Unit-2. 26/10/13

Data Preprocessing

Major steps in DataPreprocessing:-

The steps involved in the data preprocessing are

1. Data cleaning
2. Data Integration
3. Data Reduction
4. Data Transformation.

Data cleaning:-

Data cleaning routines to clean the data. by filling the missing values, Smoothing Noise data, Removing inconsistences, identifying or removing outliers.

Data Integration:-

Collecting the data from multiple sources like databases, datacubes, flatfiles can be called as Data Integration.

Ex:- DataCubes

Attribute for customer identification may be refered as customer id in the database. CU-Id, c-id in other databases. which can create a confusion (or) redundancy in analyzing the data.

Having large amount of redundant data may decrease the performance of knowledge discover process.

Hence data reduction is Required

Data Reduction:-

Reducing a dataset into a smaller set can be called as Data Reduction.

Data Reduction stategies includes

1) Dimensionality Reduction
2) Numerosity Reduction.

In dimensionality Reduction data compression techniques can be used and those are

1) Wavelet Transforms.
2) Principle Component Analysis.
3) Attribute Subset selection.
4) Attribute Construction.

Numerosity Reduction:-

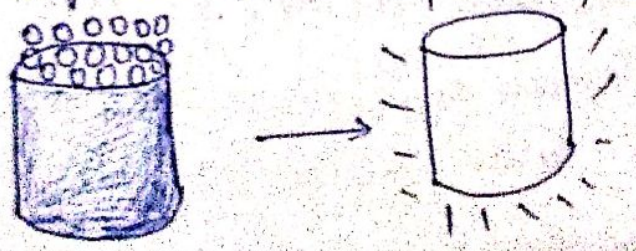
In this the data is replaced by smaller representation using either parametric models (Regression (or) loglinear models) or Non-parametric models (Histograms, Sampling, Clustering, data Integration).

Data Transformation:-

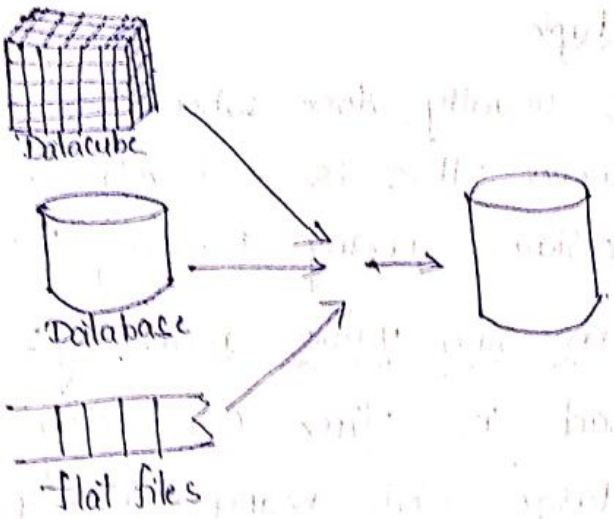
Convert the data into appropriate form for mining can be called as Data Transformation.

Forms of Data Preprocessing:-

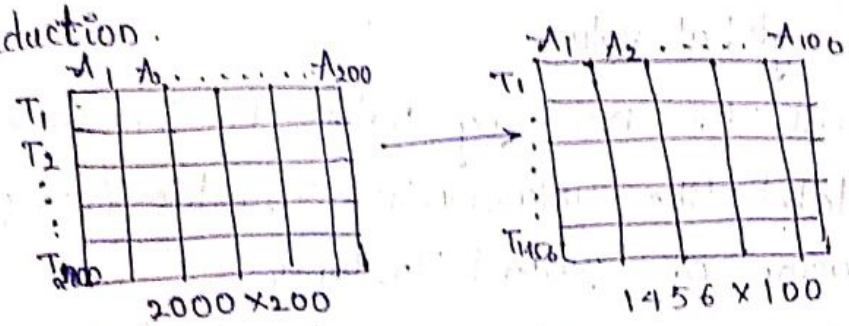
Data cleaning:-



Data Integration:-



Data Reduction.



Data Transformation:-

$-2, 48, 58, 100, 76 \longrightarrow -0.02, 0.48, 0.58, 1.0$

—

27/12/18

Data cleaning:-

Data cleaning routines to clean the data while filling the missing values, smoothing noisy data, Removing (or) identifying outliers.

Missing values:- In the dataset can be filled by using different ways so, the ways are:-

1. Ignore the tuple.
2. Filling in the missing values manually
3. place the global value.
4. Use the measure of central tendency to fill the missing values.
5. Use the attribute mean (or) median for all the samples belonging to the same class.
6. Use the most probable value to fill in the missing value.

Ignore the tuple:-

This is usually done when the class label is missing. However this is not effective unless the attribute contains many missing values.

Filling the missing values manually:-

This approach is time consuming whether the given dataset is large with many missing values.

Place the global value:-

Replace all the missing values by the same constant such as label like unknown (or) infinitive. (∞, ?)

Use a measure of central Tendancy:-

In this missing values can be filled by using the mean (or) median value.

If the data is Symmetric then use the mean value to fill in the missing value.

Skewed - median.

6)

This may be determined with regression or decision tree induction.

Eg:- Use regression to fill the missing value.

x	y	$x-\bar{x}$	$y-\bar{y}$	xy	x^2	y^2	
10	15	-3	-2	6	9	4	$\Sigma x = 49$
12	16	-1	-1	1	1	1	$\bar{x} = 13$
13	19	0	2	0	0	4	$\bar{y} = 16$
14	17	1	0	0	1	0	$\Sigma y = 67$.
15	?	2		7			$\Sigma xy = 7$.

$$y - \bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x - \bar{x})$$

$y - 17 = \frac{7}{11}(x - 13)$

$11(y - 17) = 7(x - 13)$

$11y - 187 = 7(x)$

$11y = 187 + 14$

$y = \frac{201}{11} = 18.27 \cong 19.$

29/12/18.

Noisy Data:-

Noise is a random error (or) variance in a measured variable.

Noisy data can be smooth end by using diffuent processors by binning, Regression, Outlier Analysis.

Binning:-

Binning methods smooth a sorted data values.

Diffuent Binning methods are:-

1. Bin by means
2. Bin by median.
3. Bin by boundaries.

Apply the binning methods for the following data.

4, 8, 15, 21, 21, 24, 25, 28, 34.

Sol) Partitions into equal frequencies

Bin1: 4, 8, 15

Bin2: 21, 21, 24

Bin3: 25, 28, 34.

1. Bin by means

Bin1: 9, 9, 9

Bin 2: 22, 22, 22.

Bin 3: 29, 29, 29.

2. Bin by median:-

Bin1: 8, 8, 8

Bin 2: 21, 21, 21

Bin 3: 28, 28, 28

3. Bin by boundaries:-

Bin1: 4, 4, 15

Bin2: 21, 21, 24

Bin 3: 24, 25, 34.

Bin by mean:-
Replace each bin value with its corresponding value

Bin by median:-
Replace each bin value with its median value.

Bin by boundaries:-
1. Write the boundaries of a bin.
2. Fill the middle values of the bin with the nearest boundary

*Apply the binning procedure for the following elements:
13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35,
35, 36, 40, 45, 46, 52, 70. Use the bin depth is 3.

Partition the data into equal frequencies.

Bin 1: 13, 15, 16, 16, 18, 20, 20, 21, 22
Bin 2: 22, 25, 25, 25, 25, 30, 33, 33, 35
Bin 3: 35, 35, 35, 36, 40, 45, 46, 52, 70.

Bin by means:-
Bin 1: 18, 18, 18, 18, 18, 18, 18, 18, 18
Bin 2: 28, 28, 28, 28, 28, 28, 28, 28, 28
Bin 3: 44, 44, 44, 44, 44, 44, 44, 44, 44.

Bin by median:-
Bin 1: 19, 19, 19, 19, 19, 19, 19, 19, 19
Bin 2: 25, 25, 25, 25, 25, 25, 25, 25, 25
Bin 3: 40, 40, 40, 40, 40, 40, 40, 40, 40.

Bin by boundaries:-
Bin 1: 13, 13, 13, 13, 22, 22, 22, 22, 22.
Bin 2: 22, 22, 22, 22, 22, 35, 35, 35, 35.
Bin 3: 35, 35, 35, 35, 35, 35, 35, 35, 70.

Regression:- 2/01/19

The regression can also be used for smoothing the data.

Regression can be performed in 2 ways

1. linear Regression
2. Multiple Regression.

linear Regression:-

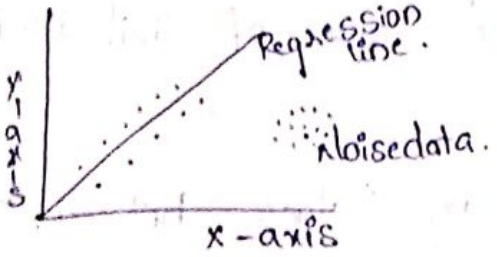
linear Regression involves finding the best line to fit two attributes. So that one attribute can be used to predict the other.

Multiple Regression:-

Multiple Regression can be used when there are more than two attributes.

(Regression is Completed.)x

Ex:-

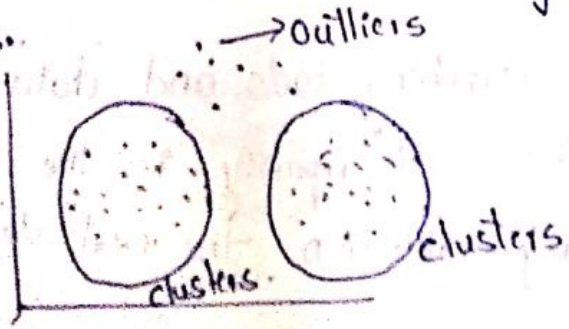


Outlier:-

Outlier can be detected by using clustering.

Objects that are similar to each other can form a group.

Objects that fall outside the groups can be called as outliers.



Data cleaning as a Process:-

The first step in the data cleaning as a process is discrepancy detection.

Discrepancy can be caused by several factors.

1) Poorly designed forms with many fields as optional.
2) Human error in data entry.
3) data decay (outdated Address).

The data should also be examined by regarding the following rules

1. Unique Rule
2. Consecutive Rule.
3. Null Rule.

Unique Rule:-

Each value of the given attribute must be different from all the other values of that attribute.

Consecutive Rule:-

There can be no missing values between the lowest and highest values of that attribute.

Null Rule:-

Null Rule specifies the use of blanks, question marks, special characters or other strings that may indicate Null condition and how such values should be handled.

There are data scrubbing tools and data auditing tools for finding the discrepancy in the data.

After the discrepancy detection the next step is data transformation.

1. Data migration tools.
2. ETL tools (Extraction Transformation and Loading)

3/01/19

Data Integration:-

Merging or Combining the data from multiple sources.

There are some important issues in the data Integration

1) Redundacy & co-relation Analysis.
2) Entity identification problem.
3) Tuple duplication.
4) Data value Conflict detection & Resolution.

Redundacy and co-relation Analysis:-

Redundacy can occur if one attribute is derived from another attribute or a set of attributes.

Inconsistency can also occurs due to the redundac in the data.

(Redundacy may be detetted by) some of the redundancie may be detected by co-relation analysis.

Co-relation Analysis:-

Given two attributes, the analysis of how strongly one attribute implies the other attribute based on the attribute data.

For the nominal data - chisquare test; applied.

Numerical data - Correlation Coefficient and Co-variation can variance be used.

ChiSquare Test:-

Given two attributes the gender and prefered reading are given in the table.

	Gender: Male	female	
Fiction	250 (90)	200 (360)	450
Non-fiction	50 (210)	1000 (840)	1050
	300	1200	

Test that is the genders.

Sol- Null hypothesis: There is a preference between the gender

tabulated value for 1 of x^2 is = 10.828.

$$\text{Cal } x^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Expected Frequencies:-

$$e_{11} = \frac{\text{count(male)} \times \text{count (fiction)}}{\text{total}}$$

$$e_{11} = \frac{300 \times 450}{1500} = 90$$

$$e_{12} = \frac{C(f) \times \text{count (fiction)}}{\text{total}} = \frac{1200 \times 450}{1500} = 360.$$

$$e_{21} = \frac{C(nf) \times C(m)}{\text{total}} = \frac{1050 \times 300}{1500} = 210.$$

$$e_{22} = \frac{C(f) \times C(nf)}{\text{total}} = \frac{1200 \times 1050}{1500} = 840.$$

$$\text{Cal } x^2 = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$

$$= \frac{25600}{90} + \frac{25600}{360} + \frac{25600}{210} + \frac{25600}{840}.$$

$$= 507.9.$$

∴ degree of freedom is $(c-1)(r-1) = (2-1)(2-1) = 1$

If the calculated chisquare is greater than the tabulated chisquare then reject null hypothesis.

Otherwise,

if calculated chisquare is less than the tabulated chisquare then accept the null hypothesis

$507.9 > 10.828$ is true, hence reject the Null hypothesis

Hence there is no preferences between the genders.

Ex:- In an investigation, all the machine performances the following results are obtained.

	No of units Inspected	No of defects	
M_1	375 (378.6)	17 (17.6) (455.9)	392
M_2	450 (446.4)	22 (21.5)	472
	825	39	864

Test whether there is any significant performance of two machines.

Sol:- There is a significant performance of two machines

The tabulated value for 1 of x^2 is = 10.828.

Expected Frequencies:-

$$e_{11} = \frac{825 \times 392}{864} = 378.6$$

$$e_{12} = \frac{39 \times 392}{864} = 17.6$$

$$e_{21} = \frac{472 \times 825}{864} = 455.9$$

$$e_{22} = \frac{39 \times 472}{864} = 21.5.$$

$$\text{Cal } x^2 = \frac{(375-378.6)^2}{378.6} + \frac{(450-455.9)^2}{455.9} + \frac{(17-17.6)^2}{17.6} + \frac{(22-21.5)^2}{21.5}$$

$$\text{Cal } x^2 = 0.215$$

$\text{Cal } x^2 < \text{tab } x^2$

$0.215 < 10.828$ hence accept the Null hypothesis.

* On the basics of information given below about the treatment of 200 patients suffering from a disease, State whether the new treatment is Comparatively Superior to the Conventional treatment.

	favourable	Non-favourable	Total
New	60 (45)	30 (45)	= 90
Conventional	40 (55)	70 (55)	= 110
	100	100	= 200

New Treatment is the Superior to the Conventional Treatment (or) new and Conventional treatment are independent

The x^2 value for the degree of freedom of 1 is 10.828.

Expected values:-

Cal $x^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$

$e_{11} = \frac{c(n) \times c(f)}{n} = \frac{90 \times 100}{200} = 45$

$e_{12} = \frac{c(n) \times c(nf)}{n} = \frac{90 \times 100}{200} = 45$

$e_{21} = \frac{c(c) \times c(f)}{n} = \frac{110 \times 100}{200} = 55$

$e_{22} = \frac{c(c) \times c(nf)}{n} = \frac{110 \times 100}{200} = 55$

Cal $x^2 = \frac{(60-45)^2}{45} + \frac{(30-45)^2}{45} + \frac{(40-55)^2}{55} + \frac{(70-55)^2}{55}$

Cal $x^2 = 18.18$

$Cal(\chi^2) > tab(\chi^2)$ so reject the hypothesis.

∴ new and Conventional Treatment are dependent.

Calculating the Co-relational for the numerical data:-

For the Numerical data Co-relation analysis can be done by co-relation coefficient.

Consider two numeric attributes A & B and a set of n observations $\{(a_1, b_1), \ldots, (a_n, b_n)\}$

The mean values of A and B respectively are also known as expected values of A & B.

$$\bar{A} = \sum_{i=1}^{n} \frac{a_i}{n} \qquad \bar{B} = \sum_{i=1}^{n} \frac{b_i}{n}$$

The co-variance between A and B defined as

$$COV(A,B) = \sum_{i=1}^{n} \frac{(a_i - \bar{A})(b_i - \bar{B})}{n}$$

The co-relation coefficient of A,B can be defined as $r_{A,B}$

$$r_{A,B} = \frac{COV(A,B)}{\sigma_A \, \sigma_B}$$

Note:-

If the value of r is limited to the $-1 < r < 1$

If $r_{A,B} = 0$ then there is no co-relation between A,B.

If $r_{A,B} > 0$ then there is +ve Corelation between AB

If $r_{A,B} < 0$ then there is −ve corelation between AB

*Let the two attributes temperature and icecream values are given below.

$(b_i - \bar{B})^2$	Temperature (A)	Icecream (B)	$a_i - \bar{A}$	$b_i - \bar{B}$	$(a_i - \bar{A})(b_i - \bar{B})$	$(a_i - \bar{A})^2$
84.64	6	20	2	9.2	18.4	4
0.64	5	10	1	−0.8	−0.8	1
10.24	4	14	0	+3.2	0	0
33.64	3	5	−1	−5.8	5.8	1
33.64	2	5	−2	−5.8	11.6	4
162.8	$\Sigma A = 20$	$\Sigma B = 54$	0		35	10

$$\bar{A} = \frac{20}{5} = 4, \quad \bar{B} = \frac{54}{5} = 10.8$$

Let us Assume that temperature is A and icecream is B

$$COV(A,B) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n} = 35/5 = 7$$

$$\sigma_A^2 = \frac{\sum_{i=1}^{n} (a_i - \bar{A})^2}{n} = 10/5 = 2.$$

$$\sigma_B^2 = \frac{162.8}{5} = 32.56.$$

$$\sigma_A = \sqrt{2} = 1.42.$$

$$\sigma_B = 5.70$$

$$r_{A,B} = \frac{COV(A,B)}{\sigma_A \sigma_B} = \frac{7}{(1.42)(5.70)} = 0.86.$$

Hence the two attributes temperature and icecream are in positive.

8/01/19.

* In One database height in inches and weight in kgs are given. Find if there is any significant corelation between the heights and weights for the given data.

