Appayya Baalige.

CS181005

data Mining

June - 2016

1a) What is Data ~~Mining~~ mining & Explain any four data visualization technique with Examples.

Data mining

Data Mingling is the process of discove-ring patterns in large data sets involving methods at the intersection of machine Learning Statistics and data base System Data Mining is the analysis step of the Knowledge discovery in data bases process or KDD

* Four data visualisation Techniques :-

Data visualisation is the process of Conveying information in a way that can be quickly information in a way that f easily digested by the viewer Examples are every where, and see the daily charts, graphs digital images & movies.

Techniques :-

1) Box plots. 2) Histograms. 3) Heat maps.
4) charts. 5) Tree maps.

1) Box plots:-

A method for graphically depicting group of numerical data through their quarti the distribution of data based on 5 features.

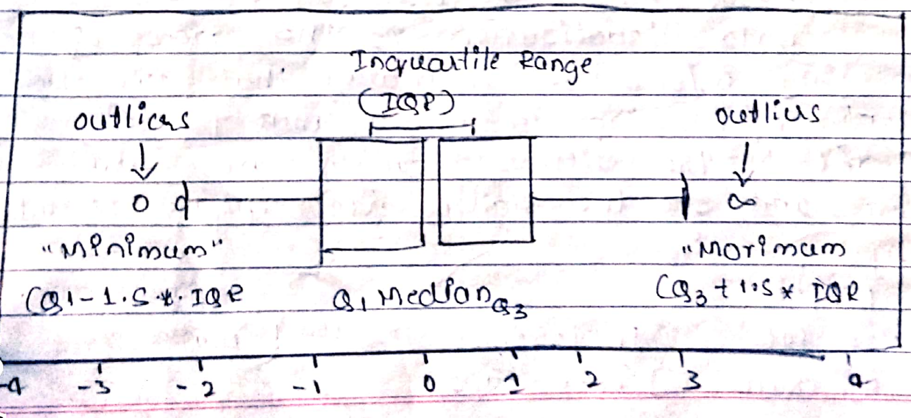
1) Minimum - minimum value in the dataset

2) First quartile - the Middle val between the smallest value & the Median of the dataset

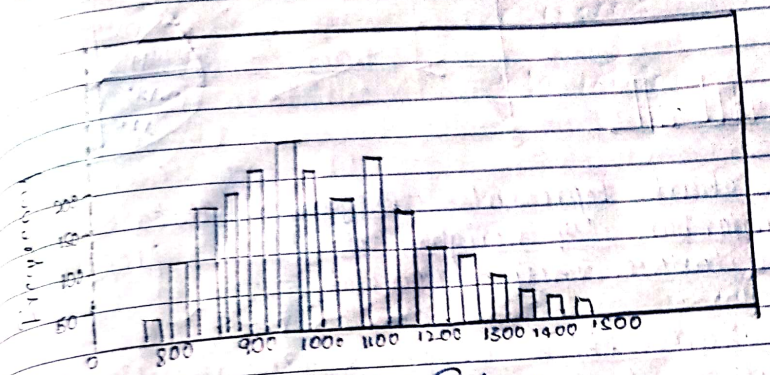
3) Median - the middle value between the of the dataset

4) Third quartile :- The middle value between median & the highest value of the data Set

5) Maximum : The maximum value in the data set also the lower & upper quartile are shown as horizontal lines of the Rectangle It has verticle line inside to indicate the median value



Histograms Histograms of monthly salary



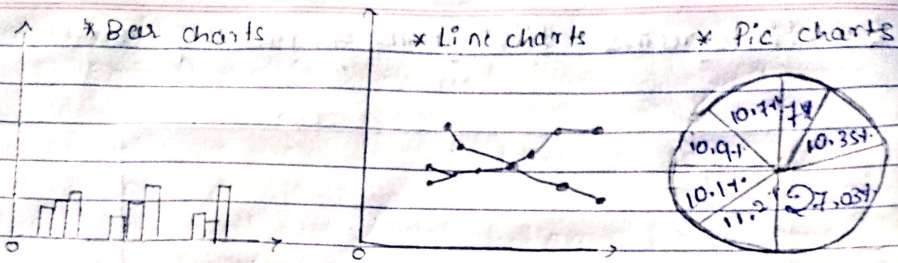
Gross monthly Salary

Is an accurate Representation of the distribution of Numerical data it Relates only one variable Includes bin or bucket the Range of values that is divide the entire Range of values that into a series of intervals of that count how many values fall in to each interval

