DOMS | Page No.
Date / /

23/9/21

Class I (Online)

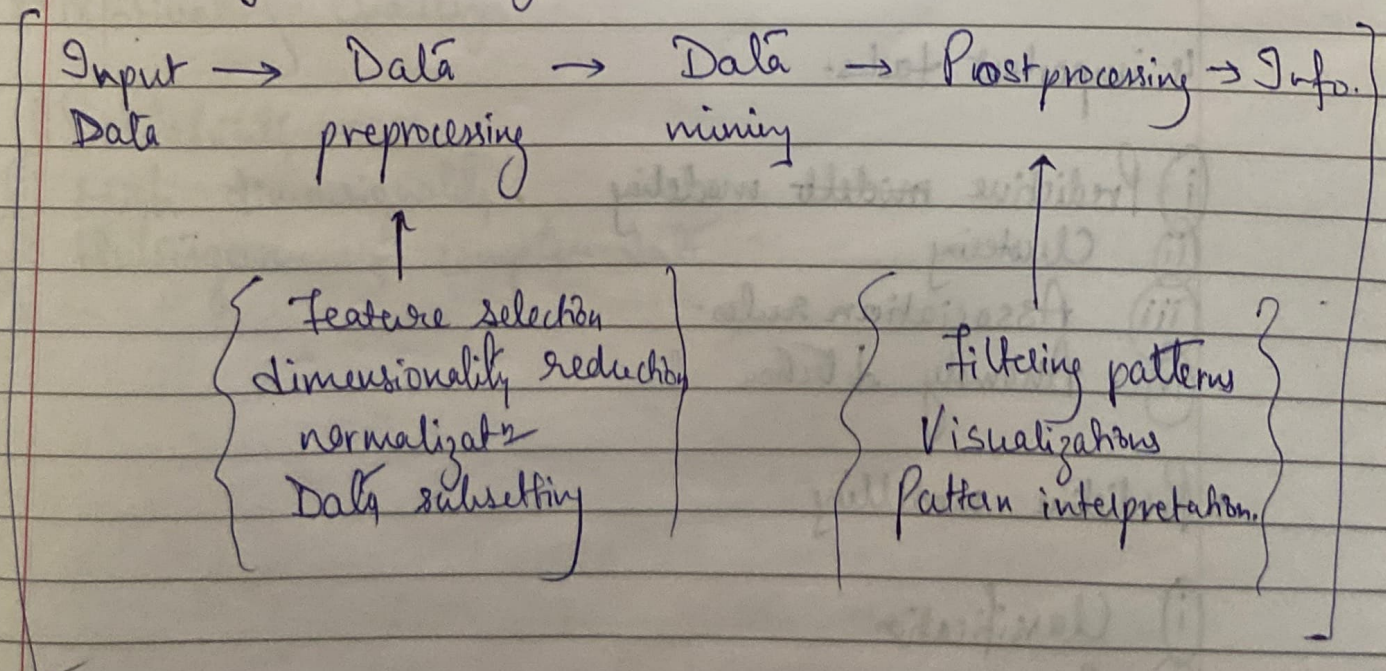
* What is data mining?

→ Technology that blends traditional methods with modern algorithms to process large amount of data.

→ Viewpoints :
(i) Commercial
(ii) Scientific
(iii)

→ Process of automatically retrieving useful information from large set of data.

→ Knowledge discovery in databases: (KDD)



* Origin of DM.

→ ML/AI
→ statistics
→ pattern recognition
→ Database s/s.
→ hypothesis
→ Sampling

* Tasks

(i) Predictive tasks → Explanatory / independent variables → target / dependent variable (value where value is predicted)
(ii) Descriptive tasks → (correlations, trends, clusters, trajectories, anomalies)
Summarize relationships in data

4 important tasks:

(i) Predictive ~~models~~ modeling
(ii) Clustering
(iii) Association rules.
(iv) Anomaly detection !!

* Predictive modelling

(i) Classification
(ii) Regression.

(i) Classification

→ Target variable should take discrete values

→ Examples in ppt.

(ii) Regression

→ Predict value of given continuous valued variable based on the value of other variables, assuming linear or non-linear model of dependency.

→ Clustering.
→ Association Rule Discovery.
→ Deviation/Anomaly/detection

* Motivating challenges

(1) Scalability
(2) High dimensionality
(3) Heterogeneous & complex data
(4)

Outline

Chpt - 2.

24/9/2021

* Outline
→ Attributes & Objs
→ Types of data
→ Data quality → Noise, outliers → Missing / inconsistent, biased data
→ Similarity & distance
→ Data preprocessing.

→ Data (What is data?)

→ Collection of data objects & their attributes
↓
Property or characteristic of an object. Collection of attributes describe an object.
→ AKA variable, field, characteristics, dimension, feature.

→ Attribute values: Numbers or symbols assigned to an attribute for a particular object.

→ Measurement of length
→ The way, we measure an attribute may not match the attribute properties

→ Types of attributes

(i) Nominal (=, ≠), any permutational value
(ii) Ordinal (<, >), an order preserving change of values
(iii) Interval (+, −), new_value = a * old_value + b
(iv) Ratio (*, /), new_value = a * old_value.

new_value = f(old_value)
↑ f = monotonic

→ Definition of attribute types is cumulative.

→ Discrete & continuous attributes

→ Asymmetric attributes

→ Types of dataset
→ Characteristics
→ Dimensionality
→ Sparsity
→ Resolution.
→ Size.

→ Record data: Data that consists of a collection of records, each of which consists of fixed set of attributes.

27/9/2002

offline lecture

Data Preprocessing

*Data Aggregation :-

→ Company XYZ

2018
2019
Year 2018 2020

Quantity	Sales
Q1	$100,000
Q2	$50,000
" 3	$100,000
" 4	$200,000

Year	Sales
2020	$560,000
2019	$500,000
2018	$510,000

→ Combining 2 or more objects/attributes into a single object/attribute.
→ Lesser storage, lesser processing time, smaller datasets.
→ Allows use of more expensive data mining algorithms
→ Allows change of scope/scale by providing high level views of data
→ Statistical observation – aggregated data is more stable than that of individual objects/attributes. (lesser variability)

→ Disadvantages :- Key details may be lost in aggregation process

Some aggregations :- Min, max, sum, avg.

* Sampling

⇒ (1) Simple random sample without replacement (SRSWOR)
(2) Simple random " with replacement (SRSWR).

Examples :-

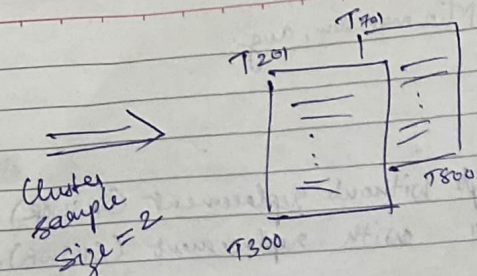
(1) T1, T2, T3, " 4, " 5, " 6, " 7, " 8

SRSWOR
S = 4
T5, T1, T4, T2

SRSWR
S = 4
T4, T1, T4, T5

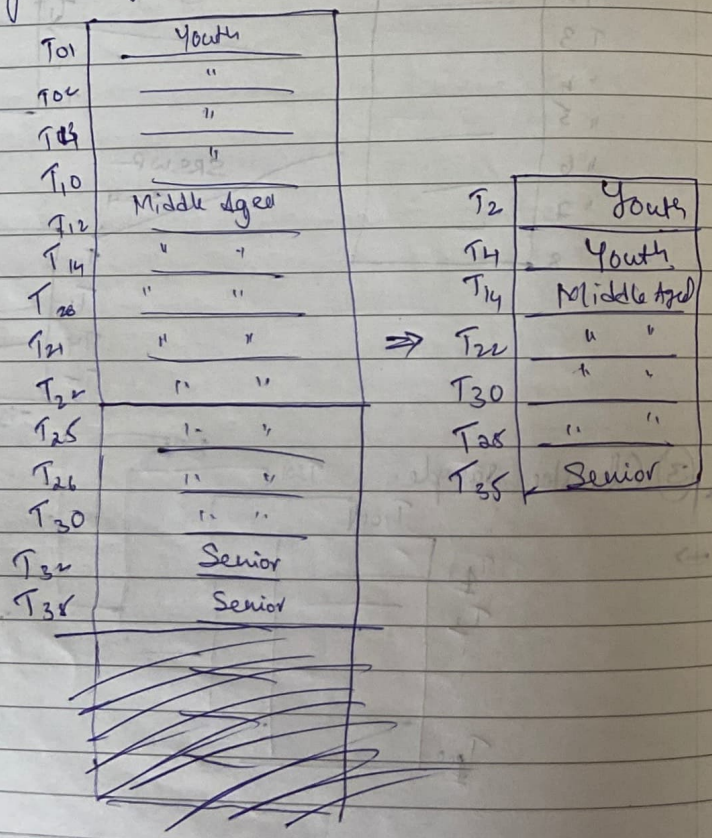
(3) Cluster sample.

⇒ T101, T200, T201, T300, T901, T1000



(4) Stratified Sample:

according to age:

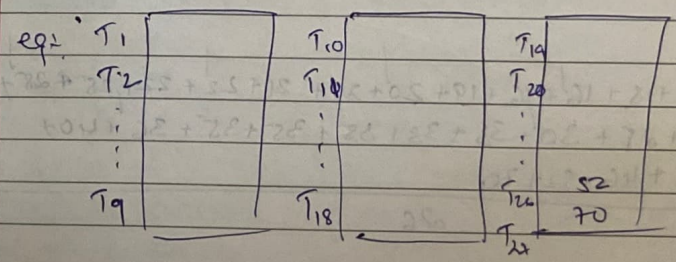


Q] For the following data, for the attribute age = 13, 15, 16, 19, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Sketch examples of each of the following sampling techniques

1. SRSWOR
2. SRSWR
3. Cluster Sampling
4. Stratified Sampling

Use Sample size = 5, Strata = "Youth" (18-25), "Middle Aged" (26-60), "Senior" (>60)



Online

30/9/21

Q] Suppose that the data for analysis includes attribute 'age'. The age values for the data tuples are
(in increasing order) : 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

1) Use Smoothing by bin means to smooth above data using bin depth of 3

Comment on effect of this technique for given data.

⇒ Mean = $\frac{13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70}{26}$

= 28.42 ⇒ 28

$W = \frac{\text{min - max}}{\text{no. of bins}}$

no. of bins = $\frac{28-0}{3}$

= 9.3

=

13, 15, 16 22, 25, 25
16, 19, 20 25, 25, 30
20, 21, 22 35

Mean.

1. 13, 15, 16 ⇒ 13+15+16/3 = 14.6 ≈ 15
2. 16, 19, 20 ⇒ = 18
3. 20, 21, 22 = 21
4. 22, 25, 25 = 24
5. 25, 25, 30 = 27
6. 33, 33, 35 = 34
7. 35, 35, 36 = 35 ≈ 35.3
8. 40, 45, 46 = 47 ≈ 43.66
9. 52, 70 = 41 ≈ 40.66

(b) Use smoothing by bin boundaries. using bin depth of 3

1. 13, 16, 16
2. 16, 20, 20
3. 20, 22, 22
4. 22, 25, 25
5. 25, 25, 30
6. 33, 33, 35
7. 35, 35, 35
8. 36, 36, 45
9. 46, 46, 70

(c) What other methods are there for data smoothing?

Online

30/9/21

Q] Suppose that the data for analysis includes attribute 'age'. The age values for the data tuples are
(in increasing order) : 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

1) Use Smoothing by bin means to smooth above data using bin depth of 3
Comment on effect of this technique for given data.

⇒ Mean = $\frac{13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70}{26}$

= 28.42 ⇒ 28

$W = \frac{\text{min-max}}{\text{no. of bins}}$

no. of bins = $\frac{28-0}{3}$

= 9.3

=

13, 15, 16 22, 25, 25
16, 19, 20 25, 25, 30
20, 21, 22 33

Mean.

