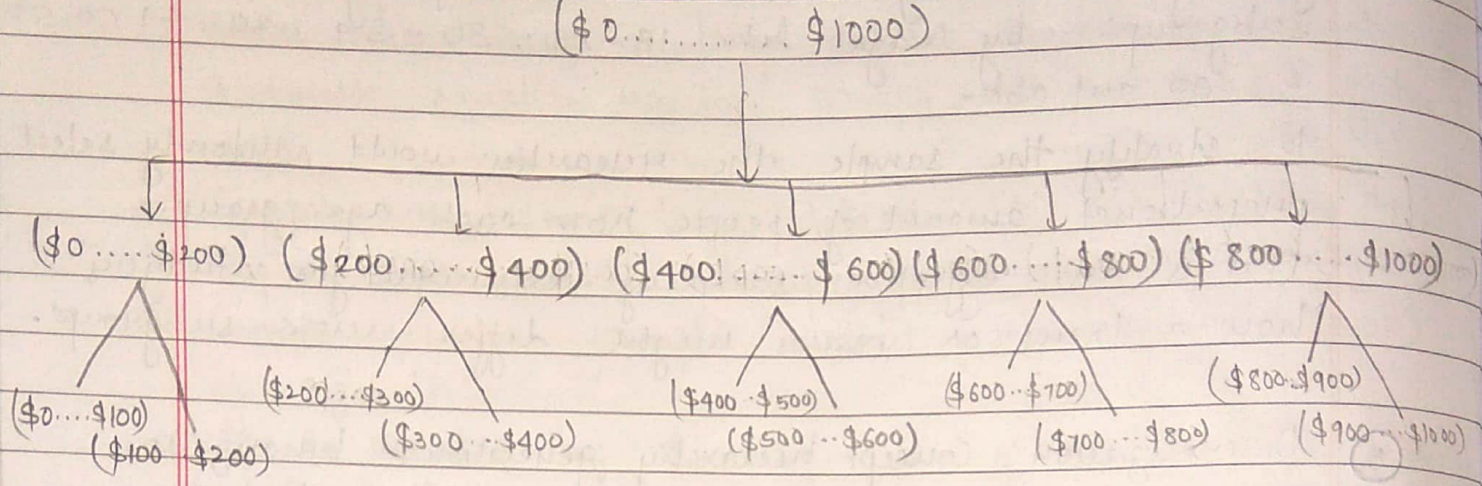
This helps to ensure a representative sample, specially when

the data are skewed
for eg One might divide the sample of adult into subgroups by edge like 18-29, 30-39, 40-49, 50-59 & 60 and above
to stratify the sample the researcher would randomly select proportional amount of people from each age group.
This is an effective sampling technique for studying how a trend or issues might differ across subgroups.

(5) Discretization, concept hierarchy generation & binarization.
Discritization techniques can be used to reduce the number of values for a given continuous attributes by dividing the range of the attributes into intervals
Interval levels can be used to be replace actual data values their by reducing the number of values for an attribute
- This methods are recursive where large amount of time is spend on sorting the data at each step.
- They also provide a hierarchical or multi resolution partitioning of attributes attribute values known as concept hierarchy shown below.
A concept hierarchy for a given numeric attribute defines discritization of the attribute.
They can be used to reduce the data by collecting & replacing low level concept such as numeric value for attribute age by higher level concept such as young, middle age or seniors.

Date 23/08/19

Concept hierarchy for attribute price



Concept hierarchies for numeric attributes can be constructed automatically based on data analysis distribution.

- The 5 different methods used for numeric concepts hierarchy generation are:-

i) Binning
ii) Histogram analysis
iii) Cluster analysis
iv) Entropy based discritization
v) Data segmentation by natural partitioning.

- Categorical data are discrete data. Categorical attributes have a finite number of distinct values with no ordering among the values.
- Example include geographic location, job category and item type.
- There are several methods for generation of concept hierarchies for categorical data.

① Specification of partial ordering of attributes explicitly at schema level by user or experts.

② Specification of a portion of hierarchy by explicit data grouping.

③ Specification of a set of attributes but not of their partial ordering.

④ Specification of only partial set of attributes.

Binarization.

- It is used to transform both the discrete attributes and the continous attributes into binary attributes in data mining.
- In other words it is a simple technique to binarize a categorical attribute. For eg. if there are m categorical values then uniquely assigned each original value to an integer in the interval [0, m-1]. If attribute is ordinal then order must be maintained by the assignment. Next, convert each of this integers into a binary number since $n = [\log_2(m)]$ binary digits are required to represent this integers, represent this binary number using n binary attribute.

To illustrate a categorical variable with 5 values, {awful, poor, ok, good, great} would require 3 binary variables x_1, x_2, x_3 as shown in the figure below.

Categorical attribute	Integer value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
ok	2	0	1	0
good	3	0	1	1
great	4	0	0	0

Table 1: Conversion of categorical attributes to three binary attributes.

Data Quality

(9) Measurement and data collection issues.

Measurement and data collection problems

- Outliers
- missing value
- Inconsistent values
- duplicate data

Problems involved in measurement Error

- Name artifacts
- Preprocessor
- Bias
- Accuracy

Standard boundary weight
mass = 1 kg

Precision = ? (0.013 = 6)
bias = ? (0.001)

weight mas 5 times
{1.015, 0.990, 0.013, 1.001, 0.986}

mean = 1.001
100, 110, 120, 95, 90, 10, 50

DBCE

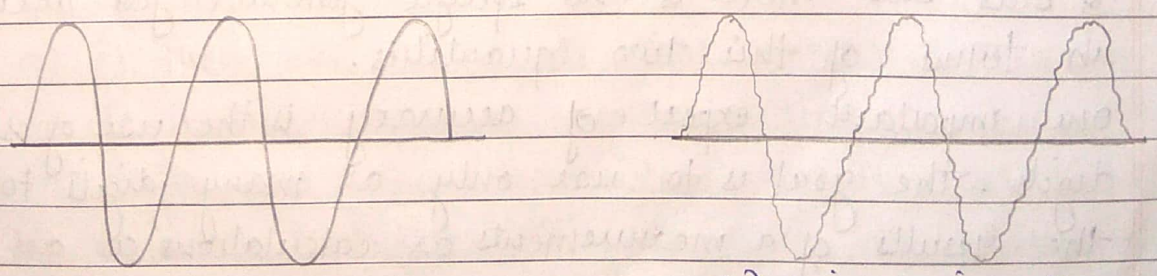
- Data quality refers to condition of set of values of qualitative and quantative variables
- Data derived of high quality if it correctly represents the real world construct to which it refers to.

Measurement and data collection issues.

- It is unrealistic to expect that data will be perfect. There may be problems due to human error, limitations of measuring devices or flaws in data collection process, or values or even entire data objects may be missing.
- The term measurement error refers to any problem resulting from measurement process such as value recorded differs from true value to some extent.
- The numerical difference of the measured and true value is called error.
- Data collection error refers to errors such as omitting data objects or attributes values or inappropriating including a data object.
- Variety of problems that involve measurement error are:

(I) Noise and end artifacts

- Noise is the random component of measurement error. It involves distribution of a value or the addition of spurious objects as shown below.



(a) Time series (b) Time series with noise

fig:- Noise in time series context.

- Data errors maybe result of more deterministic phenomenon such as strack in the same place on a set of photographs such deterministic distaution of data are referred as artifacts.

② Precision
- It is the closeness of repeated measurements (of the same quality) to one another.
- It is measured by standard derivation of set of values as shown in the example above.

③ Bias
- It is the systematic variation of measurements from the quatities being measured.
- It is measured by taking the difference between the mean of set of values and the known value of the quantity being measured.
- Bias can only be determined for objects whose measured quantity is known by means external to current situation as shown in the example above.

④ Accuracy
- It is the closeness of measurements to the time value of the quantity being measured. It depends on precision & bias but there is no specific formula for accuracy in terms of this two quantities.
- One important expect of accuracy is the use of significant digits. The goal is to use only as many digits to represent the results of a measurements or calculations as are justified by precision of data.

Data quality issues that may involve both measurement and data collection problem are:-

i) Outliers

- They are either

a) Data objects that, in same sense, have characteristics that are different from most of the other data objects in the data sets.

b) values of an attributes that are unusual with respect to the typical values for that attribute.

c) They can be legitimate data objects or values ~~and~~ they maybe sometimes of interest.

ii) Missing values (write up already given)

(iii) Inconsistent value "

(iv) Duplicate data

- A data set may include data objects that are duplicates
- Many people receive duplicate mailings because they appear in a database multiple times under slightly different names.
- To detect and eliminate such duplicates to main issues must be addressed.

a) If there are 2 objects that represent a single object then the value of corresponding attributes may differ and this inconsistent values must be resolved.

b) care needs to be taken to avoid accidently combining objects that are similar but not duplicates. such as 2 distinct people with identical name.

- The term / de-duplication is often used to refer to the process of dealing with this terms

Issues related to applications.

- As with quality issues at the measurement and data collection there are many issues that are specific to the particular applications and fields.

i) Timelines

- Some data starts to age as soon as it has been collected.
- In particular if data provides a snapshots of some on going process such as purchasing behaviour of customers or web browsing patterns then this snapshots represents a reality for only limited time.

ii) Relevance

- The available data must contain information necessary for the application.
- Consider task of building a model that predicts accidents rates for drivers. If informations of age and gender of driver is omitted then it is likely that model will have limited accuracy.

iii) Knowledge about the data

- Ideally datasets are accompanied by documentation that described different aspects of data. the quality of this data can aid or hinder the subsequent analysis, for example if documentation identifies several attributes as strongly related then this attributes are likely to provide highly redundant information.

- If documentation is poor then it gives wrong analysis.

Module 2

OLAP (online Analytical processing)

* OLAP operations
1) Slice
2) Dice
3) Roll up or drill up
4) Roll down or drill down
5) Pivot (Rotate)

* OLAP schemas
① star ② constellation ③ Snowflake

OLTP	OLAP
online (transaction)	online analytical processing.