3) Charts:

It gives a graphical Representation of date has various types

* Bar charts :- * Line charts. Pie charts



* Bar chart Represents categorical data with Rectangular bars of heighty of length proportional to the values they represent

* Line chart which displays information as a Series of data points called Markers connected by Shright line segment

* pie chart -

is a circular statistical graph which decid into Slices to illustrate Numerical proportion.

④ Tree Maps It is a method of displaying hirarchical data using Nested figures like Rectangle Has a tree like Structure with each branch is given a Rectangle which the tils which Smaller Rectangle Representing Sub branches.

A leaf Node Rectangle has an area propational to Specified dimension of data

2020-5-5 09:39

Scanned with CamScanner

How do you handle missing values in a dataset of different types of attributes? Give the missing value problem for the data set given bellow

Floor Area	Rental price	Type
42	10,000	1 BHK
52	13000	?
52	11000	2 BHK
59	?	2 BHK
65	18000	2 BHK
?	13500	3 BHK
60	12600	3 BHK

Attribute discriptions Floor Area : Numeric
Rental price : double.
Type : categorical.

=> Missing values in a dataset of different types of attributes.

1. Ignore the data Row
2. Use a global constant to till in for missing values
3. Use attributes mean
4. Use attributes mean for all samples belonging to the same class.
5. Use a data mining algorithm to predict the most probable value.

2020-5-5 09:39

2020-5-5 09:38

Q.a Briefly compare the following concepts
i) Snowflake schema and Fact constellations

Snowflake schema	Fact constellations
1) Snowflake schema contains the large antral fact table dimension tables & sub dimensione tables	1) while in fact constellation schema dimension tables are shared by many fact Tebles
2) This schema saves significant storage	2) This schema does not save storage
3) This schema consistes of one star schema at a time	3) This schema consists of more than one star schema at a time
4) Tables can be maintained easily	4) Tables are tough to maintrain
5) This schema is a normalized from of star schema	5) This schema is a normalized form of snowflake schema & star schema
6) This is easy to operate as compared to fact constellation schemaas it has less number of joins between tables	6) In this schema is not easy to operate as compared to snowflak schema as it has multiple noros joins between the tables.

6) In this to access the data from database. simple & less complex query is used.	7) In this to access the date from database. heavier complex query is used.
8) Discovery driven cube is used for guiding the user for data analysis process By making use of the precompted data & visual cues to indicate exaption at all level aggregation	Virtual ware house. i) Virtual ware house is a set of view containing data ware house over an opertional schema schas dimensions & aggregated measures.

2.b) Regarding the computation of measures in a data cube enmerate three categories of measures. With an example for each category?

Data cube computation is an essential task in data warehouse implementions

The cube is used to represent data along some measure of interest Although called a "cube" it can be 2-dimensional 3-dimension or higher dimensional

Ex:- We have a data base that contains transaction information Relating company Sales of a part to a customer at a store location

Enumerate Three categories of Measures based.

on the kind of aggregate functions used in computation a data cube they are.

1) Distributive 2) Algebraic 3) Holistic.

1) Distributive :-

It the Result derived by applying the function to n aggregate By applying the function to n values is the same as that derived By applying the function on all the data Without partitioning.

Examples :- counts, sums, mins, maxs,

2) Algebraic :-

It it can be computed by an algebraic function With M arguments (Where M is a bounded integer) each of which is obtained By applying a distributive aggregate function

Examples :- avg(), min-N(), standard-deviations()

3) Holistic :-

It there is no constant bound on the storage size needed to describe a subaggregate

Examples :-

Medians, modes, Rank(),

With respect to the frequent item set mining, write the following terms with an Example.

1. Support :-

It is one of the measure of interesting. This tells about usefulness & certainty of Rules. 5% support means total 5% of transactions in database follow the Rule.