1 ✓ 13, 15, 16 ⇒ 13+15+16/3 = 14.6 ≈ 15
2 ✓ 16, 19, 20 ⇒ = 18
3 ✓ 20, 21, 22 = 21
4 ✓ 22, 25, 25 = 24
5 ✓ 25, 25, 30 = 27
6 ✓ 33, 33, 35 = 34
7 ✓ 35, 35, 36 = 35 ≈ 35.3
8 ✓ 40, 45, 46 = 47 ≈ 43.66
9 52, 70 = 41 ≈ 40.66

(b) Use smoothing by bin boundaries. using bin depth of 3

1 13, 16, 16
2 16, 20, 20
3 20, 22, 22
4 22, 25, 25
5) 25, 25, 30
6) 33, 33, 35
7) 35, 35, 35
8) 36, 36, 45
9) 46, 46, 70

(c) What other methods are there for data smoothing?

DOMS | Page No.
Date / /

11/10/21

* Data transformation:

→ Data is transformed or consolidated into forms appropriate for mining.

Strategies

(1) Smoothing : works to remove noise from data
   → Binning
   → Regression
   → Clustering

(2) Attribute construction: (feature construction)

(3) Aggregation

(4) ~~Normalization~~ Normalization : attribute data are scaled so as to fall within smaller range such as -1.0 to 1.0 or 0.0 to 1.0

(5) Discretization : Raw values of numeric attribute (eg: Age) are replaced by interval labels (eg: 0-10, or conceptual labels (eg: youth, adult, etc (10-20))
: The labels can be recursively organized into higher level concept resulting in concept hierarchy for numeric attribute.

(6) Concept hierarchy generation for nominal data :
   → attributes such as ~~treet~~ street can be generalized to higher level concept like country or city

* Data transformation by normalization:

→ ~~Attribute~~ with This helps avoid dependence on choice of measurement units.
⇒ It gives all attributes equal weight.
→ Useful for classification algorithms & clustering algorithms
→ Many methods :
   3 methods :
   (i) Min-max normalization :
   This performs linear transformation on original data
   → Suppose min_A and max_A are min & max values of an attribute.
   ⇒ Min-max normalization maps a value v_i of A to v_i' in the range $[new\text{-}min_A, new\text{-}max_A]$ by computing

$$v_i' = \frac{v_i - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

Q] Income has min 12,000
has max 98,000

→ We would like to map income in the range [0.0, 1.0]
↗ v_i

⇒ What will £73,600 as income transform to? Using min-max normalization

Soln $new\_min_A = 0.0$
$new\_max_A = 1.0$

$$v_i' = \frac{73,600 - 12000}{98000 - 12000}(1.0 - 0.0) + 0$$

$$= 0.716$$

② Z-score normalization (zero-mean normalization)

⇒ Values for an attribute A are normalized based on mean (i.e average) & std. deviation of attribute A.

⇒ Value v_i of A is normalized to v_i' by computing

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

($\bar{A}$ ← mean; σ_A ← std. deviation of A)

Average $(\bar{A}) = \frac{1}{n}(v_1 + v_2 + v_3 + \dots)$

$$\sigma_A = \sqrt{\frac{1}{n}\sum (v_i - \bar{A})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Q] Suppose mean and std. dev. of values for income are 54,000 and 16,000, what is the values for the income 73,600 using Z-score norm.

$$v_i' = \frac{73600 - 54000}{16000}$$

$$= 1.225$$

③ Normalization by decimal scaling

→ Normalizes by moving decimal point of values of attribute A
→ No. of decimal points moved depends on max. absolute value of A.
→ A value v_i of A is normalized to v_i' by computing

$$v_i' = \frac{v_i}{10^j}$$ where j → smallest integer such that max. of $(|v_i'|) < 1$

Q] Suppose that recorded values of attribute A range from -986 to 917. The max. absolute value of A is 986.

To normalize by decimal scaling, we divide each value by 1000 (i.e j=3) so that -986 normalizes to -0.986 & 917 normalizes to 0.917

Q] Normalize the following group of data using

① Min-max normalization by setting min=0 & max=0

② Z-score normalization.

Data:- 200, 300, 400, 600, 1000
↓ min ↓ max

Soln

① $v_i' = \frac{-200}{1000-200}(1.0-0.0)+0$

$= \frac{200-200}{1000-200} = 0$ — (i) ⇒ 400 = 1/4 = 0.25

$= 300 = 1/8 = 0.125$ ⇒ 600 = 0.5

⇒ 1000 = 1

② $\bar{A} = 500$

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

$= 282.84$

$v_i' = \frac{200-500}{282.84} = -1.06$ ✓

$= \frac{300-500}{282.84} = -0.707$ ✓

$= \frac{400-500}{282.84} = -0.353$ ✓

$= \frac{600-500}{282.84} = 0.353$ ✓

$= \frac{1000-500}{282.84} = 1.767$ ✓

* Discretization & binarization

⇒ Transform a continuous (continuous) attribute into a categorical attribute [discretization)

⇒ Both continuous & discrete attributes may need to be transformed into one or more binary attributes (binarization)

⇒ Technique for binarization of categorical attribute

→ If there are n categorical values, then uniquely assign each original value to an integer in the interval $[0, m-1]$.

→ Convert each of these m integers to a binary number.

→ Since $n = \lceil \log_2 m \rceil$ binary digits are need to represent the integers the binary nos. are represented as —

Q) Categorical variable with 5 values {awful, poor, OK, good, great}

We require 3 binary variables x_1, x_2, x_3

Categorial value	Integer value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

⇒ Dislocation Convert a categorical attribute to five asymmetric binary attributes

categorical value	Integer value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

4/10/21

Q) Suppose that data for analysis includes the attribute age.
⇒ Age values for the data tuples are (increasing order)

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Q2 Q3 Q4 median

(1) What is the mean of the data?

(2) What is the median?

(2) What is the mode of the data? Comment data's modality (bimodal, trimodal, etc.)

(3) What is the mid range of the data

(4) Can you find roughly the first quartile (Q_1) & 3rd quartile (Q_3) of the data?

(5) Give the five-number summary of the data.

(1)

Mean = 13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36

$= \frac{809}{27} = 29.96$

Median: Middle value of ordered set as the no. of values in this set.
= 25.

(2) Mode : This data set has two values that occur with the same highest freq. & is therefore bimodal.

(3) Midrange

$$\text{MR} = \frac{13+70}{2} = \frac{83}{2} = 41.5$$

Avg of smallest & largest values in the dataset.

(4) Quartile

$\frac{1}{4}(N+1)^{th}$ term $\Rightarrow 23/4 \Rightarrow 13, 15, 16, 16, 19, 20, 20$

$\frac{2}{4}(N+1)^{th}$ term $\Rightarrow$

$\frac{3}{4}(N+1)^{th}$ term $\Rightarrow 21 \Rightarrow 35, 36$

First Q corr. to 25th percentile is 20
Third Q " " 75% percentile is 35.

(5) 5-Nos summ. :

↓

min value, first Q, median, third Q, max value

13, 20, 25, 35, 70

Q) Suppose a hospital tested the age & body fat data for 18 randomly selected adults ~~with~~ in for. results.

age	23	23	27	27	39	41	47	49	50	52	54	54
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8

56	57	58	58	60	61
33.4	30.2	34.1	32.9	41.2	35.7

Calculate

(1) Mean, median, std. deviation of age & % fat

(2) Normalize the two variables based on z-score

(1)

Mean (age) $= \frac{836}{18} = 46.44$ Median $= \frac{50+52}{2} = 51$

Mean (% fat) $= \frac{518.1}{18} = 28.78$. Median $= 32.9$

Std. deviation (age) $\Rightarrow \sigma = \sqrt{\sum_i^n \frac{(x_i - \bar{x})^2}{n-1}}$

$\sigma = \sqrt{\frac{(23-51)^2 + \dots}{17}}$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
23	-23.44	549.43
23	-23.44	549.43
27	-19.44	377.91
27	-19.44	377.91
39	-7.44	55.35
41	-5.44	29.59
47	0.56	0.31
49	2.56	6.55
50	3.56	12.67
52	5.56	30.91
54	7.56	57.15
54	7.56	57.15
56	9.56	91.39
57	10.56	111.51
58	11.56	133.63
58	11.56	133.63
60	13.56	183.87
61	14.56	211.99

Q) Using data for age, answer the foll.

⇒ (i) Use min-max norm. to transform value 35 for age onto the range [0.0, 1.0]

$$v_i' = \frac{(35-13)}{70-13} \times (1.0-0.0) + 0.0$$

$$= \frac{22}{57} \times 1$$

$$= 0.38$$

(ii) Use z-score norm. to transform value 35 for age where std. dev of age is 12.94

$$v_i' = \frac{35-30}{12.94}$$

$$= 0.386$$

(iii) Use normalizn by decimal scaling to transform value 35 for age.

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

$$= \frac{35}{10^2} = 0.35$$

(iv) Comment on which method you would prefer to use giving reasons as to why?

Q) What are the value ranges for foll. normalizn methods

(a) Min-max : $[0_{new\,min}, 1_{new\,max}]$

(b) z-score : $\frac{\text{Old min} - \text{mean}}{\text{std. dev}}$, $\frac{\text{old max} - \text{mean}}{\text{std. dev.}}$

(c) Norm. by decimal scaling :- (-1.0, 1.0)

11/10/21

* Similarity & dissimilarity measure

① Similarity measure
→ Numerical measure of how alike

* Euclidean Distance

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

$$d(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$
$$= \sqrt{8} = 2\sqrt{2} = 2.828$$

$P_1(0,2), P_2(2,0)$

* Minkowski Distance

⇒ Generalization of Euclidean dist.

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

r=1 → City block (Manhattan, taxicab, L_1 norm) dist
r=2 → Euclidean
r→∞ → "Supremum", (L_{max} norm, L_∞ norm).
→ Max diff. b/w any components of vectors

r = 1, $P_2(2,0)$, $P_3(3,1)$

$d(P_2, P_3) = 1 \cdot (|2-3| + |0-1|)$
$= 1 + 1 = 2.$

r = ∞.

L_∞ norm:

$P_3(3,1)$ $P_4(5,1)$

$d(P_3, P_4) = \max [|(3-5)|, |(1-1)|]$

$= 2.$

$P_1(0,2)$ $P_3(3,1)$

$d(P_1, P_3) = \max [|(0-3)|, |(2-1)|]$
$= 3.$

* Simple matching & Jaccard co-eff.

SMC = number of matches / number of attributes
$= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

Jaccard $= (f_{11}) / (f_{01} + f_{10} + f_{11})$

Q] x = 1 0 0 0 0 0 0 0 0 0
y = 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$
$f_{10} = 1$
$f_{00} = 7$
$f_{11} = 0$

$$SMC = \frac{(f_{11}+f_{00})}{f_{01}+f_{10}+f_{11}+f_{00}} = \frac{0+7}{2+1+0+7}$$

$$= \frac{7}{10} = 0.7$$

$$J = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{1+1+0} = 0$$

18/10/21

* Cosine similarity

⇒ Document vector

Document	team	coach	hockey	basketball	soccer
	(frequency)				
Doc 1	5	0	3	0	2
Doc 2	3	0	2	0	1
Doc 3	0	7	0	2	1
Doc 4	0	1	0	0	1

	penalty	score	win	loss	season	...
						→ term frequency vector
doc1	6	6	2	0	6	...
doc2	1	0	1	0	1	...
" 3	6	6	3	0	0	...
" 4	2	2	0	3	0	...

⇒ Let x & y be the vectors for comparison.

