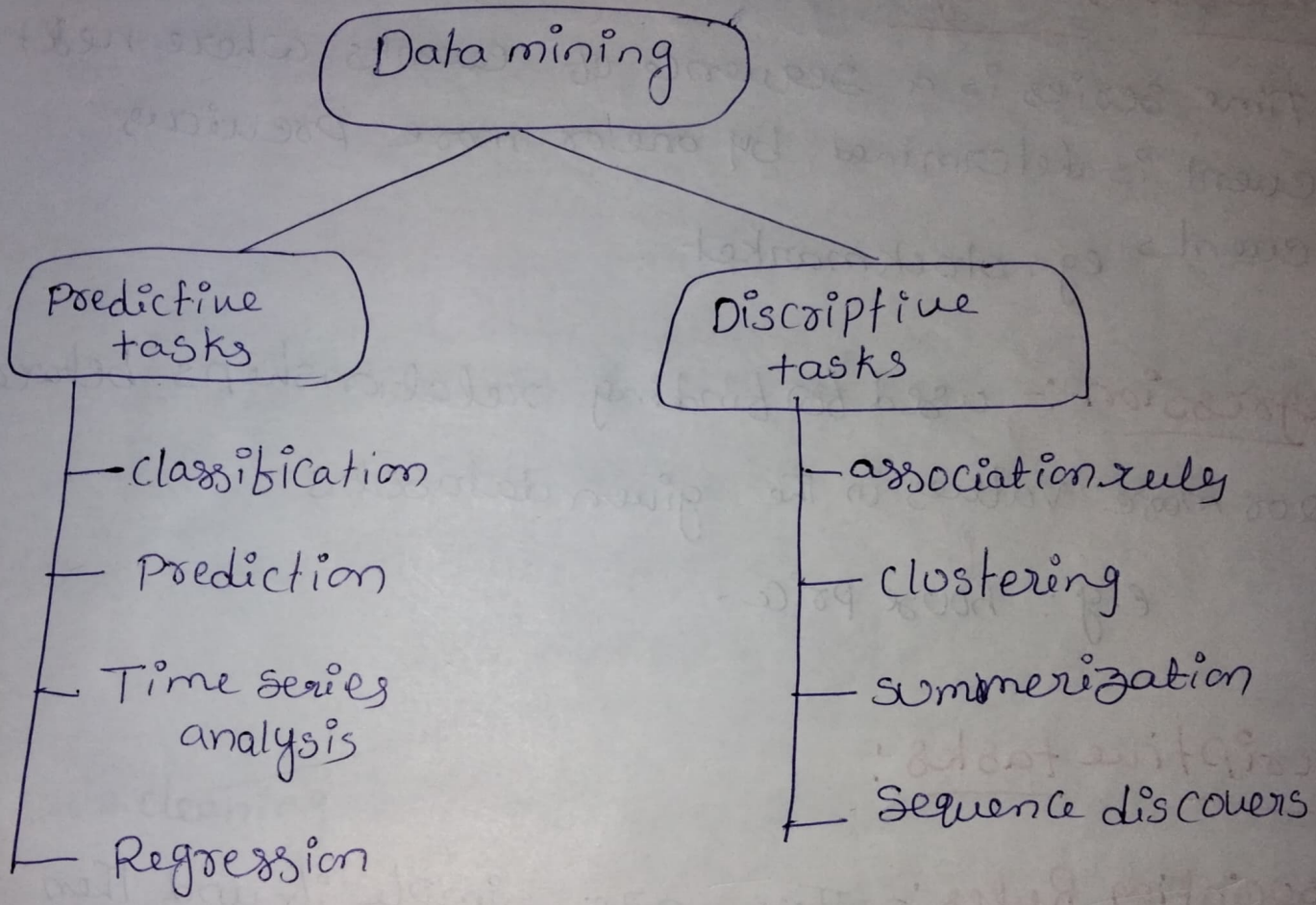
Data Mining Tasks :-



Prediction :- Prediction uses same variables or field in the database to Predict unknown or future values of other variables.