Support (A → B) = support - count (A∪B)

2. Confidence :-

A confidence of 60% means that 60% of the customers who purchased a milk and bread also bought butter.

confidence (A → B) Support - count (A∪B) / Support - count (A)

If a Rule satisfies both Minimum support and minimum confidence it is a strong rule.

Examples :

Support - count (X) : Number of transactions in which X appears. If X is A union B then it is the number of transactions in which A and B both are present

ii) Frequent Itemsets

Frequent item-set mining is an interesting branch of data mining that focuses on looking at sequences of actions or events that base data taken that born of sets of instance. for.

that each has a Number of feature.
A typical example of frequent itemset mining is market basket analysis
Examples. T_ID Item sets.
T1 {A,B,C,D}
T2 {A D}
T3 {A,E}
T4 {C,E}

Now if the minimum support threshold is 50% yes. it is present in 3 so. 75%
C, D, E are present in 2 transactions satisfying 50% so they are frequent.

iv) closed Itemsets:-
It is a frequnet itemset for which non of its immediate supersets have that same support count as itself
Examples:- of above illustration.
{A} is closed because None of its Supersent have that some support as itself
B{A}, {C}, {E}, {A,D} are closed. frequent itemsets
{D} is Not closed because its immediate supers {A,D} also has the same support count 2.

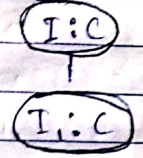
v) Association Rule:-
An Association rule is pattern that States when an event occure another event occurs with certain probability
Association Rules find all sets of items.

that have support greater than the minimum support then using the long itemsets to generate the desired Rules that have confidence greater than the minimum confidence.

vi) Frequent pattern (FP)

Frequent pattern Tree is a tree-like structure that is made with the initial item sets of the database. The purpose of the FP tree is to mine the most Frequent pattern. Each Node of the FP tree Represents an item of the item sets.

FP tree - compress data base into tree.



b) T_{ID} Item Sets.

T_1 { MONKEY }

T_2 { D,O,N,K,KY }

T_3 { M,A,K,E }

T_4 { M,Y,CK,Y }

T_5 { C,O,O,KIE }

=b

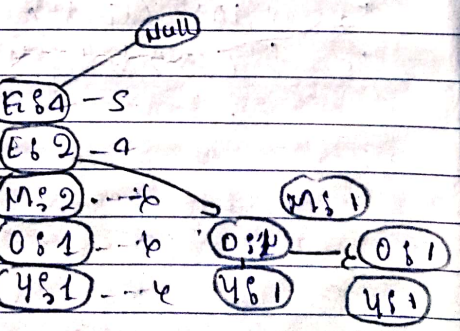
Item	Support count
M	3
O	3
N	2
K	5.
E	4
Y	3
D	1
A	1
U	1
C	2
I	1

L	
K	5
E	4
M	3
O	3
Y	3

frequint Pattern.

ordered Itemsets.
K E M O Y.
K, E, O, Y
K, E, M
K, M, Y
K E, O

Item ID	S-C	N.L
K	5	→
E	4	→
M	3	→
O	3	→
Y	3	→



2020-5-5 09:38

3.b

FP Tree

Conditional FP tree	Frequent Pattern
{K : 3}	{K, Y : 3}
{K : 3}	{K, O : 3}, {E, O : 3}
{K : 3}	{O, E, K : 3}
{K : 3}	{M, K, M : 3}
{K : 4}	{E, K : 3}

3.b) How do you construct FP-tree effectively to use in the FP-growth algorithm? Discuss with an Example

FP Growth Algorithm:

This algorithm is an improvement to the Apriori method A frequent pattern is generated without the need for candidate generation FP Growth Algorithm Represents the data base in the from of a tree called Frequent pattern tree or FP tree

This tree structure will maintain the association between the itemsets. The database is fragment using one frequent item This fragment part is called "pattern fragment

each node of the FP tree Represents an item of the itemsets. The Root Node represents Null while the lower Nodes Represents the itemsets the association of the node with the lower nodes that is the item sets with the itemsets are Maintained while forming

Table 2.

