CLASSMATE
Date 15/07/19
Page

DATA MINING

Knowledge Discovery Database (KDD)
- Data preprocessing
- Data Mining
- Data post processing

data cleaning
data Transformation
data integration
data selection

1) pattern evaluation
2) " filtering
evaluation
3) Cluster visualization

Traditional Data Analysis
- Scalability
- Out of core algorithm
- Dimensionality
- heterogeneous

KDD - converting raw data into useful info.

Applications of DM
- Business
- Medicine, Science & Engineering

Challenges
- Scalability
- High Dimensionality
- Heterogeneous & Complex Data
- Data ownership & distribution
- Non-traditional analysis

Origins of Data Mining

Data Mining
- Database & DataWarehouse
- Statistics - sampling
- Information Retrieval - find of imp. topics
- Parallel Computing
- Mle learning — supervised learning → classification
  - unsupervised " → clustering (organize data)
  - semi-supervised "
  - active " → requesting the i/p on labels.

→ Explanatory variable & Class labels.

Data Mining Task

Predictive modelling

create
- explanatory variables to derive
- target variables

Animal name	Animal type	Habitat	class
Elephant	Terrestial	Forest	Herbivore

↳ Classification → used for discrete data
↳ Regression → continuous data (floating pt values) temp, pressure, wind

Descriptive Modelling

↳ association mining → associated values
↳ laptop (Lenovo)...
↳ Clustering (grouping closely related data)
↳ anomaly detection (significantly different data) (Credit card, IPRs)

Pattern interestingness

→ valid on new data → potentially useful
→ easily understood
→ novel, if it validates a hypothesis

→ useful.

→ interesting patters represents knowledge

~~Objective measures of pattern~~

Objective Interestingness _ bassed on structure of discoved patterns

↳ Rule support

↳ P(X∪Y)

It is a percentage of transaction that contains X∪Y

It is defined as P(X∪Y)

↳ Confidence : tells association b/n either X or Y

→ It accesses the degree of certainty of a detected association.

It is defined by P(Y|X)

Conditional probability

↳ Accuracy

Defines the percentage of transaction correctly classified by the rule. If X → Y

-Note :- Here both X and Y should be satisfied

↳ Coverage

Percentage of transaction covered by the rule

Note : Only X should be satisfied.

Subjective interestingness : based on user beliefs in data

↳ unexpected : contradicts his belief/pattern

↳ expected : validates his belief

↳ actionable : gives info. on which you can take action → taking the measurements

Data Characterisation / Data Discrimination

Target class:
Summarization of general features which describes the target class
Ex: Customer profiles (Credit rating, occupation)

Data Discrimination: Finding the discriminating features.
- class descriptn
- concept "

Data

- raw facts & observation.

Data Object (Record) / Vector

Data set: Collection of data object which are categorized by certain characteristics

Imp
- Attribute / feature / characteristics
- Measurement Scale
- Measurement process

Types of attributes

- based on number properties
  - ↳ Distinctness = $\neq$
  - ↳ Order <, >
  - ↳ addition +, −
  - ↳ multiplication *, /

Categorical / Qualitative Attributes

distinctness : ↳ nominal : distinguishes one data from another

order : ↳ ordinal : " one data object from another

↳ Product rating.

Numerical / Quantitative Attributes

additive ↳ Interval attribute : finding difference is useful eg. temp

addi & multip ↳ Ratio " : eg: age

→ Based on no. of values

↳ discrete attribute : it has finite no. of values or countably infinite no. of values

eg: roll no,

↳ Continuous attribus : contains floating pt. values

eg. stock price, weather

↳ Binary attributes :- 1 or 0

↳ assymetric attribute :- non-zero values are imp.

eg: medical profiles like diabetes measur

Characteristics of Data Set

↳ Dimensionality (attri) : Curse of dimensionality

→ if no. of dimensions ↑se, it is diffic to process data.

↳ Sparsity → feature of dataset where only non-zero valu are stored.

↳ Resolution → value of data object varies with different levels of resolution.

eg: satellite images.

eg: data varies according to time.

Types of Data Set collection of

(I) (1) Record Data : A data set with fixed no. of fields.

(1) ↳ Transaction / Market Basket Data
- Each record is transaction
- Each attribute is item

T_1	milk, eggs
T_2	Cheese, milk, flour.
T_3	Cake, detergent
T_4	jam, bread, milk

Trans Id	milk	eggs	cheese	flour	Detergent	Cake	Jam	bread
T_1	1	1	0	0	0	0	0	0
T_2	1	0	1	1	0	0	0	0

Record Data

(2) ↳ Data Matrix : used for numerical data & each pt is plotted as multidimensional

m rows x n columns

row → data object

column → attributes

Sepal length	Sepal width	petal length	petal width
7.5	3.5	2.5	3.5

(3) Sparse Data Matrix

↳ assymetric attribute

term present ↓ term not-present ↓

eg: document term matrix : represented by 1 or 0 & term frequency

Doc No	Term 1	Term 2	Term 3
1	3	1	0

→ Here only non-zero values are important.

(II) Graph Data

(i) node → data object → we show link b/w diff. data object where each object is a node
eg. webpages

(ii) Displaying the relationship b/w sub-objects
eg: represent" of chemical compound.

→ Graph → for one data object
↳ each data object → contains sub object
nodes → sub objects

(IST) Ordered Data
attributes → linked to time/space

1] Sequential Data : should have 1 attri which is related to time
eg. login access.

Login Time	Logout Time	Username

2] Sequence data : order of sequence is important.
eg: genetic info (plants) (contains seq. of individual entities)

3] Temporal data / time series data : capturing data according to the time, we are not storing the time

↳ Temporal auto correlation : rainfal in June is diff. from rainfall in Aug

4] Spatial Data (area, position) : data which is represented as latitude & longitude

↳ Spatial - auto correlation : values which are close in space is similar

Date
Page

Data Cleaning & Data Quality

aspects of data quality affects
- Measurement error : it results from measurement process
- Data Collection : human errors (missing data ...)

Noise :- Random variation in the attribute.
- attribute noise :- missing, incorrectly entered
- class noise :- Wrong class lable

Artifact :- deterministic distortion of data.

Techniques to handle noise :
(1) Binning (Data Smoothing)
(2) Regression
(3) Outliers.

(1) Binning
- Unsupervised binning :- target class info is not available
  - → equal width binning
  - → equal depth binning : no. of element frequency same
    or equal frequency binning
- * Supervised binning
  - → entropy based binning

Equal width binning
→ divide data into k intervals bins which will have equal width

* Equal Depth binning
- sort the data

eg: 0, 4, 12, 16, 16, 18, 24, 26, 28 k=3.

bin 1 0, 4, 12

bin 2 16, 16, 18

bin 3 24, 26, 28,

* Data Smoothing Technique

(1) bin means. → Find mean & replace the values as mean

bin 1 5, 5, 5

bin 2 17, 17, 17

bin 3 26 26 26

(2) bin boundaries - Closest value

bin 1 0 0 12

bin 2 16 16 18

bin 3 24 24 28

(3) bin median - Find median

bin 1 4 4 4

bin 2 16 16 16

bin 3 26 26 26

* Equal Width binning :-

$$w = \frac{max - min}{k}$$

Intervals

$[min, min+w)$

$[min+w, min+2w)$

:

$[min+(k-1)w, min+kw]$

classmate
Date
Page

binning is used in
→ noise handling
→ discretization

eg: 0, 4, 12, 16, 16, 18, 24, 26, 28

k = 3

$w = \frac{28-0}{3} = \lceil 9.33 \rceil \approx 10$

[0, min+w]

[0, 10) [0, 4]

[10, 20) [12, 16, 16, 18]

[20, 30] [24, 26, 28]

bin means

bin 1 [2 2]

bin 2 [16 16 16 16]

bin 3 [26 26 26]

bin median

bin 1 [2 2]

bin 2 [16 16 16 16]

bin 3 [26 26 26]

bin boundaries

bin 1 [0 4]

bin 2 [12 18 18 18]

bin 3 [24 24 28]

* Supervised Binning :- When target class info is available

→ Entropy based binning :- measures impurity measure

$entropy(e_i) = -\sum p_{ij} \log_2 p_{ij}$

j = no. of class
ith interval

$p_{ij} = \frac{m_{ij}}{m_i}$

m_{ij} ⟶ no. of record having class j in i^{th} interval

m_i → no. of record in the i^{th} interval

net entropy = $\sum w_i e_i$

$$w_i = \frac{m_i}{m}$$

m → total no. of records.

Gain = Total entropy of data - net entropy (attribute)

Hours Studied	Aon Test
4	N
5	Y
8	N
12	Y
15	Y

$$\text{entropy (D)} = -\left[\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)\right]$$

$$= 0.97099 \simeq 0.971$$

Take adjacent values (4, 5) → find average = (4+5)/2 = 4.5

	Y: Aon Test	N: Lower than 4
<=4.5	0	1
>4.5	3	1

$$\text{entropy } (<=4.5) = -\left[\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) + \left(\frac{0}{1}\right)\log_2\left(\frac{0}{1}\right)\right]$$

$$= 0$$

$$\text{entropy}(>4.5) = -\left[\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) + \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right)\right]$$

$$= 0.811$$

$$\text{net entropy} = \frac{mi}{m} = \left(\frac{1}{5}\right) \times 0 + \left(\frac{4}{5}\right) \times 0.811$$

$$= 0.6488$$

Gain = entropy(D) − netentropy(for 4.5)

= 0.971 − 0.6488

= 0.328

		Entropy
Split value (4.5)	Gain = 0.328	0.6488
" value (6.5)	Gain = 0.232	0.9538
" value (10)	Gain = 0.4172	0.5538
" value (13.5)	Gain = 0.171	0.8

Entropy bin 1 <=10 4.5, 8

bin 2 >=10 12.15

→ For split value (6.5)

	Y	N
<= 6.5	1	1
> 6.5	2	1

$$\text{entropy}(<=6.5) = -\left[\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right]$$

$$= 1$$

$$\text{entropy}(>6.5) = -\left[\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) + \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right]$$

$$= 0.8899$$

net entropy = $\frac{m_i}{m} = \frac{2}{5} \times 1 + \frac{2}{5} \times 0.8899 = 0.93394$

Gain = entropy(D) − net entropy(for 6.5)
= 0.971 − 0.93394
= 0.03706

→ For split value (10)

	Y	N
<= 10	1	2
> 10	2	0

entropy(<=10) = $-\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right]$

= 0.9182

entropy(>10) = $-\left[\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right]$

= 0

net entropy = $\frac{m_i}{m} = \frac{3}{5} \times 0.9182 + \frac{2}{5} \times 0 = 0.55092$

= 0.551

→ Gain = entropy(D) − net entropy(for 10)
= 0.971 − 0.9182
= 0.0528

→ For split value (13.5)

	Y	N
<= 13.5	2	2
> 13.5	1	0

$$\text{entropy}(x=13.5) = -\left[\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right]$$

$$= 1$$

$$\text{entropy}(>13.5) = -\left[\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right]$$

$$= 0$$

$$\text{net entropy} = \frac{m_i}{m} = \frac{4}{5}\times 1 + \frac{1}{5}\times 0$$

$$= 0.8$$

Gain = entropy(D) − net entropy (for 13.5)

= 0.971 − 0.8

= 0.171

Noise Reduction

(1) binning

(2) Regression : use to predict the value of $y = mx + c$

y = should contain noise

x = should not contain noise

⇒ **Linear Regression**

~~Interpolate~~

If you are trying to predict the y within known range

Interpolation: points within the curve

Extrapolation: points outside the curve

Std deviation

~~Std deviation~~

x	y	xy	x^2
1	2	2	1
2	5	10	4
3	3	9	9
4	8	32	16
5	7	35	25
$\Sigma x = 15$	$\Sigma y = 25$	$\Sigma xy = 88$	$\Sigma x^2 = 55$

$y = mx + c$

for 'c' equation, $\Sigma y = n * c + m * \Sigma x$

for 'm' equation, $\Sigma xy = c * \Sigma x + m \Sigma x^2$

$25 = 5c + 15m$ — (1)

$88 = 15c + 55m$ — (2)

$c = 1.1$

$m = 1.3$

$$\boxed{y = 1.3x + 1.1}$$

$y = \beta_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \ldots \alpha_n x_n$

(3) Outlier : pt which doesn't fit region ^ containing low density

- precision : closeness of repeated measurement taken to one another
- bias : systematic variation of measurement from known value
- accuracy : closeness of given measurement to the true value.

measurements = 1.015;
0.990
1.013
0.986
1.001

Known value = 1 kg

Mean = ~~0.18038~~ 1.001

Bias = mean − known value
= 1.001 − 1
= 0.001

Precision = std deviation

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \text{mean})^2}$$

$$= \sqrt{\frac{1}{5}\left[(1.015 - 1.001)^2 + (0.990 - 1.001)^2 + (1.013 - 1.001)^2 + (0.986 - 1.001)^2 + (1.001 - 1.001)^2\right]}$$

= 0.0117

→ bias: diff. b/w measurement & known value

Outliers

Anomalous attribute

Anomalous data object :- if characteristics of one data object is differ too large from other data objects

Unusual value = anomalous

Techniques to handle missing value

→ Missing value :-
→ info was not collected
↳ Attribute is not applicable for all records.

29/07

classmate
Date
Page

Techniques to handle missing values

1) eleminate data object / attribute

2) Estimate missing values. → Find avg
Continuous → estimate missing value

Jan	Feb	March
31	32	33
71	(x)	81
81	93	1.3

Categorical In nominal → find frequency
↓
most commonly occuring attrb.

3) Ignore missing value (which performing data mining)

Inconsistent Data (Invalid data)

Eg. -ve height

Duplicate data (Multiple entities for some object)

- Same entries : keep only one.
- different entries : keep most correct entry.

AcctNo	CustName	Address	balance
1003	John	Buston	1000
1003	John	Houston	3000

analyse which is correct

- Similar objects are not duplicates.

Deduplicating

Removal of duplicate entries from a data set

Issues related to application (Desirable properties)

→ Timeliness (Data set should be upto date
eg. Century old rainfall data will produce wrong pattern.
→ Relevance (Dataset should be useful for application)
→ Knowledge about data (Description of the data)

Variable transformation

$\log_{10}(x)$ e^x

Normalisation (ensure data fits within a range) / Standardization

* **Z-score normalisation** (Range of attributes not known)

$$V_i' = \frac{V_i - \mu_x}{\sigma_x}$$

$\mu_x \rightarrow$ mean of attribute x

$\sigma_x \rightarrow$ std deviation of attribute x

$$\mu_x = \frac{V_1 + V_2 + \dots + V_n}{n}$$

$V_1, V_2, \dots V_n$ – value of attribute

n – no. of attribute 'x'

$$\sigma_x = \sqrt{\frac{1}{n}(V_1 - \mu_x)^2 + (V_2 - \mu_x)^2 + \dots + (V_2 - \mu_x)^2}$$

* **Min-max normalisation** (Range of attribute is known)

$$V_i' = \frac{V_i - \min_A}{\max_A - \min} (\text{newmax}_A - \text{newmin}_A) + \text{newmin}_A$$

* **Decimal Scaling**

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

$j \rightarrow$ smallest integer such that $\max(|V_i'|) < 1$

$j = \log_{10}(\max(v))$ ⟸ ... to nearest integer

Q. 24, 29, 31, 32, 32 | 32, 37, 42, 42, 44 | 44, 64, 68

newmin = −2 | newmax = 8

$\mu_x = 40.07$

$(24 - 40.07)^2 = 258.244$

$(29 - 40.07)^2 = 122.54$

$(31 - 40.07)^2 = 82.26$

$(32 - 40.07)^2 \times 3 = 195.37$

$(37 - 40.07)^2 = 9.42$

$(42 - 40.07)^2 = 7.44$

$(44 - 40.07)^2 = 30.88$

$(64 - 40.07)^2 = 572.64$

$(68 - 40.07)^2 = 780.08$

2058

$$\sigma x = \sqrt{\frac{1}{13} \times (2058)} = 12.58$$

$$\sigma x = \sqrt{158.374} = 12.58$$

new value,

$$(24) = \frac{24 - 40.07}{12.58} = -1.2774$$

$$\text{new value } (29) = \frac{29 - 40.07}{12.58} = -0.8799$$

new value (31) = $= -0.7209$

(32) = $= -0.6414$

(37) = $= -0.2440$

(42) = $= 0.1534$

(44) = $= 0.3124$

(64) = $= 1.9022$

(68) = $= 2.2201$

Using decimal scaling

$\max(v) = 68$

$j = \log_{10}(68) = 1.83 \simeq 2$

$$\text{new value } (24) = \frac{24}{10^j} = \frac{24}{10^2} = 0.24$$

(31) z = = 0.31

(32) z = = 0.32

(37) z = = 0.37

(42) z = = 0.42

(44) z = = 0.44

(64) z = = 0.64

(68) = = 0.68

Using min-max

new value (24) = $\frac{24-24}{(68-24)}$ (8-(-2)) + (-2) = -2

new value (29) = $\frac{29-24}{(68-24)}$ (8-(-2)) + (-2) = -0.8636

new value (31) = = -0.4090

min-max normalization

old min = 24 new min = -2

old max = 68 new max = 8

new value = $\frac{\text{old value} - \text{old min}}{(\text{old max} - \text{old min})}$ (new max - new min) + new min

new value (24) = $\frac{24-24}{68-24}$ (8-(-2)) + (-2) = -2

new value (29) = $\frac{29-24}{68-24}$ (10) - 2 = -0.86

new value (31) = $\frac{31-24}{68-24}$ (10) - 2 = -0.409

new value (32) = $\frac{32-24}{44}$ (10) - 2 = -0.18

new value (37) = $\frac{37-24}{44}$ (10) - 2 = 0.95

new value (42) = $\frac{42-24}{44}(10)-2 = 2.09$

new value (44) = $\frac{44-24}{44}(10)-2 = 2.5$ [-2, 8]

new value (64) = $\frac{64-24}{44}(10)-2 = 7.09$

new value (68) = $\frac{68-24}{44}(10)-2 = 8$

* Data Preprocessing

1] Aggregation :→ Combining one or more data together.
→ Reduce the no. of attribute values.

Adv
→ less memory
→ less time

ex:

Storename	Itembought	Purchasedate	Item Amt
Chicago	Watch	7/07/19	800
Chicago	Wallet	17/7/19	900
Chicago	Pencil	20/7/19	200
Boston	Vase	25/7/19	1200
Boston	Cutlery set	27/7/19	2500

Categorical: Storelocation, Itembought

Store location	Item	Purchase date	Item Amt
Chicago	Giftset	7 - 20/7/19	1900
Boston	Household Items	25 - 27/7/19	3700

→ Advantage
- Less time, Less memory
- offers a high level view of data.

→ Disadvantages
→ Some data analysis → require low level view of data.

→ Sampling is done to reduce processing time

2] Sampling (picking up subset from the given set)

Nyquist theor
or Blog

Sample representativeness

Property of original set & sample set should be more or less same.

↳ Simple random sampling without replacement (SRSWOR)
↳ " " " with " (SRSWR)
↳ Cluster "
↳ Stratified "
↳ Progressive (adaptive) " :- Each iteration, addition done → checks accuracy → use when do not know actual sample size

→ SRSWOR : chances of picking any set is same (not removing)
→ SRSWOR : do not replace with original data set. (removing)
→ Cluster Sampling :- Use clusters to pickup a sample
eg: Given M Clusters
↳ choose S Clusters S<M

↳ Stratified : grouping the data

ex:

Youth	Middle Age	Senior Citizens
4	8	2

Sample size = 7

- Youth = $\frac{4}{14} \times 7 = 2$

no. of samples to be chosen from each group = $\left(\frac{\text{no. of objects in the group}}{N}\right) * \text{Sample Size}$

Middle age (no. of samples) = $\frac{8}{14} \times 7 = 4$

Senior Citizen = $\frac{2}{14} \times 7 = 1$

Sample :

Youth
Youth
Middle Age
"
"
"
Senior Citizen

Ex:

Brach	Comp	Civil	Production	Automobile	IT	ETC
No	52	38	36	46	44	64

Sample Size = 60

N = 280

- Comp = $\frac{52}{280} \times 60 = 11.1 \simeq 11$
- Civil = $\frac{38}{280} \times 60 = 8.14 \simeq 8$
- Production = $\frac{36}{280} \times 60 = 7.71 \simeq 8$
- Automobile = $\frac{46}{280} \times 60 = 9.85 \simeq 10$
- IT = $\frac{44}{280} \times 60 = 9.42 \simeq 9$
- ETC = $\frac{64}{280} \times 60 = 13.71 \simeq 14$

3] Dimensionality Reduction

Linear Algebra techniques : transforming old attr. to new attr.

