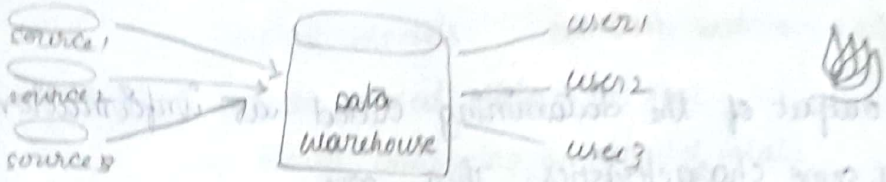
12/12/2019

The difference between datawarehouse and database:

In database if number of customers increases, the time taken to get the required information also increases. it is difficult to maintain a data in particular place.

The Datawarehouse is a collection of dissimilar data collected from various sources and stored at particular place.

Data warehouse accepts all formats of data but to store into the Data warehouse it should be transformed into standard formats.



Difference between Datawarehouse and big Data.

In Datawarehouse to store the data, it should follow some standard format but in BigData it will not follow any standar format.

Data warehouse:

Bill Inmon in 1990's introduced the data warehouse. According to the Bill Inmon, the collection of data should be called as data warehouse, only if it follows some features i.e The Datawarehouse is

1. Subject Oriented

2. Non-volatile.
   The data should not be modified or erased. The new data will be added to the collection of data without earasing the old data

3. Integrated
   The data should be collected from different sources such as pdf, doc etc..,

4. Time Varient
   The data should only be changed (updated) in certain period of time but not frequently.

14/12/2015

Data Mining:

Extracting the information from large collection of data which is unknown to the users is called as Data Mining.

The information that is extracted from data warehouse should posses some characteristics. They are

1. Non-trivial - whatever the information that is getting from the DB that should be relevent to the user (useful)
2. Novel - The o/p should be unique i.e Any type of algorithm is applied, the o/p should be same.
3. useful - The information should be used for future purpose to take decisions.

The output of the datamining called as information should follow some characteristics They are

1. Non-Trivial

The data should be extracted from warehouse, i.e the output from the Data warehouse should

2. Novel

it produces the same output for any algorithm

3. useful

The information which is extracted from Data warehouse should be used in future to take decision.

Datamining : (definition).

Extraction of intersting (Non-Trivial, Novel, useful) patterns (or) knowledge from huge collection of data.

we can use new technology such as machine learning, Bigdata, statistics, visualisation, Genetic algorithm

History of data science:

In the early 1960's Data collection and Database creation

→ file processing.

* early 1960's
  data collection and database creation
  → File processing.

* 1970's & early 198's
  → DBMS
    - ER models
    - Indexing & accessing methods
    - query language – SQL
    - query processing & optimisation

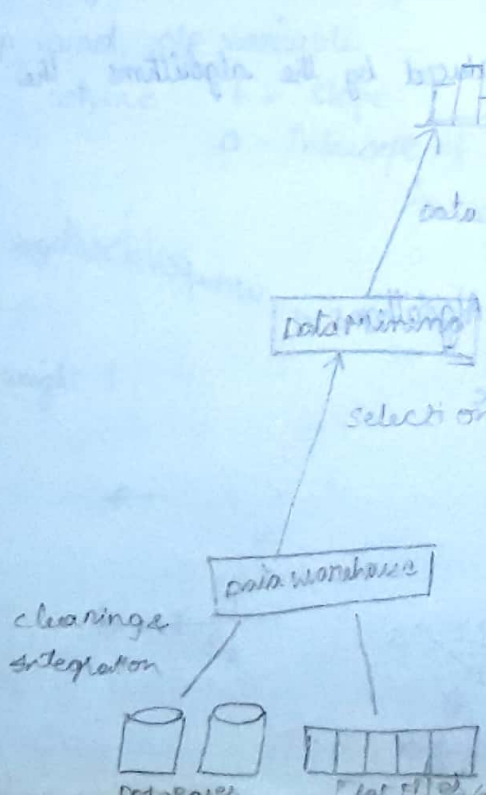
* mid 1980's to 1990's
  → EER models
  → web based data (ex, XML---)
  → cloud computing & parallel data.

* early 1990's to present
  ↳ Data warehouse
  ↳ DMT & KDD (Data mining Techniques & knowledge discovery of data
  ↳ DMT in applications.

KDD (Knowledge Discovery from Data).

It indicates the different procedures involved to get the required knowledge from large collection of data.

15/12/2015



Data cleaning:

Remove the noise, redundency data and inconsistent data.

Integration:

collect the information from different data sources which matches to the given query.

Data selection and Transformation:

By using these techniques we select the required information from data warehouse and transform to the format which is understandable by datamining algorithms.

Here we apply normalisation procedure and smoothless functions.

patterns and visualisations:-

identify the interesting information from data mining algorithms and display to the users by applying some visualisation techniques.

Datamining:

Intelligent methods applied to extract the datar patterns from data warehouse.

Basic Datamining tasks:

Depending upon the output produced by the algorithms, the DataMining is classified into two types.

1. predictive Algorithms - supervised
2. discriminative (or) Discriptive Algorithms - unsupervised

The predictive algorithms are

1. classification
2. Regression
3. Time series analysis
4. prediction.

The descriptive algorithms are classified as

1. clustering
2. summarisation
3. Association rules
4. sequence discovery.

classification :

predicting the certain output from the given input data lines. The classification consists of systematic approach and different mathematical techniques which is used to classify the data it is based upon the of output which is already known to the users.

classification algorithms.

1. Decision Trees
2. Neural N/w's
3. Rule based induction.
4. Bayesion N/W
5. Genetic algorithms.

16/12/15 **Regression :**

it is the technique that is used to predict one or more i/p variable and one o/p variable.

it shows linear relationships b/w I/P & o/P variables.

ie $y = a + bx$ is a formula to calculate relationship b/w I/P and o/P variable

where b = slope

a = intercept of y.

age	2	3	10	15	20	25
weight y	4	5	30	35	50	?