Discription :- Characterize the general properties of data and used for finding Patterns that describes the data

Classification:-

used to classify data into one of several predefined classes. eg: Loan < accept / Reject — Based on data

Prediction:- Predictive task come up with a model from the available data set that is helpful in predicting unknown or future values of another dataset

eg: Based on medical reports Doctor predict the diseases

Time series analysis:-

Time series is a sequence of events where next event is determined by one or more previous events

eg: stock market

Regression:- used for finding relationships between 2 or more values in the given datasets

e.g:- house price.

Discriptive tasks:

Association Rules:- These are simple if and then statements that help to discover relationships between datasets.

e.g: if Person buy milk Then he buy buy bread

Clustering:- It is a divison of information into groups of connected objects. Cluster is used to identify similar data.

e.g: Library books

summerization:- Used for finding compact description for subset of data.

sequence discovery:

Helps to discovery similar patterns in transactional data over time.

Data Preprocessing:-

Data preprocessing is a datamining technique which is used to transform the raw data in a useful and efficient format.

Data preprocessing

- Data cleaning
  - Missing data.
    1. Ignore the tuple
    2. Fill the missing values
  - Noicy Data
    1. Binning Methods
    2. Regression
    3. Clustering.
- Data transformation
  - Normalization
  - Attribute selection
  - Discretization
  - Concept Hierarchy Generation
- Data Reduction
  - Data Cube Aggregation
  - Attribute subset selection
  - Numerosity Reduction
  - Dimensionality Reduction.

1. Data Cleaning :-

The data can have many irrelavant and missing parts. to handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a) Missing Data :-

This situation arises when some data is missing in the data It can be handled in various ways.

1. Ignore the tuples :-

This approach is suitable only when the dataset we have is quite large and multiple values are missing with in a tuple.

2. Fill the Missing values :-

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b) Noisy Data :-

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways.

1. Binning Method :- This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task.

2. Regression:-

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables)

3. Clustering:-

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

(2) Data transformation:-

This step is taken in order to transform the data in appropriate forms suitable for mining process

1. Normalization:-

It is done in ordered to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. Attribute selection:-

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. Discretization:

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. Concept Hierarchy Generation:-

Here attributes are converted from lower level to higher level in hierarchy. For example "City" can be converted to "Country"

(3) Data Reduction:-

Data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps of data reduction are

1. Data Cube Aggregation:-

Aggregation operation is applied to data for the construction of the data cube.

2. Attribute subset selection:-

The highly relevent attributes should be used rest all can be discarded. For performing attribute selection, one can use level of significance and P-value of the attribute the attribute having P-value greater than significance level can be discarded.

3. Numerosity Reduction:

This enable to store the model of data instead of whole data For eg: Regression models

4. Dimensionality Reduction:

This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. Two effective methods of dimensionality reduction are:

→ wavelet transforms &
→ PCA (Principal Component Analysis)

Data Cleaning:

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

→ Missing values:

i) Ignore tuple:

which row having more missing values we can ignore that.

2) Fill in the missing values manually

3) Use a global constant to fill in the missing values (e.g NA)

4) use a measure of central tendency for the attribute to fill the missing value

5) use the most probable value to fill in the missing value (e.g. decision tree)

S.No	Name	occupation	Branch	Date	price	Address	Pin code
1	ramesh	govt	TPG	10-Mar		TPG	534
2	Soresh	selb	TNK	11-Mar	2500	TNK	
3	Prasad	private	TPG	11-Mar	500	TPG	534
4	Govind	private	VJWD	12-Mar	600	VJWD	520
5	Ravi						
6	nazesh	business	WJWD	14-Mar	1400	VJWD	520
7	Bhuvan	business	TPG	14-Mar	2300	TPG	
8	Kiran	govt	TNK			TNK	534
9		selb	RJY	15-Mar	1100	RJY	533
10	yogi	business	TNK	16-Mar	1800	TNK	534

→ Noisy Data :-

Noise is a random error or variance in a measured variable

Approches in Noisy data :

(i) Binning

(ii) Regression

(iii) outlier analysis

→ Binning :-

a) Partition into equal frequency bins

b) smoothing by bin means

c) smoothing by bin boundries.

eg: 6, 10, 17, 22, 22, 25, 27, 30, 36

Partition into equal frequency bins

Bin 1 : 6, 10, 17

Bin 2 : 22, 22, 25

Bin 3 : 27, 30, 36

Smoothing by bin means finding mean value.