↳ Principal Component Analysis

↳ Single Value Decomposition : AA^T

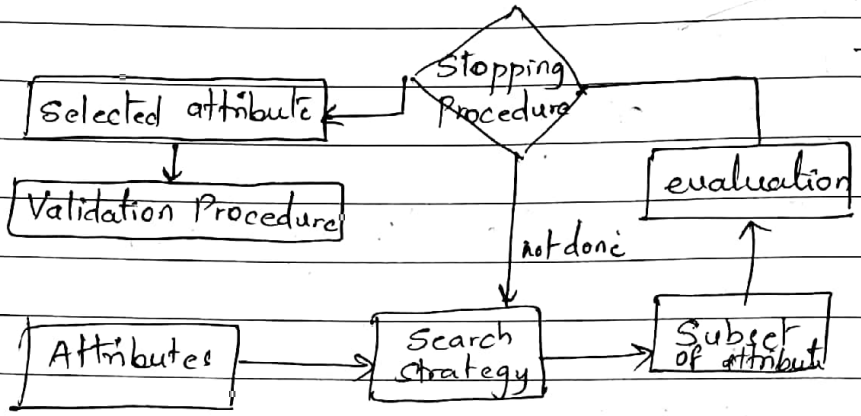
4] Attribute Subset Selection : choosing

↳ Irrelevant features

↳ redundant features : age & birthdate
eliminate one

→ embedded approach : decides which attrb. should be kept & which should be discarded.
→ filter " :- imp., not imp
→ Wrapper " :- (black-box) decides which subset should be kept & which should be discarded

Architecture for feature subset selection



Principal Component Analysis

x	y	z
7	8	13
8	9	14
11	13	18
13	10	8
Xmean	Ymean	Zmean

x − Xmean	y − ymean	z − zmean	(x−xmean)(y−ymean)	(x−xmean)(z−zmean)	(y−ymean)(z−zmean)
Con = $\frac{\Sigma(x-x_{mean})}{n}$					

$$Con(x,y) = \frac{\Sigma (x - x_{mean})(y - y_{mean})}{n}$$

$$\text{Covarience} = \begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix} \rightarrow A$$

$|A - \lambda I| = 0$

↓

quadratic equation

$\lambda_1 \quad \lambda_2 \quad \lambda_3 = 0$

eg. $\lambda_2 > \lambda_1 > \lambda_3$

$n \rightarrow k$

n dimension to k dimension

Where $k < n$

→ Stepwise forward selection :- (empty set initially) best attr is selected & added to the subset (method: gain...)

→ Stepwise backward selection: (Set has all elements) Remove the worst attrib from the set.

→ Combination of both: Putting best & removing worst

→ decision tree: - Internal nodes
- leaf nodes.

* Feature Weighting

5] * Feature Creation: understanding the data
↳ feature extraction: analyse the data ref creating new attributes
↳ mapping data to a new space: transmitting data (changing the range)
↳ feature construction: eg given birthdate we can find age (new attrib)

Identify sub-structure & create this substructure as attribute

6] (*) Discretisation: Converting continuous values into discrete
↳ Unsupervised discretization: (doesn't use class info) — Equal width
\ Equal depth / frequency

Range query

eg X
10 ≤ x < 20 low
13 " "
19 " "
20 ≤ x < 30 medium
21 " "
28 " "

↳ Supervised discretization: (use class info)

eg: of Unsupervised

bin1 11, 13, 19, 21 low

bin2 22, 26, 33, 37 high

Discrete $10 \leq x \leq 21$

$22 \leq x \leq 37$

(*) Supervised discretization: Centropy based binning)

- find split values
- Gain

7] (*) Binarisation

- If it is continuous → it has to be converted to discreate and then discrete to binary.

Categorical values: Excellent → 5 → 101

Very good → 4 → 100

Good → 3 → 011

~~bad~~ fair → 2 → 010

~~poor~~ bad → 1 → 001

Product Name	Excellent	Verygood	good	fair	bad
prod 1	1	0	0	0	0
prod 2	0	0	0	0	1

assymentric : 0 or 1

Drawback:

→ Wastage of space

→ Dimensionality ↑es

UNIT 2

10/08/19

Dissimilarities between Data Objects

Minkowski distance

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

If $r = 1$,

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

$|x_k - y_k|$ → different data objects

→ City Block / Manhattan distance / L_1 Metric

If $r = 2$

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

→ Euclidean distance / L_2 norm

If $r = \infty$,

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

→ Supremum distance / L_{max} or L_∞ metric

* Euclidean distance properties

positivity — $d(x,y) \geq 0$ for (x,y)

Symmetry — $d(x,y) = d(y,x)$

Triangle inequality — $d(x,z) \leq d(x,y) + d(y,z)$

Similarities between data object (for binary data)

Simple Matching Coefficient (SMC) = $\frac{f_{00}+f_{11}}{f_{00}+f_{11}+f_{01}+f_{10}}$

Note: Jaccard Coefficient only ~~work~~ suitable for assymetric binary attributes.

Jaccard Coefficient = $\frac{f_{11}}{f_{11}+f_{01}+f_{10}}$

Cosine Similarity (Primary used for document data)

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

$$x \cdot y = \sum_{k=1}^{n} x_k \cdot y_k$$

$$= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

$$\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2}$$

$$\|y\| = \sqrt{\sum_{k=1}^{n} y_k^2}$$

Q. $x = (0,-1,0,2)$ $y = (2,0,-1,0)$

Find the distance using Euclidean distance formula.

⇒

$$d(x,y) = \sqrt{(0-2)^2 + (-1-0)^2 + (0+1)^2 + (2-0)^2}$$

$$d(x,y) = \sqrt{4+1+1+4}$$

$$d(x,y) = \sqrt{10}$$

City block = $|0-2| + |-1-0| + |0+1| + |2-0| = 2+1+1+2 = 6$

0 → dissimilar
1 → similar

Cosine Similarity

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

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

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

$$x \cdot y = (0 \times 2) + (-1) \times 0 + 0 \times (-1) + 2 \times 0$$

$$x \cdot y = 0$$

$$Cos(x,y) = \frac{0}{\sqrt{5} \cdot \sqrt{5}}$$

$$= 0$$

Q. $x = (1,1,0,1,1,0,0,1)$

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

$\Rightarrow$ $f_{00} = 2$

$f_{11} = 3$

$f_{01} = 1$

$f_{10} = 2$

$$SMC = \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}}$$

$$SMC = \frac{2+3}{2+3+1+2} = \frac{5}{8} = 0.6$$

$$Jaccard = \frac{3}{3+1+2} = \frac{3}{6} = \frac{1}{2} = 0.5$$

Q.4) Given the foll. data set perform min-max normalization (new min = 0 & new max = 1)
2). z-score
3) Decimal Scaling

→ 200, 300, 400, 600, 1000

a) $$\text{New Value} = \frac{\text{old value} - \text{old min}}{\text{old max} - \text{old min}} (\text{new max} - \text{new min}) + \text{new min}$$

old min = 200, old max = 1000

$$\text{New value (200)} = \frac{200-200}{(1000-200)} (1-0) + (0) = 0$$

$$\text{New value (300)} = \frac{300-200}{(1000-200)} (1-0) + (0) = 0.125$$

$$\text{new value (400)} = \frac{400-200}{(1000-200)} (1-0) + 0 = 0.25$$

nv(600) = 0.5

nv(1000) = 1

b) z-score normalization

$$\frac{v' - v - \mu}{\sigma}$$

Mean (μ) = 500

Standard deviation (σ) = 282.84

$$SD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{5} [(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}$$

$$= 282.84$$

$$v' = \frac{200 - 500}{282.84} = -1.0607$$

decimal scaling

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

$$j = \log_{10}(\max(v)) = \log_{10}(1000) = 3$$

$$v' = \frac{v}{10^3} = \frac{200}{10^3} = 0.2$$

$$v'(300) = \frac{300}{10^3} = 0.3$$

Q. 5, 10, 11, 13, 35, 50, 55, 72, 92, 204, 215

k = 3.

Perform equal width & equal depth & apply data smoothing techniques for the foll.

⇒ EWB

$$W = \frac{\max - \min}{k} = \frac{215 - 5}{3} = 70$$

[5, 75)	5, 10, 11, 13, 35, 50, 55, 72
[75, 145)	92.
[145, 215]	204, 215

bin mean

bin1	31.37, 31.37, 31.37, 31.37, 31.37, 31.37, 31.37, 31.37
bin2	92
bin3	209.5 209.5

bin median

bin1 24, 24, 24, 24, 24, 24, 24, 24

bin2 92

bin3 209.5 209.5

bin boundaries

bin1 5, 5, 5, 5, 55, 72, 72, 72

bin2 92

bin3 204 215

EDB

Bin1 5, 10, 11, 13

Bin2 35, 50, 55, 72

Bin3 92, 204, 215

bin mean

bin1 9.75, 9.75, 9.75, 9.75

bin2 53, 53, 53, 53

bin3 170.3, 170.3, 170.3

bin median

bin1 10.5, 10.5, 10.5, 10.5

bin2 52.5, 52.5, 52.5, 52.5

bin3 170.3, 170.3, 170.3

bin boundary

bin1 5, 13, 13, 13

bin2 35, 55, 72, 72

bin3 92, 215, 204

UNIT-1

PCA

hours (h)	marks (m)
2	1
3	4
5	0
7	6
9	2

$h_{mean} = 5.2$

$m_{mean} = 2.6$

h - hmean	m - mmean	(h-hmean)(m-mmean)
-3.2	-1.6	5.12
-2.2	1.4	-3.08
-0.2	-2.6	0.52
1.8	3.4	6.12
3.8	-0.6	-2.28
		sum = 6.4

data adjust

$$\text{covariance}(h,m) = \frac{6.4}{(5-1)} = 1.6$$

$$\text{var}(h) = 8.2 = \frac{1}{n-1}\Sigma(h-h_{mean})^2$$

$$\text{Var}(m) = 5.8$$

Construct covariance matrix

$$\begin{bmatrix} \text{cov}(h,h) & \text{cov}(h,m) \\ \text{cov}(m,h) & \text{cov}(m,m) \end{bmatrix}$$

Scanned by CamScanner

$$A = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$$

Solve for $|A - \lambda I| = 0$

$$\left| \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} 8.2 - \lambda & 1.6 \\ 1.6 & 5.8 - \lambda \end{vmatrix}$$

$\lambda^2 - 14\lambda + 45 = 0$ → Use quadratic formula

$\lambda = 5, \quad \lambda = 9$

eigen values = [5, 9]
= [9, 5]

Solve for eigen vectors

$$\begin{bmatrix} 0.89444 & 0.4472 \\ 0.4472 & 0.8944 \end{bmatrix}$$

$$\begin{bmatrix} 0.89444 \\ 0.4472 \end{bmatrix}$$

X data adjust

$\frac{5}{2} \times \frac{2}{1} = \frac{5}{1}$ matrix.

↓

1 attribute is removed.

* Person's Correlation

$$s(x,y) = \frac{\text{covariance}(x,y)}{\text{standard\_deviation}(x) * \text{standard\_deviation}(y)}$$

n - entire population

n-1 - assume sample data set (doesn't represent entire population)

$$\text{Covariance}(x,y) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$\text{Standard\_deviation}(x) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

$$\text{Standard\_deviation}(y) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k$$

1. $x = (0, -1, 0, 2)$

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

$$\text{Covariance}(x,y) = \frac{1}{4-1}\Big[(0-0.25) \times (2-0.25) + (-1-0.25)(0-0.25) + (0-0.25) \times (-1-0.25) + (2-0.25).(0-0.25)\Big]$$

$$= -0.0833$$

Date
Page 41

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$= \frac{1}{4} \times [0 + (-1) + 0 + 2] = \frac{1}{4} = 0.25$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k$$

$$= \frac{1}{4} \times [2 + 0 + (-1) + 0] = \frac{1}{4} = 0.25$$

$$S(x,y) = \frac{\text{Covariance}(x,y)}{\text{standard\_deviation}(x) * \text{standard\_deviation}(y)}$$

$$\text{standard\_deviation}(x) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

$$= \sqrt{\frac{1}{4-1} [(0-0.25)^2 + (-1-0.25)^2 + (0-0.25)^2 + (2-0.25)^2}$$

$$= 1.258$$

$$\text{Standard\_deviation}(y) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$$

$$= \sqrt{\frac{1}{4-1} [(2-0.25)^2 + (0-0.25)^2 + (-1-0.25)^2 + (0-0.25)^2}$$

$$= 1.258$$

$$S(x,y) = \frac{-0.0833}{1.258 * 1.258}$$

$$= -0.0526 //$$

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

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$= \frac{1}{8} \times [1+1+0+1+1+0+0+1]$$

$$\bar{x} = 0.625$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

$$= \frac{1}{8} \times [0+0+0+1+1+1+0+1]$$

$$\bar{y} = 0.5$$

$$\text{Covariance } (x,y) = \frac{1}{n-1}\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$= \frac{1}{8-1} \times [(1-0.625) + (1-0.625) + (0-0.625) + (1-0.625) + (1-0.625) + (1-0.625) + (0-0.625) + (1-0.625)]$$

$$= 0$$

$$= \frac{1}{8-1} \times [(1-0.625)\times(0-0.5) + (1-0.625)\times(0-0.5) + (0-0.625)\times(0-0.5) + (1-0.625)\times(1-0.5) + (1-0.625)\times(1-0.5) + (0-0.625)\times(1-0.5) + (0-0.625)\times(0-0.5) + (1-0.625)\times(1-0.5)]$$

$$= \frac{1}{7} \times [-0.1875 + -0.1875 + 0.3125 + 0.1875 + 0.1875 + -0.3125 + 0.3125 + 0.1875]$$

$= 0.07142$

$$\text{Standard deviation}(x) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$= \sqrt{\frac{1}{7} \times [(1-0.625)^2 + (1-0.625)^2 + (0-0.625)^2 + (1-0.625)^2 + (1-0.625)^2 + (0-0.625)^2 + (0.625)^2 + (1-0.625)^2]}$$

$$= 0.5219$$

$$\text{Standard deviation}(y) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$

$$= \sqrt{\frac{1}{7} \times [(0-0.5)^2 + (0-0.5)^2 + (0-0.5)^2 + (1-0.5)^2 + (1-0.5)^2 + (1-0.5)^2 + (0-0.5)^2 + (1-0.5)^2]}$$

$$= 0.5345$$

$$S(x,y) = \frac{\text{Covariance}(x,y)}{\text{Standard deviation}(x) \times \text{Standard deviation}(y)}$$

$$= \frac{0.07142}{0.5219 \times 0.5345}$$

$$= 0.2560$$

Scanned by CamScanner

26/08/19 * Extended Jaccards Coefficient / (Tanimoto Coefficient)

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

(*) Bregman Divergences : Used to measure loss or distortion

$$D(x,y) = \phi(x) + \phi(y) - (\nabla\phi(y), (x-y))$$

↳ vector difference between (x,y)

* Issues with Similarity

1) → attribute with different scales (ranges) without correlation
↳ normalise data

→ attribute with different scale with correlation
↳ Mahalanobis distance → captures correlation
Mahalanobis

Mahalanobis distance

$$\text{Mahalanobis}(x,y) = (x-y)\,\Sigma^{-1}\,(x-y)^T$$

Σ^{-1} ↓ Inverse of Covariance matrix of [(x,y)]

2) → heterogeneous attributes : attrib. having diff. types of values
- find simmilarity b/n each attrib.
- check if it is assymetric attrib
(if it is assy. & both places have (0,0), should not be consider)

Similarities of heterogeneous objects

1] for k^{th} attribute, complete similarity $\delta_k(x,y)$ in range [0,1]

2] define an indicator variable δ_k

$\delta_k = 0$ → if k^{th} attrib → assymetric and both obj's have '0' value

or

one obj → have missing value

= 1 otherwise Wk: weight

3] Compute overall similarities

$$\text{Similarity}(x,y) = \frac{\sum_{k=1}^{n} \delta_k s_k(x,y)}{\sum_{k=1}^{n} \delta_k}$$

2) → Using weights

$$s(x,y) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(x,y)}{\sum_{k=1}^{n} \delta_k}$$

or

Using minkowski

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

* Summary statistics : measures to find diff features in data set

data exploration

Frequency & Mode

	Temp
Day1	low
Day2	high
Day3	medium
Day4	low
5	medium

$$\text{Frequency} = \frac{\text{no. of objects with attrib value 'k'}}{n}$$

$\text{freq(low)} = \frac{2}{5} = 0.4$

$\text{freq(medium)} = \frac{2}{5} = 0.4$

$\text{freq(high)} = \frac{1}{5} = 0.2$

mode = 0.4

→ Mode is the value having the highest frequency

Percentile

Given an ordinal or continuous attribute x, and a no. p between 0 and 100, the p^th percentile x_p is a value of x such that p% of the observed values of x are less than x.

			0.1 x tot no. of values	average	no. of value < xp
1	10%	10/100 = 0.1	= 0.1 x 10 = 1	$x_p = 1.5$	1
2	20%	20/100 = 0.2	= 0.2 x 20 = 2	$x_p = 2.5$	2
3	30%	0.3	= 0.3 x 30 = 3	3.5	3
4	40%	0.4	4	4.5	4
5	50%	0.5	5	5.5	5
6	60%	0.6	6	6.5	6
7	70%	0.7	7	7.5	7
8	80%	0.8	8	8.5	8
9	90%	0.9	9	9.5	9
10	100%	1.0	10	10.5	10

Sr no	x	Sr no	x
1	43	16	85
2	54	17	87
3	56	18	88
4	61	19	89
5	62	20	93
6	66	21	95
7	68	22	96
8	69	23	98
9	69	24	99
10	70	25	99
11	71		
12	72		
13	77		
14	78		
15	79		

10% = 0.1 0.1 x 25
= 25
≈ 3

average = $\frac{56+61}{2}$ $x_p = 58.5$

3 values < x_p(58.5)

3/25 = 12% values

80% = 0.8 0.8 x 25
= 20

average = $\frac{93+95}{2}$ $x_p = 94$

20 values < x_p(94)
20/25 = 80%

60% = 0.6 0.6 x 25
= 15

average = $\frac{79+85}{2}$ $x_p = 82$

15 values < x_p(82)
∴ 15/25 = 60%

Measures of locations

↳ mean
↳ median

$$\text{mean} = \frac{1}{k}\sum_{i=1}^{k} x_i$$

$$\text{median} = x_{(r+1)} \quad \text{- odd}$$
$$= \frac{x_{(r+1)} + x_{(r)}}{2} \quad \text{- even}$$

trimmed mean

1, 2, 3, 4, 5, 90
p = 40%
(p/2)% = 20%
20% of 6 values = 1.2 ≈ 1 0.20 x 6 = 1.2

trimmed data set 2, 3, 4, 5
∴ trimmed mean = 3.5

Measures of spread range and varience

range = max(x) − min(x)

$$\text{Varience}_{(x)}\ [s_x^2] = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

mean absolute deviation
= median $(|x_1 - \bar{x}|, |x_2 - \bar{x}|, \ldots, |x_k - \bar{x}|)$

absolute average deviation $= \frac{1}{m} \sum_{i=1}^{m} |x_i - \bar{x}|$

interquartile range $= x_{75\%} - x_{25\%}$

75% = 0.75	25% = 0.25
0.75 × 25	0.25 × 25
18.75	= 6.25
≃ 19	≃ 6
average = $\frac{89+93}{2}$ $x_p = 91$	average = $\frac{66+68}{2}$ $x_p = 67$
19 values < x_p (91)	6 values < x_p (67)

∴ 91 − 67 = 24

Multivariate Summary Statistics
$\bar{x} = (\bar{x_1}, \bar{x_2}, \bar{x_3} \ldots \bar{x_n})$

S_{ij} = covariance (x_i, x_j)
$= \frac{1}{m-1} \sum_{k=1}^{m} (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j})$

∴ x_{ki} → value of x_i attrib for k^{th} object
x_{kj} → " " x_j " " " "

r_{ij} = correlation (x_i, x_j)
$= \frac{\text{covariance}(x_i, x_j)}{s_i, s_j}$

OLAP → extracts the data

S_i, S_j → varience of x_i & x_j

Skewness

OLAP & Data Warehousing

(*) Warehouse :- (designed to store large amount of data)
→ A data warehouse is subject oriented, integrated (data is collected from multiple sources) time variant and non-volatile (cannot be updated) collection of data in support of the management decision making process.