Using cosine similarity, cosine simi(x,y) $= \frac{x \cdot y}{\|x\| \|y\|}$

where $\|x\|$ is the Euclidean norm of vector $x = (x_1, x_2, \ldots, x_p)$

defined by $\sqrt{x_1^2 + x_2^2 + \ldots + x_p^2}$

Conceptually, it is the length of vector

$\|y\|$ is the Euclidean norm of vector y.

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

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

$0\ 1\ 0\ 0\ 1\ 2\ 2\ 0\ 3\ 0 \Rightarrow 2+1+1+4+4+9$

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

$x \cdot y = 15 + 0 + 6 + 0 + 2 + 0 + 0 + 2 + 0 + 0 = 25$

$\|x\| = \sqrt{42} = 6.48$

$\|y\| = \sqrt{17}\ (9+4+4) = 4.12$

$Simi(x,y) = \frac{25}{\sqrt{42} \times \sqrt{17}} = 0.935$

Cosine simi(doc1, doc4)

⇒ x ⋅ y = 2.

$||x|| = \sqrt{42}$

$||y|| = \sqrt{19}$.

Simi(x,y) = 0.0748

* Disimilarity for attributes of mixed type

⇒

Object identifier	Test-1 (nominal)	Test-2 (decimal)	test-3 (numeric)
1	Code A	excellent	45
2	Code B	fair	22
3	Code C	good	64
4	Code A	excellent	28

⇒ Suppose the dataset contains 'p' attributes of mixed type, the disimilarity d(i,j) b/w objects i & j is defined as

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

where indicator $\delta_{ij}^{(f)} = 0$ if either

(1) x if or x_{jf} is missing (there is no measurement of attribute f for object i or object j)

OR

(2) if $x_{if} = x_{jf} = 0$, & the attribute f is asymmetric binary, otherwise $\delta_{ij}^{(f)} = 1$

(3) The contribution of attribute f to the disimilarity b/w i & j ($d_{ij}^{(f)}$) is completely depending on its type

(1) If f is numeric, $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$, h

where

runs over all non-missing objects of attribute f.

(2) If f is nominal or binary, $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$;

otherwise $d_{ij}^{(f)} = 1$

(3) If f is ordinal, compute the ranks r_{if} & $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

treat z_{if} as numeric.

Cosine simi(doc1, doc4)

⇒ $x \cdot y = 2$

$||x|| = \sqrt{42}$

$||y|| = \sqrt{19}$

$Simi(x,y) = 0.0748$

* Disimilarity for attributes of mixed type

⇒

Object identifier	Test-1 (nominal)	Test-2 (decimal)	test-3 (numeric)
1	Code A	excellent	445
2	Code B	fair	22
3	Code C	good	64
4	Code A	excellent	28

⇒ Suppose the dataset contains 'p' attributes of mixed type, the disimilarity d(i,j) b/w objects i & j is defined as

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

where indicator $\delta_{ij}^{(f)} = 0$ if either

(1) x if or x_{jf} is missing (there is no measurement of attribute f for object i or object j)

OR

(2) if $x_{if} = x_{jf} = 0$, & the attribute f is asymmetric binary, otherwise $\delta_{ij}^{(f)} = 1$

(3) The contribution of attribute f to the disimilarity b/w i & j ($d_{ij}^{(f)}$) is completely depending on its type

(1) If f is numeric, $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$, h where runs over all non-missing objects of attribute f.

(2) If f is nominal or binary, $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise $d_{ij}^{(f)} = 1$

(3) If f is ordinal, compute the ranks r_{if} & $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

treat z_{if} as numeric.

test 3	1	2	3	4
1	0			
2	0.55	0		
3	0.45	1	0	
4	0.404	0.14	0.857	0

$d(2,1) = \frac{45-22}{64-22}$

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

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

$(4,1) =$

$(4,2)$

$(4,3)$

test 2	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1	0.5	0

fair good excellent

0 1 2

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

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

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

$(4,1) = \frac{2-3}{2}$

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

$(4,3) = \frac{2-1}{2} = 1$

test 1	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

dissimilarity	1	2	3	4
1	0			
2	0.85 → most dissimilar	0		
3	0.65	0.83	0	
4	0.13 ↓ most similar	0.71	0.786	0

$d(3,1) = \frac{1\times0.45 + 1\times0.5 + 1\times1}{3} = 0.65$

$d(2,1) = \frac{1\times0.55 + 1\times0.5 + 1}{3} = 0.85$

$d(3,2) = \frac{1\times1 + 1\times0.5 + 1\times1}{3} = 0.83$

$d(4,1) = 0.13$

$d(4,2) = 0.71$

$d(4,3) = 0.786$

21/10/21

Correlation analysis

⇒ Measures how strongly one attribute implies the other based on available data

⇒ Chi-Square test :- (χ^2) :- for only nominal data

→ Consider two attributes a & b.
Attribute A = {$a_1, a_2, a_3, a_4 \dots a_c$}
B = {$b_1, b_2, b_3 \dots b_r$}

⇒ data tuples described by A & B can be shown as contingency table where

let (A_i, B_j) denote joint event that attribute A takes on value a_i & attribute B takes on value b_j

χ^2 value (AKA Pearson's χ^2 statistic) is computed as

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

O_{ij} : observed freq.

Where O_{ij} is the (actual count) of joint events (A_i, B_j) and e_{ij} is the expected freq. of (A_i, B_j) which can be computed as

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

Where, n → no. of data tuples.

count $(A = a_i)$ is the no. of tuples having values a_i for A, count $(B = b_j)$ is the no. of tuples having b_j for B.

The χ^2 statistic tests the hypothesis that A & B are independent i.e no correlation b/w them. Test is based on significance level with $(r-1) \times (c-1)$ degrees of freedom.

Eg: Suppose that group of 1500 ppl was surveyed gender of each person was noted. Each person was polled to find out his/her preferred type of reading material. (fiction or non-fiction)

⇒ The observed freq./count of each possible joint event is summarized in contingency table given b/w

	male	female	total
fiction	250 (90)	200 (360)	450 → how much
non-fiction	50 210	1000 (840)	1050
Total	300	1200	1500
	↓ column wise		

Calculated e_{ij}

$$e_{11} = \frac{\text{count(male)} \times \text{count(fiction)}}{n} = \frac{300}{1500} \times 450 = 90$$

$$e_{12} = \frac{\text{count(female)} \times \text{count(non-fiction)}}{1500} = \frac{1200 \times 450}{1500} = 360$$

e_{21} = count (male) × count (non-fiction) / 1500

= 300 × 1050 / 1500 = 210

e_{22} = count (female) × count (fiction) / 1500

= 1200 × 1050 / 1500 = 840

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

= 507.93

For this table, degree of freedom are
$(r-1)(c-1)$

$(2-1)(2-1) = 1$

∴ For 1 degree of freedom, χ^2 value needed to reject the hypothesis at 0.001 significance level is 10.828 (χ^2 distribution)

⇒ Since χ^2 is much higher than significance level, we reject the hypothesis that gender & preferred reading is independent, and is highly correlated.

* Correlation co-efficient for numeric data

⇒ For numeric attribute, we can evaluate correlation b/w two attributes A & B by computing correlation coefficient (Pearson's product moment coeff.)

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

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

where n is the number of tuples
a_i & b_i are respective values of A & B in tuple i,
$\bar{A}$ & $\bar{B}$ are respective mean values of A & B
σ_A & σ_B are respective std. deviations of A & B

$\Sigma(a_i b_i)$ is the sum of AB cross product

$-1 \le r_{A,B} \le +1$,

If $r_{A,B} > 0$, A & B are positively correlated when value of A ↑, B also ↑

If $r_{A,B} = 0$, attribute A & B are independent & no correlation b/w them

If $r_{A,B} < 0$, attribute A & B are negatively correlated i.e A ↓, B ↓, each attribute discourages each other.

* Covariance of numeric data

⇒ Correlation and covariance are two similar measures for assessing how much 2 attributes change together

Consider two numeric attributes A & B and set of 'n' observations

$\{(a_1, b_1), \ldots, (a_n, b_n)\}$

⇒ The mean values of A & B respectively are also known as expected values on A & B

i.e

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

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

Covariance b/w A & B is defined as

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

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

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

Also, it can be shown that $Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$

For two attributes A & B that tend to change together, if A is larger than $\bar{A}$, then B is likely to be larger than $\bar{B}$ (covariance b/w A & B is positive)

If one of the attributes is higher than expected value, and other is lower than expected (covariance b/w A & B is negative)

If A & B is independent, $E(A \cdot B) = E(A) \cdot E(B)$

$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B} = E(A) \cdot E(B) - \bar{A}\bar{B} = 0$

⇒ Consider the table given, which presents stock prices observed at five time points for AllElectronics & HighTech,

If stocks are affected by same industry trends, will their prices rise or fall together?

Timepoint	AllElectronics	HighTech
t_1	6	20
t_2	5	10
t_3	4	14
t_4	3	5
t_5	2	5

Covariance

$E(\text{AllElectronics}) = \frac{6+5+4+3+2}{5}$

$= \$4$

$E(\text{High Tech}) = \frac{20+10+14+5+5}{5}$

$= \frac{30+14+10}{5}$

$= \frac{54}{5} = \$10.80$

⇒ Covariance (AllElectronics, High Tech)

$= E(AB) - \bar{A}\bar{B}$

$= \frac{6\times20 + 5\times10 + 4\times14 + 3\times5 + 2\times5}{5}$

$- 4 \times 10.8$

$50.2 - 43.2$

$= 7$

Covariance is positive, we can say that stock prices for both companies rise together.

28/10

Q] A random sample of 395 ppl was surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table

	High School	Bachelors	Masters	PhD.	
Female	O_{11} 60	O_{12} 54	O_{13} 46	O_{14} 41	⇒ 201
Male	O_{21} 40	O_{22} 44	O_{23} 53	O_{24} 57	⇒ 194
Total = 100		98	99	98	395

Is there a relationship b/w gender of an individual & level of education they have obtained?

Two nominal attributes :-

χ^2 test :-

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

$$e_{ij} = \frac{\text{Count}(A = a_i) \times \text{Count}(B = b_i)}{n}$$

$e_{11} = \frac{100 \times 201}{395} = 50.886$

$e_{12} = \frac{98 \times 201}{395} = 49.86$

$e_{13} = \frac{99 \times 201}{395} = 50.377$

$e_{14} = \frac{98 \times 201}{395} = 49.868$

$e_{21} = \frac{100 \times 194}{395} = 49.114$

$e_{22} = \frac{99 \times 194}{395} = 48.13$

$e_{23} = \frac{99 \times 194}{395} = 48.62$

$e_{24} = \frac{98 \times 194}{395} = 48.132$

$$\chi^2 = \frac{(60-52.886)^2}{52.886} + \frac{(54-49.868)^2}{49.868}$$

$$+ \frac{(46-52.377)^2}{52.377} + \frac{(41-49.868)^2}{49.868}$$

$$+ \frac{(40-49.114)^2}{49.114} + \frac{(44-48.132)^2}{48.132}$$

$$+ \frac{(53-48.623)^2}{48.623} + \frac{(57-48.132)^2}{48.132}$$

$$= 1.632 + 0.342 + 0.380 + 1.576 + 1.691 + 0.354 + 0.394 + 1.633$$

$$= 8.002 \ (\approx 8.00)$$

Null hypothesis: Two attributes, gender & education level are independent.

D.O.F = $(r-1)(c-1) = (2-1)(4-1)$

$= 1(1)(3) = 3.$

S/P: For 3 D.O.F at 5% significance level, the χ^2 value is 7.815.

The χ^2 obtained is much higher than significance level, we reject the null hypothesis that gender & education level is independent & are highly correlated. and has an impact on educ level of an individual.

Unit 2: Simbath, dissim., Chp 3

Chp 1, Chp 2, Chp 3 uph statistics.