Item	count
I_1	4
I_2	5
I_3	4
I_4	4
I_5	2

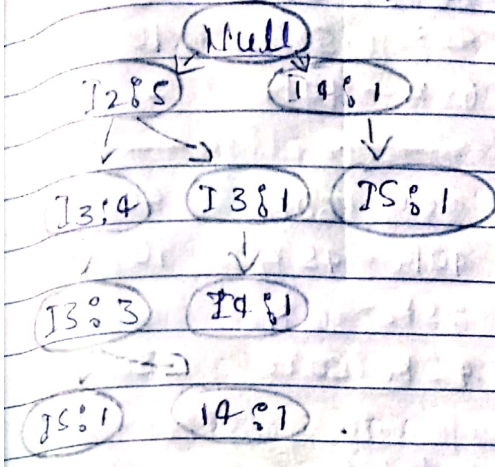
2) Sort the Itemset in descending order

Table - 3

Item	count
I_2	5

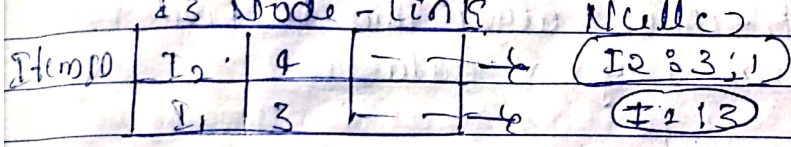
Item	conditional pattern Base	conditional FP tree
I_4	$\{I_2 I_1 : 1\}$	$\{I_2 : 2\}$
	$\{I_2 I_3 : 1\}$	$I_5 : 2\}$

Build FP-tree



Item	Frequent Patterns generated
I4	{I2 I4 : 2}, {I3 I4 : 2}, {I2, I3 I4 : 2}
I3	{I2 I3 : 4}, {I1 I3 : 3}, {I2, I1 I3 : 3}
I1	{I2 : I1 : 4}

The diagram given below depicts the conditional FP tree associated with the conditional node I3



4 a) The following table consists of training data from an employer database. The data have been generalized. For example "30-35" for age represents the age range of 31 to 35. For a given row entry count represents the No. of data tuples having the values for Department, Status, Age & Salary given in Row

Department	Status	Age	Salary	count
Sales	Senior	31-35	46K-50K	30
Sales	Junior	26-30	26K-30K	40

Systems	Senior	35 - 40	60K - 65K	15
Systems	Junior	30 - 35	45K - 50K	10
Systems	Senior	30 - 35	65K - 70K	2
~~Systems~~	—	—	—	—
Marketing	Junior	35 - 40	25K 30K	40
~~Secretary~~	Senior	46 - 50	40K - 45K	10
Secretary	Junior	26 - 50	25K - 30K	15
"	Senior	46 - 50	45K - 50K	3

Let Status be the class label construct a Design Tree from the given data

40 b] Discuss K-mean algorithm to from the clustus from a given dataset

=> K-Means algorithm is a Iterative algorithm that tries to partition the dataset into Kpre defind distinct non overlapping sub groups to only group

The way Kmeans algorithem works it as follows:-

Step 1 -> specify Number of cluster K

Step 2 -> Initialize centroids By first shuff the dataset and then randomly selecting K data points for the centroids without Replacement

Step-3 - Keep Iterating until there is no change to the centroids i.e assignment of data points to clusters isn't changing

- compute the sum of the squared distance between data points and all centroids
- assign each data point to the the closest cluster centroid)
- compute the centroid for the clusters by taking the average of the all data points that belong to each cluster

The approach K means follows to solve the problem is called Expectation maximization

This algorithm aims at minimizing an objective function know as squared error function given By

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (\|xi - vj\|)^2$$

Where $\|xi - vj\|$ is the euclidean distance between xi & vi points in ith cluster ci is the Nor of data clusters centers

5 a) What are outliers & Address the various challges of outlier detection process

An outlier is a object that deviates significantly from the rest of the objects They can be caused By measurement or execution errored to as outlier analysis or outlier mining

- Outliers ≠ Noise data.

2020-5-5 09:38

2020-5-5 09:38

Challenges of outlier Detection

1) Modeling Normal objects and outliers Properly
2) Application-specific outlier detection
3) Handling noise in outlier detection
4) Understandability

1) Modeling Normal objects & outliers properly:
Hard to enumerate all possible Normal behaviours in an application The border between Normal & outlier objects is often a gray area

2) Application-specific outlier detection
choice of distance measure among objects and the model of Relationship among objects often application dependent

Ex: clinic data a small deviation could be an outlier while in marketing analysis large fluctuations

3) Handling noise in outlier detection:-
~~Understanding ability:~~
Noise may distort the Normal objects & blur the distinction the Normal objects & outliers It may help hide outliers & Reduce effectiveness of outlier detection.

