Data Mining

11/01/2024
(1 hrs) Morning

How to process the huge amount of data?
Beyond 15-20 attribute or 1000 records.

Although, electronics is developed to process huge data but its meaningless if we don't have efficient algorithms that answer our query on such huge data. ADAM

Advanced Data Analysis Methods (ADAM)

It is the Data Mining, advanced course of DBMS.

Preprocessing - making data suitable ~~for~~ to be used in algorithms.

DM Techniques / Algorithms
- Classification
- Regression Analysis
- Clustering Algorithms
- Association Analysis

Data Presentation Tools - so Result more understandable to end-user.

Evaluation Parameters of Result
Either its appropriate or inappropriate.

Book -

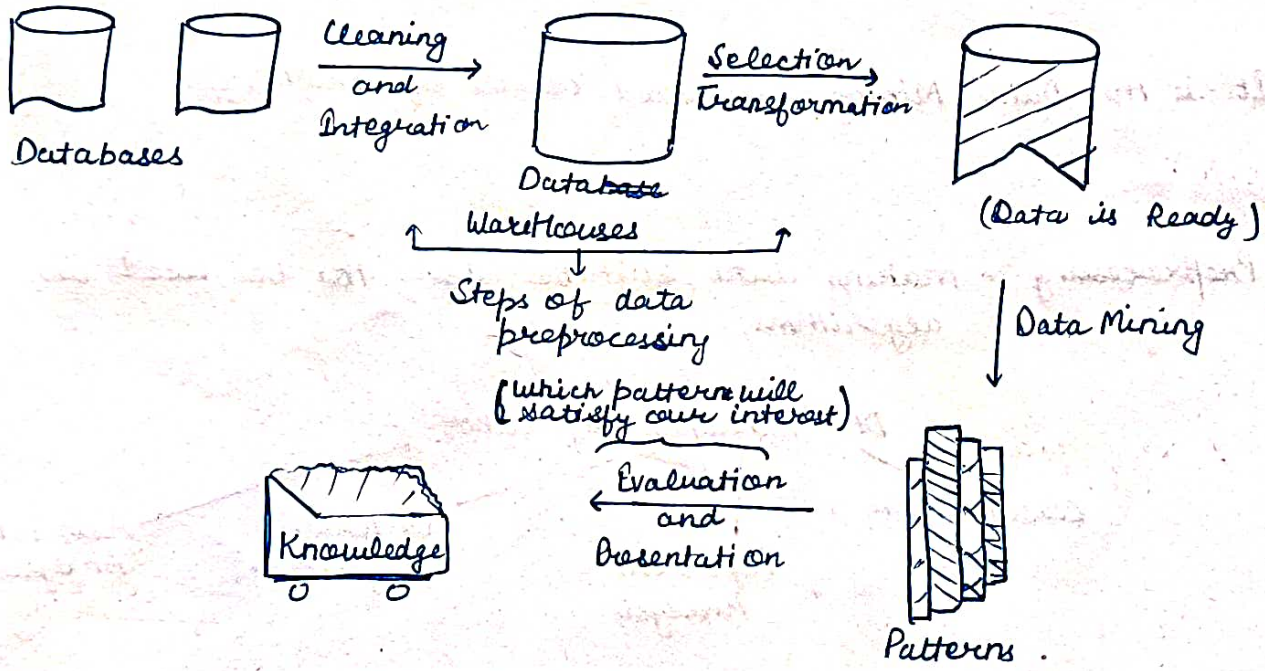
- Han & Camber — Data Mining Concepts and Techniques → Good concept & Technical language
- Vipin Kumar — Introduction to Data Mining → Easy language

DBMS	Korth ↓ Easy , Navathe ↓ Technical

Data Mining is
Process of finding interesting patterns from large amount of data.

We use data sets for storing huge data.

Representing Data Mining as a process, then it have states-



Data Mining / KDD
Process — Knowledge Discovery from Databases

Data	Information	Knowledge
Very Initial Stage	At the intermediate stage	At final stage, interesting patterns.

Types of Data

1. Relational Databases - Data stored in form of table (relational)
2. Data Ware House Data - " " " multi-dimensional format. (OLCP technol ↳querying)
   Operation ⇒ Roll up & Drill Down
3. Transactional Data - Record of what customers purchases in every visit.
4. Temporal Data / Sequence Data - Time is important constraint
5. Spatial Data - Position are important
   Temperature at different locations of world
6. Data Streams - Continuous data in various formats
   Temp. sensors, Video surveillance
7. Graph / Network Data - If feasible to draw graph that's graph data we can get useful information.
8. Text Data
9. Multimedia Data
10. WWW

Relational

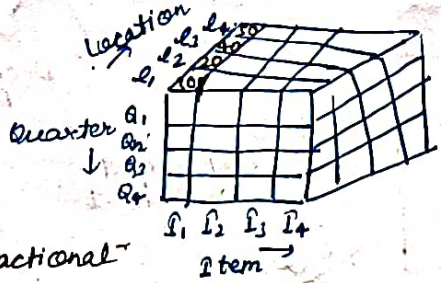
Customers (Cname, cid, Age, DOB, iatempince, income, itemp)
↳purchase

Item (itemid, itemname, itempric, itemq)

Sale (quarter, itemid, itemname, amountofsale)

Data WareHouse

Data after Data Mining, we get Knowledge.