(2) Using the data for age & body fat:

age	23	23	27	27	39	41	47	49	50
body fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

Calculate the correlation coefficient (Pearson's product moment coeff). Are these two attributes positively or negatively correlated? Compute their covariance

$$r_{A,B} = \sum_{i=1}^{n} \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

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

$n = 9.$

$\bar{A}$ (age) $= 36.22$

$\bar{B}$ (body fat) $= 22.74$

$$Cov(A,B) = \frac{1}{9}\Big[(23-36.22)(9.5-22.74) + (23-36.22)(26.5-22.74) + (27-36.22)(17.8-22.74) + (27-36.22)(7.8-22.74) + (39-36.22)(31.4-22.74) + (41-36.22)(25.9-22.74) + (47-36.22)(27.4-22.74) + (49-36.22)(27.2-22.74) + (50-36.22)(31.2-22.74)\Big]$$

$$= \frac{1}{9}(175.03 - 49.7072 + 24.0748 + 15.1048 + 137.74 + 45.54 + 50.2348 + 56.9988 + 116.578)$$

$$= 63.51$$

$$\sigma_A = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}} = \sqrt{\frac{1}{9}\left[(23-36.22)^2+(23-36.22)^2+(27-36.22)^2+(27-36.22)^2+(39-36.22)^2+(41-36.22)^2+(47-36.22)^2+(49-36.22)^2+(50-36.22)^2\right]}$$

$$= \sqrt{\frac{1}{9}\times 9.24} = 1.013$$

$$\sigma_B = \sqrt{\frac{1}{9}\times$$

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

$$\text{Cov}(A,B) = \frac{1}{9}\Big[(23-36.22)(9.5-22.74)+(23-36.22)(26.5-22.74)+(27-36.22)(7.8-22.74)+(27-36.22)(17.8-22.74)+(39-36.22)(31.4-22.74)+(41-36.22)(25.9-22.74)+(47-36.22)(27.4-22.74)+(49-36.22)(27.2-22.74)+(50-36.22)(31.2-22.74)\Big]$$

$$= \frac{1}{9}\Big[175.032 - 49.707 + 137.747 + 45.549 + 24.075 + 15.105 + 50.235 + 56.999 + 116.579\Big]$$

$$= 63.51$$

$$\sigma_A = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}} = \sqrt{\frac{1}{9}\left(519.5536 + 500.062\right)}$$

$$= \sqrt{\frac{1}{9}\times 1019.5556}$$

$$= \sqrt{113.28} = 10.6433$$

$$\sigma_A = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}} = \sqrt{\frac{1}{9}\Big((9.5-22.74)^2+(26.5-22.74)^2+(7.8-22.74)^2+(17.8-22.74)^2+(31.4-22.74)^2+(25.9-22.74)^2+(27.4-22.74)^2+(27.2-22.74)^2+(31.2-22.74)^2\Big)}$$

$$= \sqrt{\frac{1}{9}(522.0236 + 113.1788)}$$

$$= \sqrt{\frac{1}{9}(635.2024)}$$

$$= \sqrt{70.578} = 8.401$$

$$r_{A,B} = \frac{63.51}{10.643 \times 8.401} = 0.7103$$

Since, $r_{A,B} = 0.7103$ which is positive, the attributes age and body fat are positively correlated. As age increases, body fat also increases.

offline

9/11/2021 *Visual Visualization techniques

* Visualizing small no. of attributes:

(i) Stem & leaf plot
→ Insights into distrib of 1-D integer/continuous data.
→ Split the values into groups:
→ Each group contains those values that are the same except for the last digit
→ Each group becomes the stem, while the last digit of a group are the leaves.

Age attribute of employee table:

22, 22, 23, 23, 23, 25, 28, 28, 30, 31, 31, 32, 35, 40, 40, 41, 42, 45, 45, 48, 50, 52, 54, 55

2 : 2, 2, 3, 3, 3, 5, 8, 8 → leaves
3 : 0, 1, 1, 2, 5
4 : 0, 0, 1, 2, 5, 5, 8
5 : 0, 2, 4, 5
↙ stem

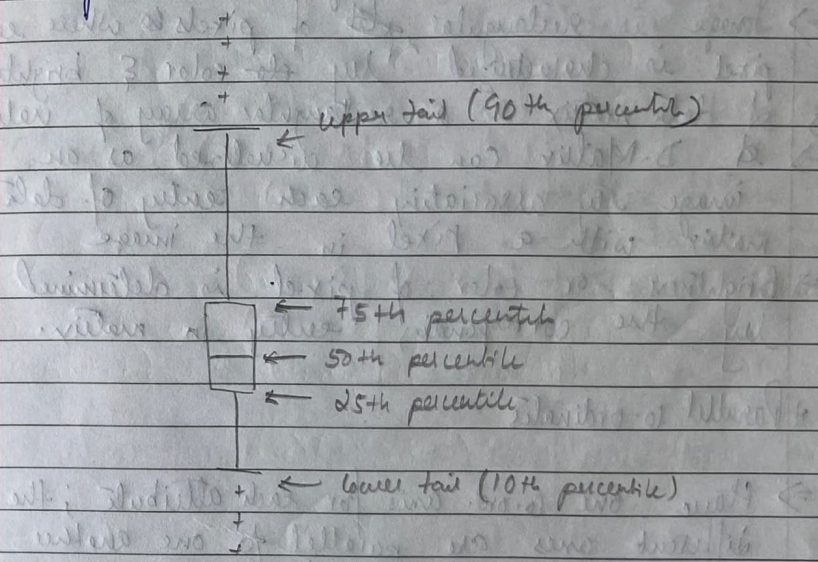
→ By plotting stem vertically & leaf horizontally visual representation of data distribution

(ii) Histogram:

→ Continuous attributes
→ Range of values is divided into bins. (equal width usually)
→ Data objects / values in each bin is counted.

→ Bar plot is constructed, wherein each bin is represented by a bar. Area of bar ∝ no. of values / objects that fall into the corr. range.

* Box plot.



* Scatter plot.

→ Used to illustrate linear correlation
→ Each data object is plotted as a point in the plane using values of two attributes : x & y co-ord
→ Attributes are either integers / real valued.

* Contour plot

⇒ Used for 3-D data of which two attributes specify position in a plane
Third attribute (continuous attribute)
Eg:- temp / elevation
⇒ This breaks the plane into regions with roughly the same attribute values.

* Higher dimension data

* Matrix plot

⇒ Image is rectangular array of pixels in where each pixel is characterized by color & brightness
⇒ A data matrix is a rectangular array of values
⇒ A D. Matrix can be visualized as an image by associating each entry of data matrix with a pixel in the image
⇒ Brightness or color of pixel is determined by the corresponding entry in matrix.

* Parallel co-ordinates

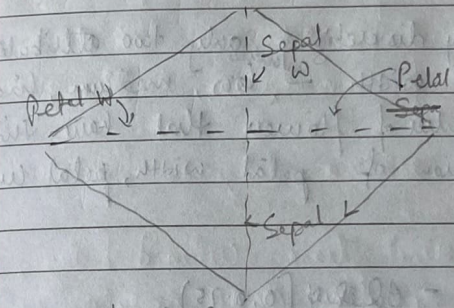
⇒ Have one co-ord. axis for each attribute; the different axes are parallel to one another instead of ⊥'r (perpendicular)
⇒ Object is represented as line instead of point.

Other techniques

* Star plots → Similar to parallel co-ords; axes radiate from central pt.; line connecting values of obj is polygon

* Chernoff faces.
→ Herman Chernoff; approach associates each attribute with characteristics of a face

Star plot



glyph / icon.

Chernoff face:

Sepal L → Size of face
" W → forehead / jaw relative axes length
Petal L → Shape of forehead
" W → " of jaw

10/4/2021

* OLAP & Multidimensional data analysis :-

OLAP

⇒ Online Analytic Processing (OLAP)

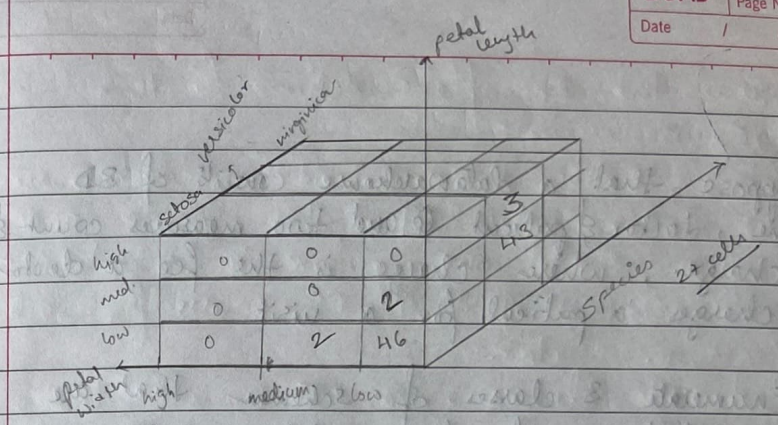
* Represent this data as multidimensional array

→ We create a table by discretizing only two attributes i.e petal length & petal width — low, medium, high. & then count number of flowers that have the a particular combination of petal width, petal length and species (class)

→ petal width — { Low - [0, 0.75), medium - [0.75, 1.75), high - [1.75, ∞) }

petal length — { low - [0, 2.5) medium - [2.5, 5) high - [5, ∞) }

petal length	petal width	species type	count
low	low	setosa	46
low	medium	setosa	2
medium	low	setosa	2
medium	medium	versicolor	43
medium	high	"	3
"	high	virginica	3
high	medium	versicolor	2
"	"	virginica	3
high	high	versicolor	2
"	"	Virginica.	44



setosa species :-

petal length \ petal width	low	medium	high
low	46	2	0
med.	2	0	0
high.	0	0	0

versicolor species :-

petal length \ petal width	low	med.	high
low	0	0	0
med	0	43	3
high	0	2	2

Virginica species

petal length \ petal width	low	med.	high
low	0	0	0
med.	0	0	3
high	0	3	44

*Write the notes of last week i.e 11th & 12th

15/11/2021

Q] Suppose that a datawarehouse consists of 3D time, doctor & patient & and two measures count & charge., where charge is the fee a doctor charges a patient for a visit

(a) Enumerate 3 classes of schemas that are popularly used for modelling data warehouses

(b) Draw a Star schema diagram for the above datawarehouse.

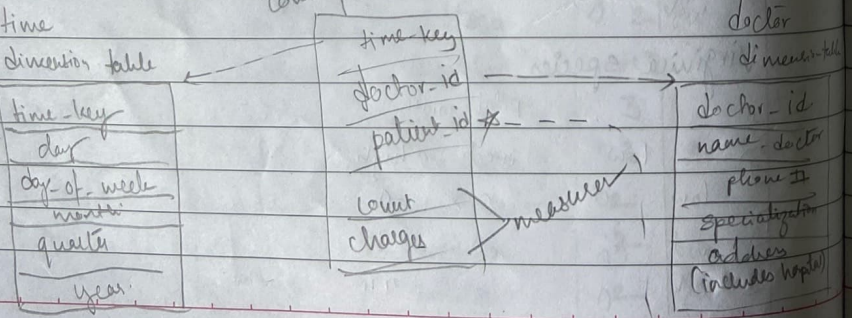
(c) Starting with the base cuboid [day, doctor, patient], what specific OLAP operation should be performed in order to list the total fee collected by each doctor in 2020.

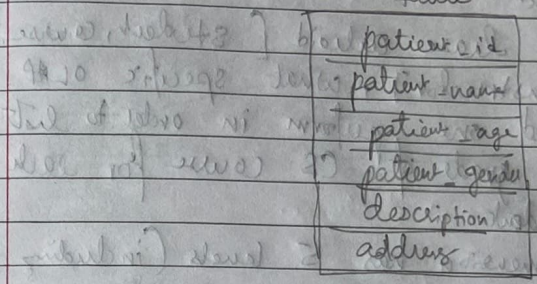
(d) To obtain the same list, write a SQL query assuming the data is stored in RDB with schema fee (day, month, year, doctor, hospital, patient, ~~charge~~, ~~count~~ count, charge)

Sol^n

(a) Star, snowflake, fact constellation schema

(b)





(c) (i) Roll-up (day to year)
(ii) Slice (on time for year = 2020)
(iii) Roll-up on patient from patient name to All.