Height in inches [A]	weight in kgs [B]
57	113
59	117
62	126
63	126
64	130
65	129
55	111
58	116
57	112
$\Sigma A = 540$	$\Sigma B = 1080$

Heights (A)	Weights (B)	$a_i-\bar{A}$	$b_i-\bar{B}$	$(a_i-\bar{A})(b_i-\bar{B})$	$(a_i-\bar{A})^2$	$(b_i-\bar{B})^2$
57	113	-3	-7	21	9	49
59	117	-1	-3	3	1	9
62	126	2	6	12	4	36
63	126	3	6	18	9	36
64	130	4	10	40	16	100
65	129	5	9	45	25	81
55	111	-5	-9	45	25	81
58	116	-2	-4	8	4	16
57	112	-3	-8	24	9	64
$\Sigma A=540$	$\Sigma B=1080$			216	102	472

$\bar{A} = \frac{540}{9} = 60 \quad \bar{B} = \frac{1080}{9} = 120$

$r = \frac{Cov(A,B)}{\sigma_A \cdot \sigma_B}$, $Cov(A,B) = \sum_{i=1}^{n} \frac{(a_i-\bar{A})(b_i-\bar{B})}{n}$

$= \frac{216}{9} = 24$

$\sigma_A = \sqrt{\frac{\sum_{i=1}^{n}(A_i-\bar{A})^2}{n}}$

$\sigma_A = \sqrt{\frac{102}{9}} = \sqrt{11.3} = 3.36.$

$\sigma_B = \sqrt{\frac{\sum_{i=1}^{n}(b_i-\bar{B})^2}{n}} = \sqrt{\frac{472}{9}} = 7.24.$

$r = \frac{Cov(A,B)}{\sigma_A \sigma_B} = \frac{24}{3.36 \times 7.24} = 0.986.$

$r = 0.986$, it means the 2 athibutes are in positive corelation.

∴ The two athibutes are strongly related. Hence one athibute can be used in the database instead of 2 attributes.

Entity Identification Problem:-

How can equivalent realworld entities from the multiple sources can be matched up?

This is called as entity identification problem.

Ex:- How the data analysist (or) a Computer be sure that the customer id in one database and customer number in other refer to same attribute.

When matching attributes from one database to another database during integration, a special attention must be paid to the structure of the data.

Tuple Duplication:-

In addition to detecting a redundancies between the attributes duplication should also be detected at Tuple level.

The use of these normalized tables is another source of the redundancy.

Data Value Conflict detection and Resolution:-

Data Integration also involves detection and resolution of data values.

For the same realworld entities attributes values from different source may differ. This may be due to difference in scaling, representation, encoding.

Example:-

A weight attributes may be stored in the metric units in one system and british imperior in another.

Variance is a special case of co-variance where the two attributes are identical.

Variance:-

The variance of n observations $(a_1, a_2, a_3, \ldots a_n)$, formulae for variance, a numeric attribute A is

$$\sigma_A^2 = \frac{\sum_{i=1}^{n} (a_i - \bar{A})^2}{n}$$

Ex:- Find the variance for the values 30, 36, 47, 50, 52, 60, 65, 70, 70, 110.

11/01/19.

Data Reduction:-

Data Reduction Techniques can be applied to obtain a reduced representation of the data set i.e much smaller in volume, yet closely maintainces the integrity of the original data.

Data Reduction Techniques involves Numerosity Reduction, Dimensionality Reduction, data Comprehension

* Under Dimensionality Reduction we have 3 Techniques

They are:-

1. Wavelet transforms 2. PCA 3. Attribute subset selection

* Numerosity Reduction
  - Parametric Reduction
    - Regression.
    - loglinear Models
  - Non-Parametric Models
    - Clustering
    - data integration
    - Cdatacube aggregation
    - Sampling
    - Histogram.

Histograms:-

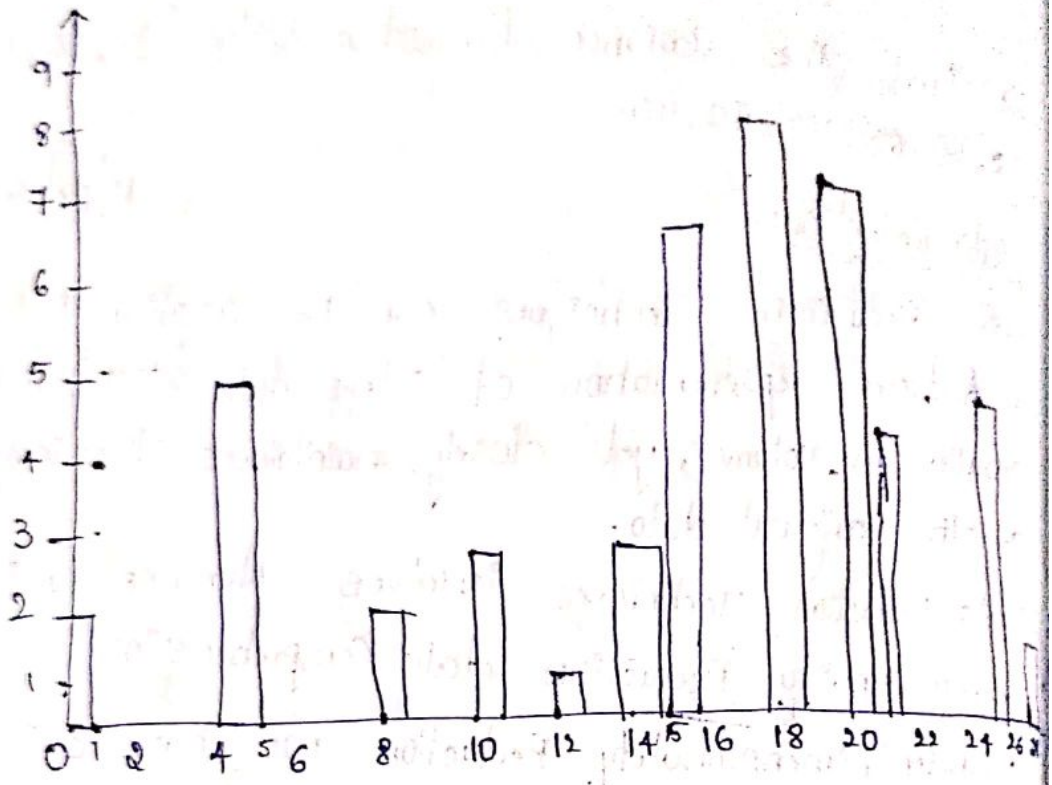
Histograms uses bining to approximate data. distribution.

A histogram for an attribute 'A' partitions the data into disjoint subset refer to as the buckets or wings bins.

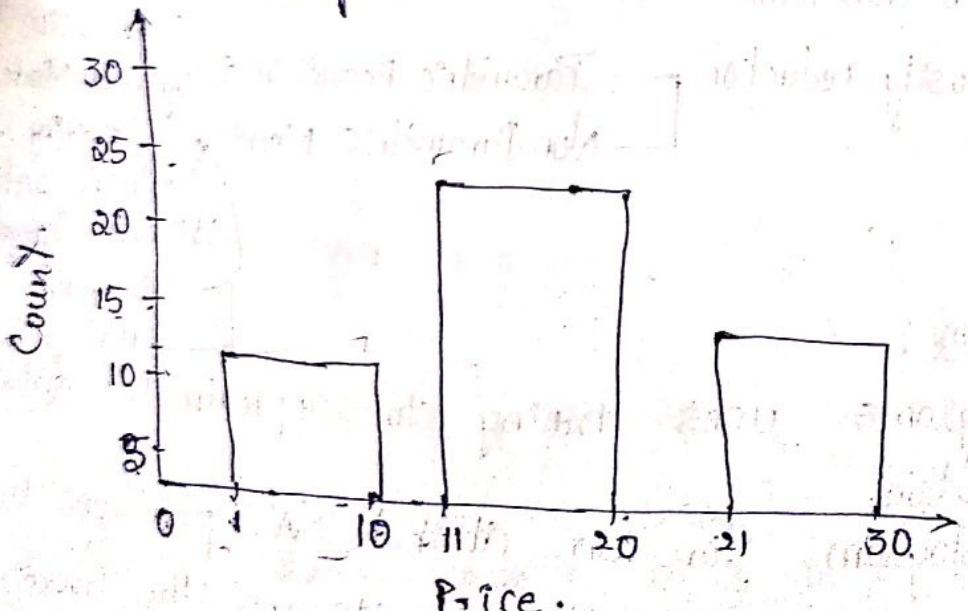
* Construct the histogram by using singleton buckets for the following data which contains

prices for Commonly sold items.

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.



Equal Width Histogram

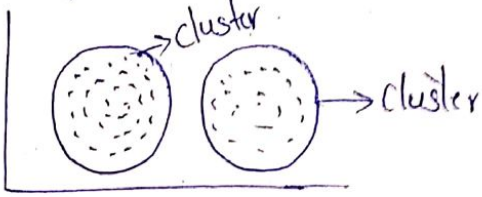


Clustering:- 29/01/19

Forming the given data into groups which are similar.

* clustering is the method of identifying similar group of data in dataset.

Objects with in the cluster are similar to Eachother and disimilar to objects of the other clusters.



*. Diameter can be used to measure the quality of a cluster.

*Centroid distance can also be used to measure the quality of a cluster.

Sampling Techniques that can be used data Reduction are as follows:-

1. Simple Random Sampling without replacement (SRSWOR)
2. Simple Random Sampling with, replacement (SRSWR)
3. Cluster Sampling
4. Stratified Sampling.

SRSWOR:-

Suppose that a large dataset 'D' contains 'n' tuples. SRSWOR can be created by drawing 's' of n tuples from D

The probability of drawing any tuple in D is 1/n ie all the tuples are equally likely to be sampled.

SRSWR:-

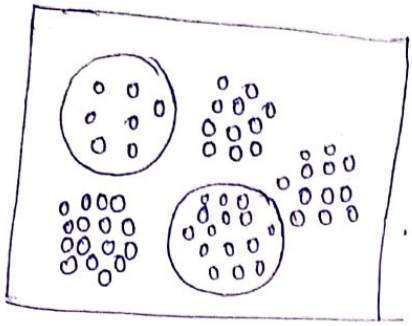
This is similar to SRSWOR except that each type a tuple is go drawn from D, it is recorded and then replaced.

Cluster Sampling:-

Divide the data into groups called clusters.

Randomly select some of the groups for sampling.

Ex:-



Stratified Sampling:-

Divide the data into groups called strata.

Select randomly from individual element of each strata.

Ex:- There are 50 states (dividing the whole population the country into states)

Randomly select 3 people from each state.

$3 \times 50 = 150$ members.

i.e Stratified Sampling count is 150 members.

Example for cluster Sampling:-

Selecting atleast two States(All the members of clust

Example for SRSWOR:-

Assume that their are 6 red, 5 blue and 4 green b in a dataset.

What is the probability of getting red ball, green ball and blue ball in one graph.

$P(RnBnG) = \frac{6}{15} \times \frac{5}{14} \times \frac{4}{13} = \frac{4}{91}$

$P(R) = 6/15$

$P(B) = 5/14$

$P(G) = 4/13$

With Replacement:-

$P(R) = 6/15$

$P(B) = 5/15$

$P(G) = 4/15$

$P(RnBnG) = \frac{6}{15} \times \frac{5}{15} \times \frac{4}{15}$

$= 8/225$

30/01/19.

Data Cube Aggregation:-

Performing aggregation operation on attribute values to analyse the data.

Ex:-

Year	2008
Q_1	$ 15000
Q_2	$ 16000
Q_3	$ 18000
Q_4	$ 24000

To know the annual sales of year 2008 then perform aggregation operation of all quarters.

Parametric Models:-

1. Regression:-

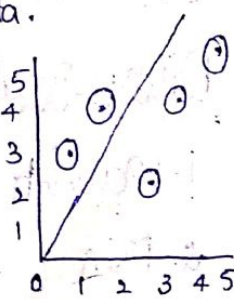
Regression and log linear models can be used to approximate the given data.

In the linear Regression, the data are modelled to fit a straight line (Regression line)

For example, A random variable 'y' (Response variable) can be modelled as a linear function of another random variable 'x' (Predictor variable) or (Independent variable) $y = mx + c$.

Find the regression line for the following data.

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	3	-2	-0.6	1.2	4
2	4	-1	+0.4	-0.4	1
3	2	0	-1.6	0	0
4	4	1	0.4	0.4	1
5	5	2	1.4	2.8	4
$\bar{x}=3$	$\bar{y}=3.6$			4	10



$$m = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{4}{10} = 0.4$$

$y = mx + c$

$3.6 = (0.4)(3) + c$

$3.6 = 1.2 + c$

$c = 3.6 - 1.2$

$\boxed{c = 2.4}$

$y = 0.4x + 2.4$ is the required equation.

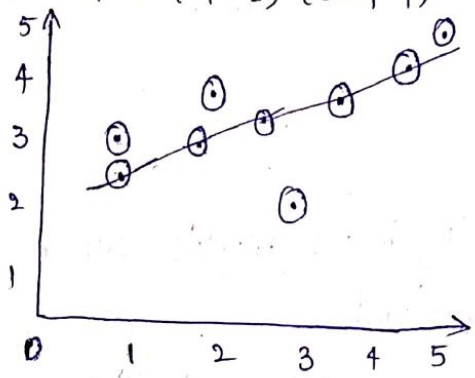
$x=1 \Rightarrow y = 0.4 + 2.4 = 2.8 \Rightarrow (1, 2.8)$

$x=2 \Rightarrow y = 0.8 + 2.4 = 3.2 \Rightarrow (2, 3.2)$

$x=3 \Rightarrow y = 1.2 + 2.4 = 3.6 \Rightarrow (3, 3.6)$

$x=4 \Rightarrow y = 1.6 + 2.4 = 4.0 \Rightarrow (4, 4.0)$

$x=5 \Rightarrow y = 2.0 + 2.4 = 4.4 \Rightarrow (5, 4.4)$



Loglinear Model:-

Loglinear Models allows a higher dimensional data space to be constructed from lower dimensional spaces.

Regression and loglinear models can be applied on a sparse data and skewed data.

31/01/19

Dimensionality Reduction:-

1. Wavelet transform.
2. PCA
3. Attribute Subset Selection.

Attribute Subset Selection:-

Extracting the subset of the attributes from the given 'n' no.of attributes can be called as

Attribute Subset Selection.

There are 4 procedures for attribute subset selection.

1. Stepwise forward selection.
2. Stepwise Backward Elimination.
3. Combination of Stepwise forward and Backward selection.
4. Decision tree induction.

Stepwise forward Selection:-

1. The procedure starts with empty set.
2. Identify the important attribute and add it to the previous set.

Ex:- $\{A_1, A_2, A_3, A_4, A_5, A_6, A_7\}$. Assume $\{A_1, A_4, A_6\}$ are required.

$S_1: \{\}$

$S_2: \{A_1\}$

$S_3: \{A_1, A_4\}$

$S_4: \{A_1, A_4, A_6\}$

Stepwise Backward Selection:-

1. Procedure starts with full set.
2. Remove the unnecessary attributes from the remaining set at each and every iteration.

Ex:- $\{A_1, A_2, A_3, A_4, A_5, A_6, A_7\}$ Assume $\{A_1, A_4, A_6\}$ are required.

$S_1: \{A_1, A_2, A_3, A_4, A_5, A_6\}$

$S_2: \{A_1, A_2, A_3, A_4, A_6\}$

$S_3: \{A_1, A_2, A_4, A_6\}$

$S_4: \{A_1, A_4, A_6\}$

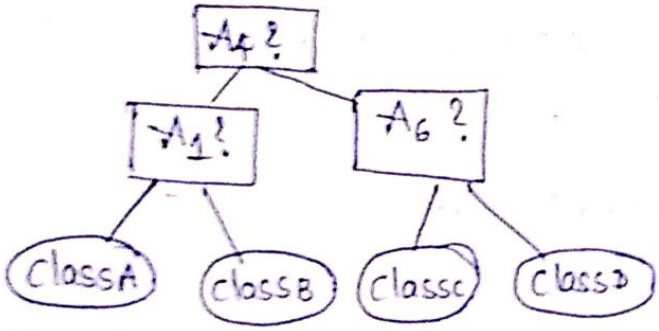
Combination of stepwise forward and Backward Sel

At each step the procedure selects the best attribute and remove the worst from among the remaining attributes.

Decision Tree Induction:-

Decision Tree Induction constructs a flow chart like a tree structure.

Ex:-



Wavelet Transforms-

Wavelet Transform is a linear Signal processing Technique when it is applied to the data vector it transform into numerically a different data vector called 'x'

The elements of 'x' can be called as wavelet co-eff Select the strongest wavelet coefficients from the x'.

Principle Components:- 2/02/19

Principle Component analysis can be used for data reduction.

Processor for identifying principle components.

Procedure for PCA:-

1. Find the co-variance matrix for the given two data points.
2. Find the co-variance matrix for the given data i.e co-variance matrix for the attributes x, y can be

$$\begin{array}{c} \\ x \\ y \end{array}\begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix} \end{array}$$ where $cov(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$

3. Find the Eigen values $|S-\lambda I| = 0$. λ_1, λ_2 can be called as eigen values.
3. Find the eigen vector namely $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$
4. Select the eign vector values which is corresponding to the highest eigen vector.

$$\begin{array}{c} 3 \\ 1 \end{array}\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

5. The selected eigen vector values can be called as principle components.

*Find the principle components for the given data.

x	2	1	0	-1
y	4	3	1	0.5

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
2	4	1.5	1.875	2.812
1	3	0.5	0.87	0.435
0	1	-0.5	-1.125	0.562
-1	0.5	-1.5	-1.62	2.43
$\bar{x}=0.5$	$\bar{y}=2.125$			$\Sigma = 6.239$

$$cov(x,y) = \frac{6.239}{3} = 2.079.$$

$$Cov(x,x) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1} = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1} = 1.60$$

$$Cov(x,x) = 5/3 = 1.66 \qquad Cov(y,y) = \sum_{i=1}^{n} \frac{(y-\bar{y})^2}{n-1} = \frac{8.16}{3} = 2.72$$

Now the covariance matrix is $S = \begin{bmatrix} 1.60 & 2.079 \\ 2.079 & 2.72 \end{bmatrix}$

Step2:- Find eigen values

$$|S - \lambda I| = 0$$

$$\begin{vmatrix} 1.66-\lambda & 2.079 \\ 2.079 & 2.72-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (1.66-\lambda)(2.72-\lambda) - (2.079)^2 = 0$$

$$\Rightarrow \lambda^2 - 4.38\lambda + 0.20 = 0$$

$$\lambda_1 = 4.34$$

$$\lambda_2 = 0.056.$$

Step3:- Find eigen vectors.