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n \cdot \Sigma x^2 - (\Sigma x)^2}$$

$$\bar{y} = \frac{\Sigma y}{n} \quad ; \quad \bar{x} = \frac{\Sigma x}{n}$$

n → no. of x variables.

no of y variables

$\Sigma x = 50$

$\Sigma y = 124$

$\Sigma xy = 1848$

$\Sigma x^2 = 738$

$(\Sigma x)^2 = 2500$

$b = \frac{5 \times 1848 - 50 \times 124}{5 \times 738 - 2500}$

$= 2.5$

$a = 24.8 - 2.5 \times 10$

$= 24.8 - 25 = -0.2$

$\bar{y} = \frac{\Sigma y}{n} = \frac{124}{5}$

$= 24.8$

$\bar{x} = \frac{50}{5} = 10$

$y = a + bx$

$= -0.2 + 2.5x$

25
25
125
50x
62.5
62.5
0.2
62.3

if $x = 25 \Rightarrow y = -0.2 + 2.5 \times 25$

$\Rightarrow y = 62.3$

Types of Regression:

1. Linear Regression
2. Multiple Regression

Association Rules:

It is used to analyse and predicts customer behaviour by using if-then statements it covers the relationships b/w unrelated data in a relational database or other information repository

The best example is

in a supermarket, the customer who buys the bread chance to buy the butter

bread $\Rightarrow$ buys (butter).

bread $\Rightarrow$ butter [20%, 40%].

20% bread

40% butter

17/12/2018

Clustering:

In this the data is divided into different sub classes called as clusters. In clusters the relationship b/w the different object is more.

In cluster the relationship b/w intra class cluster objects is more and interclass cluster relationship is less

the methods in the clusters.

1. partitioning method
2. density based method
3. model based method
4. constraint based method
5. Heirarichial method.

examples for the clusters

1. Google search Engine
2. marketing loans
3. social N/w analysis
4. credit and debit card analysis.

Time series Analysis

In this technique a series of observations and on variable recorded after successive intervals of time. The successive intervals of time are usually equal intervals in any unit of time.

Example for Time series Analysis

it is used to study the behaviour of past data and forecast the future data and it ends In bussiness planning and comparitive study.

Methods

1. univarient Time series — variate
2. Bi-varient Time series
3. Multi-varient Time series

Prediction:

Any Realworld Datamining applications can be seen as predicting future value based on past and current data.

prediction

prediction can be

the difference b/w classification and prediction is the classification only works on present data where the o/p is already known. But prediction works on past and present data & predicts future data.

Examples:

1. Flooding
2. speech recognition
3. Machine learning
4. pattern recognition.

18/12/2018

DataMining Metrics:

Metrics is used to analyse the algorithm i.e out of two algorithms which one is the better one for the given application

there are different Types of Metrics. They are

1. confusion matrix
2. Cost Matrix
3. ROC curve
4. Accuracy
5. precision
6. Recall
7. F-beta Measures
8. Loss Function
9. Square / Absolute error Methods.

(1, 2, 3, 4 — based on the confusion matrix)

1. Confusion Matrix:

it shows how many relations are correctly classified & how many are not classified correctly

it gives relationship b/w the correctly classified & not correctly classified ~~relationships~~ classes.

there are 2 classes

1. Actual class
   - The o/p of the Alg. is known in advance
2. predictive class
   - This is the o/p of the Algorithm for the same query.

a/c	predictive class		
		class = yes	class = NO
Actual class	class = yes	a	b
	class = NO	c	d

True +ve	False -ve	false +ve	True -ve

among the four regions a, d are important as they give which are the correctly classified.

Accuracy :

$$\text{Accuracy} = \frac{a+d}{a+b+c+d}$$

Example :-

	+	-
+	50	100
-	50	150

$Acc = \frac{200}{350} = 60\%$

precision :-

precision : $\frac{a}{a+c}$ (True positive Ratio).

precision (↑) ⇒ accuracy (↑)

Cost Matrix :

	predictive class	
Actual class	$C(i/j)$ class=Yes	class=NO
class = Yes	$C(Yes/Yes)$	$C(NO/Yes)$
class = NO	$C(Yes/NO)$	$C(NO/NO)$

Ex :-

	+	-
+	-1	100
-	1	0

if M1 is used

	+	-
+	150	40
-	60	250

$Acc = \frac{150+250}{500} = \frac{400}{500}$

if M2 is used

	+	-
+	250	45
-	5	200

$\Rightarrow Acc = \frac{150+200}{500} = \frac{450}{500}$

if Acc is taken ⇒ 2nd M is better.

cost1 = 150×-1 + 40×100 + 60 = 3940

cost2 = -250 + 4500 + 5 = 4255

acc to cost, 1st algorithm is better one.

Now we need to consider acc. to our application

Recall (r): Recall (r) = $\frac{a}{a+b}$ (sensitivity). (TPR).

F-beta measure: F-beta measure = $\frac{2rP}{r+P} = \frac{2a}{a+b+c}$.

weighted Accuracy:

$$\text{weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

Specificity = $\frac{d}{c+d} \Rightarrow$ (TNR)

$FPR = \frac{c}{c+d}$.

$FNR = \frac{b}{a+b}$.

ROC curve: (Reciver operating characteristics).

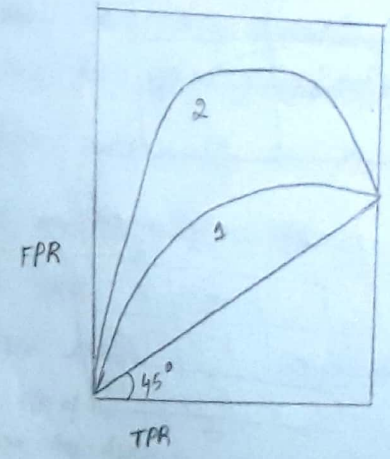
Roc curves are used for visual comparisions of classification models.

These curves came from signal detection theory.

These curves shows trade-off b/w true positive rate & false +ve rate.

The Area under Roc curve is the measure of accuracy of a model. The Area is called as AUC (Area under curve). the value is in b/w 0.5 to 1.

↓ min distance.



The area of an algorithm for which AUC is more it have more accuracy.

Loss Function:

This function generally used in Regression

Absolute error = $|y_i - y_i'| = \Sigma y_i - \Sigma y_i'$

Squared error = $(y_i - y_i')^2$

Test-error :-

$$\text{mean absolute error} = \frac{\sum_{i=1}^{d} |y_i - y_i'|}{d}$$

$$\text{Mean squared error} = \frac{\sum_{i=1}^{d} (y_i - y_i')^2}{d}.$$

$$\text{Relative absolute error} = \frac{\sum_{i=1}^{d} |y_i - y_i'|}{\sum_{i=1}^{d} |y_i - \bar{y}|}$$

$$\text{Relative-squared error} = \frac{\sum_{i=1}^{d} (y_i - y_i')^2}{\sum_{i=1}^{d} (y_i - \bar{y})^2}.$$

if we need to measure performance we use absolute error.
if there are more outlets in our dataset then we use squared error.