(d) Select doctor, sum(charge) from fee where year = 2020
group by doctor

Q] Suppose that a datawarehouse for Pkg University consists of the following four dimensions:
1. Student
2. Course
3. Semester
4. Instructor

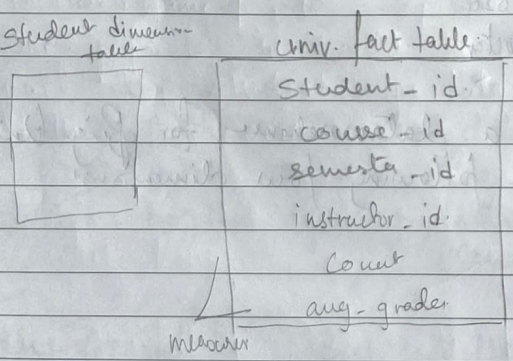
and two measures; count & avg. grade.

⇒ When at the lowest conceptual level (example; for given Student, course, semester, instructor combi), the average grade measure stores the actual course grade of student. At higher conceptual levels, avg. grade stores the avg. grade for given combin^n

(a) Draw the snowflake schema for this data warehouse.

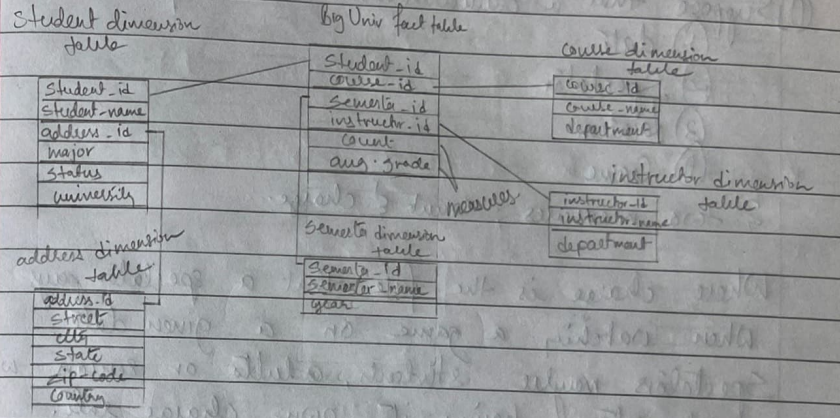
(b) Starting with base cuboid [student, course, semester, instructor], what specific OLAP operation should one perform in order to list the average grade of CS course for each BigUniv. student.

(c) If each dimension has 5 levels (including ALL) such as
"student < major < status < university < "ALL")
How many cuboids will this cube contain including the base and apex cuboid.



Soln

(a)



(b)
(i) Roll up on course from course_id to dept
(ii) Roll up on student from student_id to university
(iii) Dice up on course; dept = 'CS' & univ = "big Univ"
(iv) Drill down on student from univ to student_name

(c)(d)

No of cuboids = $(\text{Levels})^{\text{no. of dimensions}}$

Dimensions = 4
L_i = 5 Levels = 5
(L_i)

$\therefore$ No. of cuboids = $(L_i+1)^4$
$= (4+1)^4 = 625$

$\Rightarrow (L_i)^{\text{dimensions}}$
$\Rightarrow (5)^4 = 625$

17/11/2021

Q] Suppose that a datawarehouse consists of 4 dimensions
① date
② Spectator
③ location
④ Game
& two measures; count & charge.

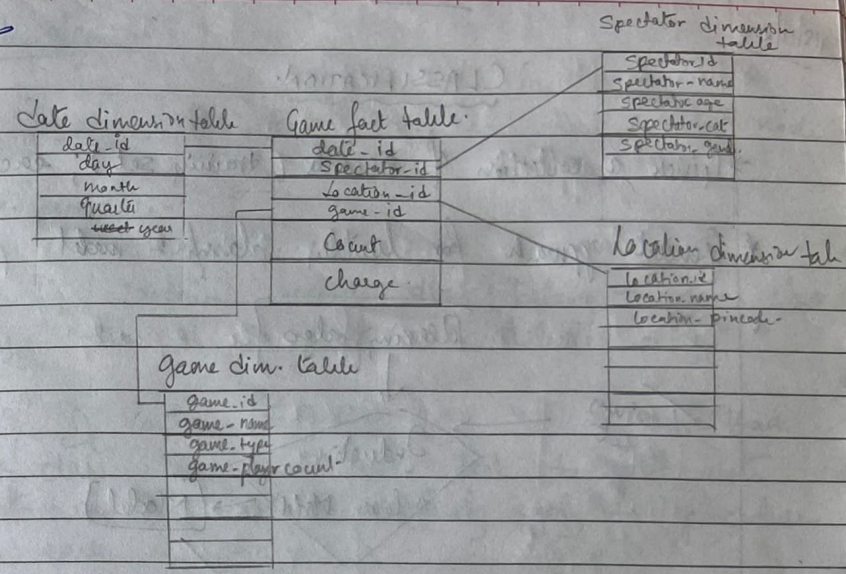
Where charge is the fare that a spectator pays when watching a game on a given date. Spectators maybe students, adults or seniors with each category having its own charge rate.

ⓐ Draw a star schema diagram for the datawarehouse.
ⓑ Starting with the base ~~cuboid~~ cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM place in the year 2016

location name

Soln

ⓐ



ⓑ cuboid [date, spectator, location, game]

① Roll up on date from date_id to year.
② Roll up on location from location_id to location_name.
③ Roll up on spectator from spectator_id to spectator_categor
④ Dice on spectator; category = student, year = 2016 & location_name = GM place.
⑤ Roll up on game from game_id to ALL.

End of Unit 2

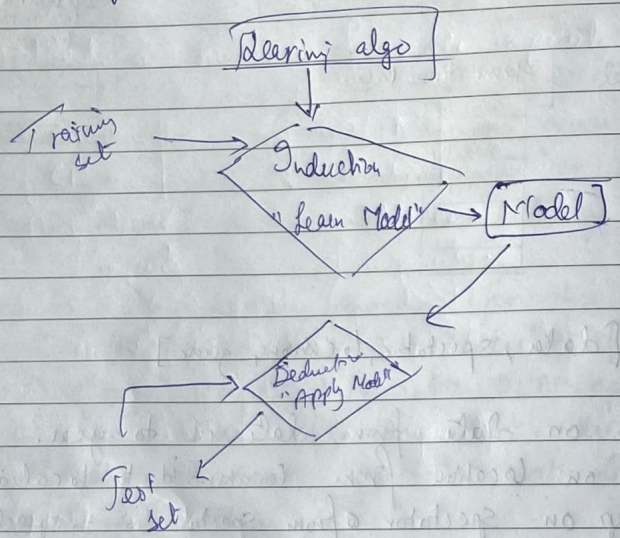
Unit 3

18/11/2021

CLASSIFICATION.

→ Given a collection of records (training set); each

* General approach for building classifier model



* Classifier tech.

→ Base classifiers

→

→ Ensemble "

* Hunt's algorithm

→ Let D_t be the set of tr. re. that reach a node t.

→ Gen. procedure.

(i) If D_t contains records that belong to same class y_t then t is leaf node labelled as y_t.

(ii) If D_t contain records " " " more than one class, use an attribute test to split data into smaller subsets. Recursively apply to procedure to each subset.

* Design issues

(1) How should training records be split?
(2) How should splitting procedure stop?

Online.

19/4/2021

* How to determine best split?

① Greedy approach

→ Nodes with purer class distribution are preferred.

→ Need a measure of node impurity:

C0 : 5
C1 : 5

High degree of impurity

C0 : 9
C1 : 1

Low degree of impurity.

↓ → equally ~~not eqn~~ divided.

≥ bias low.

→ 3 measures of node impurity.

① GINI index

$$\text{Gini Index} = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

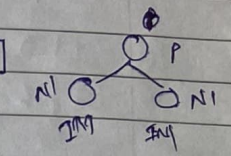
where $p_i(t)$ → freq. of class i at node t & c is total no. of classes

② Entropy

$$\text{Entropy} = -\sum_{i=0}^{c-1} p_i(t) \log_2 p_i(t)$$

③ Misclassification error

$$\text{Classification error} = 1 - \max[p_i(t)]$$



* Finding the best split

① Compute impurity measure (p) before splitting

② " " " " after "

→ Compute impurity measure of each child node

→ M is weighted impurity of ~~the~~ child nod:

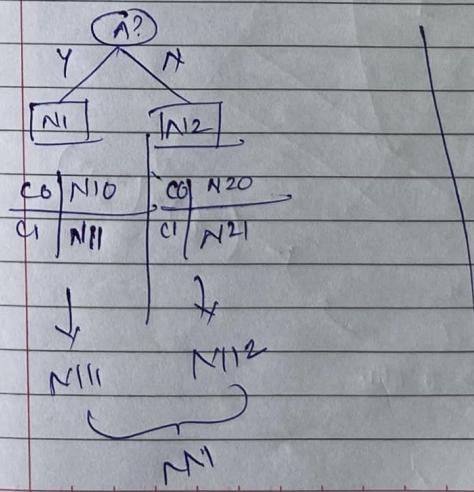
③ Choose attribute test condition that produces highest gain

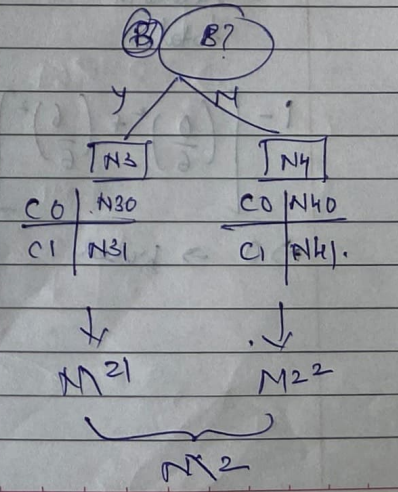
$$\text{Gain} = P - M.$$

or equivalently, lowest impurity measure after splitting.

Q]

Before splitting : C0 | N00 → P
C1 | N01





Gain = P - M1 (1) vs P - M2 (2)

Max [1, 2] is chosen for splitting.

* GINI

~~Used in tree.~~

$$1 - \sum_{i=0}^{c-1} p_i(t)^2$$

For 2-class prob (p, 1-p):

$$GINI = 1 - p^2 - (1-p)^2$$
$$= 2p(1-p)$$

(1)

C1	0
C2	6

6 records in total.

$$1 - \left[\left(\frac{0}{6}\right)^2 + \left(\frac{6}{6}\right)^2\right]$$

$\Rightarrow 0$ → biased

(2)

C1	1
C2	5

$$1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right]$$

GINI = 0.278 → Not uniformly distributed

(3)

C1	2
C2	4

$$1 - \left[\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right]$$

$= 0.444$

(4)

C1	3
C2	3

$$1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$$

$\Rightarrow 0.500$ → Equal distributed

* GINI for single node.

①

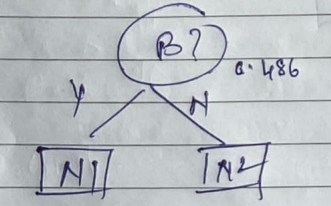
C1	0
C2	6

* GINI for collection of node.