27-08-19 (Part c)

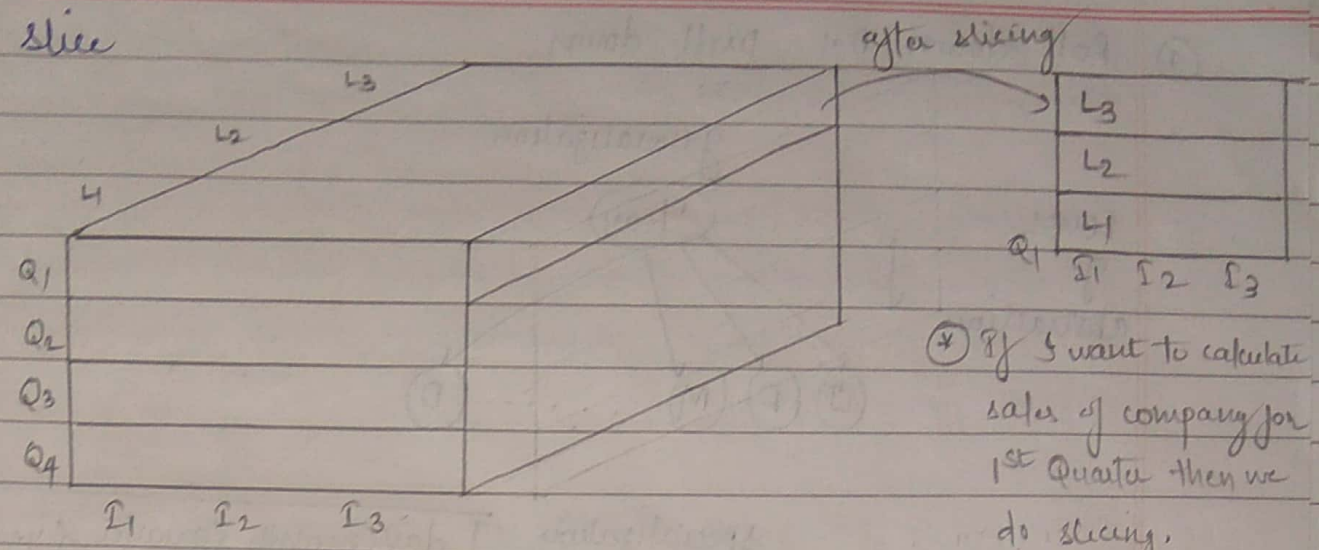
OLAP cube

- It is a multidimensional database that is optimized for datawarehouse and OLAP application.
- It is a method of storing data in a multidimensional form generally for reporting purposes.
- In OLAP cube data (measures) are categorised by dimensions.
- Query language used to interact and perform task with OLAP cube is multidimensional expressions. (MØX)
- The various OLAP operations are described below with examples.

DBCE

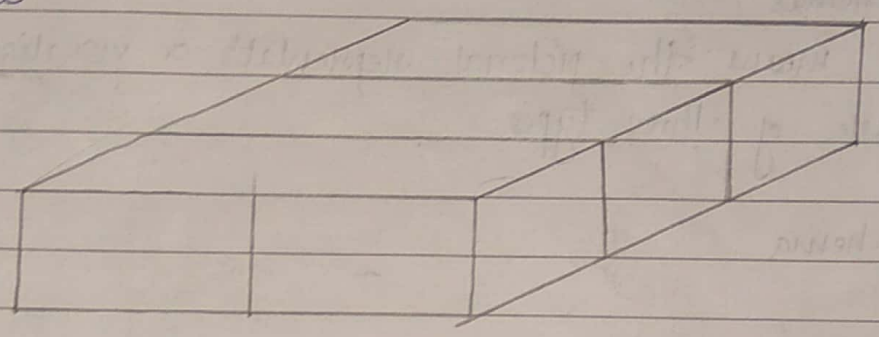
classmate
Date 27-08-19
Page

1. Slice



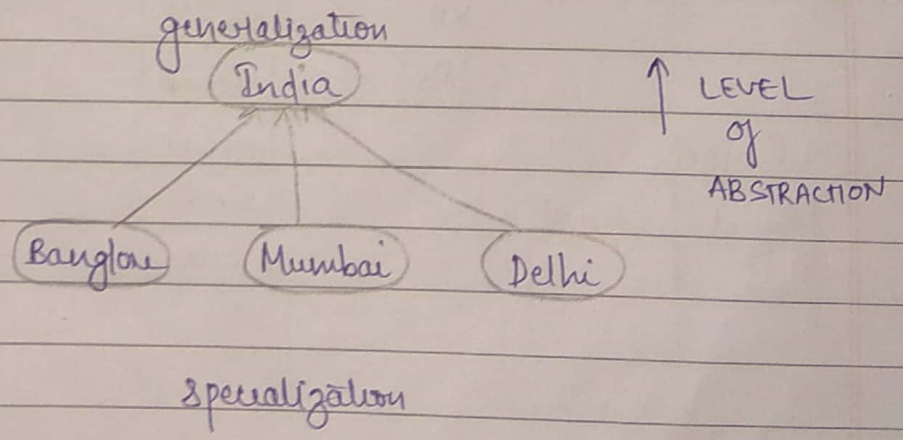
* If I want to calculate sales of company for 1st Quarter then we do slicing.

2. Dice



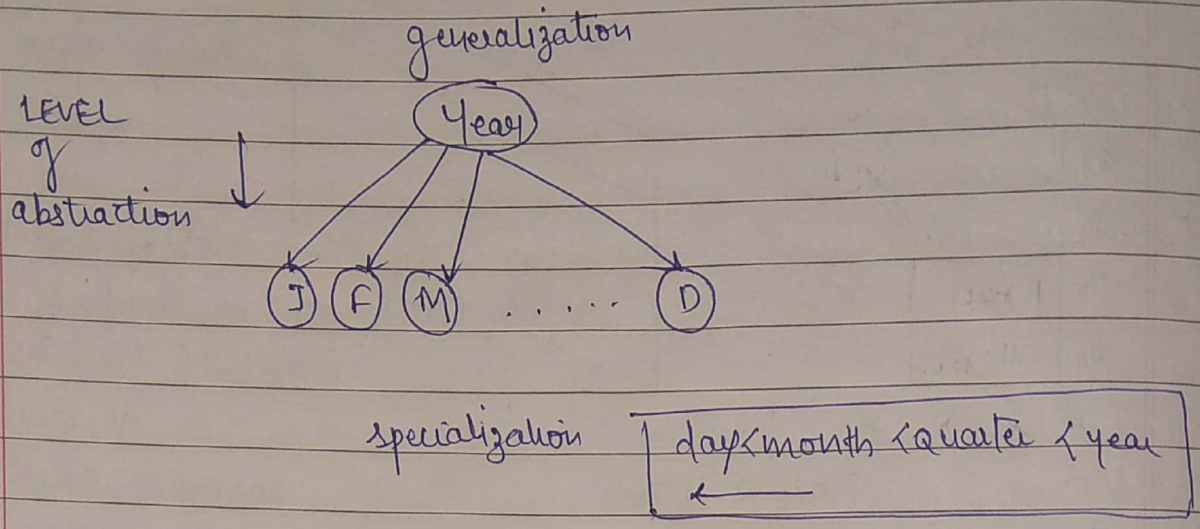
* To calculate sales of products I_1 and I_2 for Quarters 1 & 2 at all locations.

3. Roll up or Drill up



$\boxed{\text{street} < \text{city} < \text{state} < \text{country}}$

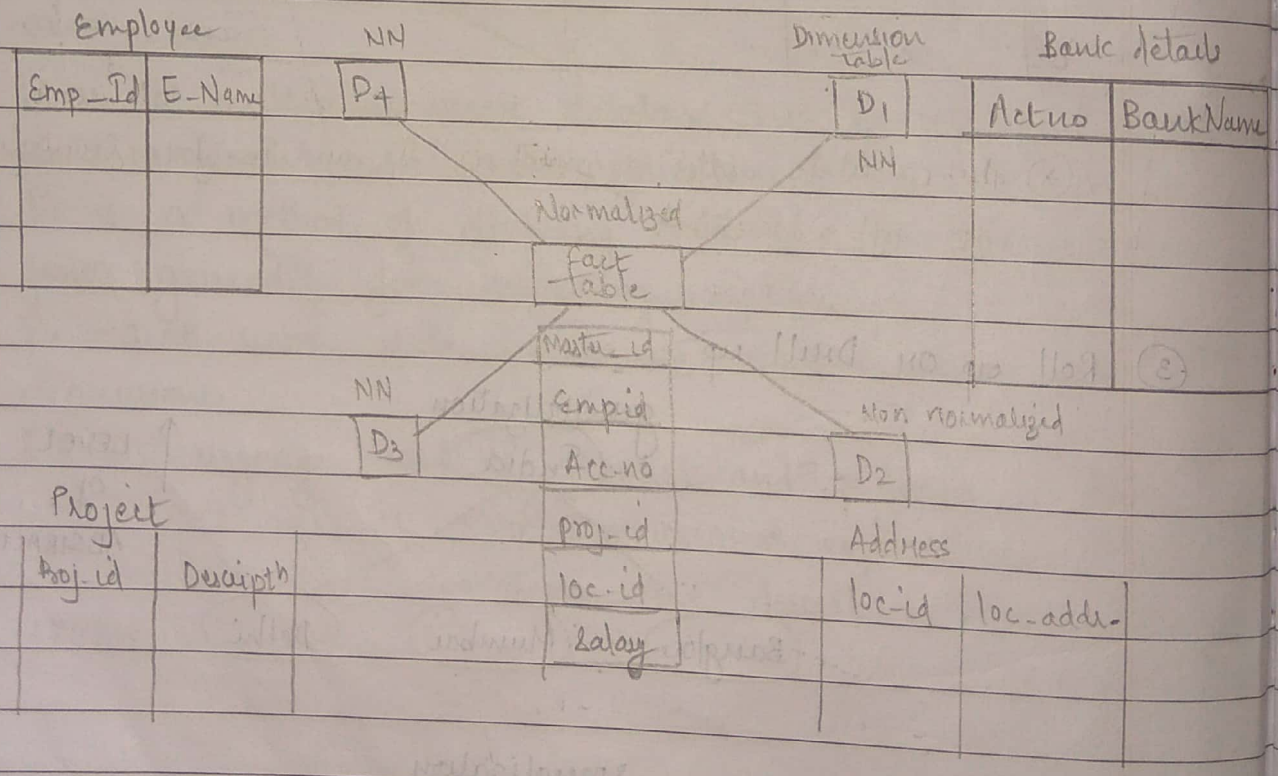
④ Roll down D4 Drill down



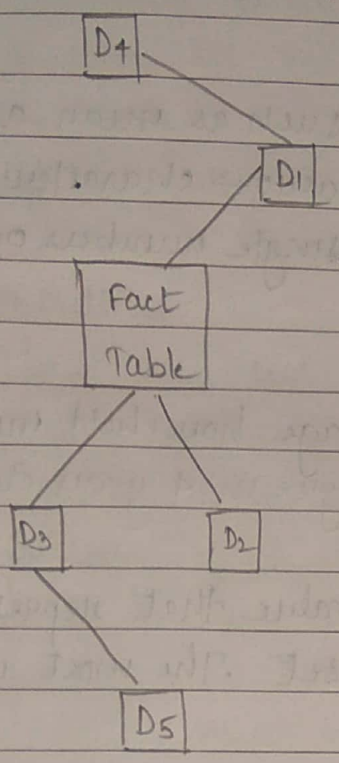
OLAP Schemas

Schemas means the pictorial representatn or visualization of data. They are of three types

① Star schema

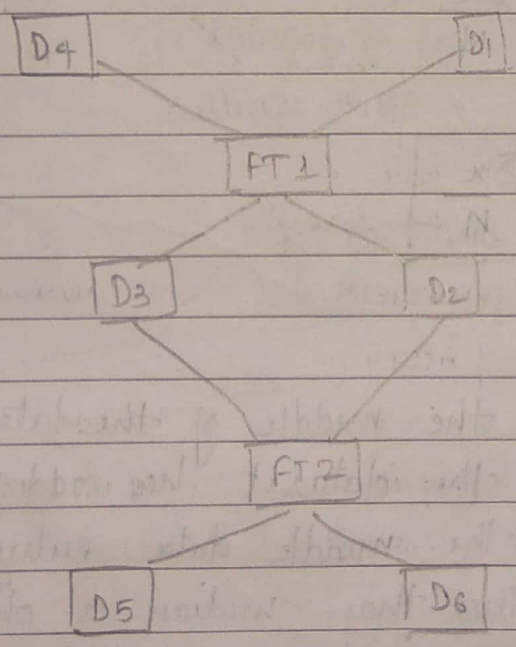


② Snowflake schema



③ Fact constellation schema

Fact table are normalize but not the dimension table



(Part B)

Summary Statistics

- Summary statistics are quantities such as mean and standard deviation that capture various characteristics of potentially large set of values with a single numbers or a small set of numbers.