Data in compact form

Transactional

T_1	Milk, Bread, Coke
T_2	Garments, Milk, Bread
T_3	Coke, Bread
T_4	Coke

↳Pattern its same

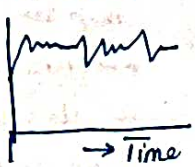
Req. Milk with Bread is high

correlation or association among data

Market-Basket Analysis → In a visit a customer purchase what things?

Temporal

Stock Market



Broad Categorization of Data

Structured Data → That have a proper format
eg. Relational data, Excel file

Unstructured Data → Not feasible to find the structure
eg Text data, PDF

Semi-structured Data → Contain both st. and unst. data.
eg. Webpage with Text & Image
↓ Unstructured ↓ Structured

Data Mining Tasks

Objective
① Predictive Tasks is to predict value of class attribute for the given inputs value
② Descriptive Tasks Identify pattern in data then on basis of it we summarize results

Represented by Random Variable $X=\{A_1, A_2, A_3\}$

Independent Variable or Normal Attribute

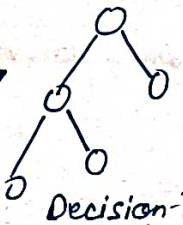
Data Set

A_1	A_2	A_3	C

C → Class Attribute or Dependent Attribute

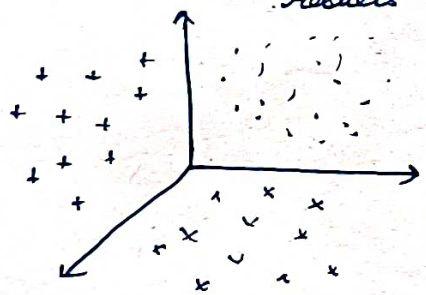
They create model → Insert Value → Get Result
How do they do it?

Productive Analysis →



Decision-Tree Model

Cluster Analysis ←



Data with similar structure/feature put in single cluster.

Association Analysis ←

Milk → Bread

Helpful for inventory preparation,

Outlier Analysis →

help to detect malicious activity as its different from regular data

Data Mining Frequent Patterns Association and correlations

- Frequent Item sets
- Frequent Subsequences
- Frequent Substructures

set of Items
↑
itemset = {Milk, Bread}

freq = 2

Frequent Itemset → Itemset has frequency greater than a threshold limit.

In Association Mining,
we identify F.I.S. then apply algo. on them and get results.

eg- {Laptop, Dig. Camera}
high probability that he will also purchase 'memory card'.

If some substructure is frequent in graph, then it must be noted as its helpful.

eg- Diff. st. of Benzene

Classification, Regression Analysis for Predictive Analysis

↓
only applicable with discrete attribute value

Prediction

Once you prepare data set then algorithm predicts the class of new instance.

Prediction with continuous attribute value.

Data Set

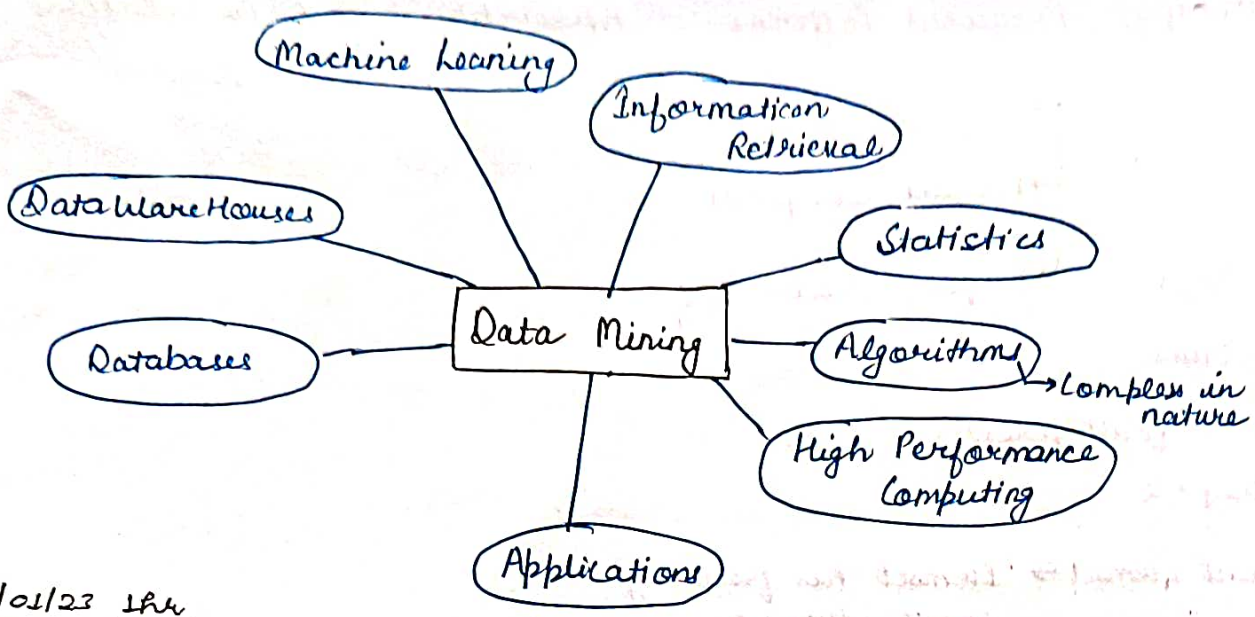
	A_1	A_2	A_3	Result	Class Attribute
→	–	–	–	Pass	→ 0.7
→	–	–	–	Pass	→ 0.3
→	–	–	–	Fail	→ 0.15
→	–	–	–	Pass	→ 0.42
→	–	–	–	Pass	→ 0.87

Classified in two groups

0-1 continuous range of values

If new student comes. on basis of grades algo predicts which class it belong.