Substitute λ_1 in $|S - \lambda I|$

$$\begin{bmatrix} 1.66-4.34 & 2.079 \\ 2.079 & 2.72-4.34 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 0$$

$$\begin{bmatrix} -2.68 & 2.079 \\ 2.079 & -1.62 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 0$$

$$-2.68\, a_{11} + 2.079 a_{12} = 0 \rightarrow (1)$$

$$2.079 a_{11} - 1.62 a_{12} = 0 \rightarrow (2)$$

Orthogonal projects on $a_{11}^2 + a_{12}^2 = 1$

$$\boxed{a_{12} = 1.299 a_{11}} \rightarrow (3)$$

$$a_{11}^2 + (1.299 a_{11})^2 = 1$$

$$a_{11}^2 + 1.664 a_{11}^2 = 1 \Rightarrow a_{11}^2 (1 + 1.664) = 1$$

$a_{11}^2 = \frac{1}{2.664} \Rightarrow a_{11} = \sqrt{\frac{1}{2.664}} = 0.64.$

$a_{12} = 1.29 \times 0.64 = 0.825.$

Step 4:- Substitute λ_2 in $|S - \lambda I|$

$$\begin{bmatrix} 1.64 - 0.05 & 2.079 \\ 2.079 & 2.72 - 0.03 \end{bmatrix} \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} = 0$$

$a_{21} = 0.743$

$a_{22} = 0.829$

$$\begin{bmatrix} 4.34 \\ 0.05 \end{bmatrix} = \begin{bmatrix} 0.64 & 0.729 \\ 0.743 & 0.829 \end{bmatrix}$$

∴ From this a_{11} & a_{12} are principal components.

Dissimilarity measures for finding numeric attributes:-

There are '4' methods for finding dissimilarity b/w numeric attributes.

1) Euclides distance.
2) Manhattan distance.
3) Minkowski distance.
4) Supremum distance.

Let $i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in})$

$j = (x_{j1}, x_{j2}, x_{j3}, \ldots, x_{jn})$

1) Euclidean distance, $d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{12} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}$

2) Manhattan distance,

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|$$

3) Minkowski distance,

$$d(i,j) = \sqrt[h]{(x_{i1} - x_{j1})^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{in} - x_{jn}|^h}$$

4) Supremum distance,

$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{1/h}$ i.e., the attribute that given the max difference in values the two objects.

Eg:- Let $x_1 = 1, 2$

$x_2 = 3, 5$ are the given two points.

1) Euclidean distance:-

Euclidean distance $d(x_1, x_2) = \sqrt{(1-3)^2 + (2-5)^2}$

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

$= \sqrt{4+9} = \sqrt{13} = 4$

2) Manhattan distance:-

$d(x_1, x_2) = |1-3| + |2-5| = |-2| + |-3| = 2+3 = 5$

3) Minkowoshi distance:-

$d(x_1, x_2) = \sqrt[2]{(|1-3|)^2 + (|2-5|)^2}$

Let $h=2$

$= \sqrt[2]{4+9} = \sqrt[2]{13}$

4) Supremum distance:-

$\max(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|)$

$= \max(|1-3|, |2-5|)$

$= \max(|1-2|, |1-3|)$

$= \max(2, 3)$.

=

Data Compression Techniques:- 5/02/19.

In data compression transformations are applied to obtain a reduced are compressed representation of the orginial data.

If the original data can be reconstructed from the compressed data without any information loss then the data reduction is called lossless.

If we can reconstruct only an approximation of the original data then the data reduction is called lossy.

Data Transformation:-

Converting the data into its appropriate form for mining can be called as data transformation.

Strategies include in data transformation. They are several strategies:-

(i) Smoothing
(ii) Attribute Construction
(iii) Aggregation.
(iv) Normalisation.
(v) Discretization.
(vi) Concept hierachy generation.

(i) Smoothing:- To remove the noisy in the data. Techniques include binning, regression & clustering.

(ii) Attribute Construction:- When the new attributes are constructed & added from the given set of attributes to help the mining process.

(iii) Aggregation:-

Where summary (or) aggregation operations are applied to the data.

Eg:- The daily sales data may be aggregated to conclude monthly & annual total amount.

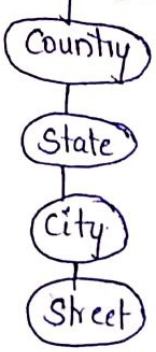
(iv) Normalisation:- where the attribute data (or) scale. so, as to form within a smaller range such as (0.0 to 1.0 (or -1.0 to 1.0)

(v) Discritization:- where the raw values of numeric attribute (age are replaced by interval labels) (0-10, 11-20, etc.,) (or) Conceptual labels (Eg:- youth, adult, senior).

(vi) Concept hierarchy generation:- where the attributes such as street can be generalised to higher level concepts like city (or) country.

street is nothing but lowerlevel concept.

Eg:-



Data transformation by Normalisation:-

Techniques include for Normalisation are:-

1) Min Max Normalisation.
2) z-score Normalisation.
3) Decimal scaling.

1) Min Max Normalisation:-

It performs a linear transformation with original data. Suppose that, min_A & max_A are the min & max values of an attribute A.

⇒ map a value v_i of A to v_i' in the range $[new\ min_A, new\ max_A]$ by computing $v_i' = \frac{v_i - min_A}{max_A - min_A} \times (new\ max_A - new\ min_A) + new\ min_A$.

Eg:- Let 'A' be the numeric attribute with observed values $v_1, v_2, v_3, \ldots\ldots, v_n$.

→ The min & max values for the attribute income are $12,000 & $98,000 respectively.

→ Map income to the range [0.0, 1.0]

→ A value v_i for income attribute is $73,600.

- $v_i = \$73,600$

$min_A = \$12,000$

$max_A = \$98,000$

$n - min_A = 0.0$

$n - max_A = 1.0$

$$v_i' = \frac{73,600 - 12,000}{98,000 - 12,000}(1.0 + 0.0) + 0.0$$

$v_i' = 0.716$.

2) Z-Score Normalization:-

Zero mean Normalisation:- The values for an attribute 'A' are normalised based on the mean & standard deviation.

Z-score normalisation transforms a value v_i of A into v_i' by computing

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

Eg:- The mean & standard deviation of the values attribute income are $54,000 & $10,000 respectively.

v_i' = when the v_i = $73,600.

$$v_i' = \frac{73,600 - 54,000}{16,000}$$

$$v_i' = 1.225.$$

Normalization by Decimal Scaling:- 6/2/19

Decimal Scaling normalises by moving the decimal point of values of attribute 'A'.

The no.of decimal points moved based on the maximum absolute value of 'A'.

A value v_i of A is normalized to v_i' by Computing $v_i' = \frac{v_i}{10^j}$ where 'j' is the smallest integer such that $\max(|v_i'|) < 1$.

Ex:- Suppose the recorded values of A' range from -986 to 917. Then find the normalized values for -986 to 917.

$$v_i' = \frac{986}{1000} = 0.986 \quad , \quad v_i' = \frac{917}{1000} = 0.917.$$

Normalized values of A is 0.986 to 0.917.

Let the attribute income contains the values

Z-Score Normalisation:-

The variation of Z-score normalization replaces the standard deviation by mean absolute deviation of A.

The mean absolute deviation of A is denoted by

$$S_A = \frac{|v_1 - \bar{x}| + |v_2 - \bar{x}| + |v_3 - \bar{x}| + \cdots + |v_n - \bar{x}|}{n}$$

$$s_i' = \frac{v_i - \bar{x}}{s_A}$$

Similarity measures cosign similarity:-

It is a measure Similarity that can be used to Compare documents.

let x, y be the two vectors for comparision

then Similarity of $x, y = \frac{x \cdot y}{\|x\| \cdot \|y\|}$

normalization $\|x\|$

where the norm $(\|x\|) = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

& norm $(\|y\|) = \sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}$

let there are '5' documents & the words that are commly appear in all the documents are appeared below.

	team	coach	hockey	baseball	soccer	penality	score	win
Doc1	5	0	3	0	2	0	0	2
Doc2	3	0	2	0	1	1	0	1
Doc3	0	7	0	2	1	0	0	3
Doc4	0	1	0	0	1	2	2	0

Sol: Assume that Doc1 is x and Doc2 is y then

$x = (5,0,3,0,2,0,0,2,0,0)$

$y = (3,0,2,0,1,1,0,1,0,1)$

$x * y = 5\times3 + 3\times2 + 2\times1 + 2\times1 = 25$.

$\|x\| = \sqrt{5^2+3^2+2^2+2^2} = \sqrt{25+9+4+4} = \sqrt{42} = 6.48$.

$\|y\| = \sqrt{3^2+2^2+1^2+1^2+1^2+1^2} = \sqrt{12} = 4.12$.

$\text{Similarity}(Doc1, Doc2) = \frac{25}{6.48 \times 4.12} = 0.986$.

7/02/19.

Dissimilarity measure for nominal attribute:-

If an attribute can take two (or) more states then that attribute can be called as nominal attribute.

The dissimilarity of an attribute can be represented by dissimilarity matrix.

Ex:-

The dissimilarity matrix for the given '4' objects and one attribute

$$\begin{bmatrix} d(1,1) & & & \\ d(2,1) & d(2,2) & & \\ d(3,1) & d(3,2) & d(3,3) & \\ d(4,1) & d(4,2) & d(4,3) & d(4,4) \end{bmatrix}$$

season
0
1
0
0

where $d(i,j) = 1$, means two objects are dissimilar

$d(i,j) = 0$ means the 'i' objects are not dissimilar

The Similarity of $sim(i,j) = 1 - d(i,j)$

The $sim(i,j) = 1$, means the 'i' objects are similar

$sim(i,j) = 0$ means the 'i' objects are nd similar.

The dissimilarity of 'i' objects for a nomial attribute.

$$d(i,j) = \frac{P-m}{P}$$

where P is no.of attributes & 'm' is no.of matches.

Problem:- Find the dissimilarity matrix for the given nominal data.

Id	Type1	Type2
1	10	A
2	30	B
3	10	A
4	20	D

Sol:

$d(1,1) = 0$

$d(2,1) = \frac{2-0}{2} = 2/2 = 1 \quad [P=2, m=0]$

$d(2,2) = 0$

$d(3,1) = \frac{2-2}{2} = 0/2 = 0 \quad [P=2, m=2]$

$d(3,2) = \frac{2-0}{2} = 2/2 = 1 \quad [m=0, P=2]$

$d(3,3) = 0$

$d(4,1) = \frac{2-0}{2} = 2/2 = 1 \quad [P=2, m=0]$

$d(4,2) = \frac{2-0}{2} = 2/2 = 1 \quad [P=2, m=0]$

$d(4,3) = \frac{2-0}{2} = 2/2 = 1 \quad [P=2, m=0]$

$d(4,4) = 0$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0 & 1 & 0 & \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

From the above data matrix all the objects are dissimilar except the objects 3, 1.

Ex:- Find the dissimilarity matrix from a given data.

Obj	Test 1.
1	code A.
2	code B
3	code C
4	code A.

$d(1,1) = \frac{1-1}{1} = 0.$

$d(2,1) = \frac{1-0}{1} = 1 \quad (P=1, m=0)$

$d(2,2) = 0.$

$d(3,1) = \frac{1-0}{1} = 1 \quad (P=1, m=0)$

$d(3,2) = \frac{1-0}{1} = 1 \quad (P=1, m=0)$

$d(3,3) = 0$

$d(4,1) = \frac{1-1}{1} = 0 \quad (P=1, m=1)$

$d(4,2) = \frac{1-0}{1} = 1 \quad (P=1, m=0)$

$d(4,3) = \frac{1-0}{1} = 1 \quad (P=1, m=0)$

$d(4,4) = 0.$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

From the above matrix all the objects attributes are dissimilar except object 4, 1.

Dissimilarity measure for ordinal attribute:-

The values of an ordinal attribute have a meaningful order or ranking above the attribute.

Ex:- 1. The sequence of small, medium, large.

2. Grade A, Grade B, Grade C.

3. Excellent, good, average.

Dissimilarity of an ordinal attribute can be computed by the following procedure.

1. Give the rank for the given states.

2. Normalize the given rankings to [0.0, 1.0] by computing

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$ M_f = no of states, r_{if} = ranking for a state.

3. Compute d(i,j) by $d(i,j) = |\text{Rank of } i - \text{Normalized Rank of } j|$.

Find the dissimilarity matrix for the given data.

Object	test
1	Excellent
2	fair
3.	good
4.	Excellent

1) Excellent - 1, good - 2, fair - 3 (ranking)

2) z_{if} for excellent $= \frac{1-1}{3-1} = 0$. $M_f = 3, r_{if} = 1$

z_{if} for good $= \frac{2-1}{3-1} = 1/2 = 0.5$ $M_f = 3, r_{if} = 2$.

z_{if} for fair $= \frac{3-1}{3-1} = 2/2 = 1$ $M_f = 3, r_{if} = 3$.

$\therefore$ [Excellent 1-0, good 2-0.5, fair 3-1]

3) $d(1,1) = 0$.

$d(2,1) = |1-0| = 1$.

$d(1,2) = 0$

$d(3,1) = |NRG - NRE)|$

$= |0.5 - 0| = 0.5$.

$d(3,2) = |NRG - NRE| = |0.5 - 1| = 0.5$

$d(3,3) = 0.$

$d(4,1) = |NRE - NRE| = 0 - 0 = 0.$

$d(4,2) = |NRE - NRF| = |0 - 1| = 1$

$d(4,3) = |NRE - NRG| = |0 - 0.5| = 0.5 \quad d(4,4) = 0$

$$\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix}\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

All the objects are dissimilar except the objects 1,4.

From the above matrix, the objects 1 and 2 are most dissimilar. and also the objects 4 and 2 are most dissimilar.

Dissimilarity measure for numeric data:- 9/02/19

The dissimilarity $d_{ij}(f) = \frac{|x_{if} - x_{jf}|}{max - min}$

Find the dissimilarity matrix for the given data.

Objects	attribute(test-3)
1	45
2	22
3	64
4	28

$d_{11} = 0$

$d_{21} = \frac{|22 - 45|}{64 - 22} = \frac{23}{42} = 0.547$

$d_{22} = 0$

$d_{31} = \frac{|64 - 45|}{64 - 22} = \frac{19}{42} = 0.452$

$$d_{32} = \frac{|64-22|}{64-22} = \frac{42}{42} = 1.$$

$$d_{33} = 0.$$

$$d_{41} = \frac{|28-45|}{64-22} = \frac{17}{42} = 0.40$$

$$d_{42} = \frac{|28-22|}{64-22} = 6/42 = 0.142$$

$$d_{43} = \frac{|28-64|}{64-22} = \frac{37}{42} = 0.857.$$

$$d_{44} = 0.$$

$$= \begin{bmatrix} 0 & & & \\ 0.547 & 0 & & \\ 0.452 & 1 & 0 & \\ 0.40 & 0.142 & 0.857 & 0 \end{bmatrix}$$

From the above data matrix Objects, all the objects are dissimilar except the objects 3 and 2 are dissimilar.

* Calculating the dissimilarity for mixed data:-

$$d_{ij} = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} \cdot d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

where $\delta_{ij}^{(f)} = 0$, if x_{if} (or) x_{jf} are missing (or) x_{if} and x_{jf} are zero otherwise it's corresponding value is 1.

Find the dissimilarity matrix for mixed data.

Object	test1 (nominal)	test2 (ordinal)	test3 (numeric)
1	codeA	Excellent	45
2	codeB	fair	44
3	codeC	good	64
4	codeA	Excellent	28

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

(nominal) (Ordinal) (numeric)

$$d_{21} = \frac{1\times1+1\times1+1\times0.55}{1+1+1} = \frac{1+1+0.55}{3} = \frac{2.55}{3} = 0.8$$

$$d_{31} = \frac{1\times1+0.5+1\times0.45}{3} = \frac{1+0.5+0.45}{3} = \frac{1.95}{3} = 0.615$$

$$d_{32} = \frac{1\times1+1\times0.5+1\times1}{3} = \frac{1+0.5+1}{3} = \frac{2.5}{3} = 0.83$$

$$d_{42} = \frac{1\times1+1\times1+1\times0.14}{3} = \frac{2.14}{3} = 0.71, \quad d_{41} = \frac{0+0+0.1}{3} = 0.$$

$$d_{43} = \frac{1\times1+1\times0.5+1\times0.86}{3} = \frac{1+0.5+0.86}{3} = \frac{2.36}{3} = 0.78$$

$$= \begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.615 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.78 & 0 \end{bmatrix}$$

Cosine Similarity:- 18/02/19

* Find the dissimilarity between the documents

	team	coach	hockey	baseball	soccer	penality	score	win	loss	seas
doc1	3	0	3	0	2	0	0	2	0	
doc2	4	0	2	0	1	1	0	1	0	
doc3	0	8	0	3	1	0	0	3	0	
doc4	1	2	0	0	1	3	2	0	4	

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

where norm $\|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$

norm $\|y\| = \sqrt{y_1^2 + y_2^2 + \ldots + y_n^2}$

Sol:- Assume that doc1 & doc2 are x & y respectively

doc1 = x = (3, 0, 3, 0, 2, 0, 0, 2, 0, 0)

doc2 = y = (4, 0, 2, 0, 1, 0, 1, 0, 1)

$x \cdot y = (3\times4 + 3\times2 + 2\times1 + 2\times1 + 0\times1) = 12 + 6 + 2 + 2 = 22$

Norm $||x|| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$

$= \sqrt{3^2+3^2+2^2+2^2} = \sqrt{9+9+4+4} = \sqrt{26} = 5.09.$

Norm $||y|| = \sqrt{16+4+1+1+1+1} = \sqrt{24} = 4.89.$

$\text{Sim}(doc1, doc2) = \frac{22}{5.09 \times 4.89} = 0.88.$

∴ The '2' doc's are more likely to be similar.

Assume that doc1 and doc3 are P, Q respectively

doc3 = Q = (0, 8, 0, 3, 1, 0, 0, 3, 0, 0)

doc1 = P = (3, 0, 3, 0, 2, 0, 0, 2, 0, 0)

PQ = (0+0+0+0+2+0+0+6+0+0) = 8

Norm $||P|| = \sqrt{3^2+3^2+2^2+2^2} = \sqrt{9+9+4+4} = \sqrt{26} = 5.09.$

Norm $||Q|| = \sqrt{8^2+3^2+1^2+3^2} = \sqrt{64+9+1+9} = \sqrt{83} = 9.11$

$\text{Sim}(doc1, doc3) = \frac{8}{5.09 \times 9.11} = \frac{8}{46.36} = 0.17.$

Similarity between binary attributes:-

Find the patience which are having similar disease.

	gender	fever	cough	test1	test2	test3	test4
jack	M	N	y	P	N	N	N
jim	M	y	N	N	N	P	N
mary	F	y	y	P	N	N	N.

Sol:- Convert the given data set into binary format.

	gender	fever	cough	test1	test2	test3	test4.
jack	M	0	1	1	0	0	0
jim	M	1	0	0	0	1	0
mary	F	1	1	1	0	0	0

$dissim(i,j) = \frac{r+s}{q+r+s}$

objecti \ objectj	1	0
1	q	r
0	s	t

$dis(jack, jim) = \frac{1+1}{0+1+1} = 1$

$q = 0$
$r = 2$
$s = 2$
$t = 2$

∴ There is no similarity between jack and jim.
There is no similar disease.