- Every eg. of summary statistics are average household income, fraction of college who complete undergraduate degree in 4 years etc.

- Measure of central tendency is a value that represents a typical or central entry of dataset. The most common measures of central tendancy are

(I)

(1) Mean (Average)

The sum of all data entries divided by no. of entries.

$$\boxed{\text{Population mean}: \mu = \frac{\Sigma x}{N}}$$ ← of population

$$\boxed{\text{Sample mean}: \bar{x} = \frac{\Sigma x}{n}}$$ ← sample

(2) Median.

The value that lies in the middle of the data when the data set is ordered. If the dataset has odd no. of entries, than the median is the middle data entry if the data has even no. entries than median is obtained by adding 2 numbers in the middle and dividing the result by 2.

DBCE

③ Mode

The data entry that occurs with the greatest frequency. A data set may have one mode, more than 1 mode or no mode.

If no entry is repeated the data has no mode.

④ Outliers

They are not just greatest and least values, but values that are very different from the pattern establish by the rest of the data.

Outliers affect the mean. when outliers are present it is best to use median as the measure of central tendency

for eg. 1, 2, 2, 3, 4, 5, 60 step 1 is sorting in ascending order.

$$\text{mean} = \frac{1+2+2+3+4+5+60}{7} = \frac{77}{7} = 11$$

median = 3

mode = 2

outlier = 60 (value which is far from other value)

Measure of Variation

① Range

(Range = max data entry − min data entry)

② Standard Deviation

Population std deviat^n "σ"

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

Sample std deviation "s"

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

③ Variance

$$V = \frac{\Sigma(x-\bar{x})}{n-1}$$

④ Frequency

Absolute freq

Relative freq

Commulative freq

1 Range

The difference betn maximum & minimum data entry in the set is called Range

2 Standard deviation

It measures variability and consultancy of the sample or population.

In most real time world applications, consultancy is a great advantage

In statistical data analysis, less variation is often better.

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

3 Variance

It is preferred as a measure of spread.

The variance of the observed value of an attribute x is given as

$$V_{population} = \frac{\Sigma(x-\mu)^2}{N}$$

$$V_{sample} = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

4 Frequency

It is defined as number of times each value occurs for a particular set of data

Following is the data of number of books read by the students and the data is 1, 2, 5, 4, 6, 3, 7, 1, 5, 6, 5, 3, 4, 5, 1

Find absolute frequency, relative frequency and cummulative frequency.

x	Absolute n(x)	Relative P(x)	Commulative P'(x)
1	3	3/16 = 0.1875	3
2	1	1/16 = 0.0625	4
3	2	2/16 = 0.125	6
4	3	= 0.1875	9
5	4	= 0.25	13
6	2	= 0.125	15
7	1	= 0.0625	16
Total	16	1	

29-08-19 Percentile

- Let 'p' be any integer between 0 and 100.
- The pth percentile of dataset is the data value at which 'p' percent of a value in the dataset are less than or equal to this value.

How to calculate percentile

Use the following steps for calculating percentile for small datasets for large datasets we use computers to find percentile.

Step 1 Sort the data in ascending order. (i.e from small to large)

Step 2 Calculate $i^{th} = \left(\frac{p}{100}\right)n$; where 'p' is the particular percentile you wish to calculate & 'n' the sample size

Step 3 If 'i' is an integer, pth percentile is the mean of data values in position i and i+1

If 'i' is not an integer, pth percentile is obtained by rounding up to the next integer. and using the value at that position.

Eg. 1 Use the following set of stock prices (in Rupees): 4, 5, 7, 8, 9, 10, 12, 12, 14, 15, 18, 20.
Find 10th percentile and 50th percentile.

→

Step 1: Write values in ascending Order

DATA	4	5	7	8	9	10	12	12	14	15	18	20
Position	1	2	3	4	5	6	7	8	9	10	11	12

Step 2: Since there are 12 values, $\boxed{n=12}$
To find 10th percentile, so $p=10$

$$i^{th} = \left(\frac{p}{100}\right) n$$

$$= \left(\frac{10}{100}\right) 12$$

$$= 1.2$$

Step 3: $i^{th} = 1.2 \approx 2$ Roundup (1.2)
So 10th percentile is 5.

for 50th

Step 2 Since there are 12 values $n=12$
To find 50th percentile, so $p=50$

$$i^{th} = \left(\frac{p}{100}\right) n = \left(\frac{50}{100}\right) 12 = \frac{1}{2} \times 12 = 6$$

Step 3 $= i^{th} = 6$ Since answer is integer we need to find the 6th & the 7th number in the dataset. so 50th percentile $= \frac{10+12}{2} = 11$

$i = 10$ $i+1 = 12$ mean of i & i+1 $\frac{10+12}{2} = \frac{22}{2} = 11$

So 50th percentile is 11

* If population word is present than only divide by n. otherwise by default divide by n-1

Correlation Coefficient (Pearson's product moment coefficient)

eg1. For the below given attributes x and y calculate correlation coefficient (Pearson's product moment coefficient). Are this 2 variables +vely or -vely correlation.

X	3	2	2	1
Y	6	3	5	2

Sol^n) Find Mean of X and Y so we get,

$$\bar{x} = M_x = \frac{\Sigma x}{n} \qquad \bar{x} = M_y = \frac{\Sigma y}{n}$$

$$M_x = \frac{3+2+2+1}{4} \qquad M_y = \frac{6+3+5+2}{4}$$

$$M_x = \frac{8}{4} \qquad M_y = \frac{16}{4}$$

$$M_x = 2 \qquad M_y = 4$$

Step 2 Create Table & find product

X	Y	X-Mx	X-My	(x-mx)*(y-my)
3	6	(3-2) 1	(6-4) 2	2
2	3	(2-2) 0	(3-4) -1	0
2	5	(2-2) 0	(5-4) 1	0
1	2	(1-2) -1	(2-4) -2	2
				$\Sigma(x-m_x)*(y-m_y) = 4$

Step 3: calculate standard deviation

0.8165 $$S_x = \sqrt{\frac{(1)^2+(0)^2+(0)^2+(-1)^2}{3}} = 0.8165$$

1.8257 $$S_y = \sqrt{\frac{(2)^2+(-1)^2+(1)^2+(-2)^2}{3}} = 1.8257$$

DBCE

Step 4 Find correlation coefficient

$$Corr(x,y) = \frac{\Sigma (x - M_x) * (y - M_y)}{(N-1)\ (S_x \cdot S_y)}$$

$$= \frac{4}{(4-1)\ (0.8165 * 1.8257)}$$

$$= \frac{4}{(3)(1.49068405)}$$

$$= \frac{4}{4.4722}$$

$\therefore Corr(x,y) = 0.8944$

20-08-19

Problem 2 For the following data point of attribute x & y in the population find or calculate correlation coefficient. are these 2 attributes +vely or -vely correlated.

x	1	3	2	5	8
y	8	6	9	4	3

→

step 1 $\bar{x} = M_x = \frac{\Sigma x}{n}$

$$M_x = \frac{1+3+2+5+8}{5} = \frac{19}{5} = 3.8$$

$\bar{y} = M_y = \frac{\Sigma y}{n}$

$$M_y = \frac{8+6+9+4+3}{5} = \frac{30}{5} = 6$$

Step 2: Create table & find product

X	Y	X - Mx	Y - My	(X-Mx) * (Y-My)
1	8	-2.8	2	-5.6
3	6	-0.8	0	0.0
2	9	-1.8	3	-5.4
5	4	1.2	-2	-2.4
8	3	4.2	-3	-12.6
				-15.2

Step 3 calculate std. deviat^n

$$S_x = \sqrt{\frac{(-2.8)^2 + (-0.8)^2 + (-1.8)^2 + (1.2)^2 + (4.2)^2}{5}}$$

$$= 2.4819$$

$$S_y = \sqrt{\frac{(2)^2 + (0)^2 + (3)^2 + (-2)^2 + (-3)^2}{5}}$$

$$= 2.28035$$

Step 4: Find correlation coefficient

$$\text{Cor}(x,y) = \frac{\Sigma (x - M_x) * (y - M_y)}{(N) \cdot (S_x \cdot S_y)}$$

$$= \frac{11.2}{5 \times (2.4819 \times 2.804)}$$

$$= -0.09186$$

So attributes x & y are -vely correlated.

+ve	1	0.1 to 0.3	SMALL Relationship
		0.31 to 0.5	MODERATE Relationship
		0.51 to 0.7	LARGE Relationship
		0.71 to above	VERY STRONG Relationship
Independent	0	0	No Relation
-ve	-1	any negetive	negetive Relationship

Covariance

eg.1 For the following data points of attributes X & Y. Find co-variance & also correlation betn attribute X & Y.

X	1	3	2	5	8	7	12	2	4
Y	8	6	9	4	3	3	2	7	7

→ Step 1: Calculate Average or mean of x and y data points

$\Sigma x = 44$, so $X_{mean} = \frac{44}{9} = 4.89$

$\Sigma y = 49$, so $Y_{mean} = \frac{49}{9} = 5.44$

Step 2:

X	Y	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$(X_i - X_{mean}) * (Y_i - Y_{mean})$
1	8	-3.89	2.56	-9.96
3	6	-1.89	0.56	-1.06
2	9	-2.89	3.56	-10.29
5	4	0.11	-1.44	-0.16
8	3	3.11	-2.44	-1.59
7	3	2.11	-2.44	-5.15
12	2	7.11	-3.44	-24.46
2	7	-2.89	1.56	-4.51
4	7	-0.89	1.56	-1.29
				-64.57

Step 3: The sum of values in product column is,

$\Sigma (x_i - X_{mean}) * (y_i - Y_{mean})$

$= -64.57$

Step 4 Number of data points for attributes X & Y are, $\boxed{n=9}$

Step 5: Using co-variance formula, ~~co-variance~~

$$\text{Co-variance } (X,Y) = \frac{\Sigma (x_i - X_{mean}) * (y_i - Y_{mean})}{(n-1)}$$

$$= \frac{-64.5}{9-1}$$

$$= -8.07$$

Step 6: Find standard deviation. S_X & S_Y

$S_X = 3.551$

$S_Y = 2.506$

Step 7: Calculate correlation coefficient betn X & Y

$$\text{Corr } (X,Y) = \frac{\text{Cov } (X,Y)}{S_X * S_Y} = \frac{-8.07}{3.551 * 2.506} = -0.907$$

Date 31-08-22

Q) Following given the table describes rate of economic growth (x) and rate of return (y) of SP graph. Using covariance formula determine, whether economic growth and rate of return has a positive or negative relationship.