So, Regression analysis algorithm needed.



18/01/23 1hr
(Morning)

Attributes / Features / Variables / Dimensions

Data objects have some characteristics which are known as attributes.

				Operations
somehow representing some category/quality. Categorical Attributes (qualitative)	Nominal (=, ≠)	Employee id		mode
	Ordinal (<, >)	T-Shirt Size (Small, Medium, large) Grades (B+, A, A+)		median, mode
Numeric (quantitative)	Interval (+, −)	app. diff. is defined	Calender date, Marks ~~Grade~~ ~30 & 45	mean, ~~median~~ median, mode
	Ratio (*, /)	weight, length	30kg, 60kg	geometric mean, harmonic mean

If a data can be represented in matrix (2-d form) or multidimensional form so that some mathematical operation can be defined on those.

45 is 15 more than 30
30 is 15 less than 45

Depend on the nature of problem in which category does the attribute.

Binary Attribute

- Symmetric Binary
- Asymmetric Binary

Gender → Male / Female.

Covid-19
+
−
−
−
−
−
−

Discrete Attribute & Continuous Attribute

{C, C+, B, B+, H, A+} Grades

Height (0-200 cm)
151.26
140.023

Binary Discrete Continuous.
qualitative (nominal or ordinal)
quantitative (~~ordinal~~ ratio or Interval.)

① Time in Terms of AM or PM (Binary, qualitative, Ordinal)
② Brightness measured by a light meter. (Continuous, quantitative,
③ Brightness measured by People's Judgement (Discrete, qualitative, ordinal)
④ Bronze, Silver & Gold Medals.
⑤ No. of patients in a Hospital
⑥ Number of Books
⑦ Military rank
⑧ Temp in °C or °F → Interval
⑨ Temp in Kelvin → Ratio → Ref. Point well-defined

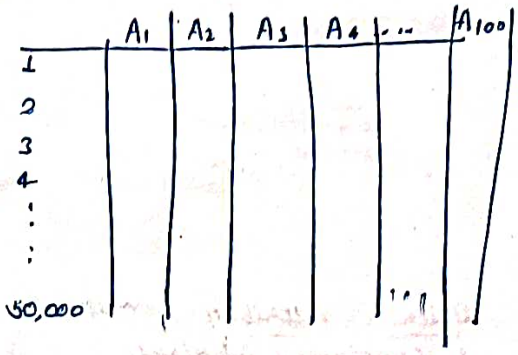
Data Similarity & Dissimilarity Measures

19/01/24
Morning (2 hrs)

Data Matrix

Dissimilarity Matrix — No. of Attr. × No. of Data Item types

After Putting/Representing data in matrix we can perform no. of operation and derive the result.



No Similarity
Similarity (0 . 1) → Both are Similar

Dissimilarity = 1 - Similarity

Id	(Nominal) Test 1	(ordinal) Test 2	(numeric) Test 3
1	Code A	Excellent	45
2	Code B	Fair	22
3	Code C	Good	64
4	Code A	Excellent	24

Dissimilarity Measurement for nominal attributes

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

p → no. of att. of nominal types
m → no. of matches in their values

	A_1	A_2	A_3
i	Code A	Code B	Code C
j	Code A	Code B	Code D

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

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

$$d(i,j) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \end{array}$$

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

Dissimilarity measurement for Binary attribute

		1	0
i	1	q	r
	0	s	t

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

$p = q+r+s+t$
↳ total no. of attribute

$\boxed{sim(i,j) = \frac{q+t}{q+r+s+t}} = 1 - d(i,j)$

↳ Simple Matching Coefficient (SMC)

Symmetric Binary will of better use

$i = (1, 1, 0, 1)$
$j = (0, 1, 1, 0)$
$r=2 \quad q=1 \quad s=1 \quad t=0$

For Asymmetric Binary Attribute

$sim(i,j) = \frac{q}{q+r+s}$ ← Both can be used depend on your req.

↳ Jaccard Coefficient

$\frac{q}{q+r+s+t}$

Covid 19 test

i	j
P	N
P	N
N	N
N	N
N	P
N	N
N	N
N	P

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

$SMC = \frac{7}{10}$

$J = \frac{0}{10}$

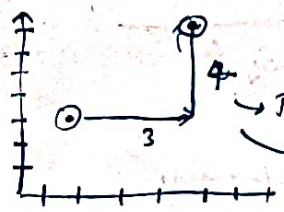
} Numerical Similarity

Dissimilarity Measurement for Numeric Attribute

- Euclidean Distance → Shortest dist. b/w two points ⇒ Shortest Dissimilarity
- Manhattan Distance → Block dist./City distance
- Supremum Distance → Max. of both axis

$i(2,3) \quad j(5,7)$

$Eucl(i,j) = \sqrt{(5-2)^2 + (7-3)^2}$



→ Total is Manhattan distance

→ Max. of both is 4 which is Supremum distance.

Normalisation is performed to reduce the huge range & differences.