19/12/2018

Data Mining in DB perspective : → data warehouse

1. Scalability
2. Real world data
3. Update
4. ~~Ea~~ Ease of use
5. security.

Scalability :
Algorithms that do not scaleup to perform well with massive real world ~~dbs~~ databases are of limited applications.
related to this is a fact that technique should work regardless of amount of available memory data.

Real world data :
The real world data are noisy data and have many missing attribute values.
Algorithm should be able to work even in the presence of these problems.

update :
Many datamining algorithm work with static datasets. this is not a realistic approach.

Ease of use :
Although some algorithms may work well, they may not be

well received by the users, if they are difficult to use and understand.

privacy & security:

we are collecting the data from different databases. Therefore may chance to get effected by database.

Data Mining Issues:

1. Human Interaction
2. over fitting
3. outliers
4. Interpretation of result
5. visualisation of results
6. Large Datasets
7. High Dimensionality
8. Multimedia Data
9. Missing data
10. Irrelevant data
11. Noisy data
12. Changing data
13. Integration
14. Application

20/12/18

Data preprocessing:

Data is nothing but collection of objects and attributes.

preprocessing means preparing raw data understandable format before applying data mining algorithms.

Importance of preprocessing:-

Now a days, we are dealing with real world data and these data in the format of

i, incomplete means lagging of attribute values or containing aggregate data

example: $a = b + c$

$b = 20, c = ?$

print(a) # incomplete.

ii, Noisy data - The data contains errors & or outliers.

example: a=10 but a=-10 #noisy (unwanted)

iii, Inconsistent data - It containing different codes or names.

example: if db has

	x_1	x_2
1	0	1
2	0	1

we give

	x_1	x_2
A	0	1
B	0	1

#inconsistent.

Tasks of data preprocessing:-

1) Datacleaning -
filling missing data, Smooth the noisy data, identify & remove outliers and remove inconsistency.

2) Data Integration -
we integrate multiple databases, dataqueues, & flat files.

3) Data Transformation -
we use normalisation & aggregation to transfer the data from one format to another format.

4) Data Reduction -
Reducing the size of data but producing the similar result

analytical results of original data.

5) Data Discretization :-

Some part of data is reduced & replacing by numerical attributes

→ 1. Data cleaning :-

i) Missing Data : Reasons to missing the data.

* Data May not included simply ~~to~~ because it was not consider important at that time of Entry.
* Relavent data may not be recovered due to misunderstanding & equipment failure.
* The attributes of interest may not always available.

21/12/18. Handling Missing Data :

1. Ignoring the tuple :

when a class label is missing we need to delete the tuple but it is an ineffective method. Unless the tuple contains more number of missing data.

2. Fill in the Missing value by Manually :

In General this method is the time consuming process & may not be feasible when given a large datasets with many datasets.

3. Use the Global Constant to Fill in the missing value :

Replace all missing attributes by some Constant, then the mining program may mistakenly think that they form an interesting pattern since, they have a value in common.

4. Use the attribute means to fill in the missing value :

5. Linear (Interpolation) & Intercorrelation : Babian

take the values of up and down of missing value & calculate average & replace it inplace of missing value.

6. use the most probable value ~~fitting~~ fill in the missing value :

this may default with inference base tool like decision tree, regression on a basian formula.

ii) Noisy Data :

To deal with the noisy data, we are applying Binning methods. The Binning method is again classified into three types. They are

1. Mean
2. Median
3. Boundaries.

1. Mean Binning Method :

step 1 : arrange the data in sorting order

step 2 : divide the given data into equal parts called as Binning depth & represented by D. and the no. of ~~Binnings~~ Bins is represented by $\frac{N}{D}$

step 3 : Calculate the mean value of each Bin by using the formula

$$\frac{1}{N}\sum_{i=1}^{N} x_i$$

step 4 : Replace each value in bin by calculate the mean value of each bin.

Example :

5, 7, 4, 6, 9, 10, 8, 6, 5

Step 1 :- Sort the data.

4, 5, 5, 6, 6, 7, 8, 9, 10

step 2 :- The depth of each bin is equal to 3

D = 3

no. of bins = $\frac{N}{D} = \frac{9}{3} = 3$.

The first bin contains 4, 5, 5

bin-2 contains 6, 6, 7

bin-3 contains 8, 9, 7

step 3 : Calculate average of each bin.

$\frac{4+5+5}{3} \Rightarrow \frac{14}{3} = 4.66 \Rightarrow 4\frac{2}{3}$

$\frac{6+6+7}{3} \Rightarrow \frac{19}{3} = 6.33 \Rightarrow 6\frac{1}{2}$

$\frac{8+9+10}{3} \Rightarrow \frac{27}{3} \Rightarrow 9$

step 4

$[4\frac{2}{3}, 4\frac{2}{3}, 4\frac{2}{3}], [6\frac{1}{3}, 6\frac{1}{3}, 6\frac{1}{3}], [9, 9, 9]$.

2. Median Binning Method : (50 percentage).

50 % (n+1).

$\frac{50}{100} \times (n+1)$ where n is no. of bins

$\frac{50}{100} \times (3+1) = \frac{1}{2} \times 4 = 2$

Replace all the values with "2".

[2, 2, 2, 2, 2, 2, 2, 2, 2]

3. Boundaries Binning Method :-

calculate the maximum value & min value in the each bin

replace all the values of the bin with the minimum value except the maximum value.

for above example:

[4, 5, 5], [6, 6, 7], [8, 8, 10].

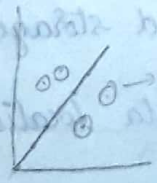
ii) Clustering:

similar values are organised into groups called as clusters. the values that falls outside of clusters called as outliers.

iii) Regression Technique:

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

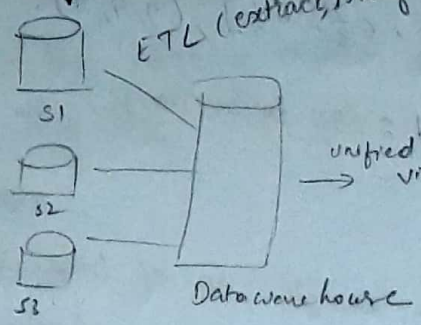
$y = a + bx$



iv) Manually we can remove the outliers. but it is not easy.