X	Economic growth	2.1	2.5	4.0	3.6
Y	Rate of return	8	12	14	10

Soln

$\Sigma x = 12.2$ $\quad$ $X_{mean} = 3.05$

$\Sigma y = 44$ $\quad$ $Y_{mean} = 11$

X	Y	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$(X_i - X_{mean})(Y_i - Y_{mean})$
2.1	8	-0.95	-3	2.85
2.5	12	-0.55	1	-0.55
4.0	14	0.95	3	2.85
3.6	10	0.55	-1	-0.55
				$\Sigma = 4.6$

$n = 4$

$$\text{Co-variance}(x,y) = \frac{(X_i - X_{mean})(Y_i - Y_{mean})}{n-1}$$

$$= \frac{4.6}{4-1} = 1.53$$

∴ It has a +ve relation.

$$\text{Correlation}(x,y) = \frac{\text{Cov}(x,y)}{S_x, S_y} = \frac{1.53}{0.816 \times 2.58} = 0.66$$

$S_x = 0.816$ $\quad$ $S_y = 2.58$

∴ $\text{Cor}(x,y) = 0.66$

UNIT 2

Part A

measures of similarity and dissimilarity

- A similarity and dissimilarity are important because they are used by number of data mining such as clustering nearest neighbour classification, anamoly detection.
- In many cases initial dataset is not needed once the similarities have been computed.
- Term proximity is used to refer either similarity or dissimilarity. Proximity between 2 objects is a function of proximity between the corresponding attributes of 2 objects.
- To measure proximity between 2 objects having only one single attribute and then consider proximity measures for objects with multiple attributes. We use measure like correlation and euclidean distance which are useful for dense data such as time series or 2 dimensional points, as well as Jaccard coefficient and cosine similarity measures which are useful for sparse data like documents.

(I) Similarity measures for Binary Data

1 Simple Matching Coefficient (SMC)

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}} = \frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}}$$

Where

f_{00} = number of attributes where $x=0$ and $y=0$

f_{01} = number of attributes where $x=0$ and $y=1$

f_{10} = number of attributes where $x=1$ and $y=0$

f_{11} = number of attributes where $x=1$ and $y=1$

* A value 1 indicates that 2 digits are completely similar, while value of 0 indicates that the objects are not at all similar.

Jaccard Coefficient (J) - used for binary data

Q) Compute Jaccard coefficient (J) for the data given below.

$x = (1,0,0,0,0,0,0,0,0,0)$; $y = (0,0,0,0,0,0,1,0,0,1)$

$$J = \frac{\text{number of matching presences}}{\text{no. of attribute not involved in 00 matches}} = \frac{f_{11}}{f_{01}+f_{10}+f_{11}}$$

$$J = \frac{0}{2+1+0} = 0$$

$f_{00} = 7$

$f_{10} = 1$

$f_{01} = 2$

$f_{11} = 0$

Q) For the following vectors x and y. calculate similarity using simple matching coefficient.

$x = (1,0,0,0,0,0,0,0,0,0)$ $y = (0,0,0,0,0,0,1,0,0,1)$

→ $f_{00} = 7$

$f_{10} = 1$

$f_{10} = 2$

$f_{11} = 0$

$$SMC = \frac{0+7}{2+1+0+7} = \frac{7}{10} = 0.7$$

Q) Compute simple matching coefficient and Jaccard coefficient for the data given below.

$x = 0101010001$ $y = 0100011000$

$f_{00} = 5$

$f_{01} = 2$

$f_{10} = 1$

$f_{11} = 2$

Simple matching coefficient $= \frac{2+5}{10} = \frac{7}{10} = \frac{2+5}{1+2+2+5} = 0.7$

Jaccard coefficient $= \frac{2}{2+1+2} = 0.4$

Similarity measures for two document vectors

① Cosine similarity

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

where x & y are two document vectors
$\|x\| \rightarrow$ length of vector x

$\|x\| \rightarrow \sqrt{x * x}$

② Extended Jaccard Coefficient (or Tanimoto Coefficient)

$$EJ(x,y) = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y}$$

Problem) For the following vector x&y, calculate cosine similarity and extended to jaccard coefficient
$x = (3,2,0,5,0,0,0,2,0,0)$ $y = (1,0,0,0,0,0,0,1,0,2)$

→

$x \cdot y = (3*1)+(2*0)+(0*0)+(5*0)+(0*0)+(0*0)+(0*0)+(2*1)+(0*0)+(0*2)$
$= 5$

$\|x\| = \sqrt{(3*3)+(2*2)+(0*0)+(5*5)+(0*0)+(0*0)+(0*0)+(2*2)+(0*0)+(0*2)}$
$= 6.48$

$\|y\| = \sqrt{(1*1)+(0*0)+(0*0)+(0*0)+(0*0)+(0*0)+(0*0)+(0*0)+(1*1)+(0*0)+(2*2)}$
$= 2.4$

$$\cos(x,y) = \frac{5}{6.48 * 2.4} = 0.31$$

$$EJ(x,y) = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y}$$

$$= \frac{5}{(6.48)^2 + (2.4)^2 - 5}$$

$$= 0.117$$

Euclidean distance

(I) E·D between two points / vectors
A (x_1, y_1, z_1) and B (x_2, y_2, z_2)

$$d(A,B) = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2 + (z_1-z_2)^2}$$

(II) E·D between 3 points or vectors
A (x_1, y_1) ; B (x_2, y_2) and C (x_3, y_3)

$$d(A,B) = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}$$

$$d(B,C) = \sqrt{(x_2-x_3)^2 + (y_2-y_3)^2}$$

$$d(A,B,C) = d(A,B) + d(B,C)$$

Determine Euclidean distance between $x(2,3,4,2)$ and $y(1,-2,1,3)$

$$d(A,B) = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2 + (z_1-z_2)^2}$$