From Table data,
Max - Min = 42

$$d(i,j) = \begin{bmatrix} 0 & & & \\ 23/42 & 0 & & \\ 19/42 & 1 & 0 & \\ 21/42 & 2/42 & 40/42 & 0 \end{bmatrix}$$

Dissimilarity measure for ordinal attribute

Convert to numeric attribute

{fair, good, excellent}
1 2 3

$$d(i,j) = \begin{bmatrix} 0 & & & \\ 2/2=1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

More standard normalisation

Formula

$$\frac{r_{i,j} - 1}{m_f - 1}$$

Fair = $\frac{1-1}{3-1} = \frac{0}{2} = 0$

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

Excellent = $\frac{3-1}{3-1} = 1$

How to integrate result -

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

ε_{ij} = Weights corresponding to dissimilarity

$= \frac{1 \cdot d(2,1) + 1 \cdot d(2,1)^2 + d(0,1)}{3}$

- 0 → missing or zero value
- 1 → otherwise

Cosine Similarity

Finds the similarity between two documents.

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

→ Very popularly used in text mining.

Document Matrix – $\theta = 0° \Rightarrow 0°$; $\theta = 90°$

or

Term Frequency Matrix

↓

(Basis of Text Mining)

	(rice) Term 1	(tea) Term 2	(bread) Term 3	...	(gas) Term 1000
doc 1	5	3	0	...	2
doc 2	20	6	0	...	12
doc 3	3	3	5	...	6
⋮	⋮	⋮	⋮		⋮
doc 100	2	6	2		1

Correlation Coefficient

$$\text{Corr}(x,y) = \frac{\text{covariance}(x,y)}{\text{std}(x) \cdot \text{std}(y)}$$

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

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

$-1 \leq \text{Corr.} \leq 1$

-ve correlation ↓ 0 ↓ +ve correlation

Neutral correlation

24th Jan '24 Morning (2 hrs)

Data Preprocessing Techniques

① Data Cleaning → inconsistency, missing values, incomplete data. After cleaning, we get req. data.

② Data Integration → Diff. sources Relational format, Non-Rel. format ⇒ Integrated in standard rel. format.

③ Data Reduction → (Below)

④ Data Transformation → Normalise our values to get them in standard format.

③

Rel. Format

A_1	A_2	..	-	A_{100}

Reduce to no. of attribute or and Reduce the no. of entries

⇒ No loss in data

Numeric			Binary
A_1	A_2	...	A_{100}
1			0
2			1
⋮			⋮
100			0

① Data cleaning

① Missing Values

A_1	A_2	A_3	A_4	A_5	A_6
10	0	25	-	-	-
20	1	26	-	-	-
30	0	-	-	-	-
40	1	25	-	-	-
50	1	27	-	-	-

- a. Ignorance Approach-

  ex 2-3 missing value out of 5,000 records so ignore those tuples (Sufficiently large)

- b. Putting Random Value -
  - May create Noisy, Outlier data

- c. Put the mean value
  - Either global mean
  - Either mean of adjacent neighbours

- d. Either selecting global constants like 0 etc.

- e. Prediction of value. {Much better, but also have some complexity}
  - Decision tree method.

    Using available values we create decision tree then the output gives us the value (unknown).

  - Linear Regression Method.

② Handling Noisy Data

Random errors in data is noise.

A_3
25
26
100
25
27

Beyond our expectation

a. Data Smoothing Method.

Smooth out the data

(i) Binning method / Basket method (Data in sorted order)

25, 26, 29, 35, 40, 42, 45, 50, 100

Baskets		Mean	Replace all value with mean value
1:	25, 26, 29	x_1	x_1 x_1 x_1
2:	35, 40, 42	x_2	x_2 x_2 x_2
3:	45, 50, 100	x_3	x_3 x_3 x_3

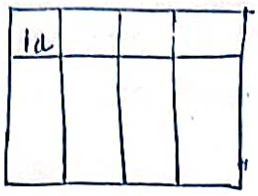
Noise entry are very less.
Mean value we get for every groups is near to filter value.

(ii) Methood of Regression.

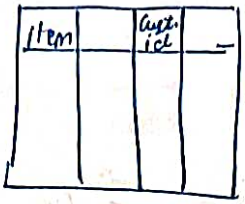
② Data Integration

① Entity Identification Problem-

Customer



Sales



Rename those attl. in common names.

② Redundancy and Correlation-

↓

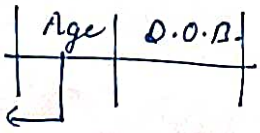
is not duplicacy

↳ Repition

Same tupes many time

Some attribute may be redundent.

ex



is redundant ←

as using current date & d.o.b. we can calc. age.

During integration, no redundancy in data.

(b) Using the correlation of attribute

If corr. coeff. very high then one attl. can be substituted by other.

Chi-square test- (χ^2)

For nominal attl.

A → c distinct values $(a_1, a_2, \dots a_c)$

B → r distinct values $(b_1, b_2, \dots b_r)$

(A_i, B_j) - Joint event that attribute A take on value a_i and attribute B take on value b_j i.e. $(A=a_i, B=b_j)$

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

$$e_{ij} = \frac{\text{count}(A=A_i) \times \text{count}(B=B_j)}{n}$$

o_{ij} = Observed freq. of the joint event (A_i, B_j)

e_{ij} = expected freq. of (A_i, B_j)

n = No. of data tuples

Reading Habits \ Gender	Male	Female	
Fiction	300 (250)	200 (240)	500
Non-fiction	100 (90)	400 (380)	500
	400	600	1000

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

$$\text{cov}(A,B) = E(A-\bar{A})\,E(B-\bar{B})$$

$$\text{cov}(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

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

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}$$

$$\text{Corr. coff. } (A,B) = \frac{\text{covar}(A,B)}{\text{std.}(A)\ \text{std}(B)}$$

If cov. is +ve, if value of A changes above mean then its more likely that B will also change. In case of -ve cov. If A is above mean, B is below mean -ve covariance.