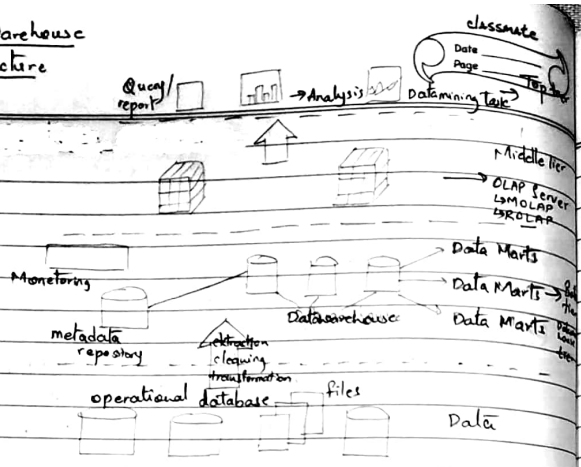
Integrated — Intergrator / wrappers → datawarehouse has its own update driver : data is first schema integrated, transformed & given to datawarehouse

→ Data is transformed & Then store into datawarehouse schema (semantic data store)

- customer profile
- repositioning of product
- sources of profit

	(System uses data warehouse) OLAP	v/s	(System uses DBMS) OLTP
Users & System Orientation	→ market oriented		→ Customer oriented
Data Contents	→ Data from historic perspective		→ Data from current database transactio
Database Design	→ specialized datawarehouse schema		→ User ER based desig
View	→ Multiple views		→ current detail/trans
Access pattern	→ Complex read-only query		→ Consist of short atomic transactio

Data Warehouse Architecture



describe command

09/09/19 Absent

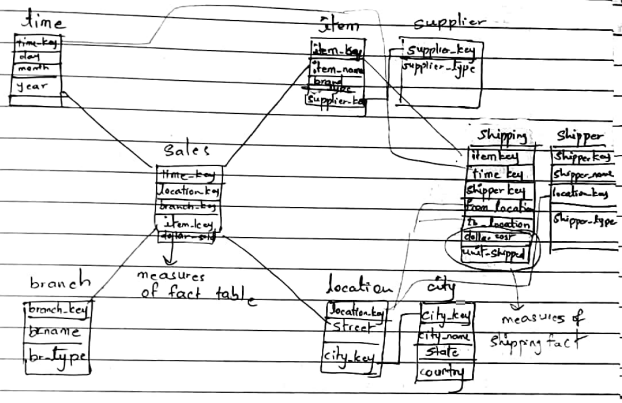
adv

Enterprise:- it capture coorporate wide data
data mart:- captures specific data wide data
eg marketing
→ impl° cost is low

12/09/19

* Snowflake Schema

↳ has a single fact table (sales)
↳ multiple dimension table. (time, item, city ...)
↳ one or more dimension table
→ normalised
↳ used for data mart



* fact constellation schema

↳ contains more than 1 fact table
↳ multiple dimension table.
↳ used for enterprise warehouse.

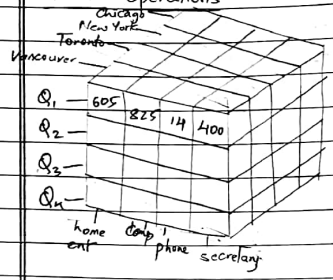
Concept Hierarchy

Concept Hierarchy is mapping of a set of low level concepts to high level concepts

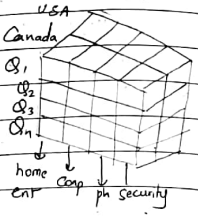
street < city < state < country

month < quarter < half-anually < yearly

OLAP Operations

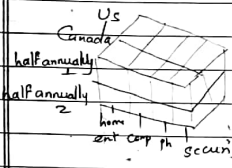


Rollup



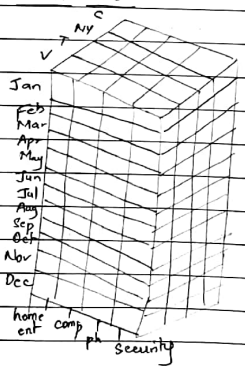
The rollup operation performs aggragation on a data cube by either climbing up a concept hierarchy or by dimension reduction

Ronce



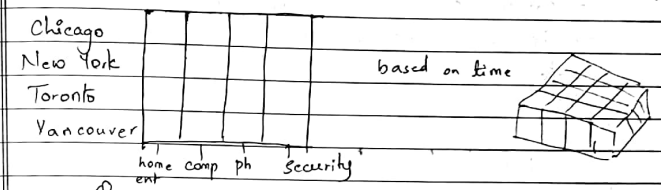
If rollup is perform using dimension reduction, one or more dimensions are removed from the given cube.

* drill-down



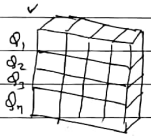
- A drill down operation navigates from less detailed data to more detailed data.
- It can be realised by stepping down a concept heirarchy for a given dimension or by introducing more dimension.

Slice Operation (selection of data based on one dimension)



→ Slice performs the selection on one dimension of a given cube resulting in a sub-cube.

slice based on one location

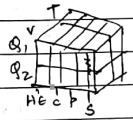


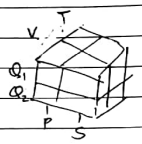
based on item

* Dice Operation

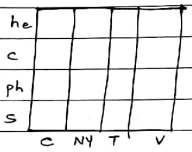
Dice Operation defines a sub-cube by performing a selection on two or more dimension.

selection based on time & ~~city~~ location.





* **Pivot Operation**
→ Rotating the axes of the diff. dimensions.
→ Its a visualisation operation that rotates the data axis with view to provide an alternate data representation.

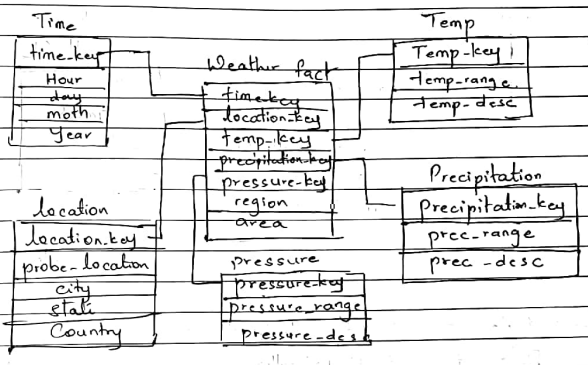


* **drill across**
→ it refers queries across more than one fact table.

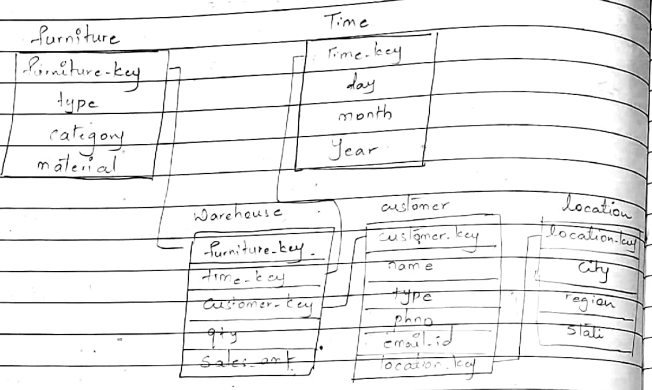
* **drill through**
is an operation that drills through the bottom level of a data cube down to its back-end relational tables.

Q. Design a datawarehouse for a regional weather bureau. The weather bureau has about 1000 probes which are scattered throughout various lands & ocean location in the region to collect basic weather data such as air-pressure, temp, precipitation at each hour. All data are sent to the central station. Which is collected such data for over 10 years. Your design should facilitate efficient query & all line analytical processing.

⇒



Q. Design a datawarehouse for the wholesale furniture company. The datawarehouse has to allow to analyse the companies situation atleast w.r.to furniture, customers & time. Moreover the company needs to analyze the furniture w.r.to its time (their, their, Wardrobe etc) form a category (kitchen, living room, office, etc) and material (wood, marble, etc). It also needs to analyze the customer w.r.to their special location by considering cities, regions & states. The company is interested in learning the quantity of furniture units sold & total sales amount generated.



✱ Data Generalisation using Attribute Oriented Induction

Data Generalisation summarises the data by replacing low level concepts/values with high level concepts or by reducing the dimensions.

→ if we cannot generalise then remove the dimensions

Generalisation can be done in 2 ways.

↳ Characterisation :- for one collection

↳ Discrimination :- two or more "

↳ find feature which discriminate ... or more collection

→ Characterization provides the concise and succint summarization of a given data collection.

→ Discrimination provides description comparing two or more data collection

OLAP → aggregation

Q. Why OLAP → not used in data generalisation.

→ ~~x~~ can't handle complex types

→ depends on end user domain knowledge.

✱ Attribute Oriented Induction

query → results are obtained & then generalised

1> data focusing

↳ Attribute Relevance analysis

→ coorelation analysis

↳ entropy, gain, gini index

data focusing

↓

query

↓

~~initial working~~ relation.

Attribute Removal

If we cannot generalise then remove the attribute

eg. rollno, aadhar number, etc.

Attribute Generalisation

~~Attribute~~ threshold :- decide how much generalisation should be done

~~Attribute threshold :- generalise~~ till the threshold value

relational threshold :- generalising the tuple

→ Output of this is → prime generalised relation.

eg. select name, gender, major, birthplace, birth-date, residence, phone, gpa

from student

where station in {"MSL", "MA", "MBA"}

→ name → attribute removal
→ gender → retain
→ major → Science, Arts, Business Mgmt
→ ~~birth-date~~ birthplace → city → stati → country.
→ birth-date → age → age ranges
→ residence → city → stati → country
→ phone → attribute removal
→ gpa → grades

1) data focusing: info.
2) query → result is called as
3) initial working relation

gpa score

- data focusing
- query

classmate Date Page

10/9/19

Data Generalisation using Attribute Oriented Induction

→ initial working relation

name	gender	major	birth-place	birth-date	residence	phone	gpa
Jim	M	CS	Vancouver, Canada	12-8-76	3511, St	6587-498	397
					Richmond		
Larsa	F	Physics	Seattle, USA	8-12-76	RS Auston	854-5233	383
					Boston		

Generalisation → Attribute Removal ← Attribute Threshold / Relation Threshold
→ Attribute Generalisation

Output of Generalisation → prime generalisation.

→ major: generalised into broad disciplined
birth-place: stati < country
birth-date: age < age-range / age-groups
residence: city < stati < country
gpa: grades.

gender	major	birth-country	age-group	residence-country	grade	count
M	Sc	US	20-25	US	First	16
F	Eng	Canada	20-25	Canada	First	20

Graduate Student Table

Gender	Major	Count
M	Sc	20
F	Sc	30
M	Eg	10
F	Eng	40
		100

UG Student table

Gender	Major	Count
M	Sc	15
F	Sc	40
M	Engg	10
F	Engg	10
		75

Q. Explain attribute oriented induction. Consider the following tables, from the given data, compute the gain of gender, compute the gain of major & analyse which attribute is more relavant

⇒ Target Class = Graduate, UG

$$\text{Entropy}(D) = -\left[\left(\frac{100}{175}\right)^{G} \times \log_2\left(\frac{100}{175}\right) + \left(\frac{75}{175}\right)^{UG} \times \log_2\left(\frac{75}{175}\right)\right]$$

$$= 0.985$$

Gender

Gender = M

$$\text{entropy} = -\left[\left(\frac{30}{55}\right)^{G} \times \log_2\left(\frac{30}{55}\right) + \left(\frac{25}{55}\right)^{UG} \times \log_2\left(\frac{25}{55}\right)\right]$$

$$= 0.9940$$

Gender = F

$$\text{entropy} = -\left[\left(\frac{70}{120}\right)^{G} \times \log_2\left(\frac{70}{120}\right) + \left(\frac{50}{120}\right)^{UG} \times \log_2\left(\frac{50}{120}\right)\right]$$

$$= 0.979$$

$$\text{net entropy(gender)} = \frac{55}{175} \times (X_1) + \frac{120}{175} \times X_2$$

$$= X_3 =$$

$$= 0.983$$

$$\text{Gain(Gender)} = \text{entropy}(D) - X_3$$

$$= 2 \times 10^{-2}$$

Major

Major = Sc

$$\text{entropy} = -\left[\left(\frac{50}{105}\right)^{G} \times \log_2\left(\frac{50}{105}\right) + \left(\frac{55}{105}\right)^{UG} \times \log_2\left(\frac{55}{105}\right)\right]$$

$$= 0.998$$

Major = Engg

$$\text{entropy} = -\left[\left(\frac{50}{70}\right)^{G} \times \log_2\left(\frac{50}{70}\right) + \left(\frac{20}{70}\right) \times \log_2\left(\frac{20}{70}\right)\right]$$

$$= 0.863$$

$$\text{net entropy(major)} = \frac{105}{175} \times 0.998 + \frac{70}{175} \times 0.863$$

$$= 0.944$$

$$\text{Gain(major)} = 0.985 - 0.944$$

$$= 0.041$$

Zoology — Attribute Oriented Induction

→ Description / Class Comparison

1] Data Collection

→ query

→ partition data → target class : high risk diabetes

↳ set of contrasting classes : low & medium risk

* generalising data for more than one data collection → descrimination

2] Dimension relivance analysis : finding most imp. attributes using impurity measure & correlation analysis.

3] Synchronous Generalisation : generalising more than one data collection simultaneously.
   ↳ prime target class relation.
   ↳ prime contrasting class relation.

4] Data presentation.
   ↳ can be presented in charts, tables, etc.

eg. find the characteristics of students pursuing UG & PG courses.

20/09/19 MODULE 3

Classification : used for discrete values.

model / learning function
explanatory variable → target variable.
attribute set x.

- A fn that map explanatory variable to target variable is called learning function.
- explanatory variables :- describes the target class.

Training Set : class labels are given.
Test Set : class labels are not given.

* **Decision Tree**
types of nodes
→ root node
→ internal nodes: 1 incoming & many outgoing edges.
→ leaf node : 1 incoming & no outgoing edges.

* **Hunt's Algorithm**
D_t = training set
$y = \{y_1, y_2, \ldots y_n\}$
→ set of class labels

1] D_t → leaf node
2] D_t → more than one class label
use attribute test condition
↳ create child node for each outcome of the chosen attribute (D_t)
↳ partition data among child nodes.
↳ recursively apply the algorithm for each child node.

Test Conditions :
1] If child node is empty :
↳ depends on the parent node.
↳ creates a leaf node → with majority class of parent node.

2] Identical attribute values.

Habitat	Mammal	Class
Land	yes	herbivore
Land	yes	herbivore
Land	yes	carnivore ~~herbivore~~

→ we cannot split

↳ create a leaf node
↳ majority class of given child node.

* Design issues → decision tree

→ attribute test condition
→ an objective measure : used to check the performance of test condition.
→ stopping condition : 1) when all records have same class lab
2) when identical attributes, we have atlea one class label in majority.

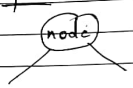
* **Performance of Classification model**
→ confusion matrix.

	No. of records classified as '0'	No. of records classified as '1'
Class 0	f_{00}	f_{01}
Class 1	f_{10}	f_{11}

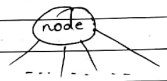
$$\text{Accuracy} = \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}}$$

$$\text{Error rule} = \frac{f_{01} + f_{10}}{f_{00} + f_{11} + f_{01} + f_{10}}$$

* Types of splits
  - Binary Split

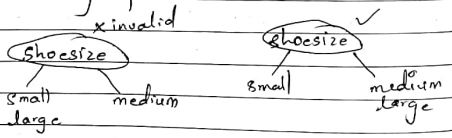


  - Multiway Split



* Nominal Attributes
  ↳ binary / multiway split

Ordinal Attribute : order should be preserved
↳ binary / Multiway split



* Continous Attribute
  ↳ binary / Multiway split

$A > V$ or $A \leq V$
or

range query
$V_i \leq A < V_{i+1}$

Measures of determine best split
impurity measures

→ entropy $= -\sum p(i|t) \log_2 (p(i|t))$
Probability of a given class at note t

→ gini index $= 1 - [\sum (p(i|t))^2]$

→ Classification error $= 1 - \max_t [p(i|t)]$

Nominal Attribute

cartype

	Sports, Luxury (N_1)	Family (N_2)
C0	9	1
C1	7	3

Gini index, $N_1 = 1 - \left(\frac{9}{16}\right)^2 - \left(\frac{7}{16}\right)^2$

$N_1 = 0.49218$

Gini index, $N_2 = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$

$N_2 = 0.375$

Weighted Gini index $= \frac{16}{20} \times 0.49218 + \frac{4}{20} \times 0.375$

$= 0.468$

	Sports N_3	Family Luxury (N_4)
C0	8	2
C1	0	10

Gini Index $N_3 = 1 - \left(\frac{8}{8}\right)^2 - \left(\frac{0}{8}\right)^2$

$N_3 = 0$

Gini Index $N_4 = 1 - \left(\frac{2}{12}\right)^2 - \left(\frac{10}{12}\right)^2$

$= 0.2777$

Weighted Gini Index $= \frac{8}{20} \times 0 + \frac{12}{20} \times 0.277$

$= 0.1666$

Q.

Node n1	Count
class 0	1
class 1	5

entropy $(N_1) = -\left[\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) + \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right)\right]$

$= 0.650$

Gini index $(N_1) = 1 - (1/6)^2 - (5/6)^2$

$= 0.278$

Classification error $= 1 - \max((1/6), (5/6))$

$= 1 - \max(0.1667, 0.833)$

$= 1 - 0.833$

$= 0.167$

Annual Income	default borrowers
60	no
70	no
75	no
85	yes
90	yes
95	yes
100	no
120	no
125	no
220	no

97

<=97 >97

avg $= \frac{60+70}{2} = 65$

- least entropy value is the split point
- Continue the same procedure
- Sort if not sorted

	yes	no
<=65	0	1
>65	3	6

entropy $(<=65) = -\left[\left(\frac{0}{1}\right)\log_2\left(\frac{0}{1}\right) + \left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right)\right]$

$= 0$

entropy $(<65) = -\left[\left(\frac{3}{9}\right)\log_2\left(\frac{3}{9}\right) + \left(\frac{6}{9}\right)\log_2\left(\frac{6}{9}\right)\right]$

$= 0.9182$

Weighted entropy $= \left(\frac{1}{10}\right) \times 0 + \left(\frac{9}{10}\right) \times 0.9182$

$= 0.8263$

Gain : Difference of impurity measure

$$\Delta \quad I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

$N(v_j)$: No. of records associated with childnode v_j

N: No. of records at parent node

k No. of attribute values.