$$= \sqrt{(2-1)^2 + (3+2)^2 + (4-1)^2 + (2-3)^2}$$

$$= \sqrt{(1)^2 + (5)^2 + (3)^2 + (-1)^2}$$

$$= \sqrt{1 + 25 + 9 + 1}$$

$$= \sqrt{36}$$

$$= 6$$

DBCE

Determine the Euclidean distance between $x(1,0,0)$; $y(1,4,5)$ and $z(10,0,0)$

$$d(A,B) = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}$$

$$= \sqrt{(1-1)^2 + (0-4)^2 + (0-5)^2}$$

$$= \sqrt{0+16+25}$$

$$= \sqrt{41}$$

$$= 6.40$$

$$d(B-C) = \sqrt{(x_2-x_3)^2 + (y_2-y_3)^2}$$

$$= \sqrt{(1-10)^2 + (4-0) + (5-0)}$$

$$= \sqrt{81+16+25}$$

$$= \sqrt{122}$$

$$= 11.04$$

$$d(A,B,C) = d(A,B) + d(B,C)$$

$$= 6.40 + 11.04$$

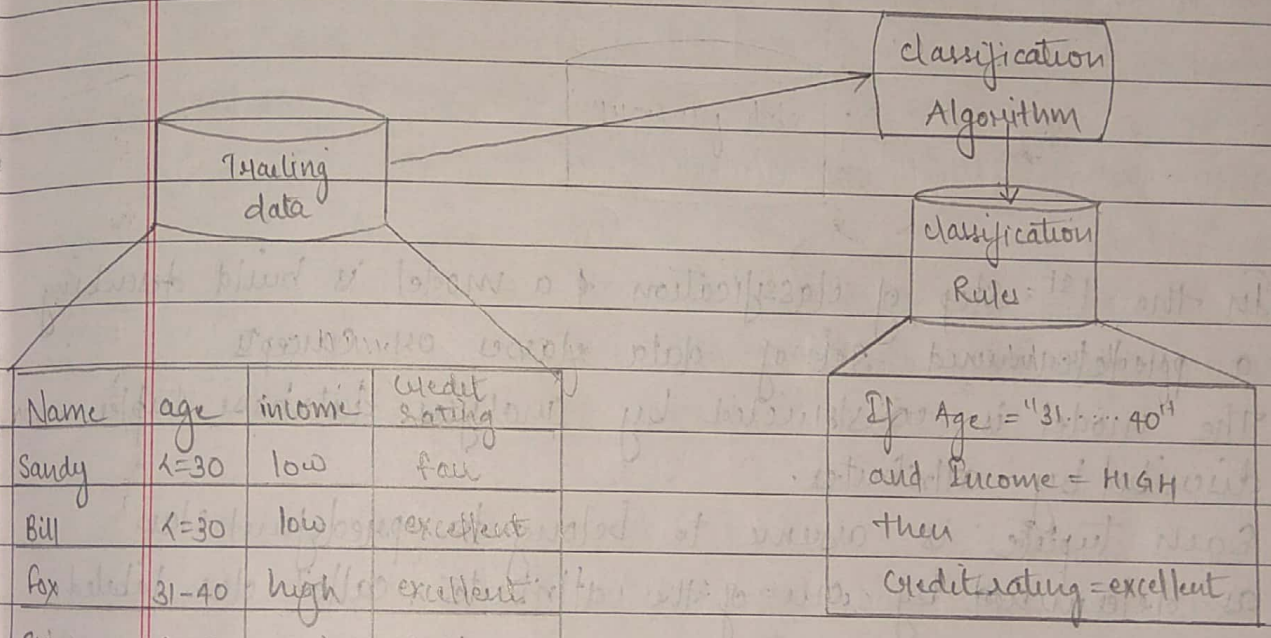
$$= 17.44$$

UNIT 3

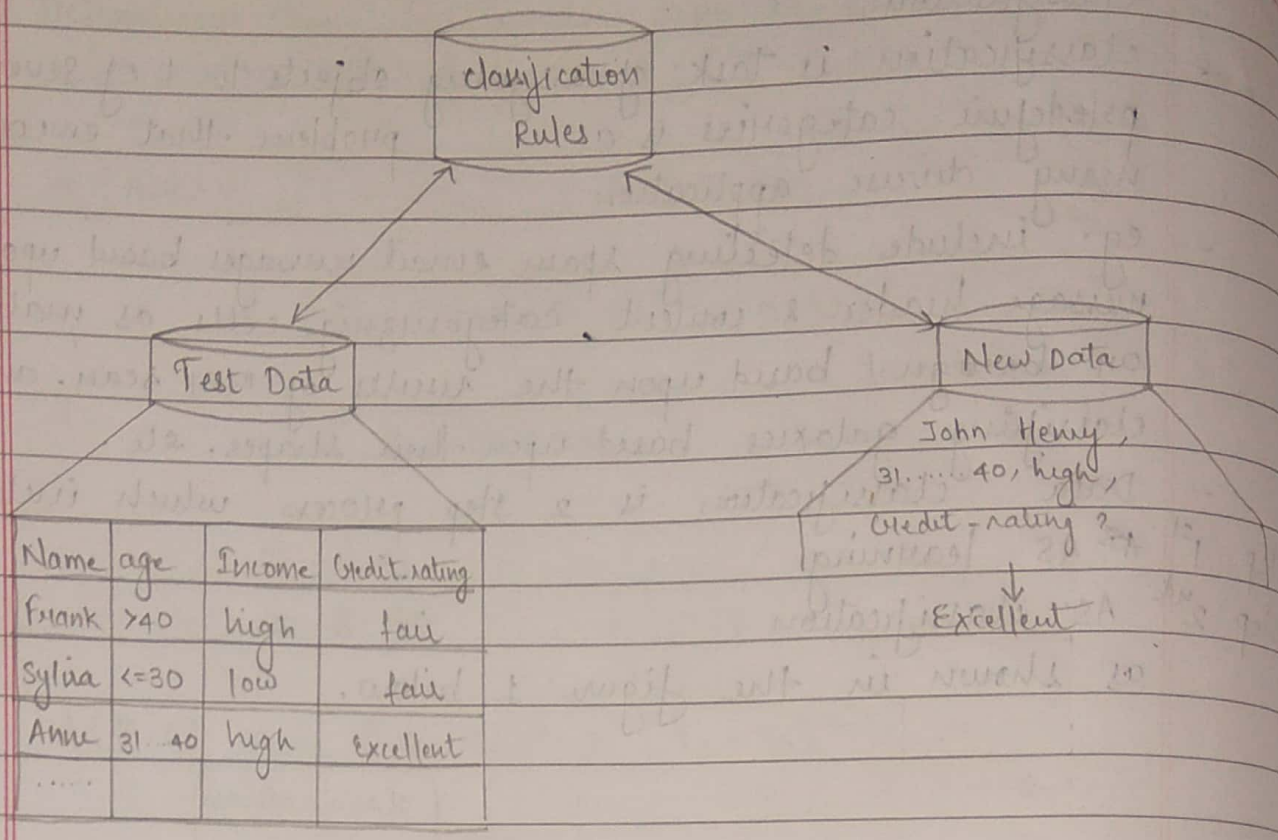
Classification

- classification is task of assigning objects to 1 of several predefine categories is a problem that encompasses many diverse application
- eg. include detecting spam email messages based upon message header & content, categorizing cells as malignant or benign based upon the results of MRI scan. and classifing galaxies based upon their shapes. etc.
- Data classification is 2 step process which includes

step 1st AS learning
step 2nd AS classification
as shown in the figure 1 below.



Name	age	income	credit rating
Sandy	<=30	low	fair
Bill	<=30	low	excellent
Fox	31-40	high	excellent
Susan	>40	med	fair
claire	>40	med	fair
Andre	31 ... 40	High	excellent
.....	.....	.....	.....



26-09-19

7 marks explanation + diag

- In the 1st step of classification a model is build describing a predetermined set of data classes or concepts
- The model is constructed by analysis database tuples discribed by attributes.
- Each tuple is assume to belong to predefined class as determined by one of the attribute called class labled attributes
- Data tuples are also referred to as samples, examples or objects
- Data tuples analys to build the model collectively form the training data sets.
- * Individual tuples making up training set are referred as training samples and are randomly selected from sample population. Since class label of each training sample is provided this step is known as <u>supervised learning</u>

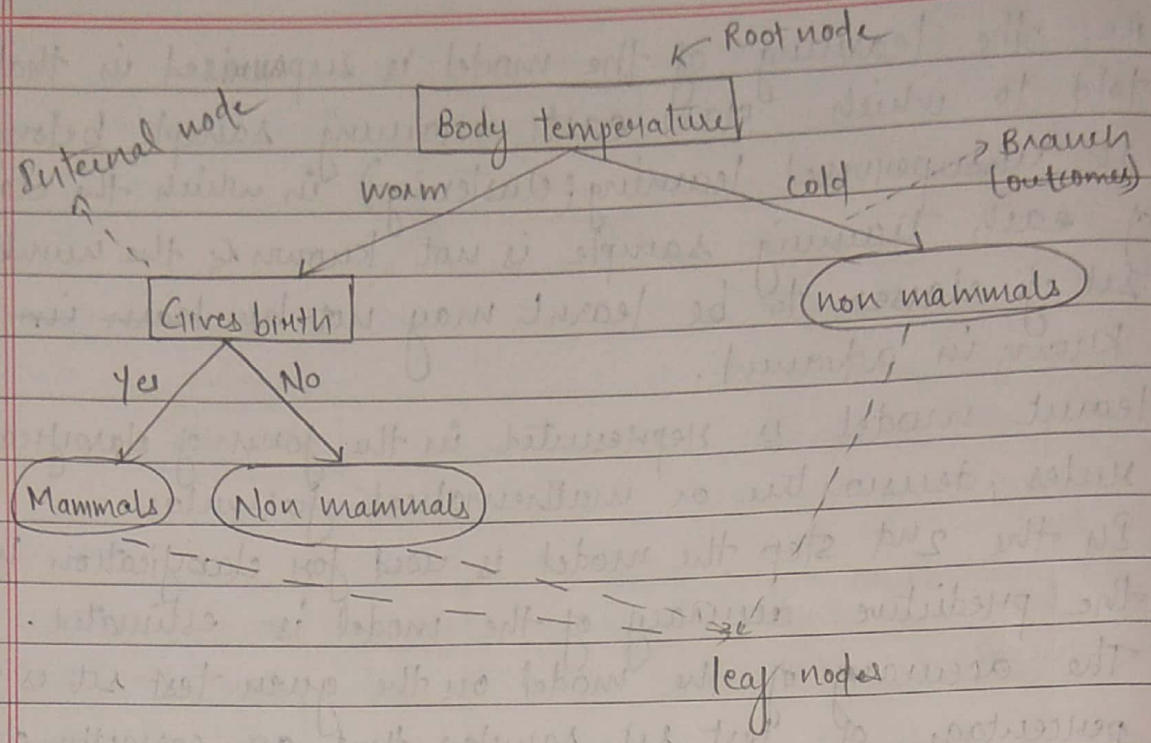
i.e the learning of the model is supervised in that it is told to which class each training sample belongs.

- In unsupervised learning (clustering) in which the class label of each training sample is not known & the number or set of classes to be learnt may not be ~~learn in~~ known in advance.
- learnt model is represented in the form of classification rules, decision tree or mathematical formulae.
- In the 2nd step the model is used for classification here 1st the predictive accuracy of the model is estimated.
- The accuracy of the model on the given test set is the percentage of test set samples that are correctly classified by the model.
- For each test sample known class label is compared with learned model, class prediction for that sample.
- If accuracy of model is considered acceptable than the model can be used to classified future data tuples or object for which the class label is not known.
- The hold out method is a simple technique that uses a test set of class label samples.

Decision tree Induction

- Decision tree is a flow chart like tree structure where internal nodes denotes a test on an attribute, each branch represents an outcome of a test and leaf nodes represents classes or class distribution.
- The topmost node in a tree is the root node.
- A typical decision tree is shown below.

6 marks Explain strategies for decision tree.



Example of Decision tree

The basic algorithm for a decision tree induct^n is a greedy algorithm that constructs a decision tree in a topdown recursive divide & conquer manner.
The algo is given below.

- The basic strategies as follows.

1 The tree starts as a single node representing the training samples.
If d samples are all of the same class than node becomes a leaf. and it is labelled with that class.

3 Otherwise the algorithm uses an entropy based method. measure known as information gain as a heuristic for selecting the attribute that will best seperate the samples into individual classes.

4. This attribute becomes the test or decision attribute at the node.

- In this version of algorithm all attributes are categorical (discrete value) continuous valued attributes must be discretized.

DBCE

5. A branch is created for each known ~~node~~ value of the test attribute and sample are partitioned accordingly
6. The algorithm uses same process to form a decision tree for the samples at each partitioned.
Once the attribute as occured at the node it may not be considered in any of the nodes descendend.
7. The recursive partition stops only when one of the following condition is true
a) All samples for a given node belong to the same class.
b) There are no remaining attributes on which the samples may be further partitioned. It involves converting given node into leaf and labeling it with a class in majority among samples.
c) There are no samples for the branch test attributes
test attributes = a_i. In these case a leaf is made with the majority class in samples.

- Attributes selection measures

1) Information Gain

$$I(s_1, s_2, \ldots\ldots s_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i)$$

2) Entropy

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots\ldots + s_{mj}}{s} \cdot I(s_{1j}, s_{2j}, \ldots\ldots, s_{mj})$$

3) Gain

$$\text{Gain}(A) = I(s_1, s_2, \ldots\ldots, s_m) - E(A)$$

where s_i = number of samples of 's' in class 'c_i'
s_{ij} = number of samples of class 'c_i' in a subset s_j.
p_i = probability that are arbitrary f^n belong to class 'c_i'.

V.V.V.V.
Imp

eg.1 Table given below presents a training set of data tuples taken from ABC customer database. The class label attribute Buys_computer has 2 distinct values i.e yes & No.
∴ there are 2 distinct classes ($m=2$). Let class 'C_1' corresponds to 'Yes' and class 'C_2' corresponds to 'No'. There are 9 samples of class 'Yes' and 5 samples of class 'No'. Compute information gain of attribute. Also find gain of all the attributes.

	0.694	0.029	0.151	0.048	
RID	age	income	student	credit rating	class: Buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31...40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31....40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31....40	medium	no	excellent	yes
13	31...40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Soln) Compute 1st expected Info^n Gain
Heath

$$I(S_1, S_2) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right)$$

yes NO

$$= 0.94$$

Step2 = G(Age) G(Inc) G(student) G(CR)

income = 0.029
student = 0.
credit =

* Then calculate Entropy of each attribute CONSIDER ATTRIBOTE AGE

for Age = "<=30"

$S_{11} = 2$, $S_{21} = 3$; $I(S_{11}, S_{21}) = 0.971$

for age = "31..... 40"

$S_{12} = 4$, $S_{22} = 0$; $I(S_{12}, S_{22}) = 0$

for age = ">40"

$S_{13} = 3$, $S_{23} = 2$; $I(S_{13}, S_{23}) = 0.971$

Entropy for attribute Age is

$$E(Age) = \frac{5}{14} I(S_{11}, S_{21}) + \frac{4}{14} I(S_{12}, S_{22}) + \frac{5}{14} I(S_{13}, S_{23})$$

$$= \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971)$$

$$= 0.694$$

So Gain in information from such a partitioning would be

$$Gain(Age) = I(S_1, S_2) - E(age)$$

$$= 0.94 - 0.694$$

$$= 0.246$$

gain (income)

for income = "high"

$S_{11} = 2$; $S_{21} = 2$; $I(S_{11}, S_{21}) = 1$

for income = "medium"

$S_{12} = 4$; $S_{22} = 2$; $I(S_{12}, S_{22}) = 0.918$

for income = "low"

$S_{13} = 3$; $S_{23} = 1$; $I(S_{13}, S_{23}) = 0.81$

$$E(Income) = \frac{4}{14}(01) + \frac{6}{14}(0.918) + \frac{4}{14}(0.81) = 0.9105$$

Gain (income) = $I(s_1, s_2) - E(\text{income})$

$= 0.94 - 0.9105$

$= 0.0295$

gain (student)

for Yes

$s_{11} = 6 \quad s_{21} = 1$

$I(s_{11}, s_{21}) = 0.59$

for No

$s_{12} = 3 \; ; \; s_{22} = 4$

$I(s_{12}, s_{22}) = 0.98$

$E(\text{student}) = \frac{7}{14}(0.59) + \frac{7}{14}(0.98)$

$= 0.785$

Gain (student) =

$= 0.94 - 0.785$

$= 0.155$

gain (credit_rating)

for fair

$s_{11} = 6 \quad s_{21} = 2$

$I(s_{11}, s_{21}) = 0.8112$

for excellent

$s_{12} = 3 \quad s_{22} = 3$

$I(s_{12}, s_{22}) = 1$

$E(\text{Credit_rating}) = \frac{8}{14} \cancel{\log}(0.8112) + \frac{6}{14}(1)$

$= \cancel{0.04}\ 0.89211$

Gain (Credit_rating) = $0.94 - 0.89211$

$= 0.048$

Gain (age)) = 0.246 ✓

Gain (Income) = 0.0295

Gain (student) = 0.155

Gain (credit_R) = 0.048

step2: ∵ Attribute "age" is having highest info^n gain it is selected as TEST Attribute and we get decision tree as

A.GE

<=30

Income	student	Credit rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

>40

Income	student	Credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no.