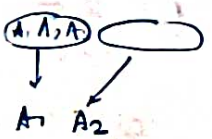
③ Data Reduction

① Feature Selection or Feature Subset Method

② Feature Reduction

↓

Merging — combine similar att. & merge them



↳ Selecting some important att. we prefer the Ranking Method.

← Principal Component Analysis (PCA).

$$f(A') = \frac{w_1 A_1 + w_2 A_2 + w_3 A_3}{w_1 + w_2 + w_3}$$

31st Jan '24 (2 hr) Morning

Data Reduction Methods

Principal Component Analysis

$f_1, f_2, f_3, \ldots, f_{100}$ → Attribute / Feature / Value

Sorted order → $PC_1, PC_2, PC_3, \ldots PC_{10}$

PC_1 → Higher / Max. Variance → Principal Component

Every feature mapped to some PC.

Arrang PC acc. to their Data Variance value / level.

Data Matrix = $\begin{matrix} \\ r_1 \\ r_2 \\ \vdots \\ r_{100} \end{matrix} \begin{bmatrix} f_1 & f_2 & \ldots & f_{100} \\ & & & \\ & & & \\ & & & \end{bmatrix}$

Variance is less ⇒ Not contributing in the classification

eg.

Data Matrix.

	f_1	f_2	f_3	f_4	f_5
r_1	5	6	7	9	2
r_2	2	7	2	7	9
r_3	7	7	1	7	3
r_4	4	7	2	7	1
r_5	1	7	5	7	6
		x		x	

Less Variance.

* How to map feature to P.C. ?

Eigen vectors and Eigen values

$A \cdot v = \lambda v$

Symm. Mat.

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

$$\text{Covariance}(f_1, f_2) = \frac{1}{n}\sum_{\substack{i=1 \\ j=1}}^{n}(f_i - \bar{f_1})(f_j - \bar{f_2})$$

Prepare Covariance Matrix from Data Matrix

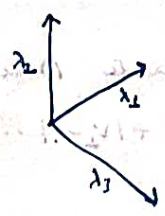
Inc. one, other Inc ⇒ +ve Cov.

Inc. one, other dec ⇒ -ve

Cov. Matrix = $\begin{bmatrix} \Sigma f_{11} & \Sigma f_{12} & \Sigma f_{13} \\ \Sigma f_{21} & \Sigma f_{22} & \Sigma f_{23} \\ \Sigma f_{31} & \Sigma f_{32} & \Sigma f_{33} \end{bmatrix}$

Map our feature ⇓ Eigenvector.

Highest eigen value ⇓ Highest PC. Max. Variance.



λ → represent amount of variance in that axis

$$D = \begin{array}{c} \begin{matrix} f_1 & f_2 \end{matrix} \\ \begin{bmatrix} 2 & 3 \\ 4 & 4 \\ 3 & 6 \\ 5 & 4 \\ 6 & 2 \\ 6 & 3 \\ 8 & 4 \\ 6 & 6 \end{bmatrix} \end{array}_{8 \times 2}$$

Cov. Matrix =

In Assignment copy

Normalization -

- Min-Max Normalization - Preserve relationship in your prev. data.

1 → 25
100 → 40

$$v_i' = \frac{v_i - min_a}{max_a - min_a}(new\text{-}max_a - new\text{-}min_A) + new\text{-}min_A$$

eg. 10, 28, 42, 76, 91

10 → 25
91 → 40

- Z-score Normalization

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

$\sigma_A \rightarrow$ standard deviation

Is applicable when the real boundaries are not known. There may be noise & outlier data.

- Mean Absolute Deviation

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

$S_A \rightarrow$ Absolute Deviation

$$S_A = \frac{1}{n}(|v_1 - \bar{A}| + |v_2 - \bar{A}| \dots + |v_n - \bar{A}|)$$

Normalization by decimal scaling.

- Normalization by decimal scaling

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

Classification 2nd Feb '24 Monday (Live)

Identify / Classify some group of entity or the classify entity into groups is known as classification problem.

We have some data, from that we create model which is used to classify the new instance.

These models are algorithms.

Requirement in data is it should be labelled. (→ we know the class attribute.) In General class attribute tells about the class of each instance

Two phases –

- Training Phase → with given data → Prepare Model
- Testing Phase → with new inputs → we do prediction.

Model

$y = f(x)$

↓ Outcome ↓ Input

You can use different models.
But the accuracy should be very high.
for good prediction.

* Value of class attribute should be discrete.

Supervised - The class of all data inst. are known. we known either we are correct or not.

Unsupervised - Class of data instances are unknown.
Training with unlabelled data.