* Using IDs (Iterative Dichostomiser)

Gender	height	Class
F	1.6	Short
F	1.6	Short
F	1.7	Short
M	1.7	Short
F	1.75	Medium
F	1.8	Medium
F	1.8	Medium
M	1.85	Medium
F	1.88	Medium
F	1.9	Medium
M	1.95	Medium
M	2	Medium
M	2.1	Tall
M	2.2	Tall

Q. Divide the height attribute into ranges as follows. & construct the decision tree for the following dataset

(0, 1.6]
(1.6, 1.7]
(1.7, 1.8]
(1.8, 1.9]
(1.9, 2.0]
(2.0, 5.0]

(0, 1.6] $0 < \text{height} \leq 1.6$

Gender	height	Class
F	(0, 1.6]	Short
F	(0, 1.6]	Short
F	(1.6, 1.7]	Short
M	(1.6, 1.7]	Short
F	(1.7, 1.8]	Medium
F	(1.7, 1.8]	Medium
F	(1.7, 1.8]	Medium
M	(1.8, 1.9]	Medium
F	(1.8, 1.9]	Medium
F	(1.8, 1.9]	Medium
M	(1.9, 2.0]	Medium
M	(1.9, 2.0]	Medium
M	(2.1, 5.0]	Tall
M	(2.1, 5.0)	Tall

$$\text{entropy}(D) = -\left[\left(\frac{4}{15}\right) \times \log_2\left(\frac{4}{15}\right) + \left(\frac{9}{15}\right) \times \log_2\left(\frac{9}{15}\right) + \left(\frac{2}{15}\right) \log_2\left(\frac{2}{15}\right)\right]$$

$$= 1.3381$$

Gender

$$\text{entropy}(\text{Gender}=M) = -\left[\left(\frac{1}{6}\right) \times \log_2\left(\frac{1}{6}\right)^{\text{short}} + \left(\frac{3}{6}\right) \log_2\left(\frac{3}{6}\right)^{\text{medium}} + \left(\frac{2}{6}\right) \log_2\left(\frac{2}{6}\right)^{\text{tall}}\right]$$

$$= 1.4591$$

$$\text{entropy}(\text{Gender}=F) = -\left[\left(\frac{3}{9}\right) \times \log_2\left(\frac{3}{9}\right)^{\text{short}} + \left(\frac{6}{9}\right) \log_2\left(\frac{6}{9}\right)^{\text{medium}} + \left(\frac{0}{9}\right) \log_2\left(\frac{0}{9}\right)\right]$$

$$= 0.9182$$

~~net entropy~~
Weighted entropy (Gender) $= \frac{6}{15} \times 1.459 + \frac{9}{15} \times 0.9182$

$= 1.1345$

Gain(gender) = entropy(D) − weighted entropy(gender)

$= 1.13381 - 1.1345$

$= 0.20354$

Height

entropy(height = $(0, 1.6]$) $= -\left[\left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) + \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) + \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right)\right]$

$= 0$

entropy(height = $(1.6, 1.7]$) $= -\left[\left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) + \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) + \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right)\right]$

$= 0$

entropy(height = $(1.7, 1.8]$) $= -\left[\left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right) + \left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right) + \left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right)\right]$

$= 0$

entropy(height = $(1.8, 1.9]$) $= -\left[\left(\frac{0}{4}\right)\log_2\left(\frac{0}{4}\right) + \left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) + \left(\frac{0}{4}\right)\log_2\left(\frac{0}{4}\right)\right]$

$= 0$

entropy(height = $(1.9, 2.0]$) $= -\left[\left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) + \left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) + \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right)\right]$

$= 0$

entropy(height = $(2.0, 5.0]$)

$= -\left[\left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) + \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) + \left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right)\right]$

$= 0$

Weighted entropy (height) $= \left(\frac{2}{15}\right)\times 0 + \left(\frac{2}{15}\right)\times 0 + \left(\frac{3}{15}\right)\times 0 + \left(\frac{4}{15}\right)\times 0 + \left(\frac{2}{15}\right)\times 0 + \left(\frac{2}{15}\right)\times 0$

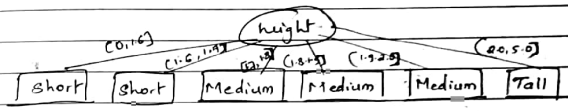
$= 0$

Gain(height) = entropy(D) − weighted entropy(height)

$= 1.3381 - 0$

$= 1.3381$

Since gain(height) is highest → split based on height



Gender	CarOwnership	TravelCost	Income level	TransportationMode
M	0	Cheap	low	bus
M	1	Cheap	medium	bus
F	1	Cheap	medium	train
F	0	Cheap	low	bus
M	1	Cheap	medium	bus
M	0	Standard	medium	train
F	1	Standard	medium	train
F	1	expensive	high	Car
M	2	expensive	medium	Car
F	2	expensive	high	car.

entropy(D) $= -\left[\left(\frac{4}{10}\right)\log_2\left(\frac{4}{10}\right) + \left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) + \left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right)\right]$

$= 1.5709 = 1.571$

Gender

entropy (Gender = F) = $-\left[\left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right) + \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) + \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)\right]$

= 1.5219

entropy (Gender = M) = $-\left[\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) + \left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right) + \left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right)\right]$

= 1.3709

Ownership

entropy (ownership = 0) = $-\left[\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) + \left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right)\right]$

=

net entropy / weighted entropy (gender) = $\frac{5}{10} \times 1.3709 + \frac{5}{10} \times 0.522$

= 1.447

Gain (gender) = entropy (D) − net entropy (gender)

= 1.571 − 1.447

= 0.1246

CarOwnership

entropy (cos = 0) = $-\left[\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right]$

= 0.9182

entropy (cos = 1) = $-\left[\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right) + \left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right)\right]$

= 1.5219

entropy (cos = 2) = $-\left[\frac{0}{2}\log_2\frac{0}{2} + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right]$

= 0

Travel Cost

entropy (TC = cheap) = $-\left[\frac{5}{5}\right.$

net entropy (cos) = $\frac{3}{10} \times 0.918 + \frac{5}{10} \times 0.522 + 0$

= 1.0364

Gain (cos) = entropy (D) − net entropy (cos)

= 1.571 − 1.0364

= 0.5346

Travel Cost

entropy (TC = cheap) = $-\left[\frac{4}{5}\log_2\left(\frac{4}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{0}{5}\log_2\left(\frac{0}{5}\right)\right]$

= 0.7219

entropy (TC = standard) = $-\left[\left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) + \frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right]$

= 0

entropy (TC = expensive) = $-\left[\frac{0}{2}\log_2\frac{0}{2} + \frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}\right]$

= 0

net entropy (Travel Cost) = $\frac{5}{10} \times 0.7219 + \frac{2}{10} \times 0 + \frac{2}{10} \times 0$

= 0.36095 = 0.361

Gain (Travel Cost) = entropy (D) − net entropy (TC)

= 1.571 − 0.36095

= 1.21005 = 1.210

Income Level

entropy (IL = low) = $-\left[\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2} + \frac{0}{2}\log_2\frac{0}{2}\right]$

= 0

entropy (IL = medium) = $-\left[\frac{2}{6}\log_2\frac{2}{6} + \frac{3}{6}\log_2\frac{3}{6} + \frac{1}{6}\log_2\frac{1}{6}\right]$

= + 1.459

entropy (IL = high) = $-\left[\frac{0}{2}\log_2\frac{0}{2} + \frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}\right]$

= 0

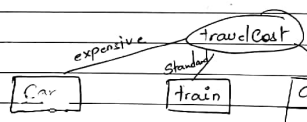
net entropy (Income level) = $\frac{2}{10}\times 0 + \frac{6}{10}\times 1.459 + \frac{2}{10}\times 0$

= 0.8754

Gain (Income level) = entropy(D) − net entropy (Income level)

= 1.571 − 0.8754

= 0.6956



(D_1)

Gender	Car ownership	Income level	transportation mode
M	0	low	bus
M	1	medium	bus
F	1	medium	train
F	0	low	bus
M	1	medium	bus

2nd Iteration

entropy $(D_1) = -\left[\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right]$ = 0.722

entropy of D_1

Gender

entropy (gender = F) = $-\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right]$

= 1

entropy (gender = M) = $-\left[\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right]$

= 0

net entropy (gender) = $\frac{2}{5}\times 1 + \frac{3}{5}\times 0$ = 0.4

gain (gender) = entropy (D_1) − net entropy (gender)

= 0.722 − 0.4 = 0.322

Car ownership

entropy (CO = 0) = $-\left[\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right]$

= 0

entropy (CO = 1) = $-\left[\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right]$

= 0.9182

net entropy (Car Ownership) = $\frac{2}{5}\times 0 + \frac{3}{5}\times 0.9182$

= 0.5509

gain (Car Ownership) = entropy (D_1) − net entropy (Car Ownership)

= 0.722 − 0.5509

= 0.1711

Income level

$$\text{entropy}(IL=\text{low}) = -\left[\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right]$$

$$= 0$$

$$\text{entropy}(IL=\text{medium}) = -\left[\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right]$$

$$= 0.9182$$

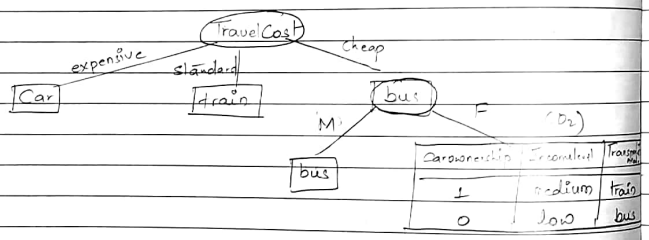
$$\text{net entropy}(\text{Income level}) = \frac{2}{5}\times 0 + \frac{3}{5}\times 0.9182$$

$$= 0.5509$$

$$\text{gain}(\text{Income level}) = \text{entropy}(D_1) - \text{net entropy}(\text{Income level})$$

$$= 0.722 - 0.5509$$

$$= 0.1711$$



3rd iteration

$$\text{entropy}(D_2) = -\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right]$$

$$= 1$$

Carownership

$$\text{entropy}(CO=0) = -\left[\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right] = 0$$

1T3

- Cosine Similarity (Document Data)
- Any association (FP Tree, Apriori, vertical Data Format)
- Decision Tree (C4.5, Entropy)

$$\text{entropy}(CO=1) = -\left[\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right]$$

$$= 0$$

$$\text{net entropy}(\text{carownership}) = 0$$

$$\text{gain}(\text{Carownership}) = \text{entropy}(D_2) - \text{net entropy} = 1 - 0 = 1$$

Income level

$$\text{entropy}(IL=\text{medium}) = -\left[\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right]$$

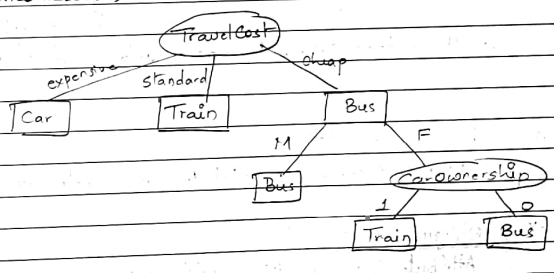
$$= 0$$

$$\text{entropy}(IL=\text{low}) = -\left[\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right]$$

$$= 0$$

$$\text{net entropy}(\text{Income level}) = 0$$

$$\text{gain}(\text{Income level}) = 1 - 0 = 1$$



Gain ratio = $\frac{\text{Gain}}{\text{Split Info}}$

Split Info = $-\sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \log_2 \frac{|Dj|}{|D|}$

v → no. of outcomes

Dj → records associated with each split

Alternative for Gain Ratio is Binary Split.

* Characteristics of a decision Tree
  - → non-parametric approach
  - → NP complete problem (Non-deterministic Polynomial)
  - → Computationally inexpensive
  - → easy to interpret → small decision tree.
  - → it is a top-down recursive partitioning approach
    - → data fragmentation :- handle by using gain ratio, binary splits, prunning.
  - → it divides attribute space into disjoint region (with defined classes)

* Decision Border
  - ↳ rectilinear → (parallel to co-ordinate axis)

* Oblique decision Tree
  - Attribute test condition with more than one attribute.

* Constructive Induction
  - Only one attribute
  - → uses composite attribute.

* the impurity measure does not affect performance of Decision Tree.

* → quite robust to noise and redundant data.

Q.

Outlook	Temperature	humidity	Wind	PlayTennis
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
Overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
Overcast	mild	high	strong	yes
Overcast	hot	normal	weak	yes
rain	mild	high	strong	no

$$\text{entropy(D)} = -\left[\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)\right]$$

$$= 0.940$$

Outlook

$$\text{entropy(Outlook = sunny)} = -\left[\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)\right]$$

$$= 0.970$$

$$\text{entropy(Outlook = overcast)} = -\left[\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right]$$

$$= 0$$

$$\text{entropy(Outlook = rain)} = -\left[\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right]$$

$$= 0.9709$$

$$\text{net entropy} = \frac{5}{14} \times 0.970 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.970$$

$$= 0.6928$$

$$\text{gain(Outlook)} = \text{entropy(D)} - \text{net entropy(Outlook)}$$

$$= 0.940 - 0.6928$$

$$= 0.248.$$

$$\text{Split Info (Outlook)} = -\left[\left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) + \left(\frac{4}{14}\right)\log_2\left(\frac{4}{14}\right) + \frac{5}{14}\log_2\frac{5}{14}\right]$$

$$= 1.577$$

$$\text{Gain ratio} = \frac{\text{Gain(Outlook)}}{\text{Split Info (Outlook)}} = \frac{0.248}{1.577} = 0.157$$

Temperature

$$\text{entropy(T = hot)} = -\left[\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right]$$

$$= 1$$

$$\text{entropy(T = mild)} = -\left[\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right]$$

$$= 0.918$$

$$\text{entropy(T = cool)} = -\left[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right]$$

$$= 0.811$$

$$\text{net entropy(Temp)} = \frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811$$

$$= 0.9108$$

gain (Temp) = 0.940 − 0.9108 = 0.0292 = 0.030

$$\text{Split Info (Temp)} = -\left[\frac{4}{14}\log_2\frac{4}{14} + \frac{6}{14}\log_2\frac{6}{14} + \frac{4}{14}\log_2\frac{4}{14}\right]$$

$$= 1.55665$$

$$\text{Gain ratio} = \frac{0.030}{1.55665} = 0.019.$$

* **Humidity**

$$\text{entropy (H = high)} = -\left[\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right]$$

$$= 0.98522$$

$$\text{entropy (H = normal)} = -\left[\frac{6}{7}\log_2\frac{6}{7} + \frac{1}{7}\log_2\frac{1}{7}\right]$$

$$= 0.59167$$

$$\text{net entropy (Humity)} = \frac{7}{14} \times 0.98522 + \frac{7}{14} \times 0.59167$$

$$= 0.788$$

gain (Humidity) = 0.940 − 0.788445 = 0.15155

$$\text{Split Info (Humidity)} = -\left[\frac{7}{14}\log_2\frac{7}{14} + \frac{7}{14}\log_2\frac{7}{14}\right]$$

$$= 1$$

$$\text{Gain ratio} = \frac{0.15155}{1} = 0.15155$$

Wind

$$\text{entropy (Wind = Weak)} = -\left[\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right]$$

$$= 0.8112$$

$$\text{entropy (Wind = Strong)} = -\left[\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right]$$

$$= 1$$

$$\text{net entropy (Wind)} = \frac{8}{14} \times 0.8112 + \frac{6}{14} \times 1$$

$$= 0.89211$$

gain (Wind) = 0.940 − 0.89211 = 0.04789

$$\text{Split Info (Wind)} = -\left[\frac{8}{14}\log_2\frac{8}{14} + \frac{6}{14}\log_2\frac{6}{14}\right]$$

$$= 0.98522$$

$$\text{Gain ratio} = \frac{0.04789}{0.98522} = 0.0486$$

03/10/19 Classification

- training error: no. of records which are incorrectly classified during training phase
  - Exact value for training value.
- generalization error: - errors which occurs during Testing phase
  - only be aproximated

→ Model overfitting (More records)
- → low training error
- → relatively high generalization error.

→ Model overfitting takes place b'coz of noise, lack of representative samples, multiple comparison procedure (multiple test conditions)

→ Model overfitting occur when you have detailed decision tree.

- Model Underfitting (less records)
  - → high training error
  - → high generalization error

Techniques to find approximate generalisation error

(1) Resubstitution Error

→ training error is considered as approx. value for general" error

(2) pessimistic estimate

↳ sum of training error + penalty term

$$e_g(T) = \frac{\sum_{i=1}^{k} [e(t_i) + \Omega(t_i)]}{\sum_{i=1}^{k} n(t_i)}$$

$$= \frac{e(T) + \Omega(T)}{N_t}$$

training error = 4
no. of nodes = 7 $\quad \Omega = 0.5$
no. of training records = 24

$$\therefore = \frac{e(T) + \Omega(T)}{N_t} = \frac{4 * 7 \times 0.5}{24}$$

* Minimum Description Length Priciple

Cost = Cost(model) + cost(data|model)

* Occam's Razor / Principle of Parsimony

Choose model which is ~~smaller~~ / simpler one

(3) Statistical Correction : The generalization is upperbound of the statistical correction applied to the training error.

$$e_{upper}(N, e, \alpha) = e + \frac{Z^2_{\alpha/2}}{2N} + Z_{\alpha/2}\sqrt{\frac{e(1-e)}{N} + \frac{Z^2_{\alpha/2}}{4N^2}} \Big/ 1 + \frac{Z^2_{\alpha/2}}{N}$$

(standardized normal value)

confidence level

Using Validation set

break training into 2 parts → a few records are kept for testing

07/10/19

* Evaluating Performance of Classifier

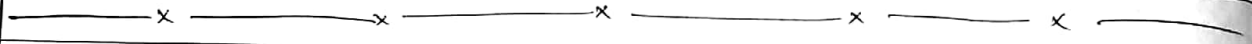
→ holdout method

Data set — training set
— Test set

Random Subsampling:
K times.