$sim(i,j) = 1 - dissim(jack, jim) = 1 - 1 = 0$

$dis(jack, mary) = \frac{0+1}{2+0+1} = 1/3 = 0.3$

$q = 2$
$r = 0$
$s = 1$
$t = 3$

$sim(i,j) = 1 - dis(jack, mary) = 1 - 0.3 = 0.7$

There is similarity between jack and mary. They have similar disease.

Q.

Unit-3

14/02/19.

Association Rules:-

Find the frequent item sets that are frequently purchased together for the given dataset.

TransID	List of Items IDs
Tid	
T_1	I_1, I_2, I_5
T_2	I_2, I_4
T_3	I_2, I_3
T_4	I_1, I_2, I_4
T_5	I_1, I_3
T_6	I_2, I_3
T_7	I_1, I_3
T_8	I_1, I_2, I_3, I_5
T_9	I_1, I_2, I_3

Apriori Algorithm:-

1. Scan the dataset 'D' for count of each candidate.

Scan D for count of each candidate → C_1

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

Compare Sup.count with min Sup count → L_1

Itemsets	Support count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

Sup = Support.

Let min. sup count = 2

2. Generate C_2 from L_1

$C_2 = L_1 \bowtie L_1$.

Generate C_2 from L_1 → C_2

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

Scan D for count of each candidate. → C_2

Itemsets	Sup.count
I_1, I_2	4
I_1, I_3	4
I_1, I_4	1
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2
I_3, I_4	0
I_3, I_5	1
I_4, I_5	0

L_2

Itemset	Support count
I_1, I_2	4
I_1, I_3	4
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2

Compare with the Sup. count with min Sup. count

Step3:- Generate C_3 from L_2. C_3

C_3

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

Generate C_3 from L_2.

Scan D for count of each candidate.

Itemset	Sup. count
I_1, I_2, I_3	2
I_1, I_2, I_5	2
I_1, I_2, I_4	1
I_1, I_3, I_5	1
I_2, I_3, I_4	0
I_2, I_3, I_5	1
I_2, I_4, I_5	0

Compare the Sup. count with min Sup. count

L_3

Item set	Sup count
I_1, I_2, I_3	2
I_1, I_2, I_5	2

Step4:- Generate C_4 from L_3.

C_4

I_1, I_2, I_3, I_5

Generate C_4 from L_3

Scan D for count of each candidate

Itemset	Sup. count
I_1, I_2, I_3 I_5	1

From the given dataset, the frequent purchased items are I_1, I_2, I_3 and I_1, I_2, I_5.

*

TID	List of Items
T_1	I_1, I_2, I_3, I_4, I_5
T_2	I_4, I_5
T_3	I_2, I_4
T_4	I_1, I_3
T_5	I_1, I_3, I_5
T_6	I_1, I_4, I_5

min count = 2.

1. Scan the dataset 'D' for count of each candidate.

C_1

Itemset	Support count
I_1	5
I_2	2
I_3	3
I_4	3
I_5	4

Scan 'D' for count of each candidate → Compare the Supportcount with minsupcount →

L_1

Itemset	Support count
I_1	5
I_2	2
I_3	3
I_4	3
I_5	4

2. Generate C_2 from L_1

Generate C_2 from L_1 →

C_2

L_1, L_2
L_1, L_3
L_1, L_4
L_1, L_5
L_2, L_3
L_2, L_4
L_2, L_5
L_3, L_4
L_3, L_5
L_4, L_5

Scan 'D' for count of each candidates →

C_2

Item set	Support count
L_1, L_2	1
L_1, L_3	3
L_1, L_4	2
L_1, L_5	4
L_2, L_3	1
L_2, L_4	2
L_2, L_5	1
L_3, L_4	1
L_3, L_5	2
L_4, L_5	2

Compare →

L_2

Itemset	Support count
L_1, L_3	3
L_1, L_4	2
L_1, L_5	4
L_2, L_4	2
L_3, L_5	2
L_4, L_5	2

3.

C_3

L_1, L_3, L_4
L_1, L_3, L_5
L_1, L_4, L_5
L_2, L_4, L_5
L_1, L_2, L_4
L_3, L_4, L_5

Generate L_3 from L_2 → Scan 'D' for count of each candidate →

C_3

Itemset	Support count
L_1, L_3, L_4	1
L_1, L_3, L_5	2
L_1, L_4, L_5	2
L_2, L_4, L_5	1
L_1, L_2, L_4	1
L_3, L_4, L_5	1

Compare →

L_3

Itemset	Support count
L_1, L_3, L_5	2
L_1, L_4, L_5	2

4.

Generate C_4 from L_3 →

L_1, L_3, L_4, L_5

Scan 'D' for count of each candidate →

Itemset	Support count
L_1, L_3, L_4, L_5	1

From the given dataset, the frequent purchased items are L_1, L_3, L_5 and L_1, L_4, L_5.

Generating Association Rules from frequent Statements are:-

$\{I_1, I_2, I_5\}$

$\{I_1, I_2, I_3\}$

Association Rule for I_1, I_2, I_5.

	Support count	Confidence	Confidence percentage
$\{I_1, I_2\} \rightarrow I_5$	2	2/4 = 0.5	50%
$I_1, I_5 \rightarrow I_2$	2	2/2 = 1	100%
$I_2, I_5 \rightarrow I_1$	2	2/2 = 1	100%
$I_1 \rightarrow I_2, I_5$	2	2/6 = 0.33	33%
$I_2 \rightarrow I_1, I_5$	2	2/7 = 0.28	28%
$I_5 \rightarrow I_1, I_2$	2	2/2 = 1	100%

$$\text{Confidence}(A \rightarrow B) = P(B/A) = \frac{\text{Support count}(A \cup B)}{\text{Support count}(A)}$$

$$\text{Confidence}(I_1, I_2 \rightarrow I_5) = \frac{\text{Support count}(I_1, I_2, I_5)}{\text{Support count}(I_1, I_2)} = 2/4.$$

$$\text{Confidence}(I_1, I_5 \rightarrow I_2) = \frac{\text{Support count}(I_1, I_5, I_2)}{\text{Support count}(I_1, I_5)} = 2/2 = 1$$

$$\text{Confidence}(I_2, I_5 \rightarrow I_1) = \frac{\text{Support count}(I_2, I_5, I_1)}{\text{Support count}(I_2, I_5)} = 2/2 = 1$$

$$\text{Confidence}(I_1 \rightarrow I_2, I_5) = \frac{\text{Support count}(I_1, I_2, I_5)}{\text{Support count}(I_1)} = 2/6 = 0.33$$

$$\text{Confidence}(I_2 \rightarrow I_1, I_5) = \frac{\text{Support count}(I_1, I_2, I_5)}{\text{Support count}(I_2)} = 2/7 = 0.28$$

$$\text{Confidence}(I_5 \rightarrow I_1, I_2) = \frac{\text{Support count}(I_1, I_2, I_5)}{\text{Support count}(I_5)} = 2/2 = 1$$

2) Given the itemsets milk(m), coke(c), jam(j), pepsi(p), Bread(b)

Find the strong association rules when the minimum Support count = 3 and confidence = 70%.

TID	Itemsets
T_1	M, C, j
T_2	M, P, j
T_3	M, j
T_4	C, B
T_5	B, j, C
T_6	M, B, j
T_7	C, B, j
T_8	B, C
T_9	M, C, B, j
T_{10}	M, P, C, j

Sol:- Step 1:- Scan the C_1 D dataset for count of each candidate.

Scan D for count of each candidate.

C_1

Itemset	Support count
M	6
C	7
j	8
P	2
B	6

Compare

L_1

Itemset	Support count
M	6
C	7
j	8
B	6

2. Generate C_2 from L_1.

Generate C_2 from L_1

C_2

Itemset
M, C
M, j
M, B
C, j
C, B
j, B

Scan D for count of each Candidate.

C_2

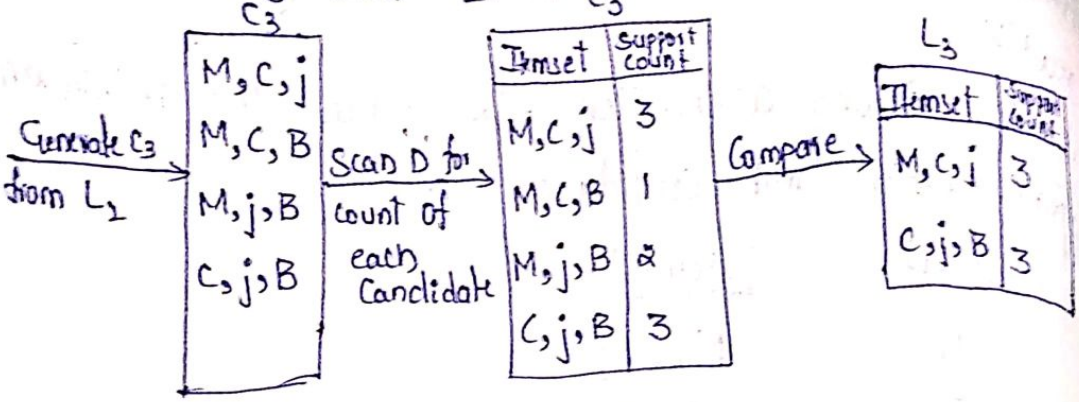
Itemset	Support count
M, C	3
M, j	6
M, B	2
C, j	5
C, B	3
j, B	4

Compare

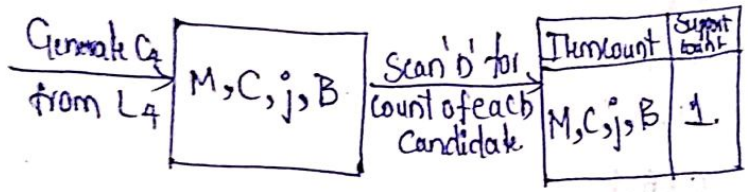
L_2

Itemset	Support count
M, C	3
M, j	6
C, j	5
C, B	3
j, B	4

3. Generate C_3 from L_2.



4. Generate C_4 from L_4.



From the given dataset, the frequent purchased items are M, C, j and C, j, B.

19/02/19.

Partitioning Algorithm for frequent itemsets:-

Transaction	items
T_1	Bread, Jelly, peanut, Butter (PNB)
T_2	Bread, peanut, Butter (PNB)
T_3	Bread, Milk, peanut, Butter (PNB)
T_4	Beer, Bread.
T_5	Beer, Milk.

1. Divide the given transactions into '2' equal parts.
2. Solve the equal parts by applying apriori algorithm
3. Combine all the transactions.

Sol:- 1. Divide the transactions T_1 & T_2 in one group and T_3, T_4, T_5 in another group.

Item	Support count
Bread	2
Jelly	1
PeanutButter	2

Compare → the MinSupp count with Supp count

Item	Support count
Bread	2
Jelly	1
Peanut Butter	2

Step2:-

itemset	Support count
Bread, Jelly	1
Bread, Peanut Butter	2
Jelly, Peanut Butter	1

Compare →

itemset	Support count
Bread, Jelly	1
Bread, Peanut Butter	2
Jelly, Peanut Butter	1

Step3:-

Itemset	Support count
Bread, Jelly, Peanut Butter	1

Compare →

itemset	Support count
Bread, Jelly, PeanutButter	1

$L_4 = \{\{B\},\{J\},\{PNB\}, \{B,J\},\{B,PNB\},\{J,PNB\},\{B,J,PNB\}\}$

Perform the apriori Algorithm for T_3, T_4, T_5.

Step1:-

Scan D' for count of each candidate

Itemset	Support count
Bread	2
Milk	2
Beer	2
Peanut Butter	1

Compare →

Itemset	Support count
Bread	2
Milk	2
PeanutButter	1
Beer	2

Step2:-

Itemset	Support count
Bread, Milk	1
Bread, PNB	1
Bread, Beer	1
Milk, PNB	1
Milk, Beer	1
PNB, Beer	0

Compare Support count with minsupcount

Itemset	Support count
Bread, Milk	1
Bread, PNB	1
Bread, Beer	1
Milk, PNB	1
Milk, Beer	1

Step3:-

Itemset	Support count
Bread, Milk, PNB	1
Bread, Milk, Beer	0
~~Bread~~	
Milk, PNB, Beer	0

Compare →

Itemset	Support count
Bread, Milk, PNB	1

$L_2 = \{\{B\},\{M\},\{PNB\},\{Beer\},\{B,M\},\{B,PNB\},\{B,Beer\},\{M,PNB\},\{M,B$
~~$\{PNB,Beer\}$~~, $\{B,M,PNB\}$, ~~$\{B,M,Beer\}$~~, ~~$\{M,PNB,Beer\}$~~$\}$.

Measuring the quality of Association Rules:- 20/02/18

The rules for measuring the quality of an Assoc rules are as follows:

1. Support
2. Confidence
3. Interest (Lift)
4. Conviction
5. Chi-square.

Support:-

$$\text{Support}(A \Rightarrow B) = \text{Support count}(A \cup B)$$

Confidence:-

$$\text{Confidence}(A \Rightarrow B) = P(B/A) = \frac{\text{Support count}(A \cup B)}{\text{Support count}(A)}$$

Interest (or) lift:-

$$\text{interest}(A \Rightarrow B) = \frac{P(A,B)}{P(A).P(B)}$$

Conviction:-

$$\text{Conviction}(A \Rightarrow B) = \frac{P(A)\,P(\neg B)}{P(A,\neg B)}$$

If the conviction value of A,B is 1 then A & B are not related.

Chi-Square:-

Chisquare for the itemset x is calculating

$$X^2 = \sum_{x \in I} \frac{(O(x) - E(x))^2}{E(x)} \quad \text{where } I = \{I_1, I_2, \ldots I_n\}$$

The chisquare value is 'zero' if all the values are independent.

	B	$\bar{B}$	Total
A	15	10	25
$\bar{A}$	55	20	75
Total	70	30	100

$$e_{11} = \frac{70 \times 25}{100} = \frac{70}{4} = 17.5$$

$$e_{12} = \frac{30 \times 25}{100} = \frac{30}{4} = 7.5$$

$$e_{21} = \frac{70 \times 75}{100} = \frac{210}{4} = 52.5$$

$$e_{22} = \frac{30 \times 75}{100} = \frac{90}{4} = 22.5$$

$$X^2 = \frac{(15-17.5)^2}{} + \frac{(10-7.5)^2}{} + \frac{(55-52.3)^2}{} + \frac{(20-22.5)^2}{}$$

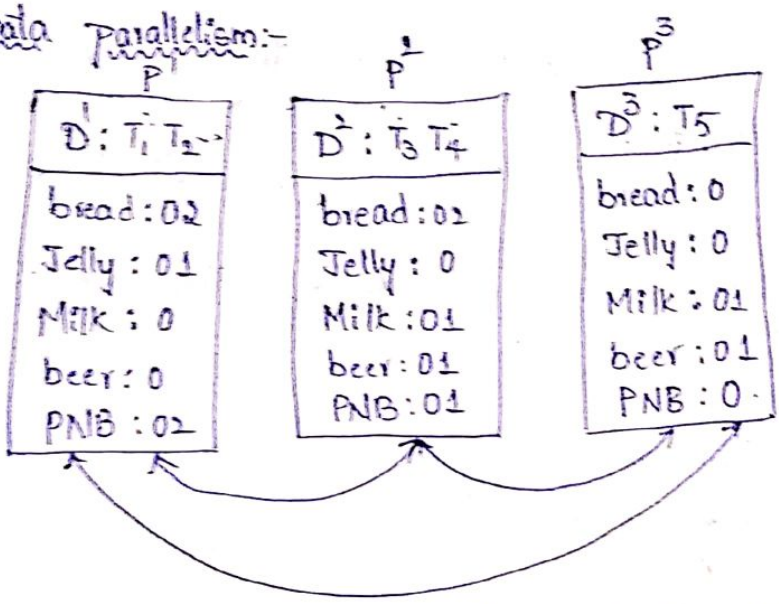
25/02/19

Parallel and Distributed Algorithm:-

Database:

T_1	bread, Jelly, PNB
T_2	bread, PNB
T_3	bread, Milk, PNB
T_4	beer, bread
T_5	beer, Milk.

Data parallelism:-



Most parallel distributed association rule algorithms try to parallies either the data known as data parallelism (or) the candiates refer to as process parallelism.

The data prallelism is also called as count distribution algorithm.

The database is divided into 3 partitions and is allocated to one for each processor.

Each processor counts the candidates for its data and then broadcast its counts throughto all other processors.

Each processor then determines the global count. These are used to determine the large itemsets.

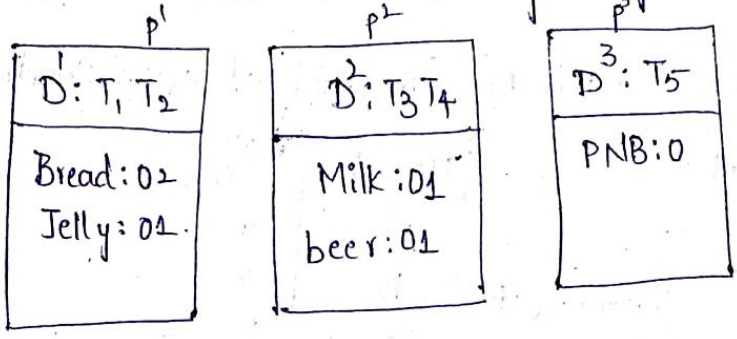
From the above diagram, There are 3 processors. First 2 transactions are counted as P^1 and next 2 transactions at P^2 and next 1 transaction at P^3.

When the local counts are obtained, they are broadcas to other processors, so that global counts can be generate

Task parallelism:-

In the task parallelism, the candidates are partitioned and counted seperately at each processor.

Eg:-



From the above diagram, There are 3 processors. P^1 is counting bread and Jelly, P^2 is counting milk and beer P^3 is counting PNB.

The first two Transactions are counted at P^1 and next 2 at P^2 and next 1 at P^3.

When the local counts are obtained, they are broadcast to other processors, so that global counts can be generat

*Apply the Apriori Algorithm for the given set of transactions. And find what are the 2 itemsets, 3 item sets when the minimum support is 60% and minimum confidence is 80%.

T_1	F, A, D, B
T_2	D, A, C, E, B
T_3	C, A, B, E
T_4	B, A, D.

Note:-
Frequent item set — minimum Support count
Association Rules — minimum Support count
minimum confidence.