Semi-supervised - Mix of both type data

Decision Tree

Labelled Data
Discrete Class Attribute

Customer Name	Age	G	Income	Defaulter
Ram	56	M	2000	Yes
Shyam	53	M	3000	Yes
Sita	52	F	2500	No
Shiv	42	M	3200	No
Abhi	47	M	2700	Yes

Generate_Decision_Tree

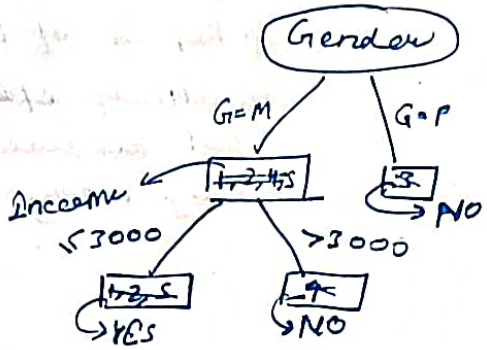
Input: Data Partition D
attribute_list, attribute_selection method

Output: Decision Tree.

Method:
1. Create a Node N
2. If tuples in N are all of the same class C then
3. return N as a leaf node labelled with the class C;
4. If attribute_list is empty then
5. return N as a leaf node labelled with the majority class in D
6. Apply attribute_selection_method (D, attribute_list) to find the best splitting criterion;
7. Label Node N with splitting criterion
8. if splitting attribute is discrete_value and multiway split allowed then
9. attribute_list ← attribute_list ← splitting_attribute
10. for each outcome j of splitting criterion // Partition the tuples and grow subtrees for each partition
11. let D_j be the set of data tuples in D satisfying outcome j
12. if D_j is empty then,
13. attach a leaf node labelled with the majority class in D to node N
14. else attach the node returned by Generation_decision_tree (D_j, attribute_list) to Node N
15. end for
16. Return N

Characteristics

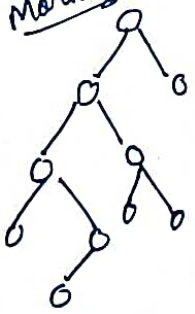
- Based on Greedy Algorithm
- Recursive in nature
- Divide and Conquer Approach



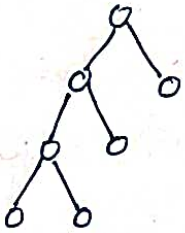
2 problems

- Overfitting
- Underfitting

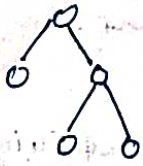
07 Feb '24 Morning (2 hrs)



Overfitted Decision Tree



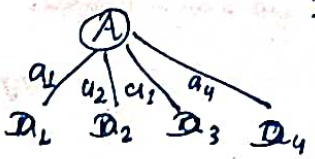
Best Suitable



Underfitted Decision Tree

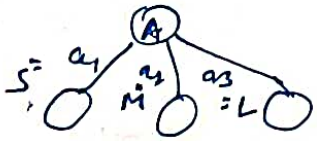
So, we have to manage the tradeoff.

Discrete Attribute



If multiway split is allowed.

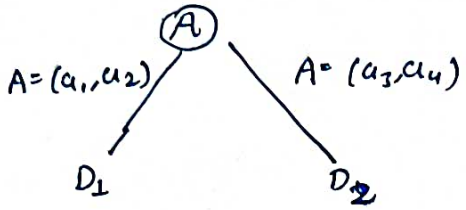
Ordinal



If multiway split is allowed.

Discrete Attribute and Binary Tree

(a_1, a_2) (a_3, a_4)



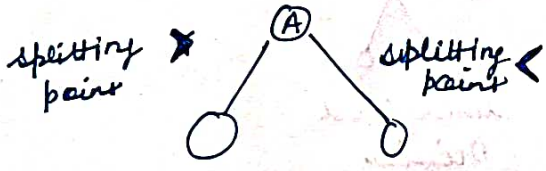
If Req. is of Binary Tree, multiway split is not allowed. Only Binary splitting

In case of ordinal attr.

{S, M}	{L}
{S, L}	{M}
{S}	{M, L}

Continuous Attribute

Splitting point.



Attribute Selection Measure

① Information Gain → Based on concept of Information Theory
② Gain Ratio
③ Gain Index

↓

Entropy

Measurement of information in some words or sentences

$$= -\sum_{i=1}^{n} P_i \log(P_i)$$

① Measure the amount of information

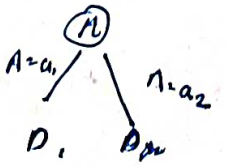
③ Min value of Gini i.e error.

Dinesh Singh

D=1 1/11
i=2 2/11
n=2 2/11
e=1 1/11
s=2 2/11
g=1 1/11
h=2 2/11
11

We put in formula & get some result.

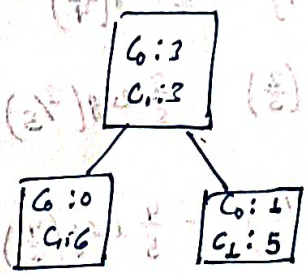
Info Gain = Info (D) - $Info_A$ (D)



$$Info_A(D_i) = \frac{|D_i|}{|D|} \; info(D_i)$$

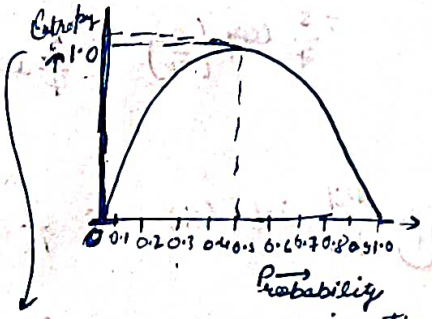
Our objective is to select attribute that gives max. info. gain..