Cross-Validation
→ 2 fold validation: divide into training & test set, perform validation & for next iteration swap
→ K fold validation: - data divide into K partition: one part for testing
→ leave one approach. (K=N)
N: size of the data set
→ N-1 → used for training
N^{th} record → testing

Note Continuous data can be represented as range query
i.e $x < A < y$
$A < V$ or
$A \geq V$



* Rule Based Classification

$A \rightarrow y$ $\quad A = C_1 \vee C_2 \vee C_3 \ldots \vee C_n$

antecedent/ Precondition — consequent

→ Rule gets triggered when antecedent satisfies the condition
$A \rightarrow y$

eg: r_1 : (Gives birth = no) ∧ (Aerial Creature = yes) → Birds
r_2 : (Gives birth = no) ∧ (Aquate Creature = yes) → Fishes
r_3 : (Gives birth = yes) ∧ (Body temp = warm blooded) → Mammals
r_4 : (Gives birth = no) ∧ (Aerial Creature = no) → Reptiles
r_5 : (Aquate Creature = semi) → Amphibians

|D| : No. of records in a data set

|A∩y| : No. of records satisfying the antecedent & consequent

|A| : No. of records satisfying the antecedent

$$\therefore \text{Accuracy} = \frac{|A \cap y|}{|A|}$$

$$\text{Coverage} = \frac{|A|}{|D|}$$

r3:

|A| = 5

|A∩y| = 5

|D| = 15

∴ Accuracy = 5/5 = 100%

Coverage = 5/15 = 33.33%

r_1:

|A| = 2

|A∩y| = 2

|D| = 15

Accuracy = 2/2 = 100%

Coverage = 2/15 = 13.33%

r2:

|A| = 2

|A∩y| = 2

|D| = 15

Accuracy = 2/2 = 100%

Coverage = 2/15 = 13.33%

r4:

|A| = 8

|A∩y| = 3

|D| = 15

Accuracy = 3/8 = 37%

Coverage = 8/15 = 53%

r5:

|A| = 4

|A∩y| = 2

|D| = 15

Avg = 2/4 = 50%

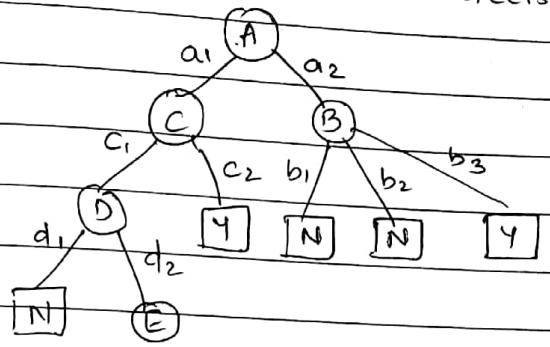
Coverage = 4/15 = 26.66%

Continuation of decision Tree

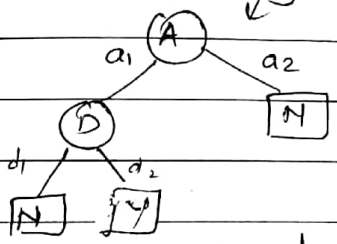
1] Pre-prunning :- finds gain & if gain<threshold, it halts the growing of tree. (Doesn't allow to construct the entire Tree)

2] Post-prunning : Allow to construct the entire tree

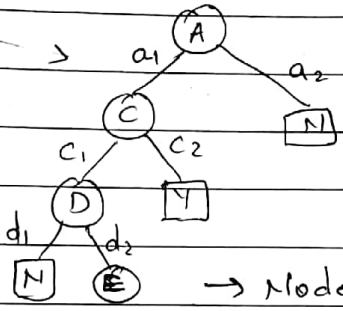
prunning :- is done to handle model over fitting
→ reducing branches in decision Tree



- Subtree Replacement
- Subtree Raising



→ Node is replaced by a subtree



→ Node is replaced by a leaf node

* **Mutually Exclusive Rules**

The rules in a rule set are said to be mutually exclusive, if no two rules are triggered by the same set of records.

* **Exhaustive Coverage**

The rules in a rule set are said to have exhaustive coverage, if there is a rule for each combination of attribute value.

training error = 0

{antecedent : condition
{consequent : class label

* Default Rule $r_d(x) \rightarrow y_d$
→ Majority class
→ When none of the rules are satisfied

* 2 Types of Rules

1] Ordered Rules: Rules which are ordered with priority
→ rules → ordered → (according) priority
↳ decision list
Adv of Ordering Rules :

2] Unordered Rules :

Types of Rule Ordering
→ Rule by Rule ordering
→ Class by Class Ordering : rules are grouped having same class → then group the class

Rule based Classifier
1] Direct method for rule extraction :- analyse → apply algo and extract rule.
2] Indirect method for rule extraction :- extracting from other classifier

1] Direct method

→ Sequential Covering Algorithm
→ D → dataset
→ AttVals → set of attribute & values.

1] Rule set = { }
2] for each class C do

3] repeat
4] Rule = Learn_One_Rule (D, AttVals, C)
5] remove tuples from Discovering rule.
6] Ruleset = Ruleset + Rule
7] until terminating condition
8] end for
9] return Rule set

* Rule growing strategies
→ general to specific :
Contain empty set & go on adding

→ specific to general :
↳ choose one +ve example
↳ Eliminates condition & check the accuracy.
eg: If it is a mammal, we need only body-temp and gives birth
remove has_legs x, aerial creature ---

* Rule Quality Measure
R ratio / Likelihood ratio

$$R = 2 \sum_{i=1}^{k} f_i \log_2 (f_i / e_i)$$

f_i is observed freq. of class i covered by the rule and e_i is the expected freq. of class i covered by the rule

$r_1 = (C_1 \wedge C_2 \rightarrow m_1)$
$r_1' = (C_1 \wedge C_2 \wedge C_3 \rightarrow m_1)$
$R(r_1)$ $R(r_1')$

10/10/19

Laplace Measure

$$= \frac{f_+ + 1}{n+k}$$

f_+ : no. of +ve tuples which are covered by the rule

n : total no. of tuple covered by the rule.

k : no. of classes

p_+ : Prior probability for +ve classes

m-estimate

$$= \frac{f_+ + k p_+}{n+k}$$

$$\text{FOIL Gain } (R_0, R_1) = p_1 \times \left(\log_2 \frac{P_1}{P_1 + n_1} - \log_2 \left(\frac{P_0}{P_0 + n_0} \right) \right)$$

R_0: Original rule

R_1: Modified rule

P_1: +ve examples covered by R_1

P_0: +ve tuples coved by R_0

n_0: no. of -ve tuples coved by R_0

Q. Training set 60 +ve examples
100 -ve examples.

$r_1 \rightarrow$ covers 50 +ve examples & 5 -ve examples.

$r_2 \rightarrow$ covers 2 +ve examples & 0 -ve examples.

r_1 : $e_i = 55 \times 60 / 160 = 20.625$
(for +ve examples)

e_i (for -ve examples $= 55 \times 100 / 160 = 34.375$

$$R = 2 \left[50 \times \log_2 \left(\frac{50}{20.625} \right) + 5 \times \log_2 \left(\frac{5}{34.375} \right) \right]$$

$R = 99.9$

Prerna

r_2

e_i (for +ve examples) $= 2 \times 60/160 = 0.75$

e_i (for -ve examples) $= 2 \times 100/160 = 1.25$

$$R = 2\left[-2 \times \log_2\left(\frac{2}{0.75}\right) + 0 \times \log_2\left(\frac{0}{1.25}\right)\right]$$

$R = 5.66$

Laplace Measure

$$(\text{for } r_1) = \frac{50+1}{55+2} = 0.894 = 89.4\%$$

$$(\text{for } r_2) = \frac{2+1}{2+2} = 0.75 = 75\%$$

$p_+ = 0.2$

r_1:

$$\text{FOIL Grain}(R_0, r_1) = 50\left[\log_2\left(\frac{50}{50+5}\right) - \log_2\left(\frac{60}{60+100}\right)\right]$$

modified rule ←

original rule ←

$$\text{FOIL Grain}(r_1, r_2) = 2 \times \left[\log_2\left(\frac{2}{2+0}\right) - \log_2\left(\frac{50}{50+5}\right)\right]$$

Q. training set 100 +ve examples
400 -ve examples

$r_1 \rightarrow$ covers 4 +ve examples & 1 -ve examples.

$r_2 \rightarrow$ covers 30 +ve examples & 10 -ve examples.

$k = 2 \qquad p_+ = 0.2$

r_1

$$m\text{-estimate} = \frac{4 + 2 \times 0.2}{5+2} = 62.857\%$$

r_2

$$m\text{-estimate} = \frac{30 + 2 \times 0.2}{40+2} = 72.380\%$$

* RIPPER Algorithm (general to specific)

→ for datasets - with imbalanced class destributions.
→ suitable for noisy data sets
→ uses validation / prune set.

2 class problem

↳ select class with a majority → default class. (Assign)
↳ find rules for class (minor)

multiclass problem

$y = \{ y_1, y_2 \ldots y_n \}$

$y_1 \rightarrow$ less frequent class

$y_c \rightarrow$ default class. / max. freq.

y_1 = least majority class.

y_c = max. majority class.

1st iteration

y_1

+ve examples are examples belonging to class y_1

-ve examples are examples not belonging to class y_1

R_0: ( ) → y_1 } Empty Rule

r_1

FOIL gain (R_0, r_1)

① r_1 → Rule Pruning ②

FOIL Prune value = $\frac{p-n}{p+n}$

eg. $r_1(C_1 \wedge C_2 \wedge C_3) \Rightarrow y_1$

FOIL Prune value(r_1)

$r_1'(C_1 \wedge C_2) \Rightarrow y_1$

FOIL Prune Value (r_1')

add rule to rule set

remove tuples covered by r_1

Stopping Condition

→ d bits (d = 64 bits)

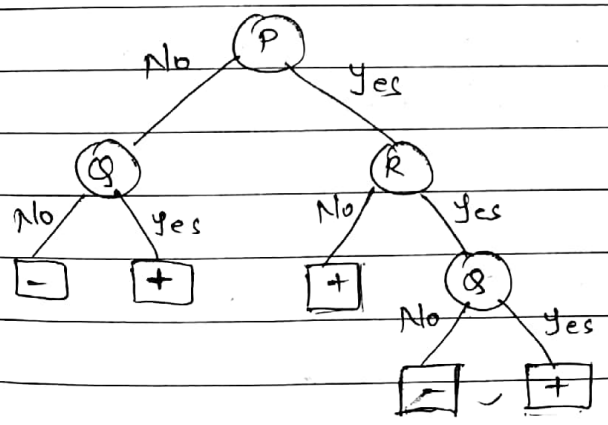
→ If error rate > 50% on Prune set

* consequent = class

11/10/19

Indirect method for rule extraction

eg:



P = No ∧ Q = No ⇒ −

P = No ∧ Q = Yes ⇒ +

P = Yes ∧ R = No ⇒ +

P = Yes ∧ R = Yes ∧ Q = Yes ⇒ +

P = Yes ∧ R = Yes ∧ Q = No ⇒ −

sort

P = No ∧ Q = No ⇒ −

P = Yes ∧ R = Yes ∧ Q = No ⇒ −

P = No ∧ Q = Yes ⇒ +

P = Yes ∧ R = No ⇒ +

P = Yes ∧ R = Yes ∧ Q = Yes ⇒ +

classmate
Date
Page

Reduce Analysis

$\therefore$ Q = Yes $\rightarrow$ +

P = Yes $\wedge$ R = No $\rightarrow$ +

Q = No $\rightarrow$ −

} AND

* Characteristics

→ The expensiveness of a rule base classifier is equivalent to the decision tree since the decision tree can be represented as a set of mutually exclusive & exhausted rules.

→ They are used to produce descriptive models that are easier to interpret but gives the comparison perfo. mac. to the decision tree classifier.

→ The class based ordering approach is suitable for data sets with imbalanced class distribution.

* Nearest neighbour based classification

Rote classifier : class for tuples are assigned to test records
- searches for exact match then assign class.
drawback: doesn't get exact match

* K- nearest neighboure
- called as Lazy - learning Technique : do not perform training

↳ majority voting

↳ inverse distance weighted approach

~~Acid Durability~~

X_1 = Acid Durability	X_2 = Acid Strength	Y = Acid Quality
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

test record = ($X_1 = 3, X_2 = 7$)

X_1	X_2	Y	dist
7	7	Bad	$d = \|7-3\| + \|7-7\| = 4$ ✓ ✓
7	4	Bad	$d = \|7-3\| + \|4-7\| = 7$
3	4	Good	$d = \|3-3\| + \|4-7\| = 3$ ✓ ✓
1	4	Good	$d = \|1-3\| + \|4-7\| = 5$ ✓

$k = 1$

$X_1 = 3, X_2 = 7$ = Good

$k = 2$

$X_1 = 3, X_2 = 7$ = Bad

$k = 3$

$X_1 = 3, X_2 = 7$ = Good

* Inverse distance weighted voting $y' = \underset{(x_i, y_i \in D_z)}{\text{argman}} \sum w_i \times I(v = y_i)$

for class bad

$= \frac{1}{(4)^2} \times 1 + \frac{1}{(3)^2} \times 0 = 0.0625$

for class good

$= \frac{1}{(4)^2} \times 0 + \frac{1}{(3)^2} \times 1 = 0.1111$

$\therefore y' = \underset{\text{greater}}{\text{argman}}(\overset{\text{bad}}{0.0625}, \overset{\text{good}}{0.111}) = \text{good}$

12/10/19

$k=1, k=3, k=4$ $x'=5$

	X	Y	distance from x'
1)	0.1	−	d = 4.9
2)	0.7	+	4.3
3)	1	+	4
4)	1.6	−	3.4 ✓
5)	2	+	3
6)	2.5	+	2.5 ✓
7)	3.2	−	1.8 ✓
8)	3.5	−	1.5 ✓
9)	4.1	+	0.9 ✓
10)	4.9	+	0.1 ✓

$k=1$ ∴ Record 10 is nearest (class = '+')

$x'=5 \Rightarrow y'=+$

$k=3$

$x'=5 \Rightarrow y'=+$

Record 8 → −
9 → +
10 → +
using majority voting

$k=4$

$x'=5 \Rightarrow y'=+$

Record 7 → −
8 → −
9 → +
10 → +

Inverse distance weighted measure

for '+'

$$= \frac{1}{(1.8)^2} \times 0 + \frac{1}{(1.5)^2} \times 0 + \frac{1}{(0.9)^2} \times 1 + \frac{1}{(0.1)^2} \times 1$$

w_i

$$= 101.23$$

for '-'

$$= \frac{1}{(1.8)^2} \times 1 + \frac{1}{(1.5)^2} \times 1 + \frac{1}{(0.9)^2} \times 0 + \frac{1}{(0.1)^2} \times 0$$

$$= 5.56$$

+ −

y' = argman (101.23, 0.730)

y' = +

Characteristics of k-nearest neighbour.

- k-nearest neighbour is part of a more general technique known as instance based learning which uses specific training instance to make predictions without having to maintain and abstraction (all models) derived from the data
- We do not require model building, hence classifying a test example is an expensive approach because we need to compute the proximity value individually b/n the test & the training example.
- Nearest neighbours make the predictions based on local information whereas decision tree are rule base classifier attempt to find the global model that fits the entire i/p space.
- They can produce arbitarily shape decision boundary.
- Its highly dependent on the promity measures and can produce wrong predictions unless the correct proximity measures & data free processing steps are taken.

12/10/19 MOD 4

* CLUSTERING : grouping closely related data
- Unsupervised learning technique.

↳ In the context of understanding data clusters are potential classes & cluster analysis is the study of automatically finding the classes.

* Applications of Clustering
→ information retrieval
→ climatic data
→ medicine

→ Utility & Summarisation

→ Compression : instead of using entire dataset it uses clusters
-reduce data

→ finding nearest neighbours:

* Types of Clusters
→ Well seperated Clusters : are ideal clusters
- inter similarity should be small value

→ prototype based ← centroid : avg. of all points
medoid : one point taken from clusters

→ Graph based clusters : every object is node

→ density based clusters : if points are in high density → clusters
: " " " " low " → outliers

→ shared / conceptual clusters : they cluster together if they have some common property.

* Types of Clustering Techniques

→ Partitioning : kmeans, k medoid

→ hierarchical
  ↳ nested clusters

→ exclusive clustering : single turn clusters
  : one cluster contain only one record

→ overlapping :-

→ Fuzzy clustering : soft clustering eg: 40% correct & 60% wrong
  ↳ membership function 0-1

→ Probabilistic Clustering :-

* Kmeans Clustering

1] K → no. of clusters
2] randomly choose k points from the data set.
3] Compute the similarities of remaining points from the k points.

1) means are same
2) clustering ... } stopping condition.
3) doesn't convert

Then stop after certain no. of iterations

Q.

Object	x	y
A	1	1
B	2	1
C	4	3
D	5	4

Initial centroids A(1,1) & B(2,1)

no. of clusters $k = 2$

	Cluster 1 distance from Centroid A	cluster 2 dist. from Centroid B
A	(0)	1
B	1	(0)
C	3.61	(2.83)
D	5	(4.24)

$dist(A,B) = \sqrt{(1-2)^2 + (1-1)^2}$

$= 1$

$distance(A,C) = \sqrt{(4-1)^2 + (3-1)^2}$

$= 3.61$

$dist(B,C) = \sqrt{(4-2)^2 + (3-1)^2}$

$= 2.83$

$dist(A,D) = \sqrt{(5-1)^2 + (4-1)^2}$

$= 5$

$dist(B,D) = \sqrt{(5-2)^2 + (4-1)^2}$

$= 4.24$

cluster 1 (A)

cluster 2 (B, C, D)

cluster mean for cluster 1 - A(1,1) → Centroid 1

cluster mean for cluster 2 - $\left(\frac{(2+4+5)}{3}, \frac{1+3+4}{3}\right)$

$= 3.66, 2.66$ → Centroid 2

	dist from Centroid 1	dist. from centroid 2
A	(0)	3.14
B	(1)	2.36
C	3.61	(0.47)
D	5	(1.89)

dist (A, Centroid 2) $= \sqrt{(1-3.67)^2 + (1-2.67)^2}$

$= 2.07$

dist (B, Centroid 2)

$= \sqrt{(2-3.67)^2 + (1-2.67)^2}$

$= 2.36$

dist (C, Cent 1)

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

$= 3.61$

cluster 1 A, B

cluster 2 C, D

cluster 1 mean $= \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1.5, 1) \rightarrow$ centroid 1

2 mean $= \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4.5, 3.5) \rightarrow$ centroid 2

	Centroid 1	Centroid 2
A	0.5	4.30
B	0.5	3.5
C	3.20	0.71
D	4.61	0.71

dist (A, Cent 1) $= \sqrt{(1-1.5)^2 + (1-1)^2}$

$= 0.5$

dist (A, 2) $= \sqrt{(1-4.5)^2 + (1-3.5)^2} = 0.5$

dist (C, 1) $= \sqrt{(4-1.5)^2 + (3-1)^2} = 3.20$

dist (D, 1) $= \sqrt{(5-1.5)^2 + (4-1)^2} = 4.609$

dist (D, 2) $= \sqrt{(5-4.5)^2 + (4-3.5)^2} = 0.71$

cluster 1 mean = (1.5, 1) Same clusters so stop

cluster 2 mean = (4.5, 3.5)

Q.

Object	x	y	
1	20	10	
2	30	20	✓
3	30	30	
4	35	35	
5	40	40	
6	50	45	~

Object 2 (30,20) and Object 5 (40,50) → initial centroids.