Sol:- Min Supportcount = 60%

$= \frac{60}{100} \times 4 = 0.6 \times 4 = 2.4$

Min.supcount = 2

Step1:- Scan the dataset 'D' for count of each candidate in C_1

Itemset	Support count
A	4
B	4
C	2
D	3
E	2
F	1

Compare the minsupcount with Supportcount

L_1

Itemset	Support count
A	4
B	4
C	2
D	3
E	2

Step2:- Generate C_2 from L_1

C_2

Scan

A, B
A, C
A, D
A, E
B, C
B, D
B, E
C, D
C, E
D, E

C_2

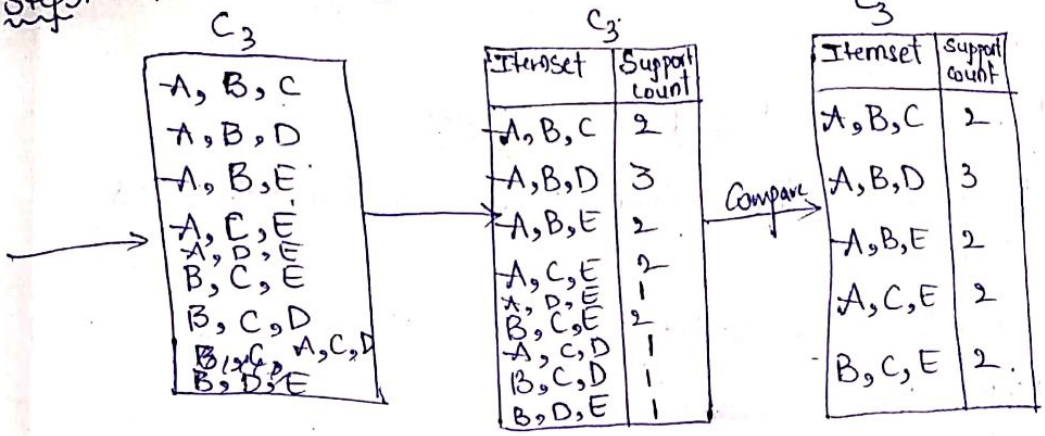
Item set	Support count
A, B	4
A, C	2
A, D	3
A, E	2
B, C	2
B, D	3
B, E	2
C, D	1
C, E	2
D, E	1

Compare

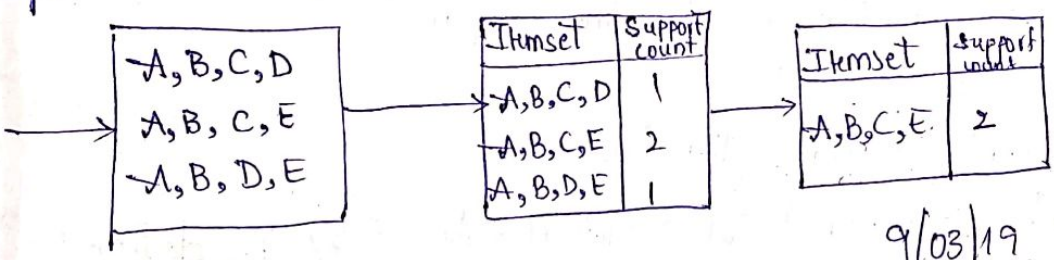
L_2

Item set	Support count
A, B	4
A, C	2
A, D	3
A, E	2
B, C	2
B, D	3
B, E	2
C, E	2

Step3:- Generate C_3 from L_2



Step4:- Generate C_4 from L_3



9/03/19

Incremental Rules:-

The algorithms like Apriori, partition, data parallelism, task parallelism, FPtree works for the static databases. However in reality the dataset will be a dynamic in the markets.

To address this issue, several approches has been proposed are:-

1. Incremental Approach.

→ The incremental Approach Concentrate on determining the large itemset for DUdb when 'D' is a Database and 'db' is the updates of the database.

One Incremental Approach is. fast Update.

This fast update is based on the apriori algorithm. In this each iteration 'k' scans both 'D and db with the candidates generated from the

iteration k-1 based on large itemset of that scan.

Advance Association Rule Technique:-

Techniques for Advanced Association Rules are as follows:

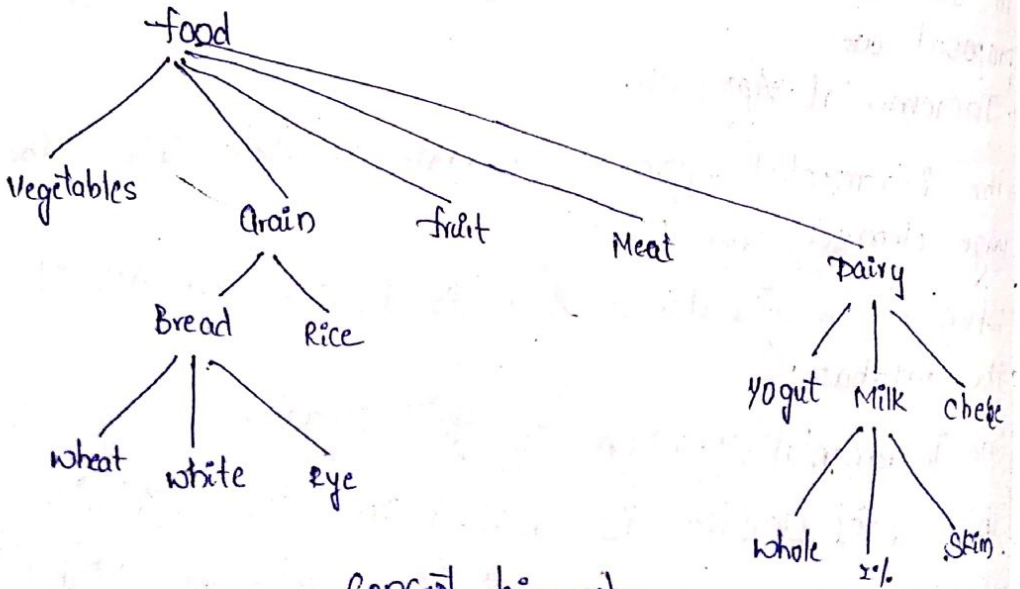
1. Generalized Association Rule.
2. Multiple level Association Rules.
3. Quantitative Association Rules.
4. Using Multiple minimum Supports.
5. Correlation Rules.

Generalized Association Rule:-

→ A Generalized Association Rule $x \rightarrow y$ is defined like an regular association Rule with the restriction that no item in 'y' may be above any item in 'x'.

Several Algorithm have been proposed to generate the association rules are:

→ The Simplest is Expand Each transaction by adding all the items above it in any hierarchy.



Concept hierarchy.

Multiple level Association Rules:-

→ A variation of Generalized Rules are Multiple level Association Rules.

In this itemset may occur from any level in the hierarchy.

Using a variation of Apriori algorithm, the concept hierarchy is traversed in top down and large itemset are formed.

The minimum support count for all the nodes in the hierarchy at the same level is Identical.

If Alpha 'i' is the minimum level of the level i in the hierarchy then $\alpha_i - i$ is the level of the level $i-1$ in then hierarchy then

$\alpha_{i-1} > \alpha_i$

Quantitative Association Rules:-

→ A Quantitative Association rule is one that is Categorical and quantitative data

Example:- A customer buys wine for between $30 and $50 of a bottle also buys a caviar.

Multiple minimum Supports:-

Using one minimum support value for a large database can be a problem.

This problem is also called as rare item problem.

If the minimum support is too high then rules involving items that appear rarely will not be generated.

If it is too low too many rules may be generated which are not important.

To handle this issue, one approache is partitioning the data based on the support & generate association rules for each partition Seperately.

Correlation Rule:-

→ Correction Rule is defined as a set of items that are correlated.

This can be useful for negative co-relations.

formula:- Correlation of $(A \rightarrow B) = \frac{P(A,B)}{P(A)\,P(B)}$ if the correction value < 1 then it indicates a negative correlation between A & B.

Classification:-

Examples of classification include image and pattern recognisation, medical diagnasis, loan approval, detecting faults in industry application.

Classification can be done by using different algorithms:

1. Statistical based Algorithms.
   (i) Regression, (ii) Bayes.
2. Distance based Algorithms
   K-nearest neighbour.
3. Decision Tree based Algorithms.
   (ID3, C4.5, CART)
4. Neural network based Algorithms.
5. Rule based Algorithms.
6. Combining Techniques.

Statistical based Algorithms:-

Baye's Algorithm:-

Consider the given weather dataset with atributes outlook, temperature, windy, humidity and classlabel play with values yes (or) No.

Find person that he can play (or) not with weather conditions.

Outlook=Sunny, temp=cool, humidity=high, windy=strong

The dataset for weather is:

	Outlook	Temperature	Humidity	Windy	play
1.	Sunny	hot	high	weak	No
2.	Sunny	hot	high	strong	No
3.	Overcast	hot	high	weak	Yes
4.	Rainy	mild	high	weak	yes
5.	Rainy	cool	nominal	weak	Yes
6.	Rainy	cool	nominal	strong	no
7.	Overcast	cool	nominal	strong	Yes
8	Sunny	mild	high	weak	No
9.	Sunny	cool	nominal	weak	Yes
10.	Rainy	mild	nominal	weak	yes
11.	Sunny	mild	nominal	strong	Yes
12.	Overcast	mild	high	strong	yes
13.	Overcast	hot	nominal	weak	yes
14.	Rainy	mild	high	weak strong	No.

probability of 'yes' = 9/14 = P(Yes)

probability of 'No' = 5/14 = P(No)

Outlook.	Yes	No	P(Yes)	P(No)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0
rainy	3	2	3/9	2/5

Temperature	Yes	No	P(Yes)	P(No)
Hot	2	2	2/9	2/5
mild	4	2	4/9	2/5
cool	3	1	3/9	1/5

Humidity	Yes	No	P(yes)	P(No)
High	3	4	3/9	4/5
Nominal	6	1	6/9	1/5

Windy	Yes	No	P(yes)	P(No)
weak.	6	2	6/9	2/5
Strong	3	3	3/9	3/5

probability that the person can play the game.

$$P(x/play = yes) \cdot P(play = yes)$$

$$= 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14$$

$$= 0.0053//$$

probability that the person cann't play the game.

$$P(x/play = No) \cdot P(play = No)$$

$$= 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14$$

$$= 0.0206//$$

$$P(C) = 0.0053 + 0.0206 = 0.0259$$

$$P(A/C) = \frac{P(C/A) \cdot P(A)}{P(C)}$$

$$P(play = yes/x) = 0.0053/0.0259 = 0.206.$$

$$P(play = No/x) = 0.0206/0.0259 = 0.795$$

The person.

∴ P(No) is having the maximum values when it is Compared with P(Yes).

Hence the person cannot go outside to play based on the given Sample 'x' i.e., weather condition outlook = sunny, temperature = cool, windy = strong and humidity = high.

13/05/19

Baye's classification:-

For the given dataset patient income with the attributes cold, running nose, headache, fever and the class label flu. Find a patient having flu will not with the given symptoms cold = yes, runnynose = no, headache = mild, fever = yes.

	Cold	runnynose	headache	fever	flu
1.	Yes	No	Mild	Yes	No
2.	Yes	Yes	No	No	Yes
3.	Yes	No	strong	yes	Yes
4.	No	Yes	Mild	yes	Yes.
5.	No	No	No	NO	No
6.	No	Yes	strong	yes	Yes
7.	No	Yes	strong	No	No
8.	Yes	Yes	Mild	Yes	Yes.

Sol:-

probability of yes = 5/8

probability of No = 3/8

Cold	Yes	No	P(yes)	P(No)
Yes	3.	1	3/5	1/3
No	2	2	2/5	2/3.

runnynose	Yes	No	P(yes)	P(No)
Yes	4	1	4/5	1/3
No	1	2	1/5	2/3

headache	Yes	No	P(yes)	P(No)
Mild	2	1	2/5	1/3
No	1	1	1/5	1/3
Strong	2	1	2/5	1/3

fever	Yes	No	P(yes)	P(No)
Yes	4	1	4/5	1/3
No	1	2	1/5	2/3

Probability that a patient person can have flu play the game

$P(X|play=yes) \cdot P(play=yes) = 5/8 \times 1/5 \times 2/5 \times 4/5 \times 3/5$

$= 0.024.$

Probability that a patient cannot have the flue

$P(X|flu=no) \cdot play\ P(flu=no) = 3/8 \times 1/3 \times 1/3 \times 1/3 \times 1/3$

$= 0.0009.$

$P(C) = 0.024 + 0.0009 = 0.0249.$

$$P(A/C) = \frac{P(C/A) \cdot P(A)}{P(C)}$$

$P(flu=yes|x) = 0.024/0.0249 = 0.963855$

$P(flu=No|x) = 0.0009/0.0249 = 0.036.$

∴ Hence P(yes) has the maximum value when it is compared with P(No)

The patient have flu with the given symptoms.

According to Baye's theorem.

But, from the database the person is having no flu.

Hence the baye's theorem classified as false positive.

Pros of Naive Baye's:-

- It is easy and fast to predict a class of dataset.
- Naive Baye's perform better prediction compare to other models.
- It performs well in the case of categorical input variables compared to the numerical variables.
- Naive baye's is also knows as bad estimator.

Applications:-

1. Medical Analysis
2. Text classification
3. Spam filtering
4. Sentiment Analysis

14/3/19

Distance based Algorithms:-

K-nearest neighbours:-

1. K is an integer value.
   Find the value of k (must be a positive integer)
2. The given data point that needs to be classified
   the algorithm computes the distance between
   the given data points and all other datapoints
   of given datasets
3. Find the 'k' nearest neighbours based on the
   distance (the least distance value is the first
   neighbour for a given data point).
4. The given data points will be placed in the
   ~~given data~~ po majority of the classes with
   respect to neighbours.

The dataset contains two attributes namely weight and height and one class variable.

Find a person 'x' belongs to which class. based on the given values weight = 57kgs and height = 170cm.

Weight	height	Class	distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4 (N_1)
57	173	Normal	3 (N_3)
55	170	Normal	2 (N_2)

$d_1 =$
$d_2 =$
$d_3 =$
$d_4 =$
d_5
d_6
d_7
d_8
d_9

$d_1 = \sqrt{(57-51)^2+(170-164)^2} = \sqrt{6^2+3^2} = \sqrt{36+9} = 6.7$

$d_2 = \sqrt{(57-62)^2+(170-182)^2} = 13.$

$d_3 = \sqrt{(57-69)^2+(170-176)^2} = \sqrt{12^2+6^2} = 13.4$

$d_4 = \sqrt{(57-64)^2+(170-173)^2} = \sqrt{7^2+3^2} = \sqrt{49+9} = \sqrt{58} = 7.6$

$d_5 = \sqrt{(57-65)^2+(170-179)^2} = \sqrt{8^2+9^2} = \sqrt{64+81} = \sqrt{145} = 8.2$

$d_6 = \sqrt{(57-56)^2+(170-174)^2} = 4.1$

$d_7 = \sqrt{(57-58)^2+(170-169)^2} = 1.4$

$d_8 = \sqrt{(57-57)^2+(170-173)^2} = 3$

$d_9 = \sqrt{(57-55)^2+(170-170)^2} = 2.$

The given datapoint 'x' belongs to normal class (consider $k=3$)

If $k=1$ (x belongs to normal class) then find the distance between all the points w.r.t 'x' and select a datapoint which is having a least distance value and place the given datapoint in the obtained class.

Remark for k:-

1. k must be a +ve integer.
2. k must not be a multiple of given number of class multiples.

* Find the person 'x' fan following with the given values age = 20 and gender = male. The given dataset is:

name	Age	gender	class/following	distance
A	32	1	X_1	12
B	40	1	X_1	20
C	16	0	X_2	4.123 (N_2)
D	14	0	X_1	6.082
E	55	1	X_2	35
F	40	1	X_1	20
G	20	0	X_2	1 (N_1)
H	15	1	X_2	5
I	55	0	X_1	35.01
J	15	1	X_2	5

Sol: Assume male = 1 and female = 0

$A = \sqrt{(20-32)^2+(1-1)^2} = \sqrt{12^2+0} = \sqrt{144} = 12$

$B = \sqrt{(20-40)^2+(1-1)^2} = \sqrt{20^2+0} = 20$

$C = \sqrt{(20-16)^2+(1-0)^2} = \sqrt{4^2+1^2} = \sqrt{16+1} = \sqrt{17} = 4.123$

$D = \sqrt{(20-14)^2+(1-0)^2} = \sqrt{6^2+1^2} = \sqrt{36+1} = \sqrt{37} = 6.082$

$E = \sqrt{(20-55)^2+(1-1)^2} = \sqrt{35^2+0} = \sqrt{35^2} = 35$

$F = \sqrt{(20-40)^2+(1-1)^2} = \sqrt{20^2+0} = 20$

$G = \sqrt{(20-20)^2+(1-0)^2} = \sqrt{0+1^2} = 1$

$H = \sqrt{(20-15)^2+(1-1)^2} = \sqrt{5^2+0} = 5$

$I = \sqrt{(20-55)^2+(1-0)^2} = \sqrt{35^2+1^2} = \sqrt{35^2+1} = 35.01$

$J = \sqrt{(20-15)^2+(1-1)^2} = \sqrt{5^2+0} = \sqrt{5^2} = 5$

∴ The persons belong to X_2.

Decision Tree:-

Each internal node is labeled with attribute a_i. Each Arc is labelled with a predicate that can be apply to the attribute associated with the parent.

Each leaf node is labelled with a class C_j.

Construct a decision tree for the weather dataset.

→ Select the root node for which attribute is having the highest gain value.