31...40

income	Student	Credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

when all i/p are "yes" ∴ we replace the whole table by yes

DBCE

AGE

<=30

Income	student	C_R	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

31...40

yes

>40

Income	student	C_R	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

iteration 2

Step 3: Expected info needed to classify given sample is

$$I_{<=30}(S_1, S_2) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right)$$

$$= 0.97$$

Step 4: Find gain for Income, student and credit-rating

Income

	high	medium	low
yes	0	1	1
no	2	1	0

$$E_{high} = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

$$E_{medium} = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$E_{low} = -\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right) = 0$$

$$\therefore E(\text{income}) = \left[\frac{2}{5} * 0\right] + \left[\frac{2}{5} * 1\right] + \left[\frac{1}{5} * 0\right]$$

$$= 0.4$$

$$\text{Gain (income)} = 0.97 - 0.4 = 0.57$$

- Student

	Yes	No
Yes	2	0
No	0	3

$$E_{yes} = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) = 0$$

$$E_{NO} = -\frac{0}{3}\log\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right) = 0$$

$$E_{student} = \left[\frac{2}{5} * 0\right] + \left[\frac{3}{5} * 0\right] = 0$$

$$\text{Gain (student)} = 0.97 - 0 = 0.97$$

- Credit_rating

	fair	excellent
Yes	1	1
No	2	1

$$E_{fair} = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.95$$

$$E_{excellent} = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

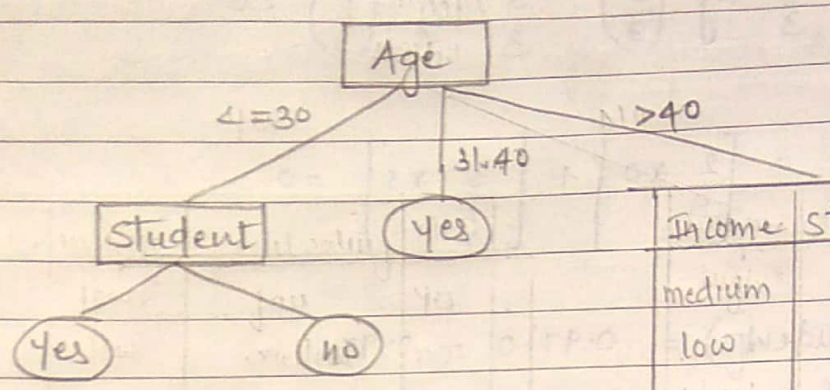
$$E(CR) = \left[\frac{3}{5} * 0.95\right] + \left[\frac{2}{5} * 1\right]$$

$$\therefore E(CR) = 0.97$$

$$\text{Gain (credit_rating)} = 0.97 - 0.97 = 0$$

Gain (income) = 0.57
Gain (student) = 0.97
Gain (CR) = 0

student is highest



Income	student	c-R	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

iteration 3:

k- Nearest Neighbour Algorithm.

Problem 1: Apply k-NN Algorithm and Predict the type of Fruit or Food Type to which Tomato (Sweet = 6, Crunch = 4) belongs.
testing data

training data

Ingredients	Sweet	Crunch	food type
Grape	8	5	fruit
Green peas	3	7	Vegetable
nuts	3	6	protein
Orange	7	3	fruit

Solution Given testing data as tomato (Sweet = 6 and crunch = 4)

Step 1: Find Euclidean distance between tomato and (Grape, Greenpeas, Nuts, Orange)

- D (Tomato, Grape) : $\sqrt{(6-8)^2 + (4-5)^2} = 2.2$
- D (Tomato, Green peas) : $\sqrt{(6-3)^2 + (4-7)^2} = 4.2$
- D (Tomato, nuts) : $\sqrt{(6-3)^2 + (4-6)^2} = 3.6$
- D (tomato, Orange) : $\sqrt{(6-7)^2 + (4-3)^2} = 1.4$

Step 2: Find minimum distance = 1.4
Since minimum distance of tomato from orange is minimum tomato will belong to fruit type.

Date 3/10/19

Problem 2 Perform K-NN classification algorithm on following dataset. and predict the class for $\underbrace{x(P_1 = 3, P_2 = 7)}_{\text{test data}}$ consider $K = 3$.

Sl. No.	P_1	P_2	Class
1	7	7	False
2	7	4	False
3	3	4	True
4	1	4	False

Solution → Given testing data as $x(P_1 = 3, P_2 = 7)$

Step 1:- $D(x,1) = \sqrt{(3-7)^2 + (7-7)^2} = 4$

$D(x,2) = \sqrt{(3-7)^2 + (7-4)^2} = 5$

$D(x,3) = \sqrt{(3-3)^2 + (7-4)^2} = 3$

$D(x,4) = \sqrt{(3-1)^2 + (7-4)^2} = 3.60$

Step 2: Find minimum distance

given $K = 3$ so 3 minimum values are 3, 3.6, 4

Since number of False are more than number of true.

The testing data $x(P_1 = 3, P_2 = 7)$ belongs to class false.

DBCE

classmate
Date 3/10/19
Page

UNIT 4

Association Rules: Apriori Algorithm

(min : 8 mks & max : 15 mks)

v.v.v.v.v.v Imp Problem) For the following given Transaction Data-set, Generate association rules using APRIORI ALGORITHM.
Consider the values as support = 50% and confidence = 75%.
– (12 mks)

Trascation Id	Items Purchased
1	Bread, cheese, Egg, Juice
2	Bread, cheese, Juice
3	Bread, Milk, yogurt
4	Bread, Juice, Milk
5	cheese, Juice, Milk

Solution

Step 1: (1-Item) frequent Item set

$$\text{Support} = \frac{\text{Frequency}}{\text{No. of Transaction}}$$

Items	Frequency	Support	
Bread	4	4/5 = 80%	
Cheese	3	3/5 = 60%	
Egg	1	1/5 = 20%	(X)
Juice	4	4/5 = 80%	
Milk	3	3/5 = 60%	
yogurt	1	1/5 = 20%	(X)

Remove Egg and yogurt because their support is less than 50% (which is given)

Step 2: (2-Item) frequent Itemset / 2-Item candidate set

Item-Purchased	Frequency	Support	
Bread, cheese	2	2/5 = 40%	(X)
Bread, Juice	3	3/5 = 60%	
Bread, Milk	2	2/5 = 40%	(X)
Cheese, Juice	3	3/5 = 60%	
Cheese, Milk	1	1/5 = 20%	(X)
Juice, Milk	2	2/5 = 40%	(X)

Remove this 2-Item candidate set as their support is 50%

Step 3 Now we have only 2 - Itemset to generate Rules.

1) Bread, Juice
- → Bread → Juice
- → Juice → Bread

2) cheese, Juice
- → cheese → Juice
- → Juice → cheese

Step 4: Using formula

(*) confidence $(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$

1.1) Confidence (Bread → Juice) $= \frac{\text{Support}(\text{Bread} \cup \text{Juice})}{\text{Support}(\text{Bread})}$

$= \frac{3/5}{4/5}$

$= \frac{3}{4} = 75\%$

⇓

Rule is valid because it's confidence is ≥ 75% which is given in the Question.

1.2) Confidence (Juice → Bread) $= \frac{S(J \cup B)}{S(\text{Juice})} = \frac{3/5}{4/5}$

$= \frac{3}{4} = 75\%$

⇓

Rule is valid ∵ confidence is ≥ 75%.

classmate
Date 3/10/19
Page

2.1) Confidence (cheese → Juice) $= \frac{S(C \cup J)}{S(C)} = \frac{3/5}{3/5}$

$= \frac{3}{3} = 100\% \rightarrow$ Rule is valid

2.2) Confidence (Juice → cheese) $= \frac{S(J \cup C)}{S(J)} = \frac{3/5}{4/5}$

$= \frac{3}{4} = 75\% \rightarrow$ Rule is valid

So all above 4 valid association rules generated can be implemented as their confidence is $\geq 75\%$.

Problem 2 Consider the dataset shown below (8 mks)

Customer_Id	Transaction_Id	Items Bought
1	1	{a, d, e}
1	24	{a, b, c, e}
2	12	{a, b, d, e}
2	31	{a, c, d, e}
3	15	{b, c, e}
3	22	{b, d, e}
4	29	{c, d}
4	40	{a, b, c}
5	33	{a, d, e}
5	38	{a, b, e}

(a) Compute the support for itemsets {e}; {b, d} and {b, d, e} by each transaction ID as a market basket.

(b) Use the results in PART (a) to compute the confidence for the association rules {b, d} → {e} and {e} → {b, d}

Problem 3: For the following given transaction Data-set, Generate Rules, using Apriori Algorithm. Consider the values as support = 22% and confidence = 70%. (12 mks)

Trasaction_Id	Items Purchase
1	I_1, I_2, I_5
2	I_2, I_4
3	I_2, I_3
4	I_1, I_2, I_4
5	I_1, I_3
6	I_2, I_3
7	I_1, I_3
8	I_1, I_2, I_3, I_5
9	I_1, I_2, I_3

Problem 2 Solution

a → Support $(\{e\}) = \frac{8}{10} = 0.8 = 80\%$

Support $(\{b,d\}) = \frac{2}{10} = 0.2 = 20\%$

Support $(\{b,d,e\}) = \frac{2}{10} = 0.2 = 20\%$

b → confidence $(\{b,d\} \rightarrow \{e\}) = \frac{S(\{b,d,e\})}{S(\{b,d\})} = \frac{0.2}{0.2} = 1 = 100\%$.

confidence $(\{e\} \rightarrow \{b,d\}) = \frac{S(\{b,d,e\})}{S(\{e\})} = \frac{0.2}{0.8} = 25\%$.