27/12/18

→ Data Integration:



problems
- redundancy
- schema integration.

It involves combining the data from different sources & desparate sources which are stored using various technologies & provide unified view of the data. We use two majo approaches to get data from different data sources.

1. tightly coupling
2. Loose coupling

1. Tight coupling:
Applies the ETL operation on different data sources & store the results in DW. from that the user will get required information.

2. Loose coupling:
In this we are not applying any ETL operations. the query submitted by the user will directly sent to the diff data sources & the results will transferred to the user.

Advantages of Data Integration:
1. Faster query processing.
2. Independence. (not depending on multiple DBs)
3. Complex query processing
4. High volume data processing.
5. Advance data summerisation and storage.

Disadvantages of Data Integration:
1. costliest operation because of data localisation, infrastructure and security
2. Latency, because data we needs to be loaded after ETL operations.

Issues In the Data Integration:
1. Schema Integration:
2. Redundancy
3. Detection & Resolution of data value conflicts.

1. Schema Integration:

The real world entities from multiple sources be matched is called as the Entity identification problem.

Example: The Data Analyst or the Computer be sured that customer_id in one DB & cust_id in another DB to represent the same Entity.

2. Redundancy:

The Redundancy is occured by 2-ways

1. object identification, means the same attribute of an object may have different names in different databases.

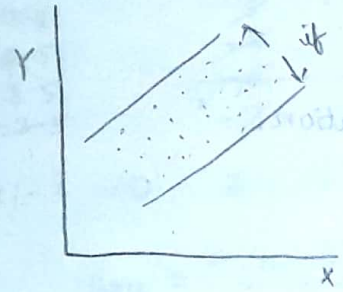
2. Derived Data, one attribute may be a derived attribute in another table.

23/12/18

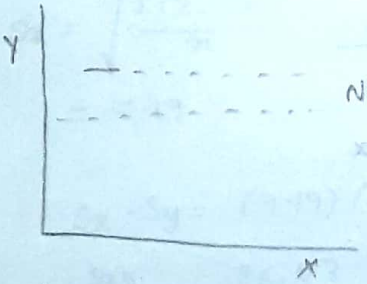
Correlation Analysis.

it is used to identify the strength of the association between two variables. In this correlation Analysis, there is no casual effect. (numeric data)

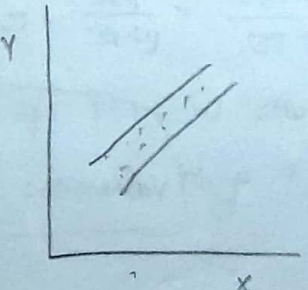
By using scatter plot, we find out the relation b/w the two variables.



if this range is large it is weak relation.



No relation



strong relation

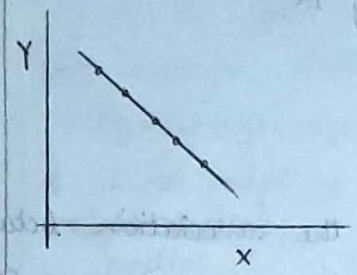
In correlation Analysis, the value of 'r' should be bt in b/w -1 & 1.

if $r_{A,B} > 0$, then there is a stronger +ve linear Relationship.

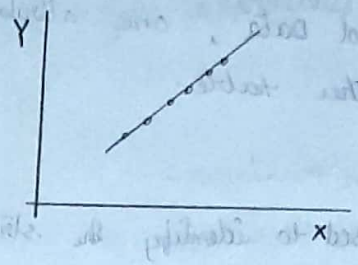
if $r_{A,B} = 0$, weaker linear relationship

if $r_{A,B} < 0$, then there is a stronger -ve linear relationship.

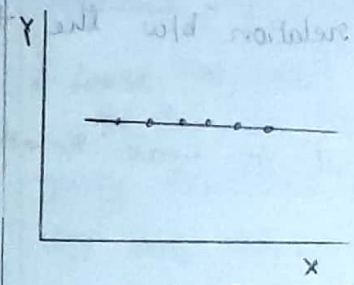
$r = -1$



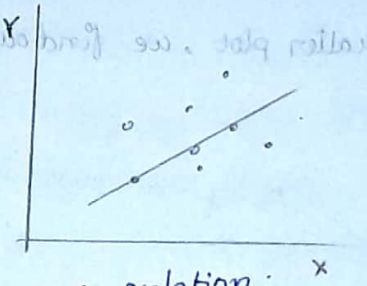
$r = 1$



$r = 0$



$r = 0.3$



The formulas to find out the co-relation;

$$\text{co-variance } s_{xy} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\text{co-relation coefficient } (r) = \frac{S_{xy}}{S_x S_y}$$

S_x - standard deviation of x

S_y - Standarad deviation of y.

$r \propto S_{xy}$.

Mean, $\bar{x} = \frac{\Sigma x}{n}$.

Standard deviation, $S \Rightarrow \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$

$\Rightarrow$ Variance $= \frac{\Sigma (x_i - \bar{x})^2}{n}$.

181)

correlation coefficient $(r) = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2}}$

Example:

Temp(x)	Icecream(Y)
66	8
72	11
77	15
84	20
83	21
71	11
65	8
70	10

73.5
66.0
-7.5

$\bar{x} = \frac{\Sigma x}{n} = 73.5$ $\quad \therefore \bar{y} = \frac{\Sigma y}{n} = 13$

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
-7.5	-5	37.5
-1.5	-2	3
3.5	2	7
10.5	7	73.5
9.5	8	76
-2.5	-2	5
-8.5	-5	42.5
-3.5	-3	10.5

$\Sigma x_i - \bar{x} = 0$ $\quad \Sigma(y_i - \bar{y}) = 0$ $\quad \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 255$

$$S_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{255}{8} = 31.87$$

$$S_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} = 7.19 \qquad S_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n}} = 5.13$$

$$S_x \cdot S_y = (7.19)(5.13) = 37.5$$

$$r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{36.43}{37.5} = 0.99 \simeq 1$$

if $r = 1$ ⇒ strongly relation.

so, if temp ↑, Icecream ↑

29/12/18

Chi-Square : (Xi-Square)

χ^2 (used to find the relation b/w two variable.

this is used to measure Nominal data. this is the 1900 data method by carl pearson).

it is developed in the year 1900 by carl pearson. The symbol χ is called as "chi".