→ Gain of attribute (A) = Information Gain - Entropy(A)

$$\text{Information gain}(P,n) = \frac{-P}{P+n}\log_2\left(\frac{P}{P+n}\right) - \frac{n}{P+n}\log_2\left(\frac{n}{P+n}\right)$$

$$\text{Entropy}(A) = \sum_{i=1}^{r} \frac{P_i+n_i}{P+n} IG(A).$$

Let P = Yes = 9

N = No = 5

$$IG(4,5) = \frac{-9}{9+5}\log_2\left(\frac{9}{9+5}\right) - \frac{5}{9+5}\log_2\left(\frac{5}{9+5}\right)$$

$$= 0.41 + 0.54$$

$$= 0.95$$

Entropy for Outlook:-

	P_i	n_i	$I_G(P_i, n_i)$
Sunny	2	3	0.970
Overcast	4	0	0
Rainy	3	2	0.970

$$IG(2,3) = \frac{-2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= 0.52 + 0.44 = 0.96$$

$IG(4,0) = \frac{-4}{4}\log_2(4/4) - \frac{4}{4}\log_2(4/4)$

$= -1\log_2(1) - \log_2(1)$

$= (-1)0 - 0 = 0$

Entropy(outlook) $= \frac{2+3}{14}(0.970) + 0 + \frac{5}{14}(0.970)$

$= \underline{0.69}$

Gain(outlook) $= IG - E(outlook)$

$= 0.94 - 0.69 = \underline{0.25}$

Entropy for temperature:-

	P_i	n_i	$IG(P_i, n_i)$
Hot	2	2	1
mild	4	2	0.91
cool	3	1	0.81

$IG(2,2) = \frac{-2}{4}\log_2(2/4) - \frac{2}{4}\log_2(2/4)$

$= -1/2\log_2(1/2) - 1/2\log_2(1/2) = 1$

$IG(4,2) = \frac{-4}{6}\log_2(4/6) - \frac{2}{6}\log_2(2/6)$

$= 0.38 + 0.52 = 0.91$

$IG(3,1) = \frac{-3}{4}(\log_2(3/4)) - \frac{1}{4}\log_2(1/4) = 0.81$

Entropy(temperature) $= \frac{2+2}{14}(1) + \frac{4+2}{14}(0.91) + \frac{3+1}{14}(0.81)$

$= 0.90$

Gain(temperature) $= 0.94 - 0.90$

$= 0.04$

Entropy for humidity:-

	Pi	ni	IG(Pi, ni)
High	3	4	0.98
Normal	6	1	0.59

$IG(3,4) = \frac{-3}{7}\log_2(3/7) - \frac{4}{7}\log_2(4/7) = 0.98$

$IG(6,1) = \frac{-6}{7}\log_2(6/7) - \frac{1}{7}\log_2(1/7) = 0.59.$

Entropy (humidity) = 0.78.

Gain (Humidity) = 0.94 − 0.78 = 0.16.

Entropy for windy:-

	Pi	ni	IG(Pi, ni)
weak	6	2	0.311
Strong	3	3	1

$IG(6,2) = \frac{-6}{8}\log_2(6/8) - \frac{2}{8}\log_2(2/8)$

$= 0.311$

Entropy (windy) $= \frac{6+2}{14}(0.31) + \frac{3+3}{14}(1) = \frac{4}{7}(0.31) + \frac{3}{7}(1) =$

Gain (windy) = 0.048.

Decision Tree Construction:-

Select the rootnode for which is having the highest Gain.

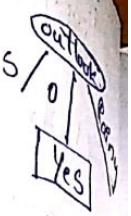
Gain (outlook) = 0.25 (✓) (Rootnode)

Gain (Temperature) = 0.04

Gain (Humidity) = 0.16

Gain (windy) = 0.048

Outlook	temperature	humidity	windy	class
Sunny	hot	high	weak	No
Sunny	hot	high	Strong	No
Sunny	mild	high	weak	No
Sunny	cool	~~high~~ nominal	weak	Yes
Sunny	mild	nominal	Strong	Yes.



Sunny:-

P = Yes = 2

N = No = 3.

IG(3,2) = 0.970.

Entropy for temperature:-

	Pi	ni	IG(Pi, ni)
hot	0	2	0
mild	1	1	1
Cool	1	0	0

Entropy(temperature) = $\frac{1+1}{2+3}(1) = 2/5 = 0.4$

Gain(temperature) = IG - Entropy(temperature)

= 0.970 - 0.4 = 0.57

Entropy for humidity:-

	Pi	ni	IG(Pi, ni)
high	0	3	0
nominal	2	0	0

Entropy(humidity) = 0

Gain(humidity) = IG - Entropy(humidity)

= 0.970 - 0 = 0.970

Entropy for windy:-

	P_i	n_i	$IG(P_i, n_i)$
weak	1	2	0.91
strong	1	1	1

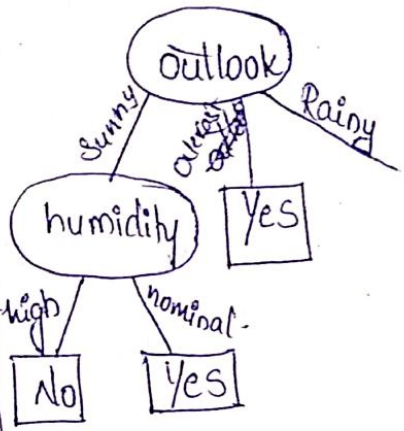
IG.

Entropy $(1,2) = \frac{-1}{3}(\log_2(1/3)) - 2/3 \log_2(2/3) = 0.91$

Entropy(windy) $= \frac{3}{5}(0.91) + \frac{2}{5}(1) = 0.94$

Gain(windy) $=$ IG $-$ Entropy(windy)

$= 0.970 - 0.94 = 0.03$



Dataset for Rainy:-

Outlook	temperature	humidity	windy	class
rainy	mild	high	weak	yes
rainy	cool	nominal	weak	Yes
rainy	cool	nominal	strong	No
rainy	mild	nominal	weak	Yes
rainy	mild	high	strong	No

P=3, N=2 IG(3,2) = 0.970.

Entropy for temperature:-

	P_i	n_i	$IG(P_i, n_i)$
mild	2	1	0.91
cool	2	1	1

Entropy(temperature) $= \frac{3}{5}(0.91) + 3/5\,(1) = 0.94$

Gain(temperature) $= 0.970 - 0.94 = 0.03$

Entropy for humidity:-

	P_i	n_i	$IG(P_i, n_i)$
high	1	1	1
nominal	2	1	0.91

E(humidity) = 0.94

G(h) = 0.97 − 0.94

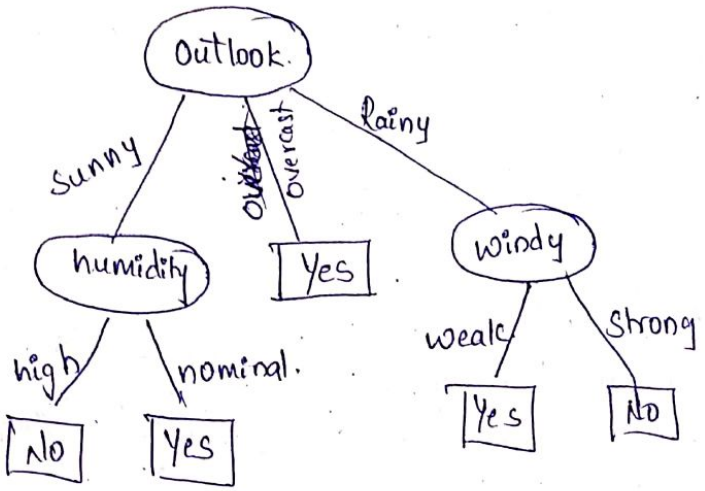
= 0.03.

Entropy for windy:-

	P_i	n_i	$IG(P_i, n_i)$
Weak	3	0	0
Strong	0	2	0

Enhopy(windy) = 0

G(windy) = 0.97 - 0
= 0.97



∴ Decision Tree for weather dataset

26/03/19

*Construct a decision tree with the attribute gender and height and class label

Gender	Height	Class
F	1.6m	Short
M	2m	Tall
F	1.9m	Medium
F	1.88m	Medium
F	1.7m	Short
M	1.85m	Medium
F	1.6m	Short
M	1.7m	Short
M	2.2m	Tall
M	2.1m	Tall
F	1.8m	Medium
M	1.8m	Medium
F	1.95m	Medium
F	1.9m	Medium
F	1.8m	Medium

Short = 4
Medium = 8
Tall = 3.

$$IG(S,M,T) = \frac{-S}{S+M+T}\log_2\left(\frac{S}{S+M+T}\right) - \frac{M}{S+M+T}\log_2\left(\frac{M}{S+M+T}\right) - \frac{T\log_2\left(\frac{T}{S+M+T}\right)}{S+M+T}$$

$$IG(S,M,T) = \frac{-4}{15}\log_2\left(\frac{4}{15}\right) - \frac{8}{15}\log_2\left(\frac{8}{15}\right) - \frac{3}{15}\log_2\left(\frac{3}{15}\right)$$

$$= 1.456.$$

Entropy for Gender:-

	S	M	T	IG(S,M,T)
Female	3	6	0	0.918
Male	1	2	3	1.459.

$$IG(3,6,0) = \frac{-3}{9}\log_2\left(\frac{3}{9}\right) - \frac{6}{9}\log_2\left(\frac{6}{9}\right) - 0$$

$$= 0.918$$

$$IG(1,2,3) = \frac{-1}{6}\log_2\left(\frac{1}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= 1.459.$$

$$\text{Entropy(Gender)} = \sum_{i=1}^{F} \frac{Si+mi+ti}{S+m+t} IG(M) + \frac{Si+mi+ti}{S+m+t} IG(F)$$

$$= \frac{1+2+3}{4+8+3}(1.459) + \frac{3+6+0}{15}(0.918)$$

$$= \frac{6}{15}(1.459) + \frac{9}{15}(0.918) = 1.1344 \ (1.12)$$

$$\text{Gain(Gender)} = IG - \text{Entropy(Gender)}$$

$$= 1.456 - 1.12$$

$$= \underline{0.32}$$

Entropy for Height:-

	S	M	T	IG(S,M,T)
≤1.7	3	0	0	0
>1.7-<2	0	6	0	0
>2	0	0	4	0

$IG(3,0,0) = \frac{-3}{3} \log_2(3/3) = 0$

$IG(0,6,0) = \frac{-6}{6} \log_2(6/6) = 0$

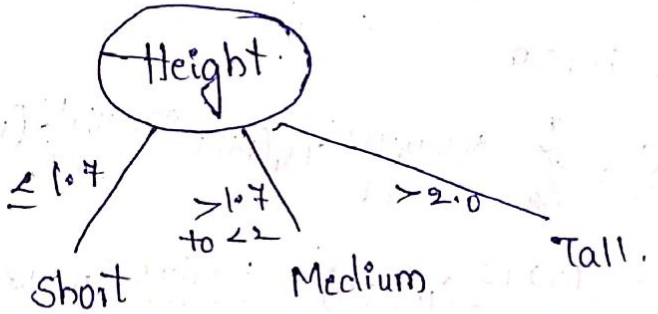
$IG(0,0,4) = -4/4 \log_2(4/4) = 0$

$\text{entropy(height)} = \frac{3}{15} IG(\leq 1.7) + \frac{8}{15} IG(>1.7-<2) + \frac{2}{15} IG(>2)$

$= \frac{3}{15}(0) + \frac{8}{15}(0) + \frac{2}{15}(0) = 0.$

$\text{Gain(height)} = IG - \text{entropy(height)}$

$= 1.456 - 0 = 1.456.$



Pruning:-

Once a tree is constructed some modifications to the tree might be required to improve the perform of the key tree during the classification.

The proning phase remove the redundant Comparisions (or) remove subtrees to acheive better Performance.

27/03/18

Neural Networks:-

A Neural Network is a set of input and output nodes, and it is connected by giving some weights for the edges.

No. of source Nodes:-

Generally, the attribute names can be used as the source nodes (or) input nodes.

Determining which attributes to use as inputs is an issue.

No. of Hidden layers:-

In the simplest case, there is only one hidden layer.

The layer which is in between Sourcenode (Input nodes) and Destination nodes (Output nodes) is called Hidden layer.

No. of Hidden Nodes:-

The nodes available in the hidden layer can be called as Hidden nodes.

Choosing the best no. of hidden nodes for hidden layer is one of the most difficult problem in neural networks.

At most (2n+1) hidden nodes can be used, where 'n' is no. of input nodes.

No. of Sinks (or) Destination nodes:-

Usually, the no. of output nodes is the same as the no. of classes.

Inter Connections:-

Each node is connected to all the nodes in the next level.

Wait:-

The wait assigned to an edge indicates the relative wait between the two nodes.

Activation functions:-

Many different types of activation functions can be used (Sigmoid, relu).

28/03/19.

Running Technique:-

The Technique for adjusting the weights is called Running Technique.

The most commonly used approach is back propagation.

The running may stop when all the training tuples has propagated through the network. May be based on time (or) may be based on error rate.

Advantages of Neural Networks:-

1. Neural Networks are more robust than the decision trees because of the weights.
2. NN improves the ~~network~~ Performance by learning.
3. NN peforms/contains no error rate and have more accuracy.

Neural Networks are more robust than the decision tree(or) in noisy environment.

Disadvantages:-

1. ANN are difficult to understand
2. Generating rules from the ANN is not a straight forward.

Input attribute values must be numeric.

- Neural Network may be quite expensive to use.

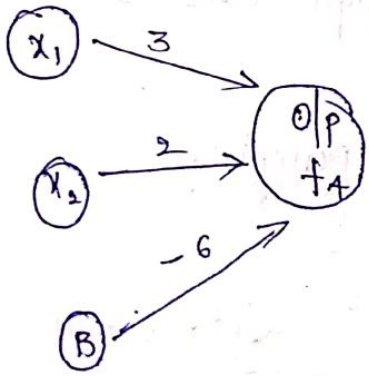
Perceptron:-

The simplest neural network is called as a perception.

A perceptron is a single neuron with multiple inputs and output.

A simple perception can be used for binary classificati Using a unipolar activation function a binary classification can be performed.

Example for A Simple Perception:-



$S = 3x_1 + 2x_2 - 6$.

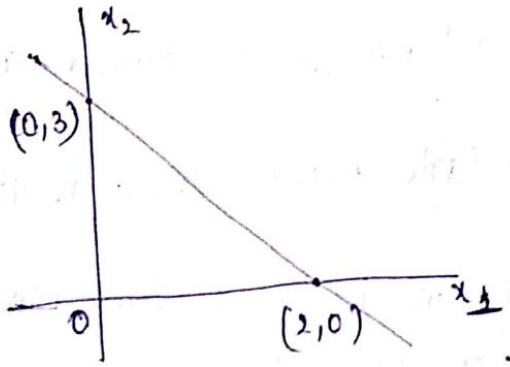
$f_4 = \begin{cases} 1 & S > 0 \\ 0 & S < 0 \end{cases}$

1 - yes
0 - No.

x_1	x_2	Class	S	predicted (f_4)
0	1	Yes	-4	No (FN)
2	4	No	8	Yes (FP)
3	6	No	15	Yes (FP)
6	7	Yes	26	Yes (TP)
2	0	No	0	Yes. (FP)

Accuracy = $\frac{1}{5}$ = 20%.

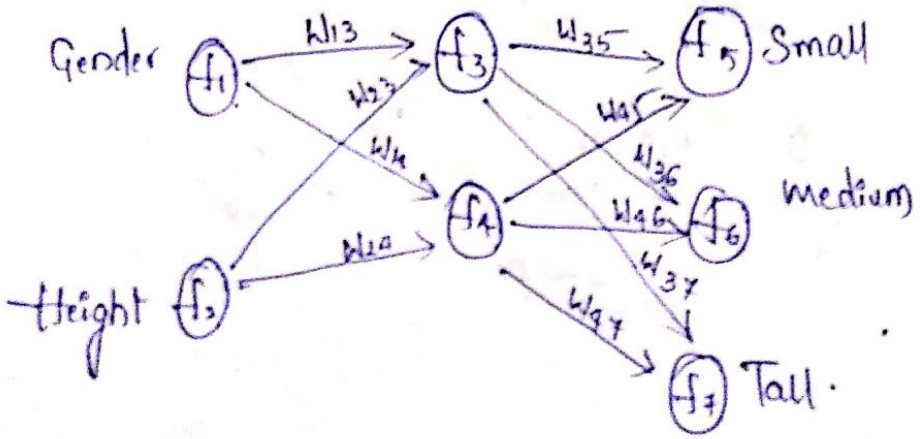
Alternative view:-



classification perception.

If only one layer exist (output) is called perception. If multiple layer exits is called multiple layer perception.

Example for Multi-layer perception.



Back Propagation:-

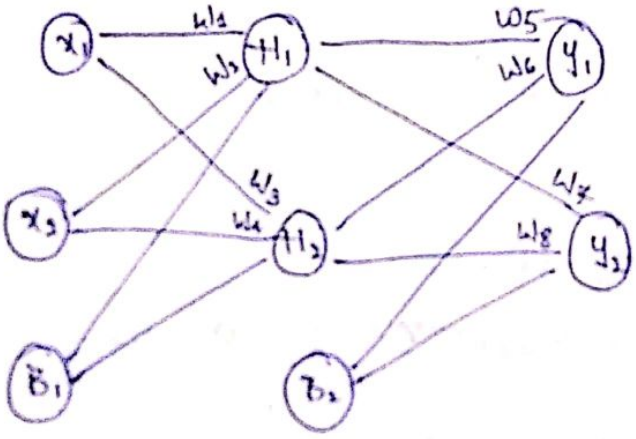
29/03/19.

Back Propagation learns by iterating processing a dataset of training tuples, Comparing the network prediction for each tuple with the actual Node target value.

For each tuple, the weights are modified so as to minimize the meansquared error between the networks prediction and the actual target value.

These modifications are made in the backward direction (from the outputlayer) through each hidden layer down to the first hidden layer. Hence the name is Back Propagation.

* Find the output values for the output nodes and find the error between the predicted and the actual for a given network.