	Centroid 1	Centroid 2
1	14.14	44.72
2	0 ✓	31.62
3	10 ~	22.36
4	15.81 ✓	15.81
5	22.36	10
6	32.01	11.18

dist(1, 1) = $\sqrt{(20-30)^2 + (10-20)^2}$ = 14.14

dist(1, Cent2) = $\sqrt{(20-40)^2 + (10-50)^2}$ = 44.72

dist(2, 1) = $\sqrt{(30-30)^2 + (20-20)^2}$ = 0

dist(2, 2) = $\sqrt{(30-40)^2 + (20-50)^2}$ = 31.62

dist(3, 1) = $\sqrt{(30-30)^2 + (30-20)^2}$ = 10

dist(3, 2) = $\sqrt{(30-40)^2 + (30-50)^2}$ = 22.36

dist(4, 1) = $\sqrt{(35-30)^2 + (35-20)^2}$ = 15.81

dist(4, 2) = $\sqrt{(35-40)^2 + (35-50)^2}$ = 15.81

dist(5, 1) = $\sqrt{(40-30)^2 + (40-20)^2}$ = 22.36

dist(5, 2) = $\sqrt{(40-40)^2 + (40-50)^2}$ = 10

dist(6, 1) = $\sqrt{(50-30)^2 + (45-20)^2}$ = 32.01

dist(6, 2) = $\sqrt{(50-40)^2 + (45-50)^2}$ = 11.18

Cluster 1 (1, 2, 3, 4)

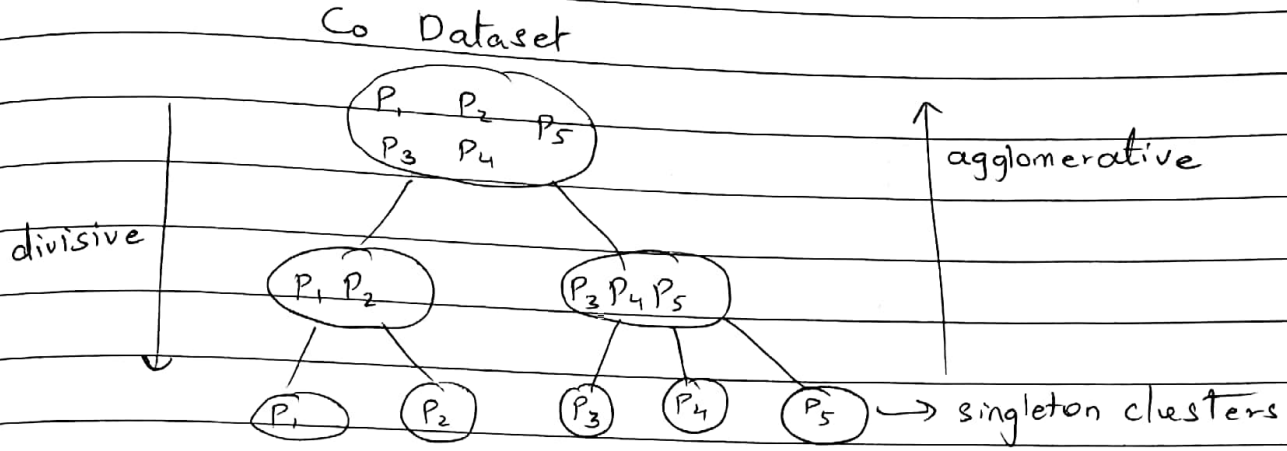
Cluster 2 (5, 6)

Hierarchical Clustering

↳ creation of nested clusters

↳ divisive

↳ agglomerative (AGNES) Clustering



Agglomerative Clustering

Proximity

↳ single linkage (min)

↳ complete linkage (max)

↳ group average

↳ Ward's distance

Q. Single Linkage (min)

Point	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	0.30

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0	0.234	0.215	0.367	0.341	0.235
P_2	0.234	0	0.143	0.194	0.143	0.243
P_3	0.215	0.143	0	0.158	0.284	0.101
P_4	0.367	0.194	0.158	0	0.284	0.219
P_5	0.341	0.143	0.284	0.284	0	0.386
P_6	0.235	0.243	0.101	0.219	0.386	0

$(P_1, P_2) = \sqrt{(0.40-0.22)^2 + (0.53-0.38)^2} = 0.234$

$(P_1, P_3) = 0.2158$

$(P_1, P_4) = \sqrt{(0.40-0.26)^2 + (0.53-0.19)^2} = 0.367$

$(P_1, P_5) = \sqrt{(0.40-0.08)^2 + (0.53-0.41)^2} = 0.341$

$(P_1, P_6) = \sqrt{(0.40-0.45)^2 + (0.53-0.30)^2} = 0.235$

$(P_2, P_3) = \sqrt{(0.22-0.35)^2 + (0.38-0.32)^2} = 0.143$

$(P_2, P_4) = \sqrt{(0.22-0.26)^2 + (0.38-0.19)^2} = 0.194$

$(P_2, P_5) = \sqrt{(0.22-0.08)^2 + (0.38-0.41)^2} = 0.143$

min dist betn P_3 and P_6

→ merge P_3, P_6

	P_1	P_2	$P_{3,6}$	P_4	P_5
P_1	0	0.234	0.215	0.367	0.341
P_2	0.234	0	0.143	0.194	0.143
$P_{3,6}$	0.215	0.143	0	0.158	0.284
P_4	0.367	0.194	0.158	0	0.284
P_5	0.341	0.143	0.284	0.284	0

dist $(P_{3,6}, P_1) = \min(d_{3,1}, d_{6,1})$

$= \min(0.215, 0.235)$

$= 0.215$

dist $(P_{3,6}, P_2) = \min(d_{3,2}, d_{6,2})$

$= \min(0.143, 0.243)$

$= 0.143$

$dist(P_{3,6}, P_4) = \min(d_{3,4}, d_{6,4})$

$= \min(0.158, 0.219)$

$= 0.158$

$dist(P_{3,6}, P_5) = \min(d_{3,5}, d_{6,5})$

$= \min(0.284, 0.386)$

$= 0.284$

min dist bln P_2 & P_5,

merge P_2 & P_5

	P_1	$P_{2,5}$	$P_{3,6}$	P_4
P_1	0	0.234	0.215	0.367
$P_{2,5}$	0.234	0	0.143	0.194
$P_{3,6}$	0.215	0.143	0	0.158
P_4	0.367	0.194	0.158	0

$dist(P_{2,5}, P_1) = \min(d_{2,1}, d_{5,1})$

$= \min(0.234, 0.341)$

$= 0.234$

$dist(P_{2,5}, P_{3,6}) = \min(d_{2,3.6}, d_{5,3.6})$ ~~$(d_{2,3}, d_{2,6}, d_{5,3}, d_{5,6})$~~ ✓ correct

$= \min(0.143, 0.284)$

$= 0.143$

$dist(P_{2,5}, P_4) = \min(d_{2,4}, d_{5,4})$

$= \min(0.194, 0.284)$

$= 0.194$

min dis bln $P_{2,5}$ & $P_{3,6}$

merge $P_{2,5}$ & $P_{3,6}$

	P_1	$P_{2,3,5,6}$	P_4
P_1	0	0.215	0.367
$P_{2,3,5,6}$	0.215	0	0.158
P_4	0.367	0.158	0

$dist(P_{2,3,5,6}, P_1) = \min(d_{2,1}, d_{3,1}, d_{5,1}, d_{6,1})$

$= \min(0.234, 0.215, 0.341, 0.235)$

$= 0.215$

$dist(P_{2,3,5,6}, P_4) = \min(d_{2,4}, d_{3,4}, d_{5,4}, d_{6,4})$

$= \min(0.194, 0.158, 0.284, 0.219)$

$= 0.158$

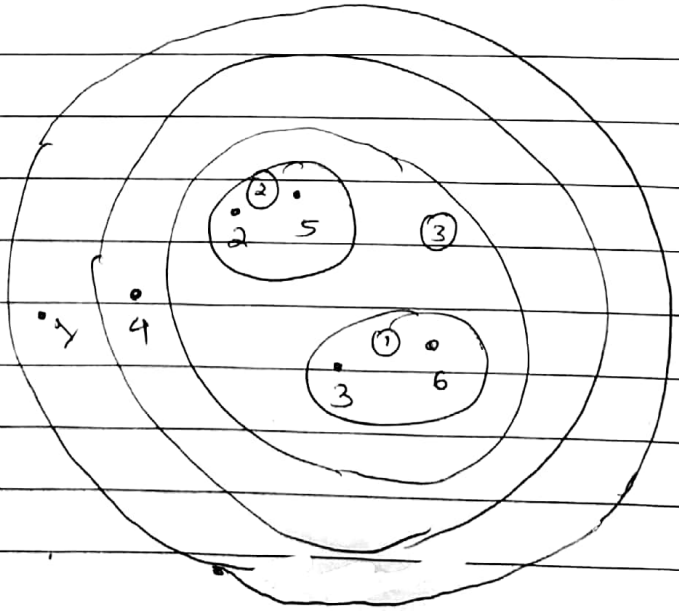
min dist b/n $P_{2,3,5,6}$ & P_4

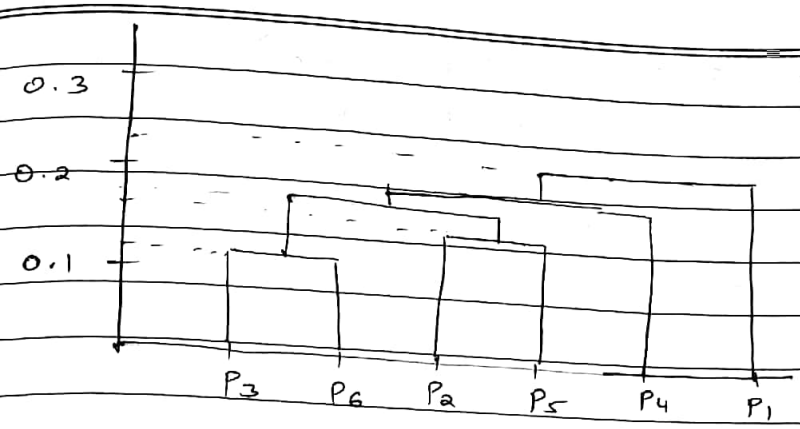
	P_1	$P_{2,3,4,5,6}$
P_1	0	0.215
$P_{2,3,4,5,6}$	0.215	0

$dist(P_{2,3,4,5,6}, P_1) = \min(d_{2,1}, d_{3,1}, d_{4,1}, d_{5,1}, d_{6,1})$

$= \min(0.234, 0.215, 0.367, 0.341, 0.235)$

$= 0.215$





Q. Complete Linkage Problem

Point	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	0.30

Soln

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0	0.24	0.22	0.37	0.34	0.23
P_2	0.24	0	0.15	0.20	0.14	0.25
P_3	0.22	0.15	0	0.15	0.28	(0.11)
P_4	0.37	0.20	0.15	0	0.29	0.22
P_5	0.34	0.14	0.28	0.29	0	0.39
P_6	0.23	0.25	(0.11)	0.22	0.39	0

min dist P_3 & P_6

→ merge P_3 & P_6

	P_1	P_2	$P_{3,6}$	P_4	P_5
P_1	0	0.24	0.23	0.37	0.34
P_2	0.24	0	0.25	0.20	(0.14)
$P_{3,6}$	0.23	0.25	0	0.22	0.39
P_4	0.37	0.20	0.22	0	0.29
P_5	0.34	(0.14)	0.39	0.29	0

$dist(P_1, P_{3,6}) = max(d_{1,3}, d_{1,6})$

$= max(0.22, 0.23)$

$= 0.23$

$dist(P_2, P_{3,6}) = max(d_{2,3}, d_{2,6})$

$= max(0.15, 0.25)$

$= 0.25$

$dist(P_4, P_{3,6}) = max(d_{4,3}, d_{4,6})$

$= 0.22$

$dist(P_5, P_{3,6}) = max(d_{5,3}, d_{5,6})$

$= 0.39$

min dist P_2 and P_5

merge P_2 & P_5

	P_1	$P_{2,5}$	$P_{3,6}$	P_4
P_1	0	0.34	0.23	0.37
$P_{2,5}$	0.34	0	0.39	0.29
$P_{3,6}$	0.23	0.39	0	0.22
P_4	0.37	0.29	0.22	0

$dist(P_1, P_{2,5}) = max(d_{1,2}, d_{1,5})$

$= 0.34$

$dist(P_{2,5}, P_{3,6}) = max(d_{2,3,6}, d_{5,3,6})$

$= 0.39$

$dist(P_4, P_{2,5}) = max(d_{4,2}, d_{4,5})$

$= 0.29$

min dist P_4 & $P_{3,6}$

merge P_4 & $P_{3,6}$

	P_1	$P_{2,5}$	$P_{3,4,6}$
P_1	0	0.34	0.37
$P_{2,5}$	0.34	0	0.39
$P_{3,4,6}$	0.37	0.39	0

$dist(P_1, P_{3,4,6}) = max(d_{1,3}, d_{1,4}, d_{1,6})$

$= 0.37$

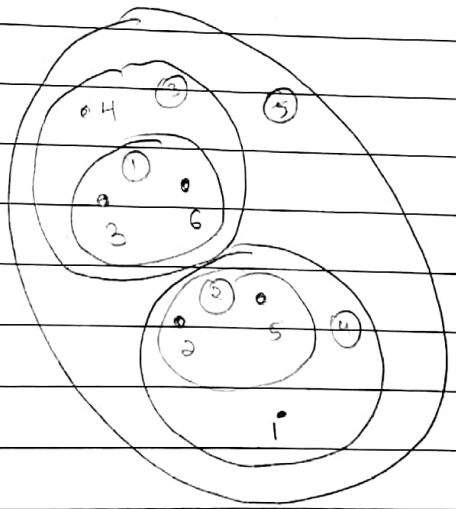
$dist(P_{2,5}, P_{3,4,6}) =$ ~~max~~ 0.39

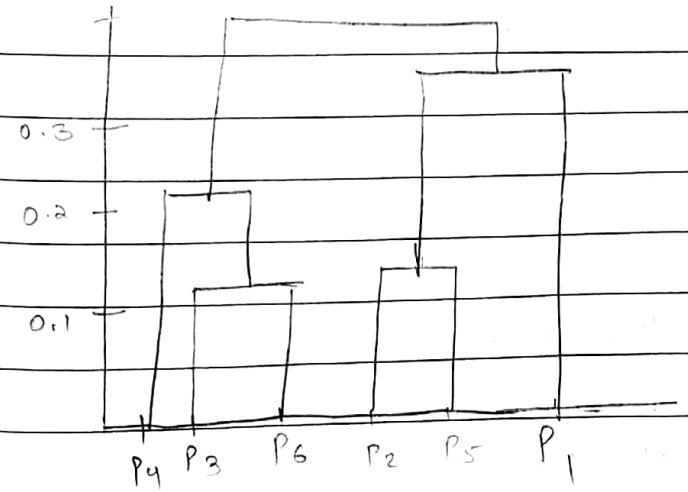
min dist P_1 and $P_{2,5}$

merge P_1 and $P_{2,5}$

	$P_{1,2,5}$	$P_{3,4,6}$
$P_{1,2,5}$	0	0.39
$P_{3,4,6}$	0.39	0

$dist(P_{1,2,5}, P_{3,4,6}) = 0.39$





* Group average method

$$\text{proximity}(C_i, C_j) = \frac{\sum_{x \in C_i} \sum_{y \in C_j} \text{proximity}(x, y)}{m_i \times m_j}$$

m_i = size of cluster C_i

m_j = size of Cluster C_j

Q.

Point	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	0.30

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0	0.24	0.22	0.37	0.34	0.23
P_2	0.24	0	0.15	0.20	0.14	0.25
P_3	0.22	0.15	0	0.15	0.28	0.11
P_4	0.37	0.20	0.15	0	0.29	0.22
P_5	0.34	0.14	0.28	0.29	0	0.39
P_6	0.23	0.25	0.11	0.22	0.39	0

min dist P_3 & P_6 → merge P_3 & P_6

	P_1	P_2	$P_{3,6}$	P_4	P_5
P_1	0	0.24	0.225	0.37	0.34
P_2	0.24	0	0.2	0.20	0.14
$P_{3,6}$	0.225	0.2	0	0.185	0.335
P_4	0.37	0.20	0.185	0	0.29
P_5	0.34	0.14	0.335	0.29	0

dist $(P_1, P_{3,6})$ = (0.22 + 0.23) / (1*2) = 0.225

dist (P_2, $P_{3,6}$) = (0.15 + 0.25) / (1 * 2) = 0.2

dist (P_4, $P_{3,6}$) = (0.15 + 0.22) / (1 * 2) = 0.185

dist (P_5, $P_{3,6}$) = (0.28 + 0.39) / (1 * 2) = 0.335

min dist bln P_2 & P_5 → merge P_2 & P_5

	P_1	$P_{2,5}$	$P_{3,6}$	P_4
P_1	0	0.295	0.225	0.37
$P_{2,5}$	0.295	0		
$P_{3,6}$	0.225		0	0.185
P_4	0.37		0.185	0

dist (P_1, $P_{2,5}$) = (0.25 + 0.34) / (1 * 2) = 0.295

18/10/19

9/10/19

Hierarchical Clustering

↳ Wards distance

SSE → Sum of squared errors.

$$\text{Cluster SSE} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$\bar{x}$ → Centroid of the cluster.

$$\text{Total SSE} = \sum_{j=1}^{k} \sum_{x_i \in C_j}^{n} (x_i - \bar{x})$$

k = no. of clusters

Increase in SSE – I_{AB}

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$

$$SSE_A = \sum_{L=1}^{n_A} (a_i - \bar{a})^2$$

$$SSE_B = \sum_{L=1}^{n_B} (b_i - \bar{b})^2$$

$$SSE_{AB} = \sum_{L=1}^{n_{AB}} (y_i - \bar{y}_{AB})^2$$

a_i – data object belonging to cluster A
b_i – data object belonging to cluster B
y_i – data object belonging to cluster AB
$\bar{a}$ – centroid of cluster A
$\bar{b}$ – centroid of cluster B
$\bar{y}_{AB}$ – centroid of cluster AB

AB = avg

Q.

	x	y
p1	0.4	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.3

Wards distance :

Q. DBSCAN Problem

	x	y
P_1	2	2
P_2	4	4
P_3	6	6
P_4	0	6
P_5	6	0
P_6	5	5
P_7	7	7
P_8	9	9

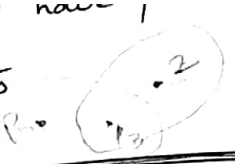
min pts = 3

$\varepsilon = 3$

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
P_1	0	(2.83)	5.65	4.47	4.47	4.24	7.07	9.90
P_2	(2.83)	0	(2.83)	4.47	4.47	(1.41)	4.24	7.07
P_3	5.66	(2.83)	0	6.00	6.00	(1.41)	(1.41)	4.24
P_4	4.47	4.47	6.0	0	8.49	5.10	7.07	9.49
P_5	4.47	4.47	6.0	8.49	0		7.07	9.49
P_6	4.24	(1.41)	(1.41)	5.10	5.10	0	(2.83)	5.66
P_7	7.07	4.24	(1.41)	7.07	7.07	(2.83)	0	(2.83)
P_8	9.90	7.07	4.24	9.49	9.49	5.66	(2.83)	0

core point should have

boundary pts: less no. of pts < min pts



	ε- neighbourhood	no. of points
P_1 -	P_1 P_2	2
P_2 -	P_1 P_2 P_3 P_6 — core point	4
P_3 -	P_2 P_3 P_6 P_7 — core point	4
P_4 -	-	-
P_5	-	-
P_6	P_2 P_3 P_6 P_7 — core point	4
P_7	P_3 P_7 P_6 P_8 — core point	4
P_8	P_8 P_7	2

P_4 & P_5 - noise points (neither core/border points)

P_1 → border point (< minpts) & lies in ε-neighbourhood of P_2

P_8 → border point (lies in ε-neighbourhood of P_7) & (< minpts)

P_1 is density reachable from P_2 (lies in ε-neighbourhood of P_2 (corepoint))

P_3 is density reachable from P_2 (lies in ε-neighbourhood of P_2 (corepoint))

P_6 is density reachable from P_2

P_2 is density reachable from P_3

P_7 is density reachable from P_3

hence P_2 & P_7 are density connected

P_3 → density reachable from P_7

P_8 → " " " P_7

P_3 & P_8 are density connected

So $P_1, P_2, P_3, P_6, P_7, P_8$ lie in the same cluster

→ P_4 P_5 → outliers / noise pts

classmate
Date
Page

Mean

Q. Differentiate bln Eigen learning & Lazy learning. (or) Why is k-nearest neighbour called as a lazy classifier. Compare its performance to a decision tree classifier

→ Eager learner:
Always build model.
Decision tree is better

Q. Explain the different rule order scheme & list its advantages & disadvantages.

Q. Explain the diff. voting strategies employed in k-nearest neighbours algorithm & illustrate each strategy.
→ majority & distance weighted approach.

Q. Explain the learn one rule function with the suitable example:

Q. Explain rule induction using sequential covering algorithm.

Q. Define accuracy & coverage of the rule. with respect to rule base classifier.

Q. What is rule pruning, How it is carried out.

Q. Explain the RIPPLE algorithm :- it is a direct algorithm.

Q. Explain the diff. types of clusters & explain the diff. clustering techniques.

Q. Explain k-means algorithm / state k-means.

Q. State agglomerative hierarchical clustering algorithm.

Q. State the DBSCAN algorithm.

- Definitions related to density based clustering: density reachability, density connected; Core points; Noise points

Clustering

selecting centroids

SSE (sum of squared errors)

- Ideally, Euclidean distance is used to find distance of each point from it's centroid.