This technique only works for the categorical data or the nominal data such as Gender means male or female, Colour means red, yellow, Green .. , & sevierity of disease mild, severe.

this chi-square technology, it consists of two types of data. i.e O → Original (or) actual (or) observed data.

E → expected data.

in order to find out whether 2 objects are strongly correlated or not, we use a degree of freedom to decide.

degree of freedom, df = (no. of rows - 1) (no. of columns - 1).

The formulas to find the expected data :

a	b	a+b	
c	d	c+d	
total	a+c	b+d	n = a+b+c+d

$$e_{11} = \frac{(a+b)(a+c)}{n} \qquad e_{12} = \frac{(a+b)(b+d)}{n}$$

$$e_{21} = \frac{(c+a)(c+d)}{n} \qquad e_{22} = \frac{(c+d)(b+d)}{n}$$

Example :

	Smoker	Non-Smokers	
suffering from lung diseases	36 $e_{11} = \frac{50 \times 66}{105} = 31.42$	14 $e_{12} = \frac{50 \times 39}{105} = 18.57$	50 (50% of 66 = 36)
not-suffering from lung diseases	30 $e_{21} = \frac{55 \times 66}{105} = 34.57$	25 $e_{22} = \frac{55 \times 39}{105} = 20.42$	55 (55% of 39)
	66	39	105 n

$$\Rightarrow \sum_{i=1}^{c} \sum_{i=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} = \chi^2$$

Calculate - numeric
Chi-square - categorical

$\Rightarrow \frac{(36-31.42)^2}{31.42} + \frac{(14-18.57)^2}{18.57} + \frac{(30-34.57)^2}{34.57} + \frac{(25-20.43)^2}{20.43}$

$\Rightarrow 0.667 + 1.124 + 0.604 + 1.022$

$\Rightarrow 3.417 \Rightarrow \chi^2$

$p < \chi^2$ (rejected) (strongly corelated)

degree of freedom = $1 \times 1 = 1$

df	0.05	0.01	0.001
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4			

$p > \chi^2$

So Accepted (Independent). (no relation)

Hence Smoking is not the only factor for lung disease.

Tuple Duplication:

if there exists one tuple multiple times in DB is called as Tuple Duplication.

delete one Repeated row to handle tuple duplication.

Data Value Conflicts & Resolution:

if there exists one data in different formats, it leads to the conflicts to find out original data (required data) because two values belongs to the same data.

this will be resolved by converting one data format into another one to scale the data.

02/01/19

→ **Data Reduction:**

it is the technique to reduce the size of the data but provide the same Analytical results if you work on original data is called as data Reduction. that closely maintains the Integrity of data

Advantages:

To improve the performance of DMT.

To reduce the space.

To reduce the N/w Bandwidth to transfer.

Types of Data Reduction Techniques:

1. Dimensionality Reduction
2. Numerosity Reduction
3. Data compression.

Dimensionality Reduction:

it is the process of reducing the no. of random Variables (or) attributes under Consideration.

methods in the dimensionality Reduction are

1. ~~wavelength~~ wavelet Transforms } transform original data
2. principle component Analysis. } into sample data (smaller space)
3. Attribute subset selection.
   ↳ irrelevant, weekly relavant (or) redundant attribute
   (or) dimensions are deleted (or) removed.

Numerosity Reduction :

it is a technique to replace the original data size by alternative smaller forms of data Representation. there are two types in numerosity reduction. they are

1. parametric numerosity Reduction Technique.
2. non-parametric

parametric

In this model, we used to estimate the data, so that the required parameters needs to be stored in memory rather than actual data. The techniques in parametric numerosity Reduction Technique are

1. Regression
2. Log-Linear models

Non-parametric.

it is used for storing reduced representation of data.

the techniques in non-parametric reduction are

1. Histograms
2. clustering
3. Sampling
4. Data cube ~~Aggregation~~.

Data Compression :

In this, transformations are applied to obtain reduced or compressed representation of data

it is a reduction in the no. of bits need to be represent the data. There are 2 types of compression techniques

1. Loss-less compression technique

   If the original data can be reconstructed from the compressed data without any information loss.

2. Lossy compression technique

   Instead of getting original data, we can reconstruct only

an approximation of original data. The techniques in the lossless are.

1. string compression
2. All the dimensionality reduction and numerosity reductio techniques are examples of data compression.

03/01/2017

7. Wavelet Transforms : it is a digital signal.

The wavelet transformation, is used for data Analysis to make more accurate data. The difference b/w ~~Foreain~~ Fourier Transformation and wavelet Transformation is

* The Fourier Transformation will not produce accurate image (or) data when compared to the wavelet Transformation
* It is not easy to work on continuous signal, But it wavelet Transformation signal, the both areas of +ve & -ve will have the common fact
* In wavelet Tra
rs. then we can apply the transformation only one side.
* There are two types of wavelet Transformation
  - Continuous wavelet Transformation
  - Discrete wavelet Transformation.

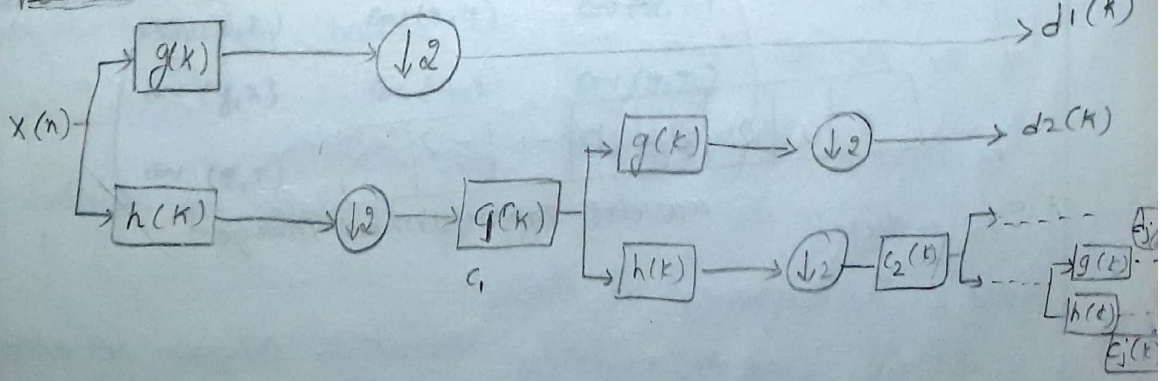
Discrete wavelet Transformation : it is a linear signal processing technique

In DWT, we are given an image (a data) of size N & divided into $(x_1, x_2, x_3 \dots\dots x_n)$ and each datapart contains the same length. we call this data parts as wavelet coefficients.