4) Understandability:
Understandability why these are outliers

and outlier :- The unlikehood of the object being generated by a Normal mechanism

Suppose a city's average temperature (in celsius) in july in the last 10 year are

28.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3 & 29.4

Indentify outliner using maximum likelihood method

$\Rightarrow \mu = \hat{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

$\hat{\mu} = \frac{29.0 + 28.9 + 28.9 + 29.0 + 29.1 + 29.2 + 29.2 + 29.3 + 29.4}{10}$

$= 28.61$

$\hat{\sigma}^2 = \frac{(29.0 - 28.61)^2 + (28.9 - 28.61)^2 + (28.9 - 28.61)^2 + (29.0 - 28.61)^2 + (29.1 - 28.61)^2 + (29.1 - 28.61)^2 + (29.2 - 28.61)^2 + (29.2 - 28.61)^2 + (29.3 - 28.61)^2 + (29.4 - 28.61)^2}{10}$

$= \frac{21.25 + 0.0841 + 0.0841 + 0.1521 + 0.2401 + 0.2401 + 0.3481 + 0.3481 + 0.4761 + 0.6241}{10}$

$= \frac{23.8469}{10}$

$\hat{\sigma}^2 = \sqrt{2.3} \quad = \hat{\sigma} = 1.51$

6 a) Briefly outline how to compute the dissimilari between object described by the following

i) Nominal attributes

A categorical variable is a generalization of the binary variable in that it can take on more than two states. The dissimilarity between two objects. i & j can be computed based on the ratio of mismatches: $d(i,j) = (p-m)/p$

where m is the No of matches (i.e the No of variables) for which i & j and are in the same state) & p is the total No of variables

Ex:

object Identifier	test-1 nominal	test-2 ordinal	test-3 numeric
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	Excellent	28

only object-identifier & the attribute test-1 are available when test-1 is Nominal

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

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

2020-5-5 09:38

ii)

Scanned with CamScanner

from this we se that two objects dissimilar exampt objects 1 & 4 (i.e) d(4,1) = 0)

ii) Asymmetric binary attributes.

A binary attributes is Asymmetric is out come of the states are Not equally important such as the positive & Negative outcome of a medical test for HIV test (e.g) The Most importa nt outcome (usually rare) is. Where as the Next one is '0'

It is all binary variables have the same weight we have the contingency Table

objects

	1	0	sum
1	q	r	q+r
0	s	t	s+t
sum	q+s	r+t	p

In computing the dissimilarity between asy mmetric binary variables the Nor of Negative maths is considered unimportant & thus is ignored in the computation that is

iii) Numeric Attributes -

use Euclidean distance Manhattan distance or supremum distance Euclidean distance & defined as.

The Manhattan (or city block) distance is defined as
$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|$

iv) Term Frequency vectors :-
To measure the distance between complex objects represented by vectors it is often easier to abandon traditional metric distance computation & introduce a nonmetric similarity function ex: similarity between two vectors x & y can be defined as cosine measure as follows

$$s(x,y) = \frac{x^t y}{\|x\| \|y\|}$$

where x^t is a transposition of vector x $\|x\|$ is the Euclidean norm of vector y & s is the essentially the cosine of the angle between vectors x & y

6 b) Explain data warehouse design process with Example

A data warehouse is a single data repository where a record from multiple data sources is integrated for online business analytical processing (OLAP) This implies a data warehouse needs to meet the requirement from all that ~~busi~~ business stages within the entire organization Thus data warehouse design is hugely complex lengthy & hence error prone process

Data warehouse design takes a method different from view materialization in the industries It sees data warehouses as data ba-

transformed & loaded (ETL) to be organized in a database as the data ware house

There are two approaches

1) "Top-down" approach
2) "bottom up" approach

1) Top down Design approach :-

In this design a data warehouse is described as a subject-oriented time variant Non-volatile & integrated data Repository for the entire enterprise data from different sources are validated reformatted saved in a Normalized (up to 3NF) data base of the data ware house.