Cluster SSE

$$\text{cluster SSE} = \sum_{\substack{i=1 \\ x_i \in \text{cluster}}}^{n} (x_i - C_0)^2$$

$C_0 \rightarrow$ centroid of the cluster

Total SSE

- Sum of all cluster SSEs.

$$\text{Total SSE} = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ x_j \in C_i}}^{n} (x_j - C_i')^2$$

$C_i' \rightarrow$ centroid of cluster i

For document data

cosine similarity

$$\text{Total cohesion} = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ x_j \in C_i}}^{n} \text{Cosine}(x_j, C_i')$$

cluster cohesion

Issues with K-means clustering

→ handling empty clusters

→ Outliers

Use outlier detection.

Postprocessing with SSE

To minimise SSE

↔ reduce no. of clusters

↳ disperse a cluster

↳ select a cluster with least SSE.

Redistribute points to remaining clusters

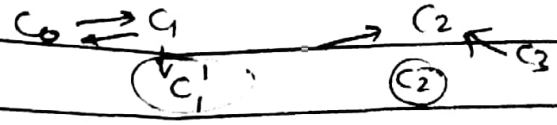
↳ merge two clusters

↳ merge two clusters having the least SSE.

→ split the clusters.

↳ split clusters having highest SSE.

Incrementally updating centroid.



Strengths and Weaknesses

Lance William Formula

$$\rho(R, Q) = \alpha_A \rho(A, Q) + \alpha_B \rho(B, Q) + \beta \rho(A, B) + \gamma |\rho(A, Q) - \rho(B, Q)|$$

	α_A	α_B	β	γ
single linkage	$1/2$	$1/2$	0	$-1/2$
complete linkage	$1/2$	$1/2$	0	$+1/2$
Group average	$\frac{m_A}{m_A + m_B}$	$\frac{m_B}{m_B + m_A}$	0	0
Ward's	$\frac{m_A + m_Q}{m_A + m_B + m_Q}$	$\frac{m_B + m_Q}{m_A + m_B + m_Q}$	$\frac{-m_Q}{m_A + m_B + m_Q}$	

→ lack of global objective function.

→ handling varying cluster size.. → weighted approach
→ unweighted approach

→ merge decisions are final. (can't reassign point to another cluster only merged)

→ outliers

single / complete linkage → last iteration may be eliminated as it is an outlier.

- Drawback of hierarchical clustering is that, it is ~~strength~~ - computationally expensive because it have to find distance between one point to another.

DBSCAN

Strength

→ Detect outliers

Weakness

→ High dimensional data can't be handled.

→ Cannot handle data of different density.

→ Computationally expensive as we have to find distance between all points.

K-medoid clustering

medoid 1, medoid 2

$K = 2$ ✓ initial medoids → 2, 4

dataset	dist from medoid 1	dist from medoid 2
2	0 ✓	2
4	2	0
10	8	6
12	10	8
3	1 ✓	3
20	18	16
30	28	26
11	9	7
25	23	21

	pt		absolute error criterion
cluster 1	2	0	$= \sum_{i=1}^{K} \sum_{j=1}^{n} \lvert x_j - c_i' \rvert$
for medoid 1 →	3	1	
cluster 2	4	0	$x_j \in c_i$
for medoid 2 →	10	6	
	12	8	$= \lvert 2-2 \rvert + \lvert 2-3 \rvert$
	(20)	16	$+ \lvert 4-4 \rvert + \lvert 4-10 \rvert + \lvert 4-12 \rvert +$
randomly select	30	26	$\lvert 4-20 \rvert + \lvert 4-30 \rvert + \lvert 4-11 \rvert$
	11	7	$+ \lvert 4-25 \rvert$
	25	21	$= 85$

old medoids 2, (4)

new medoids 2, (20)

m_1 m_2

medoid 4 is replaced

Absent - as | 10
28/10

	dist m1	dist m2		
2	0	18	cluster 1	2
4	2 ✓	16		4
10	8 ✓	10		10
12	10	8		3
3	1	17		11
20	18	0	cluster 2	12
30	28	10		20
11	9 ✓	9		30
25	23	5		25

	pt	dist
cluster 1	2	0
	4	2
	10	8
	3	1
	11	9
cluster 2	12	8
	20	0
	30	10
	25	5

absolute error criterion = 43
(new cost).

Scanned by CamScanner

10/19

Hierarchical Clustering

BIRCH (Balanced Iterative Reducing Clustering using hierarchies).

Phase 1: An initial Clustering Feature tree is built (CF Tree)

Phase 2: A conventional clustering algorithm such as k-means or k-medoid is applied to cluster the leaf nodes of the CF tree.

Clustering Feature - (n, LS, SS)

linear sum

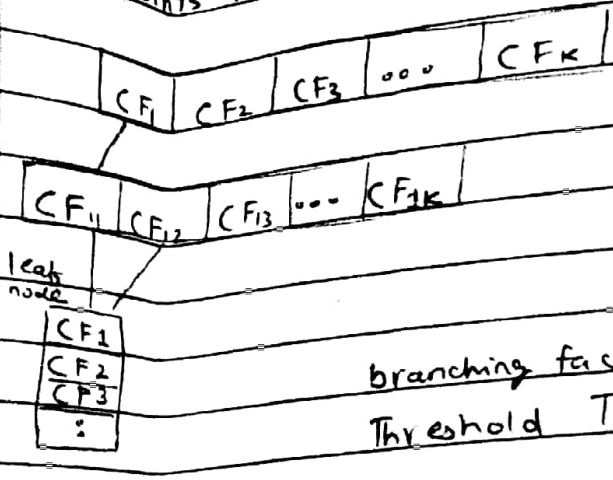
square sum

Radius $R = \sqrt{\frac{nSS - 2LS^2 + n^{LS}}{n^2}}$

Radius is the average distance of the points of the cluster to the centroid

Diameter $D = \sqrt{\frac{2nss - 2LS^2}{n(n-1)}}$

diameter is the average pair-wise distance of the points in a cluster.



branching factor b = 3

Threshold T

	A	B	C	D	E	F
cluster1	(2,5)	(3,2)	(4,3)			
cluster2	(7,3)	(9,10)	(3,8)	(2,3)		

CF1 (cluster 1) = (3, (2+3+4, 5+2+3), ($2^2+3^2+4^2$, $5^2+2^2+3^2$))

= (3, (2+3+4, 5+2+3), ($2^2+3^2+4^2$, $5^2+2^2+3^2$))

= (3, (9, 10), (29, 38))

CF2 (cluster 2) = (4, (7+9+3+2, 3+10+8+5), ($7^2+9^2+3^2+2^2$, $3^2+10^2+8^2+3^2$))

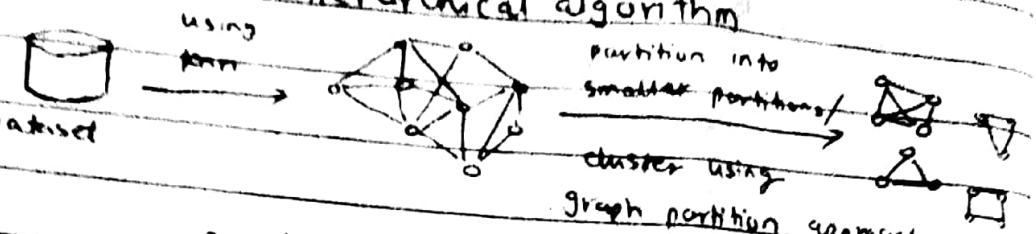
= (4, (21, 24), (143, 182))

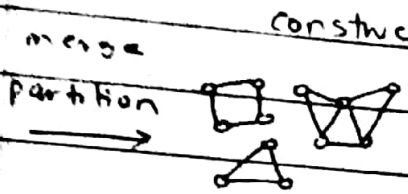
$CF_1 + CF_2 = (n_1+n_2, LS_1+LS_2, SS_1+SS_2)$

= (3+4, (9, 10) + (21, 24), (29, 38) + (143, 182))

* (7, (30, 34), (194, 220))

Chameleon → hierarchical algorithm





construct a sparse graph

Relative interconnectivity (RI).

$$RI(C_i, C_j) = \frac{EC_{|C_i - C_j|}}{\frac{1}{2}(|EC_{C_i}| + |EC_{C_j}|)}$$

EC edge cut

Relative interconnectivity is defined as the absolute interconnectivity between C_i and C_j normalised w.r.t the internal connectivity between C_i and C_j

$EC_{|C_i, C_j|}$ → absolute connectivity

↳ edge cut for cluster containing C_i and C_j

Absolute connectivity

Sum of the edges that connect vertices in C_i and C_j

EC_{C_i} → internal connectivity

↳ edge cut of min bisecter

Internal connectivity → sum of the edges that roughly partition cluster C_i and C_j into two roughly equal partitions.

Scanned by CamScanner

Relative Closeness (RC)

Relative Closeness is defined as the absolute closeness between C_i and C_j normalised w.r.t the internal closeness between C_i and C_i.

$$RC = \frac{\overline{SE}_{C|C_i, C_j|}}{\left(\frac{|C_1|}{|C_1|+|C_2|}\right)\bar{S}E_{C_i} + \left(\frac{|C_1|}{|C_1|+|C_j|}\right)S_{EC_{C_j}}}$$

$\bar{S}EC_{|C_i,C_j|}$ → average weight of the edges of the vertices that connect C_i to C_j

$\bar{S}EC_{C_i}$ → average weight of edges that belong to the min bisector of C_i and C_j.

Scanned by CamScanner

Single linkage

	X	Y
P_1	1	1
P_2	1.5	1.5
P_3	5	5
P_4	3	4
P_5	4	4
P_6	3	3.5

Proximity Matrix

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0	0.71	5.66	3.61	4.24	3.20
P_2	0.71	0	4.95	2.92	3.54	2.50
P_3	5.66	4.95	0	2.24	1.41	2.50
P_4	3.61	2.92	2.24	0	1.00	0.50
P_5	4.24	3.54	1.41	1.00	0	1.12
P_6	3.20	2.50	2.50	0.50	1.12	0

min dist bln P_4 & P_6

$\Rightarrow$ merge P_4 & P_6

	P_1	P_2	P_3	P_4P_6	P_5
P_1	0	0.71	5.66	3.20	4.24
P_2	0.71	0	4.95	2.50	3.54
P_3	5.66	4.95	0	2.24	1.41
P_4P_6	3.20	2.50	2.24	0	1.00
P_5	4.24	3.54	1.41	1.00	0

$dist(P_1, P_4P_6) = \min(d_{1,4}, d_{1,6})$

$= \min(3.61, 3.20) = 3.20$

min dist bln P_1 & P_2

$\Rightarrow$ merge P_1 & P_2

	P_1P_2	P_3	P_4P_6	P_5
P_1P_2	0	4.95	2.50	3.54
P_3	4.95	0	2.24	1.41
P_4P_6	2.50	2.24	0	1.00
P_5	3.54	1.41	1.00	0

min dist bln P_5 & P_4P_6

$\Rightarrow$ merge $P_4P_5P_6$

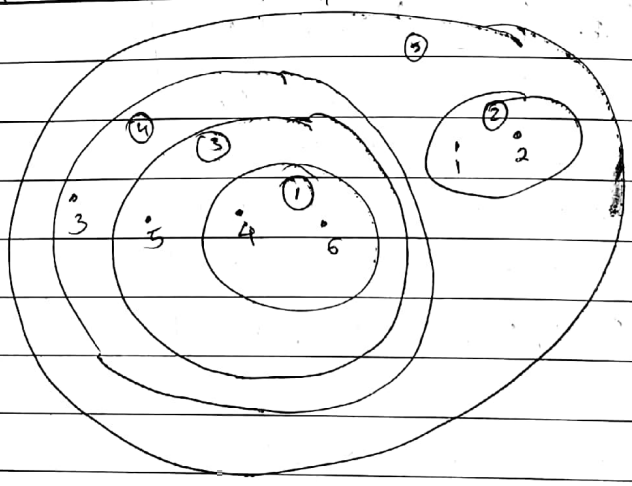
	P_1P_2	P_3	$P_4P_5P_6$
P_1P_2	0	4.95	2.50
P_3	4.95	0	1.41
$P_4P_5P_6$	2.50	1.41	0

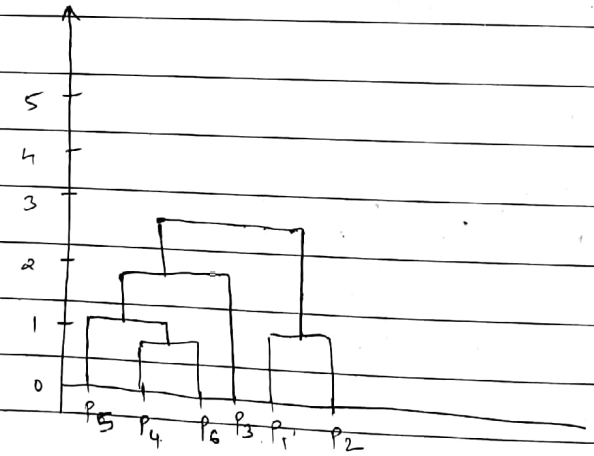
min dist b/n P_3 & $P_4P_5P_6$

⇒ merge P_3 & $P_4P_5P_6$

dist $(P_1P_2, P_3P_4P_5P_6) = 2.50$

	P_1P_2	$P_3P_4P_5P_6$
P_1P_2	0	2.50
$P_3P_4P_5P_6$	2.50	0





DBSCAN

$\varepsilon = 3$

min. points = 3

Note: circle all values which have dist less than equal to ε

Consider same proximity matrix.

P_2 & P_4 : P_6

P_2 & P_3 : P_4

	ε - neighbourhood	no. of points	Type of pt
P_1	P_1, P_2	2	border point
P_2	P_1, P_2, P_4, P_6	4	Core pts
P_3	P_3, P_4, P_5, P_6	4	"
P_4	P_2, P_3, P_4, P_5, P_6	5	"
P_5	P_3, P_4, P_5, P_6	4	"
P_6	P_2, P_3, P_4, P_5, P_6	5	"

P_2 & P_3 are density reachable from P_4, hence P_2 & P_3 are density connected

P_2 & P_4 are density " " P_6, " P_2 & P_4 " " "

P_2 & P_5 " " " " P_4 or P_6, " P_2 & P_5 " " "

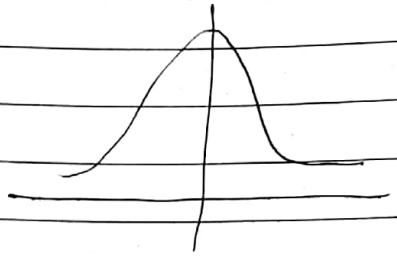
P_2 & P_6 " " " " P_4, " P_2 & P_6 " " "

01/09/19 Statistical based outlier Detection

with

If the value of $|x| \geq c$ such that probability of $(|x|) \geq c = \alpha$, where c is a constant satisfies that condition and α is a probability that a value for the given distribution may be mistakenly classified as outlier

Multivariate Data



Mahalanobis dist $(x, \bar{x}) = (x - \bar{x})\, S^{-1}\, (x - \bar{x})^T$

↑ Multivariate data

↳ mixture model

* Likelihood / log likelihood approach

Likelihood is a measure of the extend to which the sample provides support for particular values of a parameter in parametric mode.

* Likelihood - based outlier detection

1] At time $t=0$, Let $M_t \rightarrow$ contain all objects

$A_t \rightarrow$ empty

$LL_t(D) = LL(M_t) + LL(A_t) \rightarrow$ log likelihood of data set.

2] for each point x that belongs to M_t do

3] ↑ Move x from M_t to A_t to produce new Datasets A_{t+1} & M_{t+1}

4] Compute new log likelihood of D $LL_{t+1}(D) = LL(A_{t+1}) + LL(M_{t+1})$

5] Compute difference $\Delta = LL_t(D) - LL_{t+1}(D)$

6) If $\Delta > c$; c - threshold

7) x is an anomaly → M_{t+1} & A_{t+1} → become current data sets.

8) end if

9) end for.

Note:- It is suitable for data containing more than one distribution.

* Strengths & Weakness

→ not efficient for multivariate data / data set with mixed distribution.

* Proximity based Outlier Detection
↳ $O(m^2)$

It checks whether an object is outlier based on the distance to k-nearest neighbour.

* Density based outlier detection

→

An outlier score of an object is the inverse of the density around the object.

Density of the given object is a reciprocal of the average distance to the k-neirest neighbours. It is given as

$$density(x,k) = \left(\frac{\sum_{y \in N(x,k)} dist(x,y)}{|N(x,k)|} \right)^{-1}$$

* Average relative density (x, k)

$$= \frac{\text{density}(x,k)}{\sum_{y \in N(x,k)} \text{density}(y,k) / |N(x,k)|}$$

UNIT 4:

Association Mining: How two strongly the attributes are related to each other.

eg: Market Basket Analysis

List of Customer transaction (D)

↳ Transaction T

I - set of items

$T \subseteq I$

Support : % of transaction containing a given itemset

Confidence is the probability that in given item set. if item A is brought then item B is also brought. It verify the degree of certainity for given association A→B

Frequent itemset: frequency occurs in the itemset.

Eg. 10 transaction

40%

= 40% of 10 transaction

= 4

Support → min support

Support count is occurence frequency of a given item set in a list of transaction.

Support = P(AUB)

$$\text{Confidence} = P(B/A) = \frac{\text{Sup\_count}(A \cup B)}{\text{Sup\_count}(A)}$$

A → B

↳ min confidence.

classmate
Date
Page

1. Generate frequent element - min support
2. Generate strong association rules - min conflicts

frequent subsequence → common sequence occuring find the common subsequence.

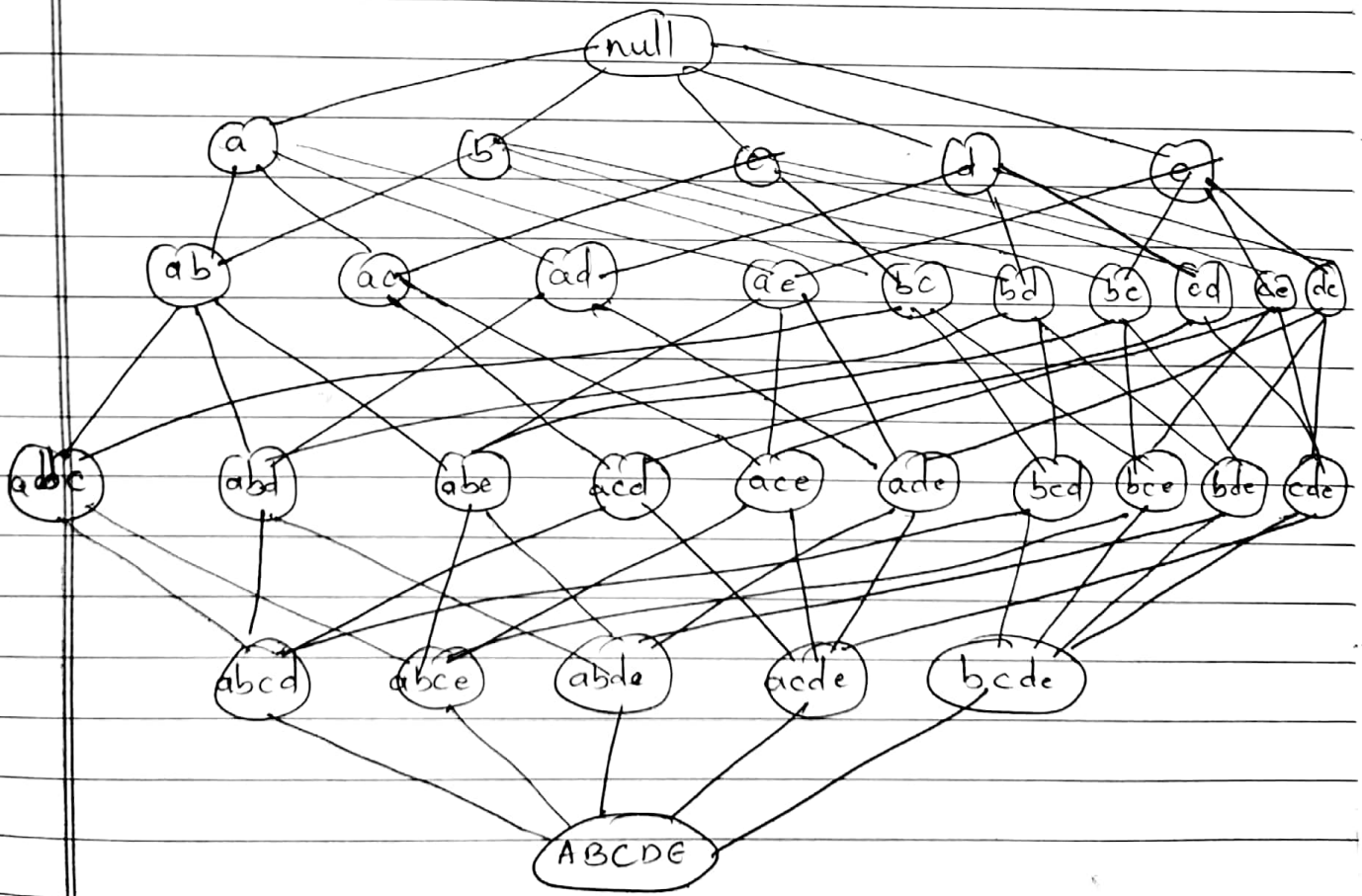
Frequent substructure → tree, graph data

Maximal freq.