	Target	
	T_1	T_2
	0.01	0.99
$x_1 = 0.05$	$w_1 = 0.15$	$w_5 = 0.40$
$x_2 = 0.10$	$w_2 = 0.20$	$w_6 = 0.45$
$B_1 = 0.35$	$w_3 = 0.25$	$w_7 = 0.50$
$B_2 = 0.60$	$w_4 = 0.30$	$w_8 = 0.55$

$H_1 = x_1 w_1 + x_2 w_2 + B_1$

$= (0.05 \times 0.15) + (0.10 \times 0.20) + 0.35$

$H_1 = 0.375.$

Activation function used is sigmoidal function is

$$y = \frac{1}{1+e^{-x}}.$$

Output of $H_1 = \frac{1}{1+e^{-H_1}} = \frac{1}{1+e^{-0.375}} = 0.592.$

$H_2 = x_1 w_3 + x_2 w_4 + B_1$

$H_2 = 0.05 \times 0.25 + 0.10 \times 0.30 + 0.35 = 0.3925.$

Output of $H_2 = \frac{1}{1+e^{-0.3925}} = 0.5967.$

For y_1:-

$y = H_1 w_5 + H_2 w_6 + B_2$

$y = 0.592 \times 0.40 + 0.596 \times 0.45 + 0.60 = 0.925.$

Output of $y_1 = \frac{1}{1+e^{-y_1}} = \frac{1}{1+e^{-0.925}} = 0.71.$

For y_2:-

$y_2 = H_1 w_7 + H_2 w_8 + B_2$

$y_2 = 0.592 \times 0.50 + 0.5967 \times 0.55 + 0.60 = 1.224.$

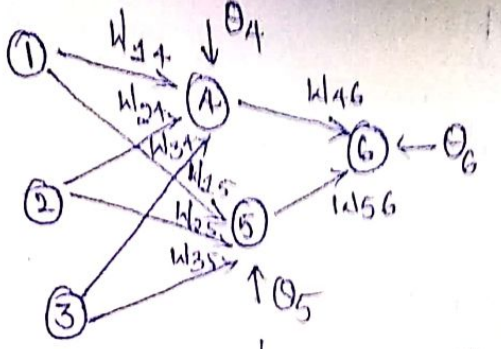
Output of $y_2 = \frac{1}{1+e^{-y_2}} = \frac{1}{1+e^{-1.224}} = 0.772.$

Error prediction is $\sum_{i=1}^{m} \frac{(\text{Target} - \text{predicted})^2}{m}.$

error $= \frac{1}{2}(T_1 - \text{output } y_1)^2 + \frac{1}{2}(T_2 - \text{output } y_2)^2$

$= \frac{1}{2}(0.01 - 0.73)^2 + \frac{1}{2}(0.99 - 0.775)^2 = 0.297$

For a given neural network. Find the output for the hidden nodes & output nodes. Find the errors of hidden nodes and output nodes and change the weights for the edges. For the bias values.



			Bias
$x_1 = 1$	$w_{14} = 0.2$	$w_{34} = -0.5$	$\theta_4 = -0.4$
$x_2 = 0$	$w_{15} = 0.3$	$w_{35} = 0.2$	$\theta_5 = 0.2$
$x_3 = 1$	$w_{24} = 0.4$	$w_{46} = -0.3$	$\theta_6 = 0.1$
	$w_{25} = 0.1$	$w_{56} = -0.2$	

Target value = 1 learning value = 0.9.

At node4 :-

$$\text{node } 4 = x_1 w_{14} + x_2 w_{24} + x_3 w_{34} + \theta_4$$

$$= 1 \times 0.2 + (0)(0.4) + (1)(-0.5) - 0.4$$

$$= 0.2 + 0 - 0.5 - 0.4$$

$$= 0.2 - 0.9 = -0.7$$

$$O_4 = \frac{1}{1 + e^{-(-0.7)}} = \frac{1}{1 + e^{0.7}} = 0.33$$

At node 5 :-

$$\text{node } 5 = x_1 w_{15} + x_2 w_{25} + x_3 w_{35} + \theta_5$$

$$= -0.3 + 0 + 0.2 + 0.2 = 0.1$$

$$O_5 = \frac{1}{1 + e^{-0.1}} = 0.52.$$

At node 6:-

Input to node 6 = $\theta_6 + I_{4,6} W_{4,6} + I_{5,6} W_{5,6}$

$= (1 \times 0.1) + (0.332 \times -0.3) + (0.525 \times 0.2) + 0.1$

$= 0.105$

$O_6 = \frac{1}{1 + e^{-0.105}} = 0.476$

The output of node 6 is not equal to the target value. Hence there is a need to adjust the weights as well as bias values.

To calculate the error at the output node is,

$Err_j = O_j(1 - O_j)(T - O_j)$

Error for hidden nodes:

$Err_j = O_j(1 - O_j) \sum Err_k w_{kj}$ where

w_{kj} is the weight of the connection from the node j to node k.

Error at output node 6:-

$Err_6 = O_6(1 - O_6)(T - O_6)$

$Err_6 = (0.476)(1 - 0.476)(1 - 0.476)$

$\boxed{Err_6 = 0.13}$

Error at node 5:-

$Err_5 = O_5(1 - O_5)(Err_6 w_{56})$

$= (0.52)(1 - 0.52)(0.13)(-0.2)$

$\boxed{Err_5 = -0.006}$

Error at node 4:-

$Err_4 = O_4(1-O_4)\, Err_6\, w_{46}$

$= (0.33)(1-0.33)(0.13)(-0.3)$

$\boxed{Err_4 = -0.008}$

Formulae for adjusting the weights is

$w_{ij} = w_{ij} + (l)\, Err_j \cdot O_i$

Formula for adjusting the bias values is

$\theta_j = \theta_j + (l)\, Err_j$

l → learning rate.

$w_{ij} = w_{ij} + (l)\, Err_j \cdot O_i$

$w_{46} = w_{46} + (0.9)\, Err_6 \cdot O_4$

$= -0.3 + (0.9)(0.13)(0.33)$

$w_{46} = -0.26$

$w_{56} = w_{56} + (0.9)\, Err_6 \cdot O_5$

$= -0.2 + (0.9)(-0.008)(0.13)(0.52)$

$= -0.13$.

$w_{14} = w_{14} + (l)\, Err_4 \cdot O_1$

$= 0.2 + (0.9)(-0.008) \cdot O_1$

$= 0.19$

$w_{24} = -0.405$

$w_{15} = -0.306$

$w_{25} = 0.1$

$w_{34} = -0.508$

$w_{35} = 0.194$

$\theta_4 = \theta_4 + (l)\, Err_4$

$= (-0.4) + (0.9)(-0.008)$

$= -0.408$.

$\theta_5 = \theta_5 + (l)Err_5$

$= 0.2 + (0.9)(-0.006)$

$= 0.1946$

$\theta_6 = \theta_6 + (l)Err_6$

$= 0.1 + (0.9)(0.13)$

$= 0.217$

At Node4:-

$node4 = x_1 \times w_{14} + x_2 \times w_{24} + x_3 w_{34} + \theta_4$

$= 1 \times 0.19 + 0 + 1 \times (0.508) + (-0.408)$

$= -0.726$

$O_4 = \frac{1}{1 + e^{(-0.726)}} = 0.32.$

At node5:-

$node5 = x_1 w_{15} + x_2 w_{25} + x_3 w_{35} + \theta_5$

$= 1 \times (-0.306) + 0 + 1 \times 0.194 + 0.2$

$= 0.079.$

$O_5 = \frac{1}{1 + e^{0.079}} = 0.51.$

Rule Based classification:- 3/04/19

Rule based classification is a set of "if-then" rules.

Syntax:- IF(condition) THEN Conclusion.

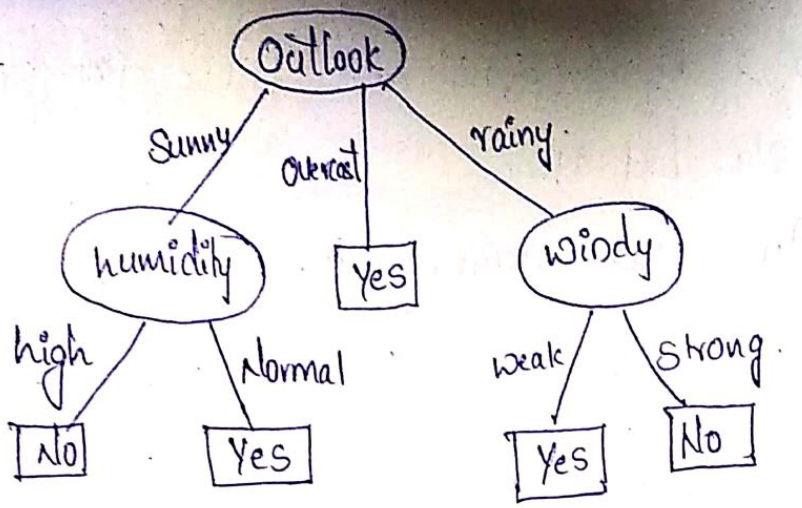
The 'IF' part is called as rule antecedent and 'THEN' part is called as rule Consequent.

Ex:- IF(outlook=Sunny) AND(Humidity=high) then play=no

Extracting IF-THEN Rule from a DT:-

For a given DT the no of if then rules can be derived based on the no of leaf nodes.

Ex:-



Rule 1:- IF (Outlook = Sunny) and (humidity = high) THEN (play = No)

Rule 2:- IF (outlook = Sunny) and (humidity = Normal) THEN (play = yes)

Rule 3:- IF (outlook = Overcast) THEN (Play = Yes)

Rule 4:- IF (outlook = rainy) and (windy = weak) THEN (play = yes)

Rule 5:- IF (outlook = rainy) and (windy = strong) THEN (play = No)

IF-THEN rules can be assessed by 1. Coverage and Accuracy. (parameters of IF-THEN)

Let 'n' Covers be the no of tuples Covered by rule 'r' for a given data set

Let 'n' Correct be the no of tuples Correctly classified by the rule 'r'.

Let |D| be the size of the dataset.

1. Coverage = $\frac{n_{covers}}{|D|}$

2. Accuracy = $\frac{n_{correct}}{n_{covers}}$.

From the above rules, rule '1' covers 3 tuples (1, 2, 3) and those three tuples are correctly classified.

Hence Coverage = 3/14 and Accuracy = 3/3 = 1 = 100%.

* rule 2 covers 2 tuples (9, 11) and those two tuples are correctly classified. Hence

Coverage = 2/14 and Accuracy = 2/2 = 1 = 100%.

Extracting IF-THEN rules from a neural network:-

Rule Extraction Algorithm:-

Input:-

D // Training data

N // Initial neural network.

Output:-

R // Derived rules.

RX Algorithm:-

Cluster output node activation values;

Cluster hidden node activation values;

generate rules that describe the output values in terms of the hidden activation values;

generate rules that describe hidden output values in terms of Inputs;

Combine two set of rules;

Combining Techniques:- 4/03/19.

Given a classification problem, one classification Technique does not yields best results.

There have been some proposals for combining Techniques.

There are two approaches for Combining Technique

1. Synthesis Approach
2. Multipleindependent Approach.

Synthesis Approach:-
It takes multiple Techniques and combine them into new approach.

Clustering

Methods:-

1. Partitioning (k means)
2. Hierarchial (Agglome ratio divisor) approaches.
3. Density based (DBSCAN)
4. Grid based.

K-means clustering:-

Let $k=2$

2, 4, 6, 7, 18, 20, 23, 30, 26, 35

Select 2 elements from the given

Let $m_1 = 7$ $m_2 = 30$

i:

$K_1 = \{2,4,6,7,18\}$ $K_2 = \{20,23,30,26,35\}$

$m_1 = 2+4+6+7+18$ $m_2 = 20+23+30+26+35$

$= 7.4 \cong 8$ $m_2 = 26.8 \cong 27$

ii:

$K_1 = \{2,4,6,7\}$ $K_2 = \{18,20,25,30,26,35\}$

$m_1 = 2+4+6+7$ $m_2 = 25.3 (26)$

$m_1 = 4.77 (5)$

$K_1 = \{2,4,6,7\}$ $K_2 = \{18,20,25,30,26,35\}$

$m_1 = 5$ $m_2 = 26.$

Here 2 previous clusters are same. So we can stop it.

K-mean Algorithm:-

I/p:-

D // Dataset

K // no of clusters

O/p:- A set of K clusters.

Method:-
Arbitrarily choose k objects from D.
repeat
Assign each object to the cluster to which the object is most similar based on the mean value.
Update the cluster means.
until no change.

K-Medoid:- This is also known as PAM (Partitioning Around Models)

Ex:- 1, 2, 3, 8, 9, 10, 25. Perform the k-medoid algorithm for given dataset. The given pairs are:

(2, 6), (3, 4), (3, 8), (4, 7), (6, 2), (6, 4), (7, 4), (7, 3), (8, 5), (7, 6)

Sol:- Let k = 2

distance (3, 4) distance from (7, 4)

If (x_1, y_1) & (x_2, y_2)

Manhattan distance = $|x_1 - x_2| + |y_1 - y_2|$

$\|2-3\|+\|6-4\| = 3$	$\|2-7\|+\|6-4\| = 7$
$\|3-3\|+\|8-4\| = 4$	$\|3-7\|+\|8-4\| = 8$
$\|4-3\|+\|7-4\| = 4$	$\|4-7\|+\|7-4\| = 6$
$\|6-3\|+\|2-4\| = 5$	$\|6-7\|+\|2-4\| = 3$
$\|6-4\|+\|4-4\| = 2$	$\|6-7\|+\|4-4\| = 1$
$\|7-3\|+\|3-4\| = 5$	$\|7-7\|+\|3-4\| = 1$
$\|8-3\|+\|5-4\| = 6$	$\|8-7\|+\|5-4\| = 2$
$\|7-3\|+\|6-4\| = 6$	$\|7-7\|+\|6-4\| = 2$

Total cost = 3+4+4+3+1+1+2+2 = 20.

Cluster 1: {(3, 4), (2, 6), (3, 8), (4, 7)}

Cluster 2: {(7, 4), (6, 2), (6, 4), (7, 3), (8, 5), (7, 6)}

select the next 2 points (3,4) & (7,3)

	distance (3,4)	distance(7,3)
(2,6)	$\|2-3\|+\|6-4\|=3$	$\|2-7\|+\|6-3\|=8$
(3,4)		
(3,8)	$\|3-3\|+\|8-4\|=4$	$\|3-7\|+\|8-3\|=9$
(4,7)	$\|4-3\|+\|7-4\|=4$	$\|4-7\|+\|7-3\|=7$
(6,2)	$\|6-3\|+\|2-4\|=5$	$\|6-7\|+\|2-3\|=2$
(6,4)	$\|6-4\|+\|4-4\|=2$	$\|6-7\|+\|4-3\|=2$
(7,3)	$\|7-3\|+\|3-4\|=5$	$\|7-7\|+\|4-3\|=1$
(8,5)	$\|8-3\|+\|5-4\|=6$	$\|8-7\|+\|5-3\|=3$
(7,6)	$\|7-3\|+\|6-4\|=6$	$\|7-7\|+\|6-3\|=3$

Total cost = 3+4+4+2+2+1+3+3 = 22.

22 > 20

(3,4) (7,3) > (3,4) (7,4)

∴ select the previous clusters as answers

∴ Answer is

Cluster for (3,4) = {(3,4), (2,6), (3,8), (4,7)}

Cluster for (7,4) = {(7,4), (6,2), (6,4), (7,3), (8,5), (7,6)}

k-medoid is more robust than k-mean in the presence of noise and outlier because a medoid is less influ by outliers (or) other extreme values than A.

The Complexity of each iteration in k-mediod algorithm is $O(k(n-k)^2)$

where n is no.of records (or) points (or) size of data points.

k = no of k mediod / no.of clusters.

A typical k medoid (PAM) works well for small dataset but doesnot scale well for large datasets.

To deal with large dataset sampling based methods called CLARA (clustering large applications) can be used

CLARA:

Instead of taking whole dataset into consideration. CLARA uses a random sample of dataset. The PAM algorithm is then applied to compute best medoid for the sample.

→ CLARA builds clustering from the multiple random samples and returns the best clustering as the output.

The Complexity of computing the medoid on a random sample is $O(ks^2+k(n-k))$

where s = size of sample
k = no. of clusters.
n = size of the total objects.

Effectiveness of CLARA depends upon the samplesize

→ The time Complexity of k-means to find the clusters is $O(n)$.

k-mediod - $O(n^2)$

k-medoid Algorithm:-

I/p:-
k = no. of clusters
D = dataset.

O/p:-
A set of k clusters.

Method:-

Orbitrarily choose k objects in D as the initial representative objects

Repeat

Assign each remaining object to the clusters with the nearest rep object.

Randomly select a non-representative object Orandom

Compute the total cost, S, of swapping rep object Oj with Orandom.

if S<0 then swap Oj with Orandom to form the new set of k rep objects.

Until no change.

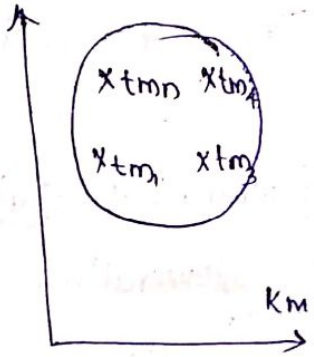
Similarity and distance measure:-

Given a cluster km of n points $\{t_{m_1}, t_{m_2}, \ldots t_{m_n}\}$

$$\text{Centroid} = C_m = \sum_{i=1}^{N} \frac{t_{mi}}{N}$$

$$\text{radius} = R_m = \sqrt{\frac{\sum_{i=1}^{N} (t_{mi} - c_m)^2}{N}}$$

$$\text{Diameter } D_m = \sqrt{\frac{\sum_{j=1}^{N} \sum_{j=1}^{N} (t_{mi} - t_{mj})^2}{N(N-1)}}$$

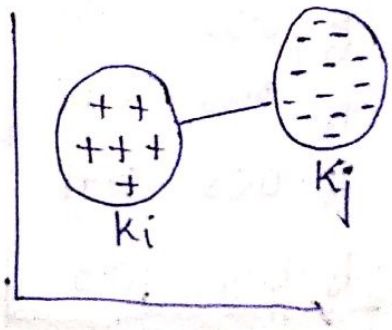


Centroid is the middle of the clusters.

Given clusters k_i & k_j there are several standard alternatives to calculate the distance between clusters.

→ The representative list is

1. Single link
2. Complete link
3. Average link



Single link:-

The smallest distance between an element in one cluster and an element in other cluster.

Thus, the distance of $(k_i, k_j) = \min(\text{distance}(t_{i1}, t_{jm}))$

$\forall t_{i1} \in k_i \notin k_j$ &

$\forall t_{jm} \in k_j \notin k_i$

Complete link:-

The largest distance between an element in one cluster and an element in other.

Thus, the distance of $(k_i, k_j) = \max(\text{distance}(t_{i1}, t_{jm}))$

Average link:-

An average distance between in one cluster an element in other cluster.

Thus, the distance $(k_i, k_j) = \text{mean}(\text{distance}(t_{i1}, t_{jm}))$

16/04/19

Hierarchical Clustering:-

1. Agglomerative.
2. Divisive.

Perform Agglomerative clustering using Single link

	x	y
P_1	0.40	0.53
P_2	0.22	0.38
P_3	0.35	0.32
P_4	0.26	0.19
P_5	0.08	0.41
P_6	0.45	0.3

Step 1:- Calculate the distance matrix using Euclidean distance.