When a node p is split into k partition

$GINI_{split}$ =

* Q]



Parent

C1	7
C2	5
	12

At parent,

$\therefore 1 - \left[\left(\frac{7}{12}\right)^2 + \left(\frac{5}{12}\right)^2 \right]$

$\Rightarrow 0.4861$

	N1	N2
C1	5	2
C2	1	4
	6	6

$1 - \left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2$

$\Rightarrow 0.278$ $\Rightarrow 0.444$

Weighted GINI : N1 N2

$\Rightarrow \frac{6}{12} * 0.278 + \frac{6}{12} * 0.444$

$\Rightarrow 0.361$

$\therefore$ Gain = 0.486 - 0.361

* GINI for categorical ~~Attributes~~

Multiway split : Two-way split

* Least GINI index to be selected to split

* refer ss one drive 19/11/2021

Family	sport	Luxury
0.375	0	0.219

Weighted gain : $\frac{4}{20} \times 0.375 + \frac{8}{20} \times 0.375 + \frac{8}{20} \times 0.219$

$= 0.163$ (0.1622)

* Continuous attribute GINI index

→ Sort attribute on values.
→ Linearly scan these values, each time updating the count matrix & computing gini index
→ Choose split position that has least GINI index.

22/11/2021

Q) Draw a full decision tree for the parity function of 4 boolean attributes A, B, C, & D. Assume even parity

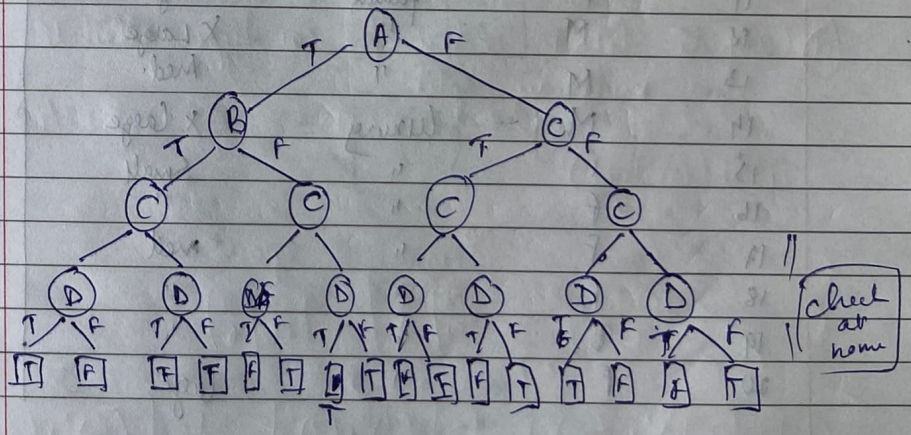
o/p

Even no of T's → 1
Odd no of T's → 0
o/p

= E[A, B, C, D]

Attribute :

A	B	C	D	class
T	T	T	T	T
T	T	T	F	F
T	T	F	T	F
T	T	F	F	T
T	F	T	T	F
T	F	T	F	T
T	F	F	T	T
T	F	F	F	F
F	T	T	T	F
F	T	T	F	T
F	T	F	T	T
F	T	F	F	F
F	F	T	T	T
F	F	T	F	F
F	F	F	T	F
F	F	F	F	T



Q) Consider the training examples shown in table below for binary classification.

Customer ID	Gender	Car type	Shirt size	Class
1	M	family	Small	C0
2	M	Sports	med	"
3	M	"	"	"
4	M	"	large	"
5	M	"	X Large	"
6	M	"	X Large	"
7	F	"	Small	"
8	F	"	"	"
9	F	"	med	"
10	F	Luxury	large	C0
11	M	family	large	C1
12	M	"	X Large	"
13	M	"	med.	"
14	M	Luxury	X Large	"
15	F	"	Small	"
16	F	"	"	"
17	F	"	med	"
18	F	"	"	"
19	F	"	"	"
20	F	"	large	C1

(a) Compute GINI index for overall collection of training examples.

$$GINI(t) = 1 - \sum_{i=0}^{c-1} [p_i(t)]^2$$

$p_i(t)$ → fraction of data records belonging to class partic.

$$GINI(t) = 1 - \left[\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2\right]$$

$$= 0.5$$

(b) Compute GINI index for customer id attribute.

Cust. Id = 1 → [C0 : 1, C1 : 0]

$$\Rightarrow 1 - \left[\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right] = 0$$

Cust. Id = 20 → [C0 : 0, C1 : 1]

$$= 0$$

∴ Overall customer id GINI index = 0.

(c) Compute GINI index for gender attribute

GINI (gender) =

Male	
C0	6
C1	4

Female	
C0	4
C1	6

$\Rightarrow 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right] \quad 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right]$

$\Rightarrow 0.48 \quad \Rightarrow 0.48$

Compute overall GINI for gender.

$\frac{10}{20} * 0.48 + \frac{10}{20} * 0.48 \Rightarrow 0.48$

(d) Compute GINI index for car type attribute using multiway split

GINI (car type)

Family	
C0	1
C1	3

Luxury	
C0	1
C1	7

Sports	
C0	8
C1	0

$1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right] \quad 1 - \left[\left(\frac{1}{8}\right)^2 + \left(\frac{7}{8}\right)^2\right] \quad 1 - \left[\left(\frac{8}{8}\right)^2 + \left(\frac{0}{0}\right)^2\right]$

0.375 0.2188 0

Overall GINI $\Rightarrow \frac{4}{20} * 0.375 + \frac{8}{20} * 0.2188 + \frac{8}{20} * 0$

$= 0.16252$

(e) Shirt ~~type~~ size attribute

GINI (shirt size)

Small	Medium	Large	XL
C0 : 3	C0 : 3	C0 : 2	C0 : 2
C1 : 2	C1 : 4	C1 : 2	C1 : 2
$1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right]$	$1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right]$	$1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$	$1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$
0.48	0.4897	0.5	0.5

Overall: $\frac{5}{20} * 0.48 + \frac{7}{20} * 0.4897 + \frac{4}{20} * 0.5 + \frac{4}{20} * 0.5$

$= 0.491395$

(f) Which attribute is better for split.

→ Car type attribute (reason: lowest GINI value).

(g) Explain why customer id should not be used as attribute test cond. even though it has lowest GINI

→ No predictive power even though GINI is 0.

24/11/2021 Tutorials

3 indexes
① Gini index
② Entropy
③ Misclassification rate (Classification error)

Q] Consider the training examples shown in the table below for binary classification problem.

Instance	a_1	a_2	a_3	target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−
				+ 4/9 − 5/9

(a) What is the entropy of this collection of training example wrt positive class

$$\text{Entropy} = -\sum_{i=0}^{c-1} p_i(t) \log_2 p_i(t) \quad \left(p_i(t) = \text{fraction of records belonging to class } i\right)$$

where
c = number of classes & $0 \log_2 0 = 0$ for entropy calculations.

$$\therefore \text{Entropy} = -\sum_{i=0}^{8}$$

Lesser entropy, purer class

⇒ 4 records − + class
5 " − − " "

$\log_2(5/9) = \frac{\log_{10}(5/9)}{0.3010}$

$P(+) = 4/9 \qquad P(-) = 5/9$

$\therefore$ Entropy of training samples $= -\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right)$

$= 0.991$

(b) What are the Info. gains of a_1 & a_2 related to this training examples.

⇒

a_1	+	−
T	3	1
F	1	4

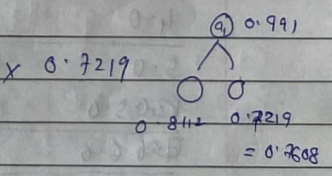
a_2	+	−
T	2	3
F	2	2

$\therefore$ Entropy $a_1(+) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.8112$

Entropy $a_1(-) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) = 0.7219$

Overall entropy $= +\frac{4}{9} \times 0.8112 + \frac{5}{9} \times 0.7219$

$= + 0.7615$



Info. gain $= 0.991 - 0.7615$
for a_1 $= 0.2295$

Entropy $a_2(+) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= 1$

Entropy $a_2(-) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$= 0.9709.$

Overall entropy $= \frac{4}{9} \times 1 + \frac{5}{9} \times 0.9709$

$= 0.9839$

$\therefore$ Info. gain $= 0.9911 - 0.9839$

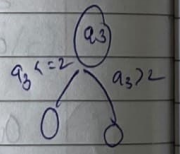
$= 8 \times 10^{-3}$ (0.0072)

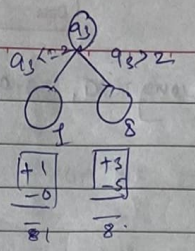
(c) For a_3 which is a continuous attribute, compute info. gain for every possible split

⇒

a_3	Class label	Split point
1.0	+	2.0
3.0	−	3.5
4.0	+	4.5
5.0	−	5.5
5.0	−	
6.0	+	6.5
7.0	−	7.5
7.0	+	
8.0	−	

(consecutive record divide by 2)



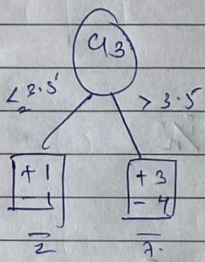


⇒ $a_3(+)$

$-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)$

⇒ $\frac{1}{9}\left[-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right] + \frac{8}{9}\left[\frac{-3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right]$

$= 0.8484.$



⇒ $\frac{2}{9}\left[\frac{-1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] + \frac{7}{9}\left[\frac{-3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$

$= 0.9885$

$\therefore$

Entropy	Inf. gain
0.8484	0.9911 − 0.8484 = 0.1427
0.9885	0.0026
0.9183	0.0728
0.9839	0.0072
0.9728	0.0183
0.8889	0.1022

Point of best split

(d) What is the best split among a_1, a_2 & a_3 according to info. gain?

⇒ a_3

25/11/21

* Algorithm for Decision Tree Induction.

* Problem with large no. of partitions

* Gain Ratio

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

$$\text{Split Info} = -\sum_{i=1}^{k} n_i/n \log_2 n_i/n$$

	C.T			C.T			C.T	
	F	S	L		S,L	F		S	F,L
C1	1	8	1	C1	9	1	C1	8	2
C2	3	0	7	C2	7	3	C2	0	10

Split Info = 1.52 | S. Info = 0.72 | S. Info = 0.97

$-\frac{4}{20}\log_2\left(\frac{4}{20}\right)$ $-\frac{8}{20}\log_2\left(\frac{8}{20}\right)$ $-\frac{8}{20}\log_2\left(\frac{8}{20}\right)$ $= 1.5219$

$-\frac{16}{20}\log_2\left(\frac{16}{20}\right)$ $-\frac{4}{20}\log_2\left(\frac{4}{20}\right)$ $= 0.7219$

$-\frac{8}{20}\times\log_2\left(\frac{8}{20}\right)$ $-\frac{12}{20}\times\log_2\left(\frac{12}{20}\right)$ $= 0.9709$

* Classification error.

⇒ Class. error at node t:

$$\text{Error}(t) = 1 - \max_i [p_i(t)]$$

↓ max frac. belonging to particular class

→ Max. of $1 - \frac{1}{c}$ when records are equally distrib. among all classes.