In DWT, there are many algorithms that are used to analyse the data, one among the is pyramid algorithm.

pyramid Algorithm

$X \Rightarrow X^1$ (wavelet)



The decision Tree is used to reduce the dimensionality by removing & not considering the attributes in the original data sets.

parametric data Reduction: There are two techniques.

↳ regression ↳ log linear methods.

we can also called as numerosity reduction.

These two models are used to reduce the size of data.

Regression:-

In linear regression, the data are modeled to fill a straight line

$$y = wx + b$$

where

x - independent variable (predictor variable)

y - dependent variable (response variable).

w, b are regression coefficients.

The data which does not fit in the line is called as outlier and this will be deleted from original dataset

Log linear

this is used to approximate discrete multi dimensional probability distribution.

estimate the probability of each point in a multi dimensional space for a set of discretised attributes based on a smaller subset of dimensional combination.

$$F = ae^{bx}.$$

where a, b are log linear coefficients.

Non parametric data Reduction:

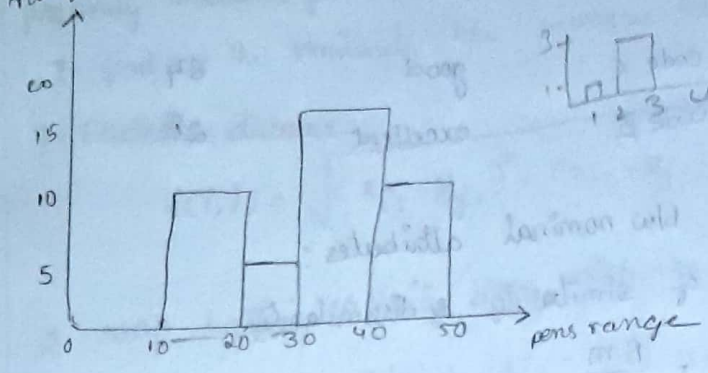
1. histogram.
2. clustering
3. Sampling

Histogram:

the graphical presentation of frequency distribution of a continuous series

example:

pens range	10-20	20-30	30-40	40-50
No of pens	10	6	15	10



2/2/19. single ton bucket:

4/2/19. Similarity Measurements: The similarity b/w the two objects will indicate the relationship b/w the two objects. (to make clusters we make similarity & dissimilarity

Ex:- $sim(i,j) = 1 - Dis(i,j)$

similarity = 0 means there is no relation b/w two objects also indicated by dissimilarity = 1.

similarity = 1 means more relation b/w two objects & also indicated by dissimilarity = 0.

To represent the similarity b/w the objects will use two datacenters

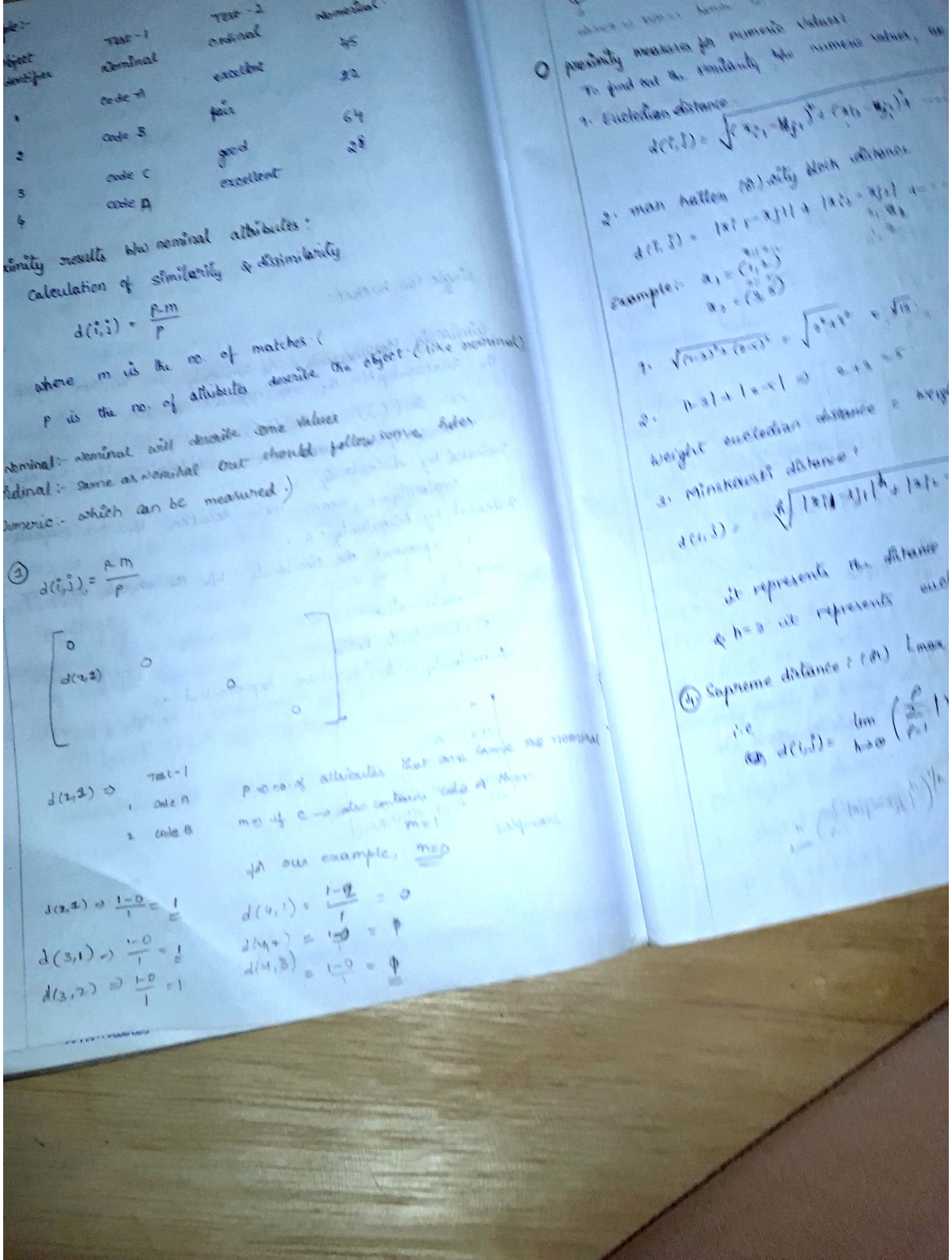
1. data matrix
2. dissimilarity matrix.