	A_1	A_2	A_3	C
1				YES
2				NO
3				YES
4				YES
5				NO
6				NO



$$C_0 = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$$

$$= 1 \quad \leftarrow \text{Maximum Value of entropy}$$

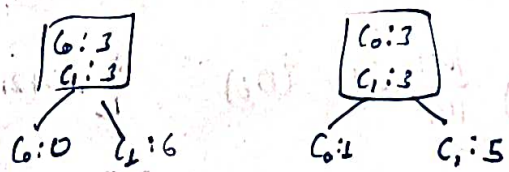


If we write in decision tree

Info = 0

Info Gain = 1 - 0

= 1 ← Max. Gain

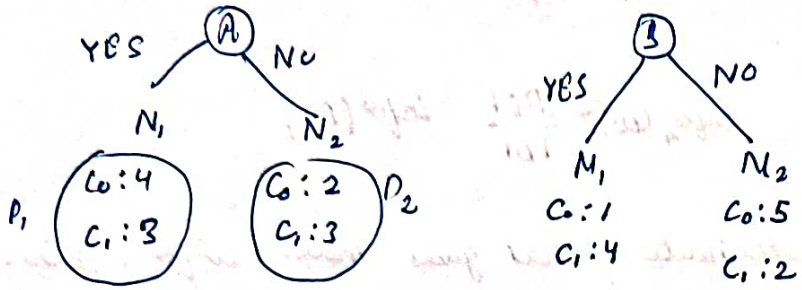


$$Info = -\frac{1}{6}\log\left(\frac{1}{6}\right) - \frac{5}{6}\log\left(\frac{5}{6}\right)$$

$$= \log 6 - \frac{5}{6}\log 5$$

$C_0 : 6$
$C_1 : 6$

info (D) = 1



P_1 $-\frac{4}{7}\log\left(\frac{4}{7}\right) - \frac{3}{7}\log\left(\frac{3}{7}\right) = 0.985$

P_2 $-\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.970$

P_3 $-\frac{1}{5}\log\left(\frac{1}{5}\right) - \frac{4}{5}\log\left(\frac{4}{5}\right) = 0.721$

P_4 $-\frac{5}{7}\log\left(\frac{5}{7}\right) - \frac{2}{7}\log\left(\frac{2}{7}\right) = 0.863$

(A) $\frac{|D_i|}{|D|}$ info (D_i) $= \frac{7}{12}$ info$(D_1) + \frac{5}{12}$ info (D_2)

$= 0.97875$

↳ Infor Gain $= 1 - 0.97875 = 0.02128$

(B) $\frac{|D_i|}{|D|}$ infer (D_i) $= \frac{5}{12} \times 0.721 + \frac{7}{12} \times 0.863$

$= 0.8038$

↳ Infor Gain $= 1 - 0.8038 = 0.1962$

Greater.

So better attribute to be chosen is B.

Income

75, 70, 85, 60, 120, 100, 95, 90, 125, 220

Sort

NO NO NO Yes's YES YES NO NO NO NO

60, 70, 75, 85, 90, 95, 100, 120, 125, 220

55 65 72 80 87 92 97 110 122 172

≤55	>55
G:0	G:3
G:0	G:7

Yes

72	
0	3
2	5

87	
1	2
3	4

≤65	>65
0	3
1	6

80	
0	3
3	4

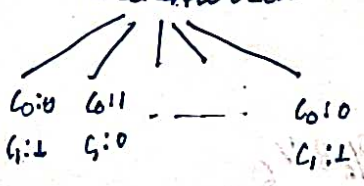
9th Feb '24

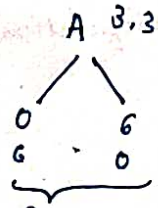
Morning (1 hrs)

Gain Ratio -

$$\text{Information Gain} = \text{info}_A(D) - \sum_{i=1}^{n} \text{info}_A(D_i)$$

If Multiway split allowed.





Purer classification

We get the max. information gain, but least/worst type partitioning. Complexity cost is very high So, no to choose such att..

We can divide on basis of distinct values the attribute have. instead of classifying in terms of att..

If purer class., then impurity is least. least Entropy least Randomness.

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

sum of weights of each Partition

$$= -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2\left(\frac{|D_j|}{|D|}\right)$$

↳ computed for all attributes. Whose is more, the more appropriate it is.

Gain Index –

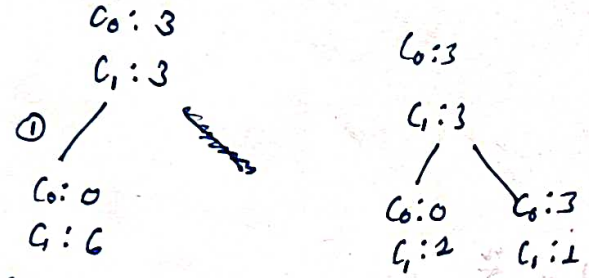
(Misclassification Error Index)

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2$$

$$p_i = \frac{|C_i, D|}{|D|}$$

↳ No. of tuples in your dataset

Q.



↳ Purer classification

Compute Misclassification error.

$p_1 = 1$

Gini(D) = 0

$p_1 = 2/6$ $p_2 = 4/6$, $p_1 = 1/6$

$$\frac{2}{6_3}\left[\frac{\frac{1}{2}}{\frac{36}{9}}\right] + \frac{4}{6_3}\left[\frac{\frac{13}{36}}{\frac{36}{9}}\right]$$

Gini = $\frac{21}{27}$

Errors in Decision Tree

16th Feb Morning (1 hrs)

Training Error. → Error in design of model.

Test Error. → How much error the model gives on unseen records?