advantages :-

1) data marts are laded from the data warehouses is very easy

disadvantages :-

the locations of the data warehou & the data marts are reversed in the bottom-up approach design

7 (a) A database has five transactions let min Support = 60% & Min-conf = 80%

TID	Item-bought
T100	{M, O N K E Y}
T200	{D O N K E Y}
T300	{M A K E}
T400	{M U C K Y}
T500	{C, O O K I E}

Find all frequent item sets using Apriori algorithm

C1

Item set	count
A	2
C	1
D	1
E	4
K	5
M	3
N	2
O	3
U	1
Y	3

Itemset	count
E	4
K	5
M	3
O	3
Y	3

C2

Item	count
Set	-a
EK	2
EM	3
EO	2
EY	3
KM	3
KO	3
KY	4
MO	1
MY	2
OY	2

Item	Count
Set	
E K	4
E O	3
K M	3
K O	3
K Y	3

C3

Item set	count
EKO	3
KMO	1
KMY	2
KOY	2

Item set	count
EKO	3

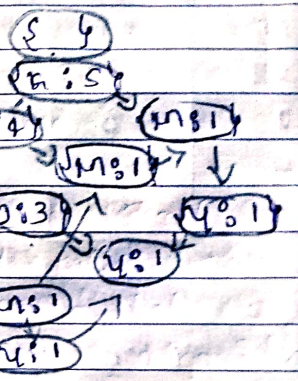
FP Growth

Item	Frequency
K	5
E	4
O	3
M	3
Y	2
N	2
C	1
D	1
A	1
V	1
I	1

TID	Item bought	Order frequ Item
T100	M,O,N,K,E,Y	K,E,O,M,Y
T200	D,O,N,K,E,Y	K,E,O,Y
T300	M,A,K,E	K,E,M
T400	M,U,C,K,Y	K,M,Y
T500	C,O,O,K,I,E	K,E,O

Header table

Item	Frequency	Head
K	5	
E	4	
O	3	
M	3	
Y	3	



Item	conditional Pattern base	F. Frequent itemset the pattern base	Frequent pattern generated
E	{K:4}	{K:4}	(K,E:4)
O	{K,E:3}	{K:3, E:3}	(K,O:3) (E,O:3) (K,E,O:3)
M	{K,E,O:1} {K,E:1} {K:1}	{K:3}	(K,M:3)
Y	{K,E,O,M:1} {K,E,O:1} {K,M:1}	{K:3}	(K,Y:3)

list all the association rules matching the following metarule where x is a variable representing customer & item denotes variably representing items (Ex - B, B' etc) for all x in transaction

performance

=6 metrics.	Formula	Evaluating Focus.
Accuracy (Acc)	$\frac{tp+tn}{tp+tp+tn+fn}$	In general the accuracy Metric measures the Ratio of correct predictions over the total number of instances evaluated.
Error Ratio	$\frac{fp+fn}{fp+fp+fn+fn}$	Misselassification error measures the Ratio of incorrect predictions over the total Nor of instance evaluate
Sensitivity	$\frac{tp}{tp+fn}$	This metric is used to Measure the fraction of Negetive patterns that are correctly classified
Precision (p)	$\frac{tp}{tp+fn}$	This metric is used to Measure the fraction of negative patterns that are correctly classified
Recall (R)	$\frac{tp}{tp+tn}$	Recall is used to measure. the fraction of +ve patterns that are correctly classified
F-Measure (FM)	$\frac{2*p*r}{p+r}$	This metric Represent the harmonic mean between Recall & problem precision values
Geometic mean (GM)	$\sqrt{+p+tn}$	The Metric is used to maximize the +p Rate & +e Rate & simultaneous keeping both Rates Relatively ablanced

Averaged Error Rate	$\sum_{i=1}^{l} \frac{fpi + fni}{tpi + fni + fpi + tni}$ / l	The average error Rate of all classes
Averaged Recall	$\sum_{i=1}^{l} \frac{tpi}{tpi + fni}$ / l	The average Per-class Recall
Averaged F Measure	$2 * \frac{Pm * rm}{Pm + rm}$	The average of per calls F - measure