① clusters.
② Google search

dissimilarity matrix can be represented as
↳ this structures stores a collection of proximities that are available for all pairs of n objects.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ d(4,1) & d(4,2) & d(4,3) & 0 & \\ d(5,1) & d(5,2) & d(5,3) & d(5,4) & 0 \end{bmatrix}$$

Examples:



Scanned by CamScanner

example:-

object identifier	Test-1 Nominal	Test-2 Ordinal	Test-3 Numerical
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

proximity results b/w nominal attributes:

Calculation of similarity & dissimilarity

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

where m is the no. of matches.(

p is the no. of attributes describe the object.(like nominal)

(Nominal:- Nominal will describe some values.
Ordinal:- same as Nominal but should follow some order.
Numeric:- which can be measured.)

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

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$

$d(2,1) \Rightarrow$ Test-1

1 code A

2 code B

p ⇒ no. of attributes that are same as nominal

m ⇒ if 2 → also contains code A then m = 1

for our example, m = 0

$d(2,1) \Rightarrow \frac{1-0}{1} = 1$

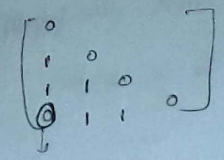
$d(3,1) \Rightarrow \frac{1-0}{1} = 1$

$d(3,2) \Rightarrow \frac{1-0}{1} = 1$

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

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

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



dis=0 ⇒ sim=1 hence i, j are similar to each other

proximity measures for numeric values:

To find out the similarity b/w numeric values, we use 3 methods

1. Eucledian distance

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots + (x_{ip}-x_{jp})^2}$$

2. man hatten (or) city block distance.

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

Example:- $x_1 = (1,2)$

$x_2 = (3,5)$

1. $\sqrt{(1-3)^2+(2-5)^2} = \sqrt{2^2+3^2} = \sqrt{13}$

2. $|1-3| + |2-5| \Rightarrow 2+3 = 5$

weight eucledian distance = weight x eucledian distance.

3. Minskowski distance:

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

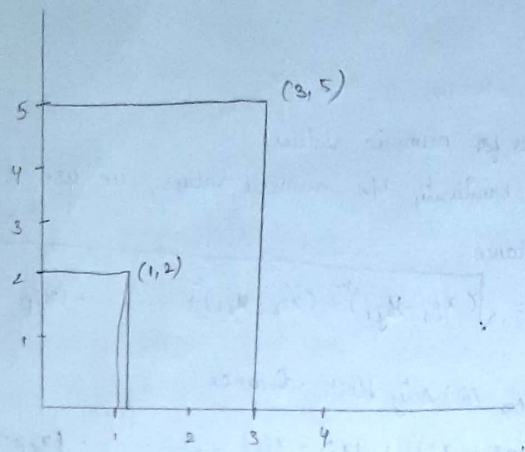
it represents the distance of the manhatten when h=1

& h=2 it represents eucledian distance.

4. Supreme distance: (or) L_{max} (or) L_∞ norm (or) chebyshev

i.e

$$d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h\right)^{1/h} = \max_{f=1}^{p} |x_{if} - x_{jf}|$$



1. Euclidian $= \sqrt{4+9} = \sqrt{13}$

2. man hatten $= 2+3 = 5$

3. supreme $= |2-3|+|5-2| = 5$

Numeric

$$d_{ij} = \frac{|x_{if} - x_{jf}|}{\max_h x_h^f - \min_h x_h^f}$$

Ex:-

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1 & 0 & \\ 0.404 & 0.14 & 0.85 & 0 \end{bmatrix}$$

$$d(2,1) = \frac{|22-45|}{64-22} = \frac{23}{42} = 0.549 = 0.55$$

$$d(3,1) = \frac{|64-45|}{64-22} = \frac{19}{42} = 0.452$$

$$d(3,2) = \frac{|64-22|}{64-22} = 1$$

$$d(4,1) = \frac{|83-45|}{64-22} = \frac{17}{42} = 0.404$$

$$d(4,2) = \frac{|128-22|}{64-22} = \frac{6}{42} = \frac{1}{7} = 0.14$$ high similarity

$$d(4,3) = \frac{|128-64|}{64-22} = \frac{36}{42} = 0.857$$

4,2 are similar

proximity results b/w two ordinal attributes.

ordinal ⇒ it contains specific order.

(3) excellent
(1) fair
(2) good
(3) excellent

For ordinal data:- $\frac{r_{if}-1}{M_f-1}$ (rank of value) (to find the disimilarity b/w ordinal attributes)

(maximum rank)

① $\frac{3-1}{3-1} = 1$

② $\frac{1-1}{3-1} = 0$

③ $\frac{2-1}{3-1} = \frac{1}{2} = 0.5$

④ $\frac{3-1}{3-1} = 1$

⇒ disimilarity matrix

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

$d(2,1)$ = |difference b/w 2 & 1| ⇒ $|0-1| = 1$

$d(3,1) = |0.5-1| = 0.5$

$d(3,2) = 0.5$

$d(4,1) = 0$

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

$d(4,3) = |1-0.5| = 0.5$

$d(4,1) = 0$ i.e similarity = 1

proximity results for binary attributes.

the binary attribute values may be zero & 1, yes & no.