Overfitting

Underfitting

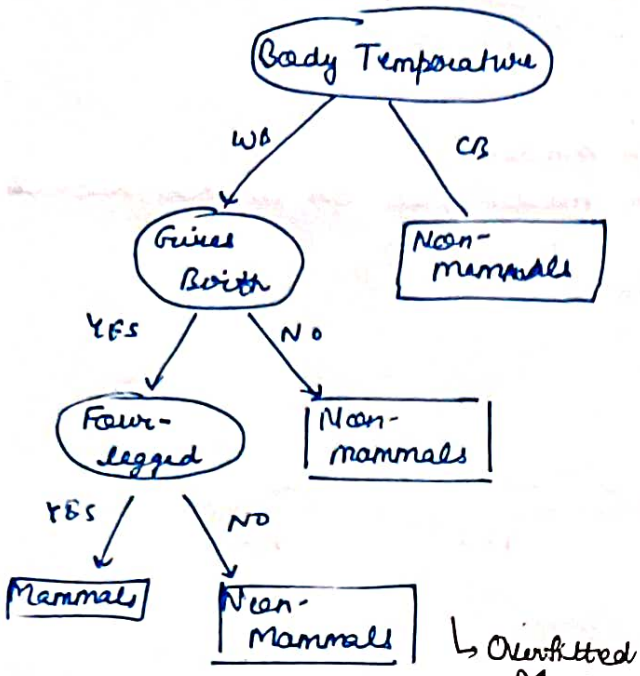
Training

	Name	Body Temp.	Gives Birth	Four legged	Hibernates	Class Label
1	Porcupine	WB	YES	YES	YES	YES
2	Cat	WB	YES	YES	NO	YES
3	Bat	WB	YES	NO	YES	NO
4	Whale	WB	YES	NO	NO	NO
5	Salamander	CB	NO	YES	YES	NO
6	Komodo Dragon	CB	NO	YES	NO	NO
7	Python	CB	NO	NO	YES	NO
8	Salmon	CB	NO	NO	NO	NO
9	Eagle	WB	NO	NO	NO	NO
10	Guppy	CB	YES	NO	NO	NO

Testing

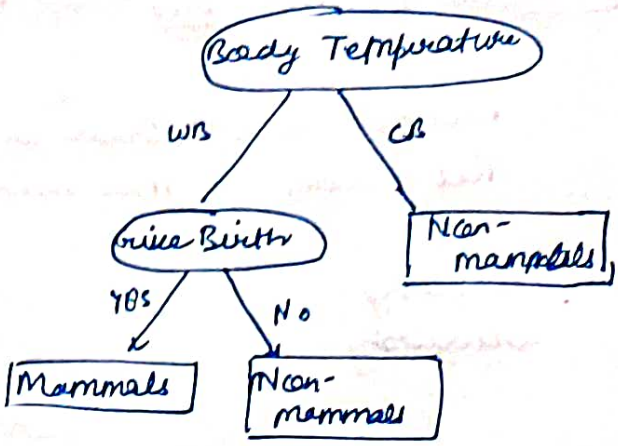
	Name	Body Tem	Gives Birth	Four legged	Hibernates	Class Label
1	Human	WB	YES	NO	NO	YES
2	Pigeon	WB	NO	NO	NO	NO
3	Elephant	WB	YES	YES	NO	YES
4	Leopard Shark	CB	YES	NO	NO	NO
5	Turtle	CB	NO	YES	NO	NO
6	Penguin	CB	NO	NO	NO	NO
7	Eel	CB	NO	NO	NO	NO
8	Dolphin	WB	YES	NO	NO	YES
9	Spiny Anteat	WB	NO	YES	YES	YES
10	Gila Monster	CB	NO	YES	YES	NO

Model-1



↳ Overfitted Model

Model-2

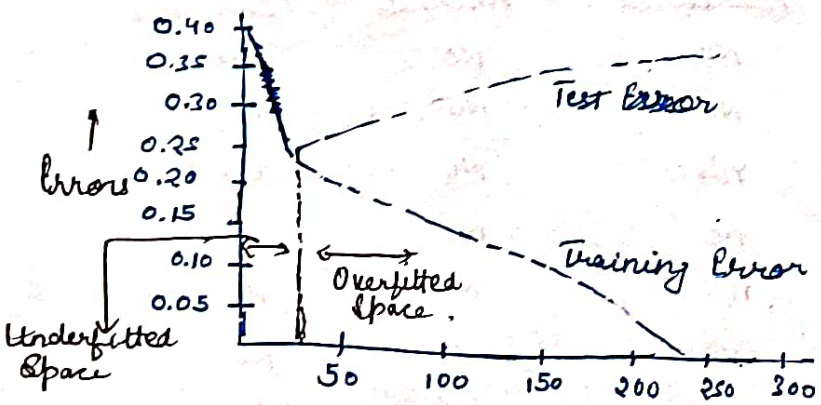


① 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Train error = 0

② 1, 2, 5, 6, 7, 8, 9, 10
Train error = 20%
3, 4

① 2, 3, 4, 5, 6, 7, 10
Test error = 30%
1, 8, 9

② 1, 2, 5, 4, 5, 6, 7, 8, 10
Test error = 10%
9



If Input size is very small, and our model is not properly trained, so high training error leads to testing error.

↳ Underfitting

No. of nodes in Decision Tree inc.

⇓

Complexity of model inc. & cost also inc.

So, balance best.