Write a shorts Note on the following

a) data Normalization

Normalization is used to scall the data of an attribute so that it falls in a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0 It is generally usefull for clasification algori

Methods of data Normalization

1) Decimal Scaling Method

It Normalized by moving the decima point of values of the data To norm the data by this technique we divide each va of the data by the Maximum absolute value of data the data value v_i of data The by us the formula below

$v_i = v_i / 10^j$

Where j is the smallest integer Such that $\max(|v_i'|) < 1$

2020-5-5 09:37

Ex: Let the input data is: -10 201, 301, 401 501, 601 701

To Normali the above data,

Step-1: Maximum absolute value in given data (m) is 701

Step 2: divid the given data by 1000 (i.e) = 3)

Result: The Normalized data is:

0.01, 0.201, 0.301, -0,401, 0.501, 0.601 0.701

2) Min-Max Normalization:—

To this technique of data Normali-zation linear transformation is perfor medon the orginal data minimum & maximum value from data is fetched & each value is Replaced according to the following formal

$v' = \frac{v - min(A)}{Max(A) - min(A)} (New - max(A) - New - min(A)) +$

Where A is the following attribute data

min(A), max(A) are the minimum & maximum absolute value of A Respectively

v' is the New value of each entry in data

v is the old values of each entry in data

New-max(A), New-min(A) is the Max & min values of the Range Respectively

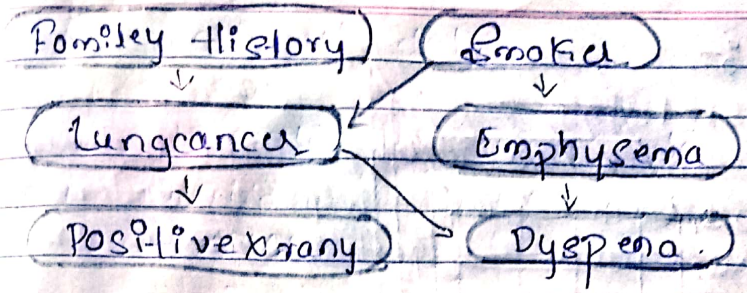
3) Z-score Normalization:—

In this technique values are Normalized Mean & standard deviation of the

Bayesian belief Networks

- Bayesian Belief Networks specify joint conditional probability distributions They are also known as Belief Networks
- * A Belief Network allows class conditional independencies to be defined between subsets of variable
- * It provides a graphical model of causal Relationship on which learning can be performed
- * We can use a trained Bayesian Network for classification

Two components that define a Bayesian Belief Network



2) A set of conditional Probability table :-

The conditional probability table for the values of the variable lung cancer (LC) showing each possible combination of the values of its parent nodes Family History (FH) Smoker (S) is as follows

	FH,S	FH,~S	~FH,S	~FH,~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

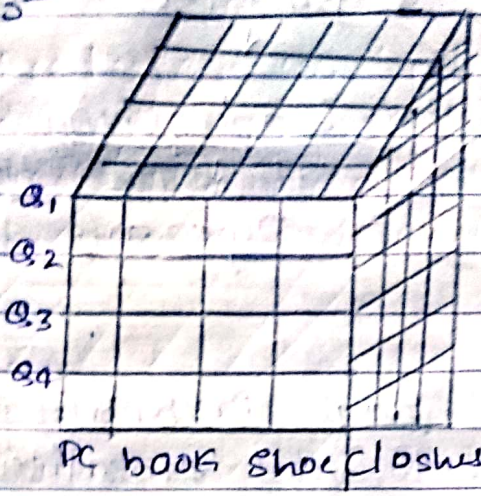
C) OLAP

Online Analytical Processing

OLAP is a category of software that allows users to analyze information from different points of view

OLAP databases are divided into one are more cubes The cubes are designed in such a way that creating & viewing Reports become easy

OLAP cube :-



At the core of the OLAP concept is an OLAP cube The OLAP cube is a data structure optimized for very quick data analysis

The OLAP cube consists of numeric facts called Measures which are categorized by dimensions OLAP cube also called a hypercube OLAP contains Multidimensional data with data usually obtained from a different & unrelated source.

2020-5-5 09:37