0 - object is absent

1 - object is present

To find the similarity b/w two objects, we use Jaccard coefficient

$P = q+r+s+t$ (for no of attributes)

q → i = 1, j = 1

r → i = 1, j = 0

s → i = 0, j = 1

t → i = 0, j = 0

	i	j
q	1	1
r	1	0
s	0	1
t	0	0

p = q + r + s + t

p → i = p, j = 1

Name	gender	fever	cough	test1	test2	test3	test4
Jack	m	Y	N	P	N	N	N
Jim	m	Y	Y	N	N	N	N
mary	F	N	N	P	N	P	N

P(Y) Y = 1

N = 0

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

r = 1, s = 1, q = 1

$$d(\text{jack, jim}) = \frac{1+1}{1+1+1} = \frac{2}{3} = 0.66$$

$$d(\text{jim, mary}) = \frac{r+s}{q+r+s} = \frac{1+2}{1+1+2} = \frac{3}{4} = 0.75$$

$$d(\text{mary, jack}) = \frac{1+0}{2+1+0} = \frac{1}{3} = 0.33$$

$$d(\text{jack, mary}) = \frac{0+1}{2+0+1} = \frac{1}{3} = 0.33$$

Proximity Results for Mixed variables

Nominal Matrix

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

Ordinal Matrix

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

Numerical

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

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

$\delta_{ij} = 1$ means if all the attributes are present ;

else $\delta_{ij} = 0$ (del-δ).

$$d(2,1) = \frac{1(1) + 1(1) + 1(0.55)}{1+1+1} = \frac{1+1+0.55}{3} = 0.85$$

$$d(3,1) = \frac{1(1) + 1(0.5) + 1(0.45)}{3} = \frac{1+0.5+0.45}{3} = \frac{1.95}{3} = 0.65$$

$$d(3,2) = \frac{1(1) + 1(0.5) + 1(1)}{3} = 0.83$$

$$d(4,1) = \frac{1(0) + 1(0) + 1(0.4)}{3} = 0.13$$

$$d(4,2) = \frac{1 + 1 + 0.14}{3} = \frac{2.14}{3} = 0.71$$

$$d(4,3) = \frac{1+0.5+0.86}{3} = 0.785$$

5/12/19

The cosine similarity :

if the data is missed in the documents, By using previous methods, we cannot find similarity. i.e the algorithm doesn't "work".

sparse matrix ⇒ the matrix with more zeroes and less data values

It is used to handle the sparse data (the most of the data values are zero). In information retrieval system most of the data objects are absent. for those documents, we will the normal similarity measures will not produce accurate similarity

The cosine ~~simple~~ similarity is used to handle sparse data. if resultant cosine value = 1, ie the angle b/w x,y is 0 then the documents are more similar.

if cosine value is nearer to zero, then documents are disimilar.

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

for example :

document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
document1	5	0	3	0	2	0	0	2	0	0
2	3	0	2	0	1	1	0	1	0	1
3	0	7	0	2	1	0	0	3	0	0
4	0	1	0	0	1	2	2	0	3	0

term frequency table. (how may time a word is repeated in a document).

$$Sim(x,y)_{(1,2)} = 5\times3 + 0\times0 + 3\times2 + 0\times0 + 2\times1 + 0\times1 + 0\times0 + 2\times1 + 0\times0 + 0\times1$$

$$= 15 + 6 + 2 + 2 = \frac{25}{2}.$$

$$|x1| = \sqrt{5^2+0^2+3^2+2^2+2^2} = \sqrt{25+9+4+4} = 6.48.$$

$$|y1| = \sqrt{9+4+1+1+1} = 4.12.$$

$Sim(x,y) = \frac{25}{6.48 \times 4.12} \approx 0.936 \approx$

Here the doc1 & doc 2 are more similar.

if the document attributes are binary attributes, then the

similarity of $x, y = \frac{x \cdot y}{x \cdot x + y \cdot y - x \cdot y}$

→ Data Transformation:

The ~~Transf~~ Data are transformed & consolidated into the required format so that, the resultant mining process may be more efficient and the patterns are easier to understand.

Types of Data Transformation:

1. Smoothing
2. Aggregation
3. Generalisation
4. Normalisation
5. Attribute Construction.

Normalisation: In Normalisation, the attribute data are scaled within a specified range [(-1 to 1) & (0 to 1)]

Types of Normalisation:

1. Min Max Normalisation
2. Z-score Normalisation
3. Decimal scaling

Min-Max Normalisation:-

$$v = \frac{v - min}{max - min} \left(new\ max - new\ min \right) + new\ min.$$

Example: marks 8 10 15 20.

new max = 1

new min = 0

$$v' = \frac{8-8}{20-8}(1-0) + 0 = 0$$

$$v' = \frac{10-8}{20-8}(1-0) + 0 = 1/6 = 0.1$$

$$v' = \frac{15-8}{20-8}(1-0)+0 = \frac{7}{12} = 0.58.$$

$$v' = \frac{20-8}{20-8}(1-0)+0 = 1.$$

Z-Score Normalisation:

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$\text{Z-score} = \frac{x - \mu(\bar{x})}{S}$$

Example:- Marks: 8 10 15 20

$$\bar{x} = \frac{8+10+15+20}{4} = \frac{53}{4} = 13.25$$

$$S = \sqrt{\frac{(8-13.25)^2 + (10-13.25)^2 + (15-13.25)^2 + (20-13.25)^2}{3}}$$

$$= 5.37.$$

$$\text{z-score for } 8 = \frac{8-13.25}{5.37} = -0.97$$

$$\text{z-score for } 10 = \frac{10-13.25}{5.37} = -0.60$$

$$\text{z-score for } 15 = \frac{15-13.25}{5.37} = 0.32$$

$$\text{z-score for } 20 = \frac{20-13.25}{5.37} = 1.25$$

Since our range is only -1 to 1 it is called as outlier.

Decimal Scaling:

$$v' = \frac{v}{10^j}$$

j is no. of digits.

$$v' = \frac{8}{10} = 0.8$$

$$v' = \frac{10}{10^2} = 0.1$$

$$v' = \frac{15}{10^2} = 0.15$$

$$v' = \frac{20}{10^2} = 0.20$$

6/2/19

Smoothing: which works to remove the noise from the data

The techniques that are used are

1. Binning
2. clustering
3. Regression.

Aggregation: where the summary & Aggregation operations are applied to the given data.

Ex:- Rollup & cube operations in DW.

Generalisation: The data where the low level or primitive data are placed by the higher level data by using the concept of Heirarchies.

Ex:- Address.

Attribute Construction: where a new attributes are constructed and added from a given set of attributes to help the Mining process.

Ex:- Feature extraction (PCA).