Item	frequency	support	
I_1	6	6/9 = 66.67 %	
I_2	7	7/9 = 77.778 %	
I_3	6	6/9 = 66.67 %	
I_4	2	2/9 = 22.22 %	
I_5	2	2/9 = 22.22 %	
I_1, I_2	4	4/9 = 44.44	
I_1, I_3	4	4/9 = 44.44	
I_1, I_4	1	1/9 = 11.11	X
I_1, I_5	2	2/9 = 22.22	
I_2, I_3	4	4/9 = 44.44	
I_2, I_4	2	2/9 = 22.22	
I_2, I_5	2	2/9 = 22.22	
I_3, I_4	0	0/9 = 0%	X
I_3, I_5	1	1/9 = 11.11	X
I_4, I_5	0	0/9 = 0%	X

DBCE

Date 12/10/19

Generate FP tree for the transact^n data set. Consider support count = 3

Transaction Id	Item Set
1	f, a, c, d, g, m, p
2	a, b, c, f, l, m, o
3	b, f, h, o
4	b, k, c, p
5	a, f, c, l, p, m, n

solution)

Create table with Items frequency and Priority.

while writing write in order

Items	Support count (freq.)	Priority	
a	3	3	3
b	3	6	4
c	4	2	1
d	1	—	
f	4	1	2
g	1	—	
k	1	—	
l	2	—	
m	3	4	4
n	1	—	
o	2	—	
p	3	5	5
h	1	—	

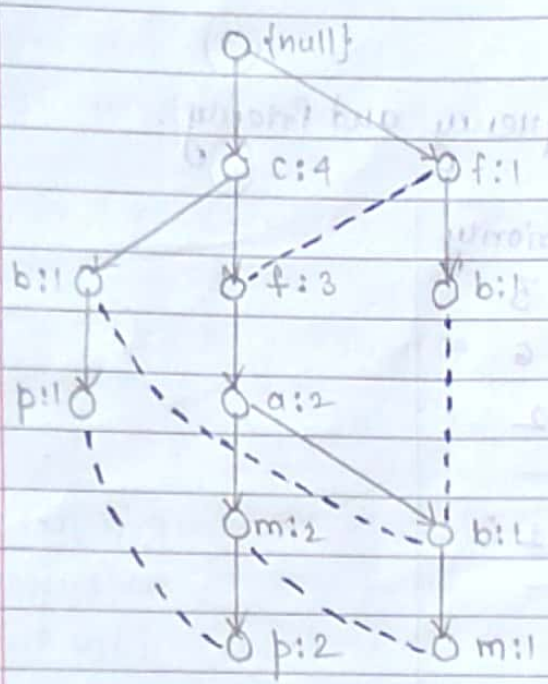
Remove the rows as support count is less than 3

before writing priority we have to remove the items whose support count is less than 3

step 2 Order the items according to Priority

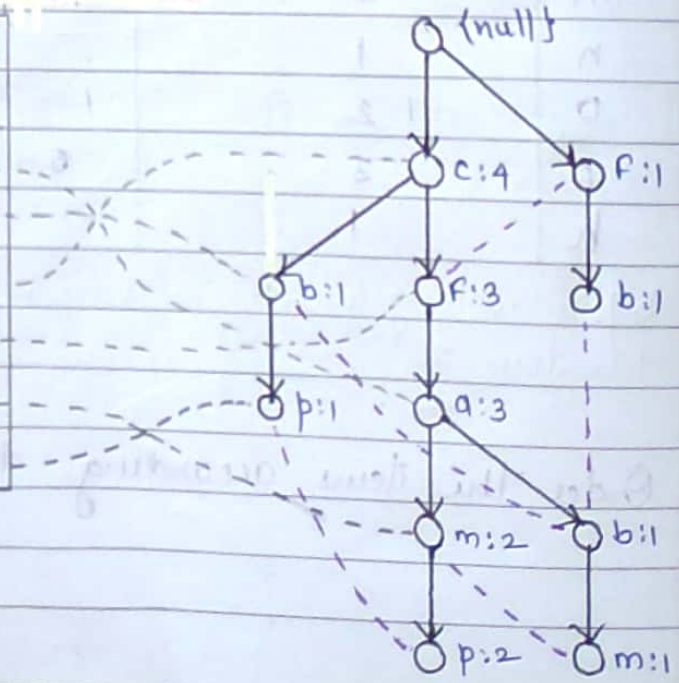
Transaction Id	Item set	
1	f, a, c, d, g, m, p	c, f, a, m, p
2	a, b, c, f, l, m, o	c, f, a, b, m
3	b, f, h, o	f, b
4	b, k, c, p	c, b, p
5	a, f, c, l, p, m, n	c, f, a, m, p

Step 3 Draw frequent pattern tree (FPT)



c : 1, 2, 3, 4
f : 1, 2, 3
a : 1, 2,
m : 1, 2
b : 1
m : 1
f : 1
b : 1
b : 1
p : 1

Items	Support count	Node link
a	3	•
b	3	•
c	4	•
f	4	•
m	3	•
p	3	•



Date 12/10/19

clustering

Divide the given sample data into 2 clusters using k-means algorithm

Sr. No.	Height(H)	Weight (W)
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

step1: Initialize 2 ~~classes~~ clusters

	H	W	Centroid
c_1	185	72	(185, 72)
c_2	170	56	(170, 56)

Assume cluster

$c_1 = \{1, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

$c_2 = \{2, 3\}$

Step2: Find Euclidean distance betn centroid & all Rows

$$E.D = \sqrt{(X_H - H_1)^2 + (X_W - W_1)^2}$$

observed values

centroid values.

For 3rd Row

* Euclidean distance of c_1 to 3rd Row is $= \sqrt{(168-185)^2 + (60-72)^2}$
$= 20.8$

* Euclidean dist. of c_2 to 3rd row is $= \sqrt{(168-170)^2 + (60-56)^2}$
$= 4.48$

Since cluster $4.48 < 20.8$, we put 3rd row in CLUSTER 2

For 4th Row

* 1st we find new centroid value of c_2 as 3rd Row is put in c_2

$\therefore c_1 = (185, 72)$ $\quad c_2 = (169, 58)$

$$c_2 = \left(\frac{170+168}{2}, \frac{60+56}{2}\right) \quad \therefore c_2 = (169, 58)$$

* Euclidean dist. of c_1 to 4th Row is $= \sqrt{(179-185)^2 + (68-72)^2}$
$= 6.32$

* Euclidean dist of c_2 ~~of~~ to 4th Row is $= \sqrt{(179-169)^2 + (68-58)^2}$
$= 14.14$

Since $6.32 < 14.14$, we put 4th Row value in CLUSTER 1

For 5th Row

1st we find new centroid value of c_1 as 4th Row is put in cluster 1

$$\therefore c_1 = \left(\frac{185+179}{2}, \frac{72+68}{2}\right) \quad \therefore c_1 = 182, 70$$

$c_2 = 169, 58$

E.D of C_1 to 5th Row is $= \sqrt{(182-182)^2 + (72-70)^2} = 2$

E.D of C_2 to 5th Row is $= \sqrt{(182-169)^2 + (72-58)^2} = 19.10$

Since $2 < 19.10$, we put 5th Row is put in CLUSTER 1

For 6th Row

1st we find new centroid value of C_1 as 5th Row is put in C_1

$C_1 = \frac{182+182}{2}, \frac{70+72}{2}$ $\therefore$ $C_1 = (182, 71)$
$C_2 = (169, 58)$

E.D of C_1 to 6th Row is $= \sqrt{(188-182)^2 + (77-71)^2}$
$= 8.48$

E.D of C_2 to 6th Row is $= \sqrt{(188-169)^2 + (77-58)^2}$
$= 26.87$

Since $8.48 < 26.87$, we put 6th Row is put in CLUSTER 1

For 7th Row

$C_1 = \frac{188+182}{2}, \frac{77+71}{2} = (185, 74) = C_1$
$(169, 58) = C_2$

E.D of C_1 to 7th Row $= \sqrt{(180-185)^2 + (71-74)^2} = 5.83$

C_2 $= \sqrt{(180-169)^2 + (71-58)^2} = 17.02$

DBCE

Date 18-9-19

Apply k-mean on the following 2D data for k=3.

$A_1=(2,10)$ $A_2=(2,5)$ $A_3=(8,4)$ $A_4=(5,8)$

$A_5=(7,5)$ $A_6=(6,4)$ $A_7=(1,2)$ $A_8=(4,9)$

→ Step 1: Draw fall table for above data

A_1	2	10
A_2	2	5
A_3	8	4
A_4	5	8
A_5	7	5
A_6	6	4
A_7	1	2
A_8	4	9

& assume centroid - $A_1 = (2,10)$

centroid - $A_2 = (2,5)$

centroid - $A_3 = (8,4)$

& ∴ clusters be $C_1 = \{A_1, A_4, A_8\}$

$C_2 = \{A_2, A_7\}$

$C_3 = \{A_3, A_5, A_6\}$

For 4th Row

E.D of C_1 to 4th row $= \sqrt{(2-5)^2 + (10-8)^2} = \sqrt{9+4} = \sqrt{13} = 3.6$

$= \sqrt{(2-5) + (5-8)^2} = \sqrt{18} = 4.24$

$= \sqrt{(8-5)^2 + (4-8)^2} = \sqrt{25} = 5$

⇒ C_4 will come in C_1.

For 5th row

update $C_1 \Rightarrow \left(\frac{2+5}{2}, \frac{10+8}{2}\right) = (3.5, 9)$

E.D of C_1 to 5th row $= \sqrt{(3.5-7)^2 + (9-5)^2} = 5.3$

$= \sqrt{(2-7)^2 + (5-5)^2} = 5$

$= \sqrt{(8-7)^2 + (4-5)^2} = 1.4$

⇒ C_5 will come in C_3

for 6th row

update $C_3 = \left(\frac{8+7}{2}, \frac{4+5}{2}\right) = (7.5, 4.5)$

ED of C_3 to 6th row $= \sqrt{(7.5-6)^2 + (4.5-4)^2} = 1.58$

C_2 $= \sqrt{(2-6)^2 + (5-4)^2} = 4.1$

C_4 $= \sqrt{(3.5-6)^2 + (9-4)^2} = 5.5$

$\Rightarrow$ C_6 will come in C_3

for 7th row

update $C_3 = \left(\frac{7.5+6}{2}, \frac{4.5+4}{2}\right) = (6.7, 4.2)$

E.D of C_4 to 7th row $= \sqrt{(3.5-1)^2 + (9-2)^2} = 7.4$

C_2 $= \sqrt{(2-1)^2 + (5-2)^2} = 3.16$

C_3 $= \sqrt{(6.7-1)^2 + (4.2-2)^2}$

$\Rightarrow$ C_4 will come in C_2

for 8th row

update $C_2 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

E.D of C_4 to 8th row $= \sqrt{(3.5-4)^2 + (9-9)^2} = 0.5$

C_2 $= \sqrt{(1.5-4)^2 + (3.5-9)^2} =$

C_3 $= \sqrt{(6.7-4)^2 + (4.2-9)} =$

$\Rightarrow$ C_3 will come in C_4 //

cluster 1 $= \{A_1, A_4, A_8\}$

" 2 $= \{A_2, A_7\}$

" 3 $= \{A_3, A_5, A_6\}$

another method (go for previous method)
↓

P.3 Apply k-means on the foll 1D dataset for k=2.
Dataset = {2, 4, 10, 12, 3, 20, 30, 11, 25}

Iteration 1

	A_1	2
$c_1 \rightarrow$	A_2	4
	A_3	10
	A_4	12
	A_5	3
	A_6	20
	A_7	30
$c_2 \rightarrow$	A_8	11
	A_9	25

$c_1 = (2, 3, 4, 10, 12, 11)$
$c_2 = (20, 30, 25)$

Assume $c_1 = 4$ & $c_2 = 11$
do it using method 1.