→

Q]

C1 0
C2 6 ⇒ $P(C1) = 0/6 = 0$, $P(C2) = 6/6 = 1$

$\therefore$ Error $= 1 - \max(0, 1)$
$= 1 - 1 = 0$

C1 1 ⇒ $P(C1) = 1/6$, $P(C2) = 5/6$
C2 5 Error $= 1 - \max(1/6, 5/6)$

$= 1 - 5/6 = 1/6 = 0.166$

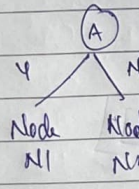
C1 2 ⇒ $P(C1) = 2/6$, $P(C2) = 4/6$
C2 4 Error $= 1 - \max(2/6, 4/6)$

$= 1 - 4/6 = 2/6 = 1/3 = 0.333$

DOMS | Page No. | Date / /

Gini

* Misclassification Error v/s Gini



	Parent
C1	7
C2	3

$\therefore$ Gini = 0.42

Gini N1 = 0 Gini N2 = 0.489

	N1	N2
C1	3	4
C2	0	3

Gini (Children) = (weighted) 0.342

$P(C1) = \frac{7}{10}$ $P(C2) = \frac{3}{10}$

$\therefore$ Error $= 1 - \left(\frac{7}{10}, \frac{3}{10}\right)$

$= 1 - \frac{7}{10} = \frac{3}{10} = 0.3$

Error at N1 $\Rightarrow$ $1 - \max\left(\frac{3}{3}, \frac{0}{3}\right)$

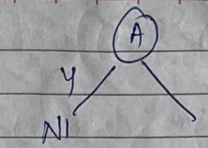
$= 1 - \frac{3}{3} \Rightarrow 0$

Error at N2 $\Rightarrow$ $1 - \max\left(\frac{4}{7}, \frac{3}{7}\right)$

$= 1 - \frac{4}{7} = \frac{3}{7} = 0.428$

$\therefore$ Overall error of children $= \frac{3}{10} \times 0 + \frac{7}{10} \times 0.42$

$= 0.3$



* Algorithm for Decision Tree Induction

1) Basic algo. (Greedy Algo.)

→ Tree is constructed in top-down recursive divide & conquer manner.

Complexity of algo : $O(n \times |D| \times \log |D|)$

where, n → no. of attributes describing tuples in D and |D| is no. of training tuples in D.

2) For the following dataset, generate the decision tree

Rid	age	Income	Student	Credit rating
1	Youth	high	no	fair
2	Youth	"	no	Exc.
3	M-Aged	high	no	fair
4	Senior	medium	no	"
5	"	low	Yes	"
6	"	low	Yes	Exc
7	M-Aged	low	Yes	"
8	Youth	medium	no	fair
9	"	low	Yes	"
10	Senior	medi	"	"
11	Youth	med.	"	Exc
12	Middle-Aged	medi	No	Exc
13	"	high	Yes	fair
14	Senior	medium	No	Exc

class: buys - computer

1	no
2	no
3	yes
4	"
5	"
6	no
7	yes
8	no
9	yes
10	"
11	"
12	"
13	"
14	no

Use entropy & info. gain

(a) Entropy of entire dataset

⇒ 9 - Yes Total = 14
5 - No

$\therefore$ Entropy of dataset $= -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

$= 0.9402$

(b) Info. gain for attribute income, age, student, credit rating

Income	Y	N
H	2	2
M	2	1
S	0	3
	4	6

Age	Y	N
Y	2	3
M	4	0
S	3	2

Entropy (Income) $= \frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)$

$+ \frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4}\right) + \frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$

$= 0.694$

Info (age) = entropy

$\therefore$ Gain (age) $= 0.9402 - 0.694 = 0.246$.

Entropy of Info (income)

Income	Y	N
H	2	2
M	4	2
L	3	1

Entropy = Info (student)

Student	Y	N
N	3	3
Y	6	2

Entropy Info (Credit rating) ⇒

Cred. rating	Y	N
fair	6	2
Exc.	3	3

Entropy (income) $= \frac{4}{14}\left(\frac{-2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)$

$+ \frac{6}{14}\left(\frac{-4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right)$

$+ \frac{4}{14}\left(\frac{-3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)$

$= 0.6792 + *$

$= 0.9110$

$\therefore$ Info gain $= 0.9402 - 0.9110 = 0.0292$

Entropy (student) $= \frac{6}{14}\left(-\frac{3}{6}\left(\log_2\left(\frac{3}{6}\right)\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right)$

$+ \frac{8}{14}\left(-\frac{6}{8}\left(\log_2\left(\frac{6}{8}\right)\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right)$

$\Rightarrow 0.4285 + 0.463$

$= 0.8921$

Info. gain $= 0.0481$

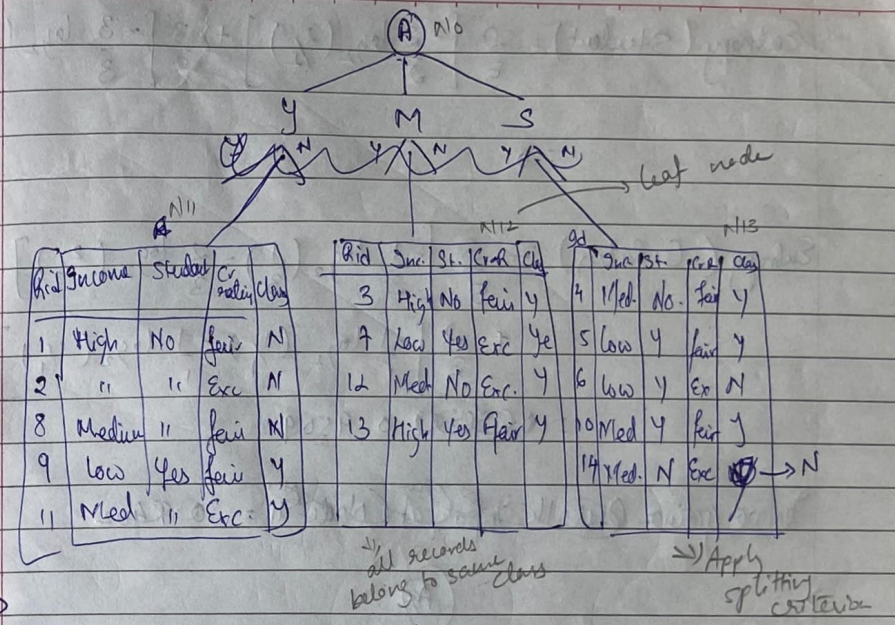
Entropy (credit rating) $= \frac{8}{14}\left(-\frac{6}{8}\left(\log_2\left(\frac{6}{8}\right)\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right)$

$+ \frac{6}{14}\left(-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right)$

$= 0.4635 + 0.4285$

$= 0.8920$

$\therefore$ Info gain $= 0.0482$.



$\Rightarrow$

N11 $\Rightarrow$ Info. gain on attributes Income, student, credit rating to determine which one to use

Income	Y	N
High	0	2
Med.	1	1
low	1	0

Student	Y	N
Y	2	0
N	0	3

Credit rating	Y	N
fair	1	2
Exc	1	1

Entropy (income) $\Rightarrow \frac{2}{5}\left(\frac{-2}{2}\log_2\left(\frac{2}{2}\right)\right) + \frac{1}{5}\left(\frac{-1}{1}\log_2\left(\frac{1}{1}\right)\right)$

$+ \frac{2}{5}\left(\frac{-1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$

$= 0.4$

Info. gain $=$

Entropy (Student) = $\frac{2}{5}\left[\frac{-2}{2}\log_2(2/2)\right] + \frac{3}{5}\left[\frac{-3}{3}\log_2(3/3)\right]$

$= 0$

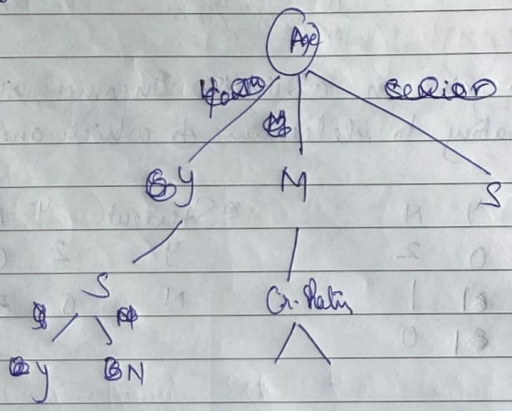
Entropy (C. Rating) = $\frac{3}{5}\left[\frac{-1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})\right]$

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

$= 0.9509.$

Info gain = ~~Gain(Node)~~ = 0.9402

⇒



Q13 → Entropy of income, student, credit rating.

Income	y	N
Med.	2	1
Low	1	1

Student	y	N
Y	2	1
N	1	0

Cr. Rating	y	N
Fair	3	0
Exc.	0	2

Entropy (Income) = $\frac{3}{5}\left(\frac{-2}{3}\log_2(2/3) - 1/3\log_2(1/3)\right)$

$+ \frac{2}{5}\left(\frac{-1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right)$

$= 0.9509.$

Entropy (Student) = $\frac{3}{5}\left(\frac{-2}{3}\log_2(2/3) - \frac{1}{3}\log_2(1/3)\right)$

$+ \frac{2}{5}\left(\frac{-1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right)$

$= 0.5509.$

Entropy (Cred. rating) = $\frac{3}{5}\left(\frac{-3}{3}\log_2(3/3)\right)$

$+ \frac{2}{5}\left(\frac{-2}{2}\log_2(2/2)\right)$

$= 0$

29/11/2021

Rule based classifier :

→ Classify records using a collection of "if .. then .." rules.

Rule : (Condition) → y
(antecedent → consequent)

(Condition) → conjunction of tests on attributes
y → class label

eg: (Blood Type = warm) ∧ (lay eggs = Yes) → Birds.
(Taxable income < 50k) ∧ (Refund = Yes) → Evade = No

Rule coverage & accuracy

⇒ only antecedent
antecedent + consequent

Q) Consider a training set that contains 60 positive examples & 100 negative examples. Suppose we have generated two rules where rule

Rule R1 : covers 50 positive examples & 5 negative examples
Rule R2 : covers 2 " " & 0 " "
Rule R3 :

Can you compute accuracy of these two rules.

⇒ Total = 160
= 100 + 60
↓ ↓
+ve −ve

∴ Accuracy :

Accuracy (r1) = 50/55 = 90.9 %.

" (r2) = 2/2 = 100 %

(Q) Likelihood ratio statistic

$$R = 2 \cdot \sum_{i=1}^{k} f_i \log_2 \left(\frac{f_i}{e_i} \right)$$

where
k → no. of classes
f_i → observed freq. of class i examples that are covered by the rule.
e_i → expected freq. of a rule that makes random prediction.

R has a chi-square distribution with k-1 degrees of freedom

Rule R1 covers 55 examples,
expected freq. of +ve class (e_+) = no. of examples Covered × total +ve eg / total eg.

= 55 × 60/160 = 20.625

Expected freq. of (-) class (e_-) = $55 \times \frac{100}{160}$

= 34.375

∴ Likelihood ratio for R1 i.e R(R1) = $2 \times \left[50 \times \log_2 \left(\frac{50}{20.625} \right) + 5 \times \log_2 \left(\frac{5}{34.375} \right) \right]$

= 99.93.

R(R2) :-

$e_+ = \frac{2}{2} \times \frac{60}{160} = 0.75$

$e_- = \frac{2}{2} \times \frac{100}{160} = 1.25$

∴ $R(R2) = 2 \times \left[2 \times \log_2 \left(\frac{2}{0.75} \right) + 0 \right]$

= 5.66

1/12/21 → Assignment 2

Q2) R1 : A → + (contains 4 +ve & 1 -ve examples)
R2 : B → + (" 30 +ve & 10 -ve ")
R3 : C → + (" 100 +ve & 90 -ve examples)

Total ⇒ 500 (100 +ve + 400 -ve)

Accuracy (R1) = $\frac{4}{5}$ = 80%.

Accuracy (R2) = $\frac{30}{40}$ = 75%.

Accuracy (R3) = $\frac{100}{190}$ = 52.631%.

Likelihood :-

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

R(R1) → $e_+ = 5 \times \frac{100}{500}$

= 1.

$e_- = 5 \times \frac{400}{500} = 4.$

∴ R(R1) = $2 \times \left[4 \times \log_2 \left(\frac{4}{1} \right) + 1 \times \log_2 \left(\frac{1}{4} \right) \right]$

⇒ 12

R(R2) → $e_+ = 40 \times 100/500 = 8$

$e_- = 40 \times \frac{400}{500} = 32$

DOMS | Page No.
Date / /

$$R(R2) = 2\left[30 \times \log_2\left(\frac{30}{8}\right) + 10 \times \log_2\left(\frac{10}{32}\right)\right]$$

$$= 80.8519.$$

$R(R3) =$

$e_+ = 190 \times 100/500 = 38$

$e_- = 190 \times 400/500 = 152.$

$$= 2\left[100 \times \log_2\left(\frac{100}{38}\right) + 90 \times \log_2\left(\frac{90}{152}\right)\right]$$

$$= 143.0923$$

offline

6/12/2021

(IV) Laplace measure

→ Evaluation metric that takes into account rule coverage.