Bin 1 : 11, 11, 11

Bin 2 : 23, 23, 23

Bin 3 : 31, 31, 31

Smoothing by bin boundaries.

~~Smoothing by b~~

Bin1 : 6, 6, 17

Bin2 : 22, 22, 25

Bin3 : 27, 27, 36

→ Regression :-

Linear regression involves finding the "best" line to fit two attributes so that one attribute can be used to predict the other.

Multiple Linear regression is an extension of Linear regression. where more than two attributes are involved and the data are fit to multidimensional surface.

→ outlier analysis :-

outlier may be detected by clustering for example, where similar values are organized into groups, or 'clusters' Intuitively, values that fall out side of the set of clusters may be considered outliers.

Data Transformation:-

It is a data preprocessing technique that transforms (or) Consolidate the data into alternate forms appropriate for mining.

(i) Smoothing:- Remove the noise from data [binning, regression, clustering]

(ii) Aggregation:- summary or aggregate function. It is used to constructing a data cube

(iii) Generalization: low level concepts are replaced with high level. eg: city to country

(iv) Normalization:-

attribute values are normalized by scaling their values so that they fall in specified range.

e.g: {2, 40, 500, 1, 3, 900}

we should change these values into 0-1 range.

That means we can normalize the attribute values

In this we have two types

(i) Min-Max Normalization:

$$v' = \frac{v - min_A}{max_A - min_A} (new\_max - new\_min) + new\_min$$

v' - new value

min_A - minimum value of attribute

max_A - maximum " " "

v - original attribute value.

(ii) Z-score normalization.

$$v' = \frac{v - \bar{x}}{\bar{\sigma}}$$

$\bar{x}$ - mean of attribute
$\bar{\sigma}$ - standard deviation
v - original value
v' - new value.

e.g: min-max normalization.

8
10
15
20 } Marks. It will scale the data into 0 & 1

$$v' = \frac{v - min_n}{max_n - min_n}$$

In the given attribute $min_n = 8$
$max_n = 20$

(i) original value $v = 8$

$$v' = \frac{8-8}{20-8} = \frac{0}{12} = 0$$

v' for 8 = 0

(ii) 10

$$v' = \frac{10-8}{20-8} = \frac{2}{12} = \frac{1}{6} = 0.16$$

(iii) 15

$$v' = \frac{15-8}{20-8} = \frac{7}{12} = 0.58$$

(iii) 20

$$v' = \frac{20-8}{20-8} = \frac{12}{12} = 1$$

new value = 1

new values are

8	0	new min = 0
10	0.16	new max = 1
15	0.58	
20	1	

eg: z score normalization:

$$v' = \frac{v - \bar{x}}{\sigma_x}$$

mean of marks = total of all attributes / no. of attributes

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

$= \sqrt{\frac{\Sigma (\text{every individual marks} - \text{mean of marks})^2}{n}}$

=

8
10
15
20
} marks

mean of marks = $8+10+15+20/4 = 13.25$

standard deviation = $\sqrt{\frac{(8-13.25)^2+(10-13.25)^2+(15-13.5)^2+(20-}{4}}$

$= \sqrt{\frac{(-5.25)^2+(-3.25)^2+(1.75)^2+(6.75)^2}{4}}$

$= \sqrt{\frac{27.56+10.56+3.06+45.56}{4}}$

$= \sqrt{\frac{86.74}{4}} = \sqrt{21.6} = 4.6$

standard deviation = 4.6

(i) $8 = \frac{8-13.25}{4.6} = -1.14$

(ii) $10 = \frac{10-13.25}{4.6} = -0.70$

(iii) $15 = \frac{15-13.25}{4.6} = 0.38$

(iv) $20 = \frac{20-13.25}{4.6} = 1.46$

Data Reduction:-

It is a PreProcessing technique that helps in obtaining reduced representation of data set from the available data set

→ Integrity of the original data should even after reduction in data volume

→ It should produce same analytics result as on original data.