10 marks

Hierarchical clustering

- Clustering is the extracth of natural grouping of similar data objects.
- Hierarchical clustering relies using clustering techniques to find hierarchy of cluster, where this hierarchy resembles a tree structure called Dendrogram. It records sequences of merges or splits.
- There are 2 top level methods for finding these hierarchical clusters

1) Agglomerative (similarity & dist)

Steps

It uses a bottom up approach where in each data pt starts in its own cluster. These clusters are then joined greedily, by taking the 2 more similar clusters together & merging them.

not there in syllabus → 2) Divisive ×

It uses a top-down approach where in all data pts start in the same cluster. You can then use a parametric clustering algo like K-means to divide the cluster into 2 clusters. For each cluster, you further divide it down to 2 clusters until you hit the desired no. of clusters.

* if diag. matrix is 0 ⟹ distance
  1 ⟹ Similarity

Q1. Generate Dendrogram for the foll. proximity matrix (distance) using hierarchical clustering or agglomerative clustering.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.6	0.93	0.20	0.67	0

0's are diagonal ∴ distance

→ Step 1: Since it is dist-matrix, we will find values close to 0 (i.e smallest value from dist matrix)

- The smallest value is 0.12

∴ We combine A & B.

	AB	C	D	E	F
AB	0				
C	0.25	0			
D	0.16	0.14	0		
E	0.28	0.70	0.45	0	
F	0.34	0.93	0.20	0.67	0

(I)

$$dist(AB, C) = Min[dist(A, C), dis(B, C)]$$
$$= min(0.51, 0.25)$$
$$= 0.25$$

$$dist(AB, D) = Min[dist(A, D), dist(B, D)]$$
$$= Min[0.84, 0.16]$$
$$= 0.16$$

$$dist(AB, E) = Min(dist(A, E), dist(B, E))$$
$$= Min(0.28, 0.77)$$
$$= 0.28$$

$$dist(AB, F) = min(dist(A, F), min(B, F))$$
$$= min(0.34, 0.61)$$
$$= 0.34$$

Step 2: check for min. dist from (I). Since 0.14 is smallest vall we combine C & D.

DBCE

	AB	CD	E	F
AB	0			
CD	0.16	0		
E	0.28	0.45	0	
F	0.34	0.20	0.67	0

II

for similarity take highest val.

$$dist(CD, AB) = \min(dis(C, AB), dist(D, AB))$$
$$= \min(0.25, 0.16)$$
$$= 0.16$$

$$dist(CD, E) = \min(dist(C, E), dis(D, E))$$
$$= \min(0.70, 0.45)$$
$$= 0.45$$

$$dist(CD, F) = \min(dist(C, EF), dist(D, F))$$
$$= \min(0.93, 0.20)$$
$$= 0.20$$

step 3 check for min dist from II since 0.16 is smallest we combine CD & AB.

	ABCD	E	F
ABCD	0		
E	0.28	0	
F	0.20	0.67	0

III

$$dist(ABCD, E) = \min(dist(AB, E), dist(CD, E))$$
$$= \min(0.28, 0.45)$$
$$= 0.28$$

$$dist(ABCD, F) = \min(dist(AB, F), dist(CD, F))$$
$$= \min(0.34, 0.20)$$
$$= 0.20$$

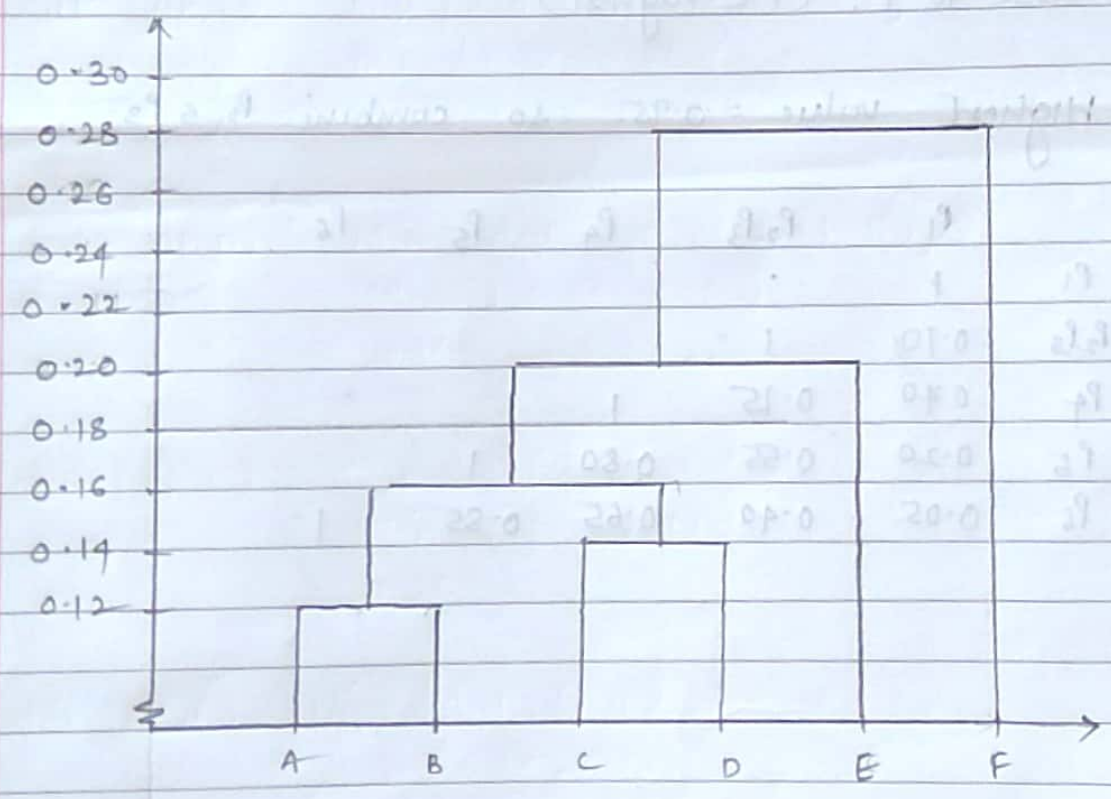
Step 4:- Min dist from (III) is 0.20 ∴ we combine ABCD & F.

	ABCDF	E
ABCDF	0	
E	0.28	0

(IV)

$$dist(ABCD, F) = min(dist(ABCD, E), dis(F, E))$$
$$= min(0.28, 0.67)$$
$$= 0.28$$

Step 5 Since only one val. remains, combine ABCDF & E & final Dendogram is



P2. Generate dendogram for the following given similarity matrix using Hierarchical clustering.

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	1					
P_2	0.70	1				
P_3	0.65	0.95	1			
P_4	0.40	0.70	0.75	1		
P_5	0.20	0.50	0.55	0.80	1	
P_6	0.05	0.35	0.40	0.65	0.85	1

step 1: since it is similarity matrix, we will find value close to 1. (i.e highest)

Highest value = 0.95 so combine P_2 & P_3.

	P_1	P_2P_3	P_4	P_5	P_6
P_1	1				
P_2P_3	0.70	1			
P_4	0.40	0.75	1		
P_5	0.20	0.55	0.80	1	
P_6	0.05	0.40	0.65	0.85	1

Dissimilarities betn data objects

Distances: The Euclidean distance measure given in equation below is generalize by Minkowski distances metric

$$d(x,y) = \sqrt[r]{\sum_{k=1}^{n} |x_k - y_k|^r} \quad \text{——— (1)}$$

where r is a parameter.

The following are the 3 most common examples of minkowski distances

1 $r = 1$, Manhattan or cityblock distance or L_1 norm

so eqn (1) becomes

$$d(x,y) = \sum_{k=1}^{n} |x_k - y_k|$$

2 $r = 2$, Euclidean distance / L_2 norm

so eqn (1) becomes

$$d(x,y) = \sqrt[2]{\sum_{k=1}^{n} (x_k - y_k)^2}$$

3 $r = \infty$, Supremum / L_{max} / L_∞ norm / chebyshev

This is the max difference between any attributes of the objects

so eqn (1) becomes

$$d(x,y) = \sqrt[r]{\lim_{r \to \infty} \sum_{k=1}^{n} |x_k - y_k|^r}$$

Problem 1 Given 2 objects represented by the attribute values 1, 6, 2, 5, 3) and (3, 5, 2, 6, 6)

i) Compute manhattan distance between 2 objects
ii) Compute Euclidean distance between 2 objects
iii) Compute supremum distance between 2 objects.
iv) Compute Minkowski distance when $r = 4$.

→

i) Manhattan distance

$$d(x,y) = \sum_{k=1}^{n} |x_k - y_k|$$

$$= |1-3| + |5-6| + |2-2| + |6-5| + |6-3|$$

$$= +2 + 1 + 0 + 1 + 3$$

$$= 7$$

ii) Euclidean distance

$$d(x,y) = \sqrt[2]{\sum_{k=1}^{n} (x_k - y_k)^2}$$

$$= \sqrt[2]{(1-3)^2 + (5-6)^2 + (2-2)^2 + (6-5)^2 + (6-3)^2}$$

$$= \sqrt[2]{4 + 1 + 0 + 1 + 9}$$

$$= 3.87$$

iii) Supremum distance (max distance)

$|1-3| = 2$
$|5-6| = 1$
$|2-2| = 0$
$|6-5| = 1$
$|6-3| = 3$

∴ supremum distance = 3

DBCE

iv)

$$d(A,B) = \sqrt[4]{\sum_{k=1}^{n} (x_k - y_k)^{4}} \quad r=4$$

$$d(A,B) = \sqrt[4]{(1-3)^4 + (6-5)^4 + (2-2)^4 + (5-6)^4 + (3-6)^4}$$

$$= 3.15$$

Rule base classifier

It is a technique for classifying records using a collection of "if ..... then ....." Rules

- Table 1 shown below is an example of model generated by rule base classifier for vertebrate classification problem
- Rules for the model r represented in a disjunctive normal form, $R = (r_1 \vee r_2 \vee r_3 \vee \cdots r_k)$ where 'R' is a rule set and 'r_i's are the classification rules or disjuncts

Rule set "R"

r_1 = (Gives birth = No) ∧ (Aerial_creature = YES) → Birds
r_2 = (Gives birth = No) ∧ (Aquatic_creature = YES) → Fishes
r_3 = (Gives_birth = Yes) ∧ (Body Temp = Warmblooded) → Mammals
r_4 = (Gives birth = No) ∧ (Aerial_creature = No) → Reptiles
r_5 = (Aquatic_creature = Semi) ⟶ Amphibian

	Name	Body temp	Skin cover	Gives birth	Aquatic creature	Aerial creature	Has legs	Hibernates
r_3	lemur	warm blooded	fur	Yes	No	No	Yes	Yes
r_4 & r_5	turtle	cold blooded	scales	No	semi	No	Yes	No
X	dogfish shark	cold blooded	scales	Yes	Yes	No	No	No

classification rule can be expressed as

$$r_i : \underbrace{(\text{condition}_i)}_{\text{Rule Antecedent}} \longrightarrow \underbrace{y_i}_{\text{Rule consequent}}$$

where condition $= (A_1 \; op \; v_1) \wedge (A_2 \; op \; v_2) \wedge \ldots (A_k \; op \; v_k)$