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	0.23	0				
P_3	0.21	0.14	0			
P_4	0.36	0.20	0.15	0		
P_5	0.34	0.14	0.28	0.29	0	
P_6	0.23	0.25	0.11	0.22	0.39	0

P_1:-

$(0.22, 0.38), (0.40, 0.53)$

$D = \sqrt{(0.40-0.22)^2+(0.53-0.38)^2} = 0.23$

:-

$0.35, 0.32), (0.40, 0.53)$

$D = \sqrt{(0.40-0.35)^2+(0.53-0.32)^2} = 0.21$

$.35, 0.32), (0.22, 0.38)$

$= \sqrt{(0.22-0.35)^2+(0.38-0.32)^2} = 0.14$

$0.26, 0.19), (0.40, 0.53)$

$= \sqrt{(0.40-0.26)^2+(0.53-0.19)^2} = 0.36$

$.26, 0.19), (0.22, 0.38)$

$\sqrt{(0.22-0.26)^2+(0.38-0.19)^2} = 0.20$

P_1P_3:-

$(0.26, 0.19), (0.35, 0.32)$

$ED = \sqrt{(0.35-0.26)^2 + (0.32-0.19)^2} = 0.15$

P_5P_1:- $(0.08, 0.41), (0.40, 0.53)$

$ED = \sqrt{(0.40-0.08)^2 + (0.53-0.41)^2} = 0.34$

P_5P_2:- $(0.08, 0.41), (0.22, 0.38)$

$ED = \sqrt{(0.22-0.08)^2 + (0.38-0.41)^2} = 0.14$

P_5P_3:- $(0.08, 0.41), (0.35, 0.32)$

$ED = \sqrt{(0.35-0.08)^2 + (0.32-0.41)^2} = 0.28$

P_5P_4:- $(0.08, 0.41), (0.26, 0.19)$

$ED = \sqrt{(0.26-0.08)^2 + (0.19-0.41)^2} = 0.29$

P_6P_1:- $(0.45, 0.3), (0.40, 0.53)$

$ED = \sqrt{(0.40-0.45)^2 + (0.53-0.3)^2} = 0.23$

P_6P_2:- $(0.45, 0.3), (0.22, 0.38)$

$ED = \sqrt{(0.22-0.45)^2 + (0.38-0.3)^2} = 0.25$

P_6P_3:- $(0.45, 0.3), (0.35, 0.32)$

$ED = \sqrt{(0.35-0.45)^2 + (0.32-0.3)^2} = 0.11$

P_6P_4:- $(0.45, 0.3), (0.26, 0.19)$

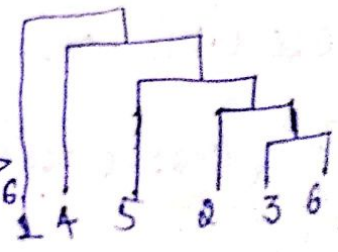
$ED = \sqrt{(0.26-0.45)^2 + (0.19-0.3)^2} = 0.22$

P_6P_5:- $(0.45, 0.3), (0.08, 0.41)$

$ED = \sqrt{(0.08-0.45)^2 + (0.41-0.3)^2} = 0.39$.

Step2:- Select the point in which is having the minimum distance.

∴ Min distance point is P_3P_6

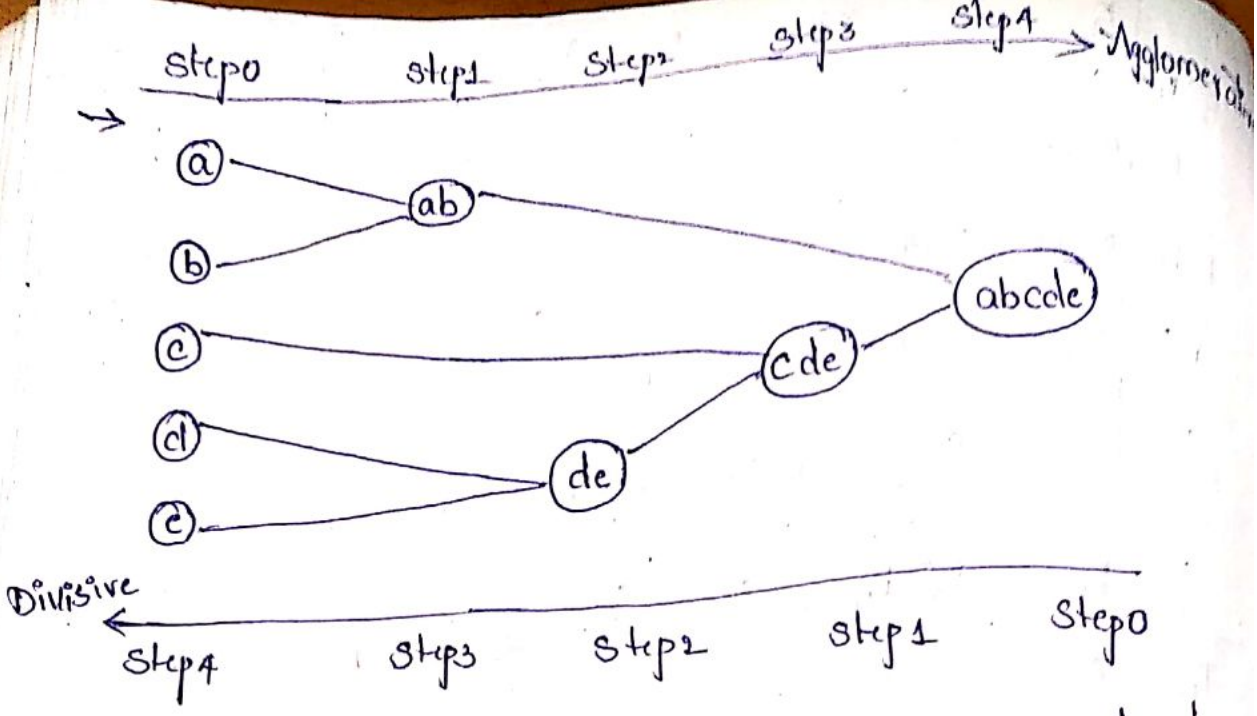


	P_1	P_2	P_3P_6	P_4	P_5
P_1	0				
P_2	0.23	0			
P_3P_6	0.21	0.14	0		
P_4	0.36	0.20	0.15	0	
P_5	0.34	0.14	0.28	0.29	0

	P_1	$P_3P_6P_4$	P_4	P_5
P_1	0			
$P_3P_6P_4$	0.21	0		
P_4	0.36	0.45	0	
P_5	0.34	0.14	0.29	0

	P_1	$P_3P_6P_2P_5$	P_4
P_1	0		
$P_3P_6P_2P_5$	0.21	0	
P_4	0.36	0.15	0

	P_1	$P_3P_6P_2P_5P_4$
P_1	0	
$P_3P_6P_2P_5P_4$	0.21	0



20/04/19

Clustering with large Databases:-

BIRCH → Balanced

Clustering With large Databases:

→ BIRCH – Balance Iterative Reducing & clustering usin Hierarchies.

→ Combination of Hierarchial & Partitioning Algorithms.

→ 2 Difficulties in Agglomerative Overcomes

1.) Scalability

2.) Undo Operation (for previous step).

→ applicable for numeric data.

2 notions:

(1) CF (clustering feature) ——— (n, LS, SS) — 3 tuples

(2) CF Tree (Height Balanced Tree)

n → no. of points.

LS → Linear sum.

SS → Squared sum.

$CF_1 = (n_1, LS_1, SS_1)$

$CF_2 = (n_2, LS_2, SS_2)$

$x = \{x_1, x_2, x_3, \ldots x_n\}$

$CF_1 + CF_2 =$

$LS = \sum_{i=1}^{n} x_i$

$(n_1+n_2, LS_1+LS_2, SS_1+SS_2)$. $SS = \sum_{i=1}^{n} x_i^2$

When cluster C_1 contains the elements

Suppose there are 3 points

(2,5) (3,2) (4,3) in cluster C_1

(4,6) (6,7) (7,8) in cluster C_2 ..

$LS_1 = (2+3+4, 5+2+3)$

$= (9, 10)$

$SS_1 = (2^2+3^2+4^2, 5^2+2^2+3^2)$

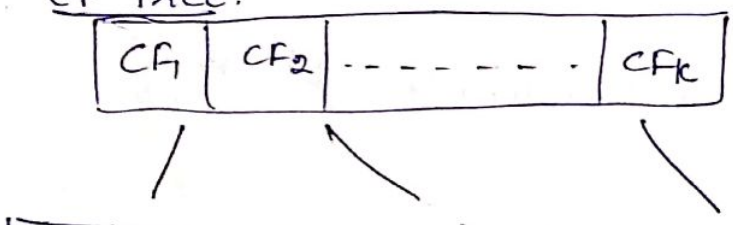
$= (29, 38)$

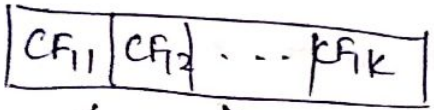
$CF_1 = (3, (9,10), (29,38))$

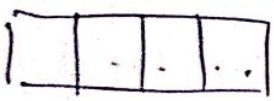
$CF_2 = (3, (17,21), (101,149))$

$CF_1 + CF_2 = (6, (26,31), (130,187))$

CF Tree:







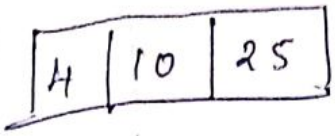
x	y	z	t	y	L	S	T	x
		13	7	4				
					3	8		
		2	10	10	10			
						3		
					3.	3		
					16.	15		

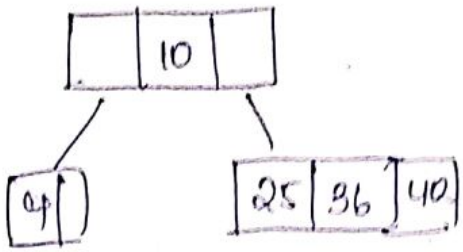
Time (1) Lyrics (2) Voice (3) (sink) 20M

It is a balanced Tree and It is also a B-Tree.

4, 10, 25, 36, 40

$n = 3$





(2) DBSCAN: Density based Spatial Clustering Applications with NOISE.

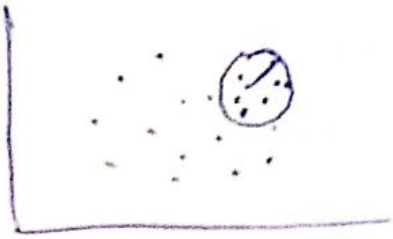
- Used for Spherical shape of clusters for finding by Hierarchial & Partitioning algorithms.
- S-shaped & Oval shaped clusters by using DBSCAN.

3 Imp points

- Density
- Core point
- Boundary Point

ϵ (Epsilon) > 0 $\Rightarrow$ +ve Integer $\epsilon = 2$ (Radius).

with Radius 2 we need to derive a circle by selecting a random point.



Density :- no of points covered within a radius of a circle is called Density.

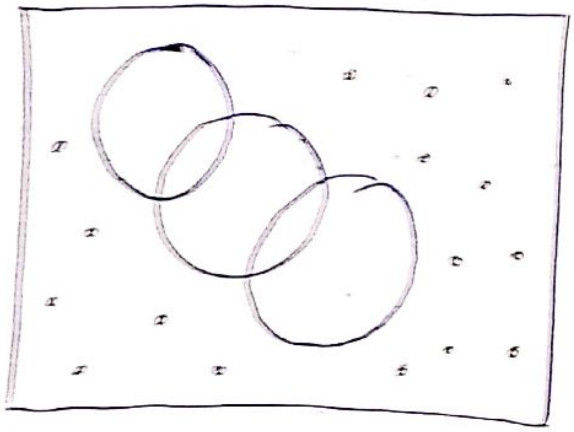
$\therefore$ Density = 7.

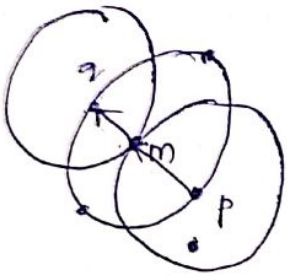
Core point :

Boundary point: Points that lie on the Boundary of a circle is called Boundary point.

Core point:

From the given point, if we can traverse any of the 3 points in that Circle. Given min points = 3.





m → core point

If min point=3; then the density should be = (0,2) > its density than the

p → core point

m → core point

q → core point (iff and only if it traverses from p → q.

From the above diagram p & m are directly core point, q is also a core point bcoz, q can be reached from p → m & m → q. Hence 'q' is a core point.

DBSCAN finds core objects i.e; the objects that have dense neighbourhoods, it connects core objects & their neighbourhoods to form dense regions as clusters.

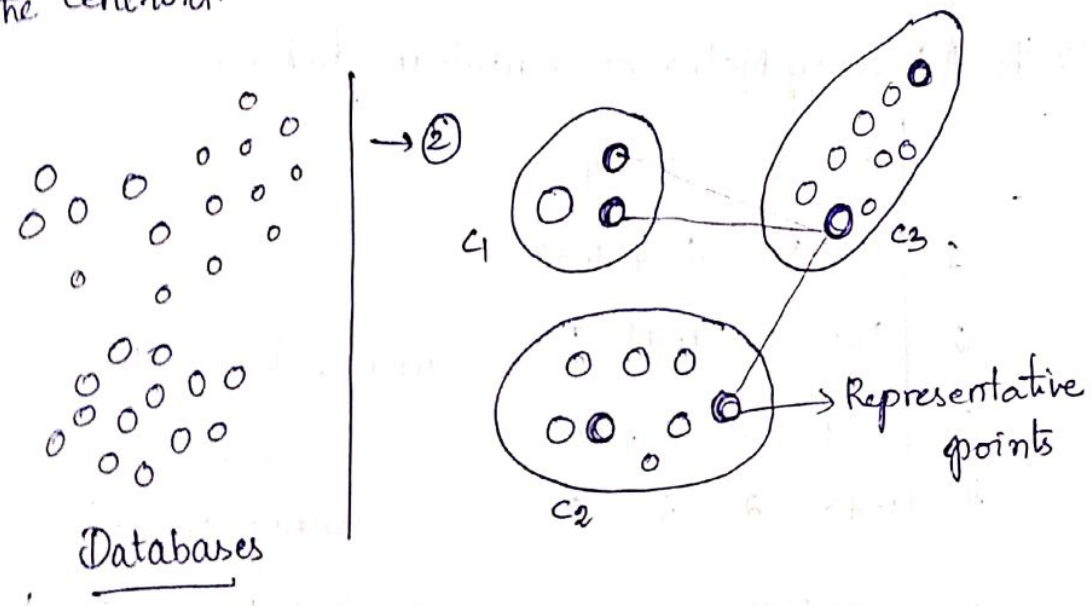
3) CURE:- Clustering Using REpresentatives

1) For a given Dataset find the no. of clusters that can be formed 2) Identify the Representative points

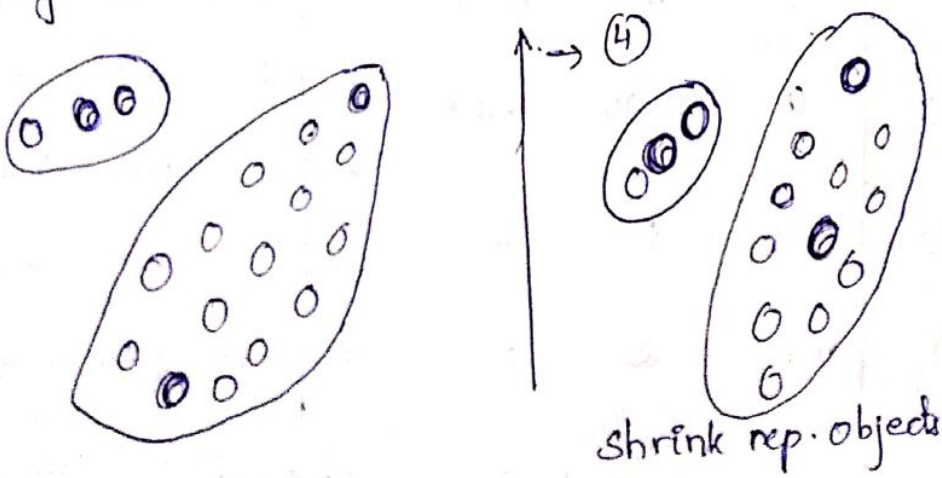
3) From the Representative points, find the Distance b/w the points from one cluster to another cluster.

4) Merge the cluster which is having the min. Distance b/w the Representative points.

5) After Merging the Representative points / shrinking towards the Centroid.



(3) Merge C_2 & C_3 (min distance b/w 2 rep. points)



Clustering with Categorical Attributes:

Ex:

→ {book, water, sun, sand, swim, read}

→ There are 4 documents

→ 1 - {books}

→ 2 - {water, sun, sand, swim}

→ 3 - {water, sun, swim, read}

→ 4 - {read, sand}

boolean Matrix

1 (1, 0, 0, 0, 0, 0)

2 (0, 1, 1, 1, 1, 0)

3 (0, 1, 1, 0, 1, 1)

4 (0, 0, 0, 1, 0, 1)

Write the Adjacency Matrix as Euclidean distance

	1	2	3	4
1	0	2.24	2.24	1.73
2	2.24	0	1.41	2
3	2.24	1.41	0	2
4	1.73	2	2	0

min value = 1.41

= (2, 3).

→ average of 2 & 3

Matrix

K = 2 → No. of clusters. → 2&3 (0, 1, 1, 0.5, 1, 0.5)

	1	23	4
1	0	2.24	1.73
23	2.24	0	2
4	1.73	2	0

	1	23	4
1	0	2.12	1.73
23	2.12	0	2.34
4	1.73	2.34	0

k-cluster = (2, 3)(1, 4)

2) ROCK: Algorithm:

✓ It can be used to find the clusters for both boolean data and categorical data.

✓ In this Identifying the similarity based on no. of links b/w the items.

✓ Similarity of items: A pair of items are said to be Neighbour, if their similarity exceeds some threshold.

✓ The no. of links b/w the items is defined as no. of Common Neighbours that they have.

✓ Instead of using Euclidean distance, use the Jacard Coefficient $JC = simi(t_i, t_j) = \frac{q}{q+r+s}$.

given threshold = 0.6.

q → 1 1
r → 1 0
s → 0 1
t → 0 0

	1	2	3	4
1	1	0	0	0
2	0	1	0.6	0.2
3	0	0.6	1	0.2
4	0	0.2	0.2	1

Sim(1,2) = 0/

Simi(1,3) = 0

Simi(1,4) = 0

Simi(2,3) = $\frac{3}{3+1+1} = \frac{3}{5} = 1/2 = 0.6$

The points that satisfy the threshold are (2,3) (2,4) & (3,4)

The points that exceeds the threshold are (1,1) (2,2) (3,3) (4,4)

- Agglomerative Algorithms may also called as Agnes (AGNES)
  Agglomerative Nesting.
- Divisive Clustering is called as DIANA (Divisive Analysis)