Is a item set where none of item set are immediate itemset are frequent.

~~Maximal freq~~

Every item above the border is frequency & every item below the border is not frequent.



26/08/19 closed itemset : ~~if none of immediate s~~
closed frequent element

* Closed itemset:
If item set x is said to be closed itemset if none of the immediate superset have the same support count as x

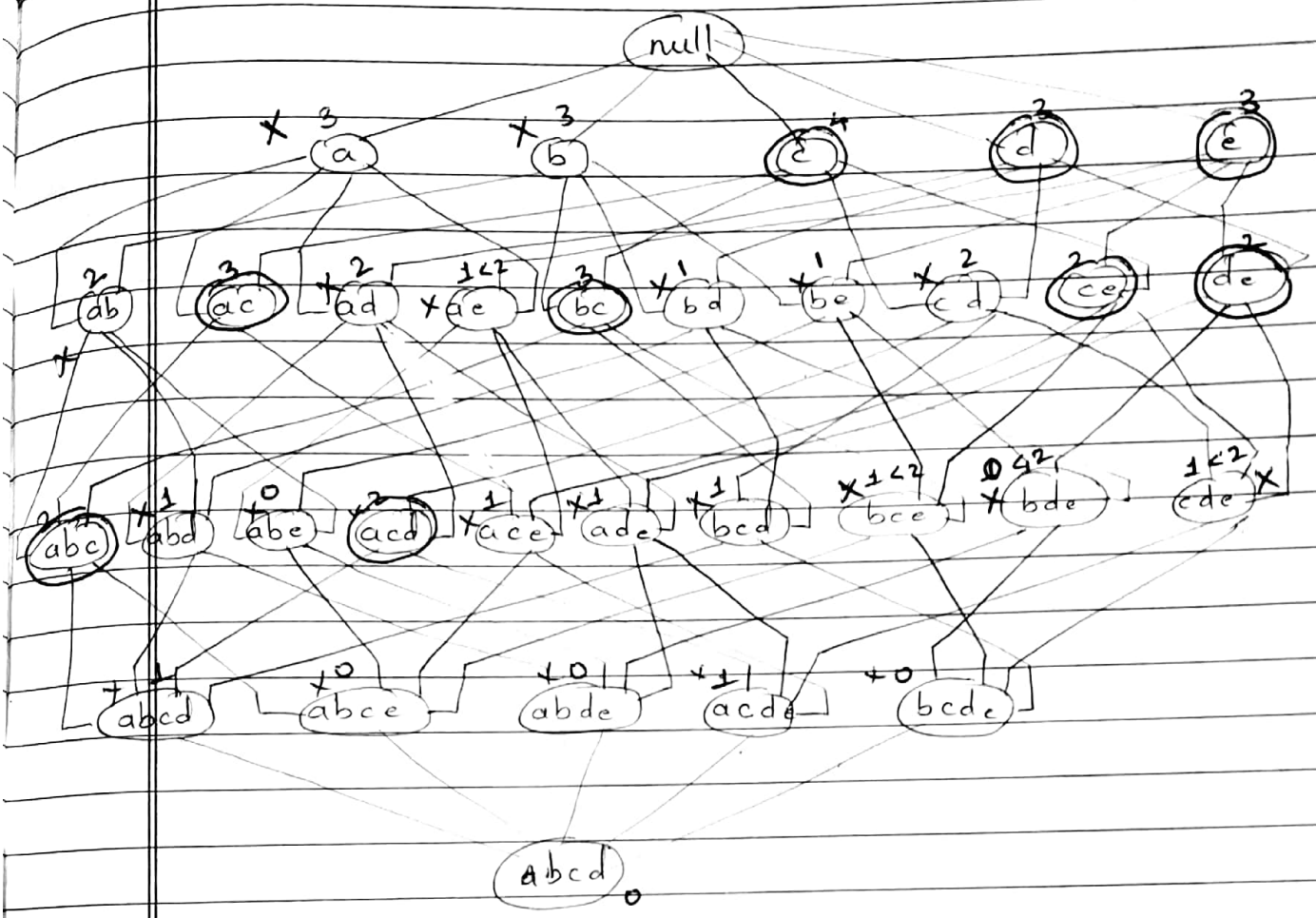
* Closed frequent Element :-
If item set x is a closed frequent itemset if x is closed (none of its imm. superset have the same support count as x) and x has a support count which is greater than equal to the minimum support count

Tid	Itemset
1	a,b,c
2	a,b,c,d
3	b,c,e
4	a,c,d,e
5	d,e

$\frac{50}{100} \times 5$

min sup = 40%

min sup_cnt = 40% of total transaction.
= 40% of 5
$= \frac{40}{100} \times 5$
$= 2$



$a^3 = \{ab^2\}\ \{ac^3\}\ \{ad^2\}\ \{ae^1\}$ ✗ Not a closed freq. itemset

$c^4 = \{ae^3\}\ \{bc^3\}\ \{cd^2\}\ \{ce^2\}$ ✓ closed freq. itemset

min sup-count = 3

min cofidence = 80%.

Tid	Itemset
1	a, d, e
2	a, b, c, e
3	a, b, d, e
4	a, c, d, e
5	b, c, e
~~7~~	b, d, e
~~8~~	c, d
8	a, b, c
9	a, d, e
10	a, b, e

Apriori

Generate

1st iteration

Candidate set C_1 → frequent itemset L_1

Item	Sup_cnt
a	7
b	6
c	5
d	6
e	8

→

item	Sup_cnt
a	7
b	6
c	5
d	6
e	8

2nd iteration

C_2 → L_2

Itemset	Sup_cnt
ab	4
ac	3
ad	4
ae	6
bc	3
bd	2 X
be	5
cd	2 X
ce	3
de	5

itemset	Sup_cnt
ab	4
ac	3
ad	4
ae	6
bc	3
be	5
ce	3
de	5

3rd iterat

C_3 → L_2

Itemset	Sup_cnt
abc	2 X
~~abd~~	1
abe	3
~~acd~~	1
ace	2 X
ade	4
bce	2 X

Itemset	Sup_cnt
~~a~~ -	
abe	3
ade	4
. .	

abc = {ab} {bc} {ac} ✓

$C_4 \Rightarrow \phi \quad L_4 = \phi$

Frequent itemsets = $\{a,b,e\}$ $\{a,d,e\}$

2] Generate Association Rules.

$\{a,b,e\}$

$\{a\}$ $\{b\}$ $\{e\}$ $\{a,b\}$ $\{a,e\}$ $\{b,e\}$

- $\{a\} \Rightarrow \{b,e\}$

$$\text{Confidence} = \frac{\text{sup\_cnt}(\{a\} \cup \{b,e\})}{\text{sup\_cnt}\{a\}} = \frac{\text{sup\_cnt}(a,b,e)}{\text{sup\_cnt}(a)} = \frac{3}{7} = 0.42 = 42\% < \text{min confidence}$$

- $\{b\} \Rightarrow \{a,e\}$

$$\text{Confidence} = \frac{\text{sup\_cnt}(a,b,e)}{\text{sup\_cnt}(b)} = \frac{3}{6} = 50\%$$

$<$ min confidence

- $\{e\} \Rightarrow \{a,b\}$

$$\text{Confidence} = \frac{\text{sup\_cnt}(a,b,e)}{\text{sup\_cnt}(e)} = \frac{3}{8} = 37.5\%$$

$<$ min confidence

- $\{a,b\} \Rightarrow \{e\}$

$$\text{Confidence} = \frac{\text{sup\_cnt}(a,b,e)}{\text{sup\_cnt}(a,b)} = \frac{3}{4} = 75\%$$

$<$ min confidence

- $\{a,e\} \Rightarrow \{b\}$

$$\text{Confidence} = \frac{\text{sup\_cnt}(abe)}{\text{sup\_cnt}(ae)} = \frac{3}{6} = 50\% < \text{min confidence}$$

✓ ad ⇒ e 100 %.
✓ de ⇒ a 80%

{be} ⇒ {a}

Confidence = $\frac{3}{5}$ = 60% < min confidence.

l = {a, d, e}

{a} {d} {e} {ad} {ae} {de}

- {ad} ⇒ e

Confidence = 100%

- {de} ⇒ a

Confidence = 80%

9/08/19

FP Tree

Tid	Items
1	$I_1 I_2 I_5$
2	$I_2 I_4$
3	$I_2 I_3$
4	$I_1 I_2 I_4$
5	$I_1 I_3$
6	$I_2 I_3$
7	$I_1 I_3$
8	$I_1 I_2 I_3 I_5$
9	$I_1 I_2 I_3$

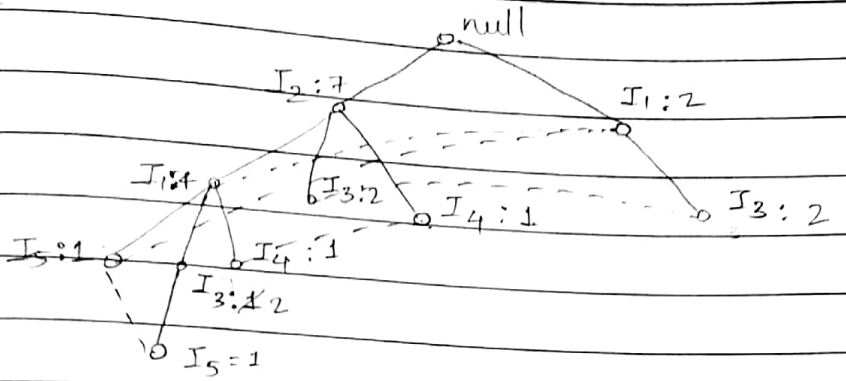
Item	sup-cnt
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

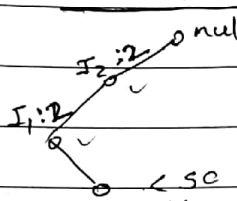
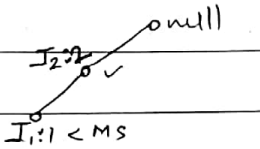
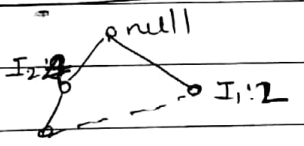
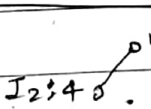
→

Item	sup-cnt
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

min sup = 2

Tid	Items
1	$I_2 I_1 I_5$
2	$I_2 I_4$
3	$I_2 I_3$
4	$I_2 I_1 I_4$
5	$I_1 I_3$
6	$I_2 I_3$
7	$I_1 I_3$
8	$I_2 I_1 I_3 I_5$
9	$I_2 I_1 I_3$



Item	Conditional pattern base	Conditional FP-tree	frequent pattern
I_5	$I_2 I_1 : 1$ $I_2 I_1 I_3 : 1$		$I_2 I_1 I_5 : 2$ $I_2 I_5 : 2$ $I_1 I_5 : 2$
I_4	$I_2 : 1$ $I_2 I_1 : 1$		$I_2 I_4 : 2$
I_3	$I_2 : 2$ $I_2 I_1 : 2$ $I_1 : 2$		$I_2 I_1 I_3 : 2$ $I_1 I_3 : 4$ $I_2 I_3 : 4$
I_1	$I_2 : 4$		$I_2 I_1 : 4$

$\therefore I_2 = 4$

Q.

Tid	Items
t_1	$I_1 I_3 I_4$
t_2	$I_2 I_3 I_5$
t_3	$I_1 I_2 I_3 I_5$
t_4	$I_2 I_5$
t_5	$I_1 I_2 I_3 I_5$

①

Item	sup-cnt
I_1	3
I_2	4
I_3	4
I_4	1
I_5	4

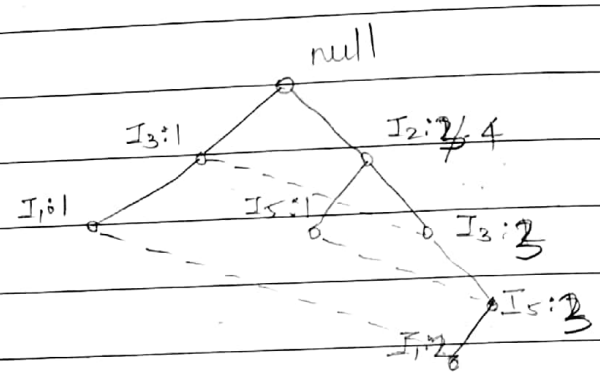
②

Item	sup-cnt
I_2	4
I_3	4
I_5	4
I_1	3
I_4	1 ✗

$\therefore$ Exclude I_4 < sc

min. sup-cnt = 2

Tid	Items
t_1	$I_3 I_1$
t_2	$I_2 I_3 I_5$
t_3	$I_2 I_3 I_5 I_1$
t_4	$I_2 I_5$
t_5	$I_2 I_3 I_5 I_1$



- association → find frequent sets

3

Classification
Hunts
types of splits (Binary multiway ...
Properties of decision tree
impurity measures
gain & gain ration

Mod 4

12/09/19

Association Mining

Using vertical data format

Tid	Itemset
1	Strawberry, litchi, orange
2	Strawberry, butter_fruit
3	butter_fruit, vanilla
4	strawberry, litchi, oranges
5	Banana, Orange
6	Banana
7	banana, butter_fruit
8	Strawberry, litchi, orange, apple
9	apple, vanilla
10	Strawberry, litchi

l, o, s	
b-f, s	
b-f, v	
l, o, s	
b, o	
b	
b, b-f	
a, l, o, s	
a, v	
l, s	min sup-cnt=3

(1)

Item	Transaction
a	8, 9 ✗
b	5, 6, 7
b-f	2, 3, 7
l	1, 4, 8, 10
o	1, 4, 5, 8
s	1, 2, 4, 8, 10
v	3, 9 ✗

(2)

Item	Transaction	
b, b-f	7	✗
b, l	–	✗
b, o	5	✗
b, s	–	✗
b-f, l	–	✗
b-f, o	–	✗
b-f, s	2	✗
l, o	1, 4, 8	
l, s	1, 4, 8, 10	
o, s	1, 4, 8	

⇒ no. of transaction, ok less than 3. ∴ discard

(3)

Item	Transaction
l, o, s	1, 4, 8

frequent item set

l, o, s

Q.

Transaction	Item	
1	Bread, Cheese, Eggs, Juice	B, C, E, J
2	Bread, Cheese, Juice	B, C, J
3	Bread, Milk, Yogurt	B, M, Y
4	Bread, Juice, Milk	B, J, M
5	Cheese, Juice, Milk	C, J, M

The min_support is 50%

min_sup_cnt = 50% of total transaction

= 50% of 5

$= \frac{50}{100} \times 5$

$= 2.5 \simeq 3$

(1)

Item	Transaction
B	1, 2, 3, 4
C	1, 2, 5
E	1, ✗
J	1, 2, 4, 5
M	3, 4, 5
Y	3, ✗

(2)

Item	Transaction
B, C	1, 2, ✗
B, J	1, 2, 4
B, M	3, 4 ✗
~~B, Y~~	
C, J	1, 2, 5
C, M	5 ✗
~~C, Y~~	
JM	4, 5 ✗

(3)

Item	Transaction
B, C, J	1, 2 ✗

Frequent item set : B, J, CJ

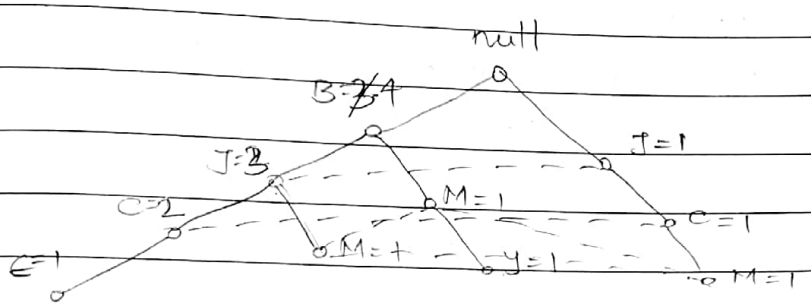
Q. FP Tree (same question 1)

Item	Sup-cnt
B	4
C	3
E	1
J	4
M	3
Y	1

Item	Sup-cnt
B	4
J	4
C	3
M	3
E	1
Y	1

↑

Tid	Item
1	B, J, C, E
2	B, J, C
3	B, M, Y
4	B, J, M
5	J, C, M



Item Conditional pattern base Conditional FP-tree Frequent pattern

FP Tree

19/09/19

Tid	Items
1	M, O, N, K, E, Y
2	D, O, N, K, E, Y
3	M, A, K, E
4	M, U, C, K, Y
5	C, O, O, K, I, E

min sup = 60%

min sup_cnt = 60% of transaction

= 3

①

Item	Sup-cnt	
M	3	✓
O	3	✓
N	2	✗
K	5	✓
E	4	✓
Y	3	✓
D	1	✗
A	1	✗
U	1	✗
C	2	✗
I	1	✗

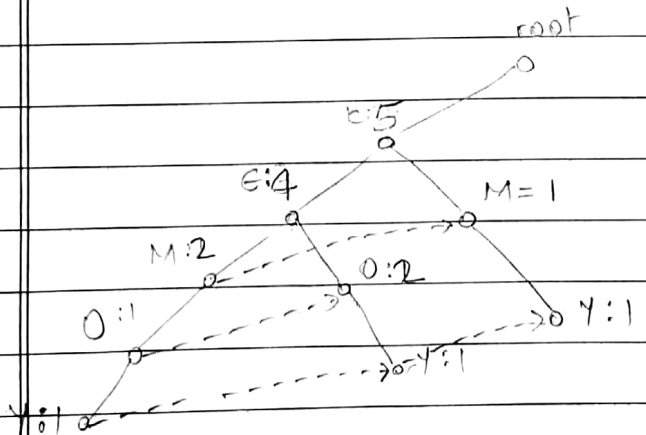
→

②

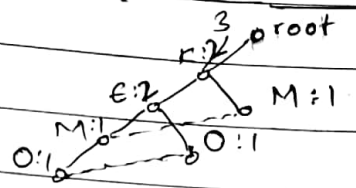
Item	Sup-cnt
K	5
E	4
M	3
O	3
Y	3

③

Tid	Items
1	K, E, M, O, Y
2	K, E, O, Y
3	K, E, M
4	K, M, Y
5	K, E, O



Note: Y should not be the first Item

Item	Conditional pattern base	Conditional FP Tree	Frequent Pattern
Y	K, E, M, O : 1		KY : 3
	→ K, E, O : 1		
	K, M : 1		

~~K:3~~ ~~E:2~~ ~~M:1~~ ~~O:1~~

~~K:3, M:1~~

~~K:3, E:2, O:1~~

~~E:2~~

~~M:2~~

~~O:2~~