→ Given by formula:

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

where n → no. of examples covered by the rule.

f_+ → no. of positive examples covered by the rule.

k → no. of classes.

(III) m-estimate measure

→ Given as

$$\text{m-estimate} = \frac{f_+ + k\,p_+}{n + k}$$; p_+ → prior probability of the class

Q] → Problem solved on 29th Nov.

(I) Laplace

f_+ (R1) → 50

$n = 55$

$k \to 2.$

$$\therefore \text{Laplace}(R1) = \frac{51}{57}\left(\frac{50+1}{55+2}\right) = 0.8947.$$

$f_+(R2) = 2 \Rightarrow \text{Laplace}(R2) = \frac{2+1}{2+2} = \frac{3}{4} = 0.75$

$n = 2$

$k \to 2$

$$R(R2) = 2\left[30 \times \log_2\left(\frac{30}{8}\right) + 10 \times \log_2\left(\frac{10}{32}\right)\right]$$

$$= 80.8519.$$

$R(R3) =$

$e_+ = 190 \times 100/500 = 38$

$e_- = 190 \times 400/500 = 152.$

$$= 2\left[100 \times \log_2\left(\frac{100}{38}\right) + 90 \times \log_2\left(\frac{90}{152}\right)\right]$$

$$= 143.0923$$

offline

6/12/2021

(II) Laplace measure

→ Evaluation metric that takes into account rule coverage.

→ Given by formula:

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

where n → no. of examples covered by the rule.

f_+ → no. of positive examples covered by the rule.

k → no. of classes.

(III) m-estimate measure

→ Given as

$$\text{m-estimate} = \frac{f_+ + k p_+}{n + k}$$; p_+ → prior probability of the class

Q] → Problem solved on 29th Nov.

(I) Laplace

f_+ (R1) → 50

$n = 55$

$k \to 2.$

$$\therefore \text{Laplace}(R1) = \frac{51}{57}\left(\frac{50+1}{55+2}\right) = 0.8947.$$

$f_+(R2) = 2 \Rightarrow \text{Laplace}(R2) = \frac{2+1}{2+2} = 3/4 = 0.75$

$n = 2$

$k \to 2$

(III) m-estimate

$p_+ = 0.2$

$$m\text{-est}(R_1) = \frac{50 + 2\times 0.2}{57} = 0.8842$$

$$m\text{-est}(R_2) = \frac{2 + 2\times 0.2}{4} = 0.6$$

First Order Inductive Learner.

(IV) FOIL's info. gain

→ Takes into account the support count of the rule.
→ Support count → No. of the examples covered by the rule.

Suppose rule r: A → + covers p_0 positive examples & no negative examples. After adding a new conjunct B, the extended rule r': A ∧ B → + covers p_1 positive egs. & n_1 negative examples.

∴ FOIL's info gain is given by

$$= p_1 \times \left(\log_2 \frac{p_1}{p_1+n_1} - \log_2 \frac{p_0}{p_0+n_0} \right)$$

Same prob:

$p_0 \to 60$
$n_0 \to 100$

$p_1 \to 50$ $\quad p_2 \to 2$
$n_1 \to 5$ $\quad n_2 \to 0$

$$\therefore \text{FOIL's}(R1) = 50\left(\log_2 \frac{50}{55} - \log_2 \frac{60}{160} \right)$$

$$= 63.8766.$$

$$\text{foils}(R2) = 2\left(\log_2 \frac{2}{2} - \log_2 \frac{60}{160} \right)$$

$$= 2.83007.$$

Q] Consider a training set that contains 100 +ve egs. & 400 negative examples.
For each of the foll. candidate rules.

R1 : A → + (covers 4 +ve & 1 -ve example)
R2 : B → + (" 30 +ve & 10 -ve ")
R3 : C → + (" 100 +ve & 90 -ve examples).

→ Determine which is the best & worst candidate rule according to (a) Rule Accuracy (b) Likelihood Ratio Statistic (c) Laplacian measure (d) M-est (e) FOIL's info. gain

Laplace:

$f_+(R1) \to 4$
$n = 5$
$k \to 2.$

$$\text{Laplace}(R1) = \frac{4+1}{5+2} = \frac{5}{7}$$

$$= 0.7142$$

Laplace $(R_2) = \frac{30+1}{40+2}$

$= \frac{31}{42}$

$= 0.7380$ ✓ B

Laplace $(R_3) = \frac{100+1}{190+2} = 0.5260$ W

---

m - estimate

$p_+ = 0.2$

m-est $(R_1) = \frac{4 + 2\times 0.2}{5+2}$

$= 0.6285$

m-est $(R_2) = \frac{30 + 2\times 0.2}{40+2}$

$= 0.7238$ ✓ B

m-est $(R_3) = \frac{100 + 2\times 0.2}{190+2}$

$= 0.5229$ W

FOIL's:

$p_0 = 100$; $p_1 = 4$; $p_2 = 30$; $p_3 = 100$

$n_0 = 400$; $n_1 = 1$; $n_2 = 10$; $n_3 = 90$

FOIL's $(R_1) = p_1 \left(\log_2 \frac{p_1}{n_1} - \log_2 \frac{100}{500} \right)$

$= 4 \left(\log_2 \frac{4}{5} - \log_2 \frac{100}{500} \right)$

$= 8$ W

FOIL's $(R_2) = 30 \left(\log_2 \frac{30}{40} - \log_2 \frac{100}{500} \right)$

$= 57.42067$

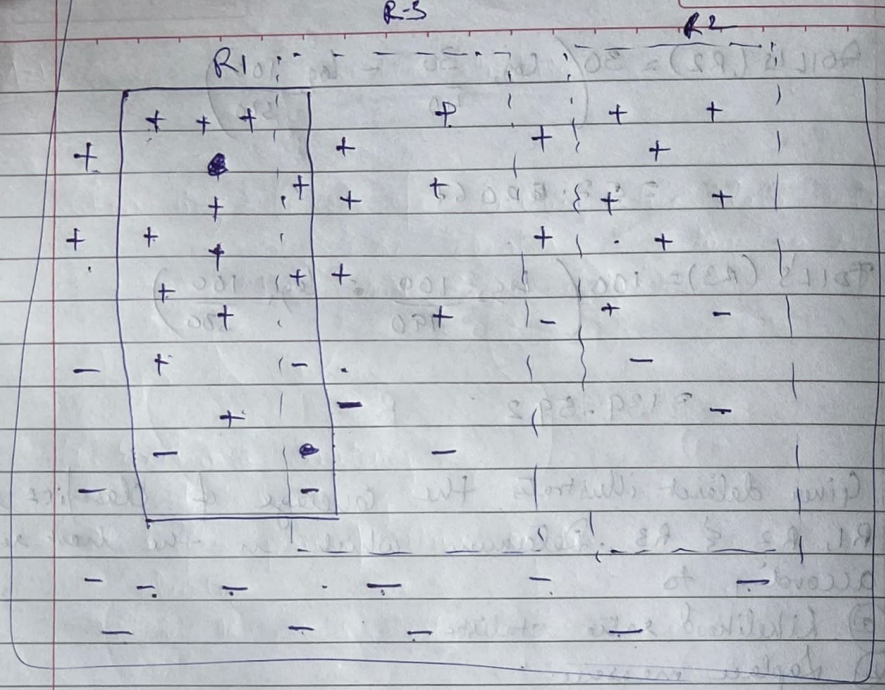
FOIL's $(R_3) = 100 \left(\log_2 \frac{100}{190} - \log_2 \frac{100}{500} \right)$

$= 139.592$ B

Q] Given dataset illustrate the coverage of classification rules R1, R2 & R3. Determine which is the best rule according to

(a) Likelihood ratio statistic
(b) Laplace measure.
(c) M-estimate
(d) Rule accuracy after R1 has been discovered where none of the examples covered by R1 are discarded.
(e) Rule " " R1 " " " " " only true examples covered by R1 are discarded.
(f) R " " " " " " " " both true & -ve examples ~~are~~ covered by R1 are discarded

P.T.O



Dataset

no. of +ve = 29

" " -ve = 21

R1

no. of positive = 12

no of -ve = 3

R2

+ve = 7

-ve = 3

R3 when none of eg. of R1 are discarded.

+ve = 8

-ve = 4

R3 when +ve eqs. are discarded

+ve = 6

-ve = 4

R3 when both +ve & -ve eqs. are discarded.

+ve = 6

-ve = 2

→ Compute measures.

8/12/21

K-means clustering.

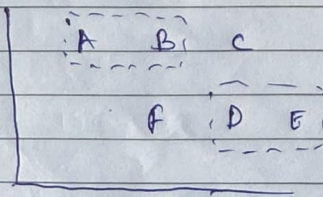
Clustering : Process of dividing the data into groups or clusters depending on relationships that exist in a data.

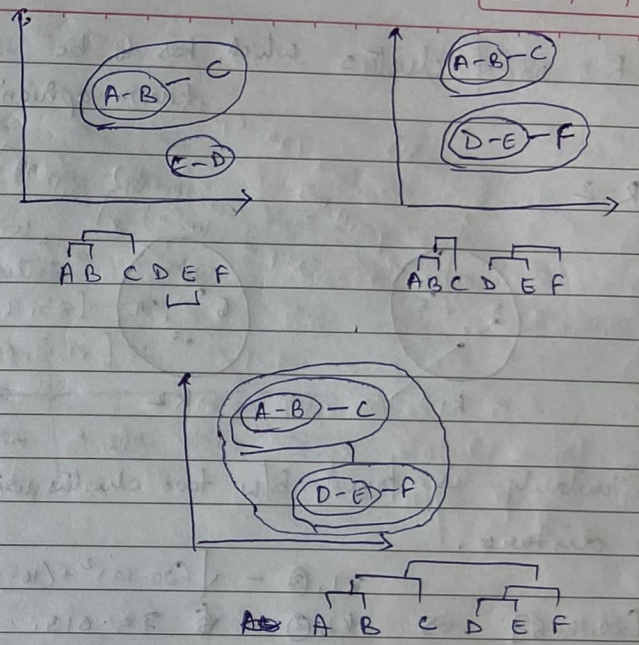
Types of clustering :

(1) Exclusive clustering
→ Hard clustering
→ Data points / items exclusively belong to one cluster.
→ Eg : K-means clustering.

(2) Overlapping clustering
→ Soft clustering
→ Data point / item can be belong to multiple clusters.
→ Eg : Fuzzy / C-Means clustering.

(3) Hierarchical clustering.
→





* K-Means clustering :

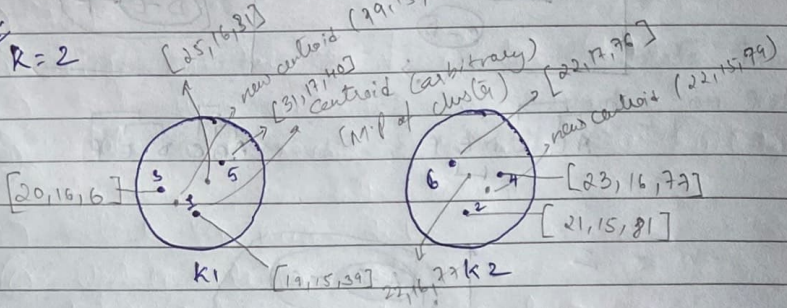
→ Iterative algo.
→ Partitions dataset into k-clusters

Mall Customer Dataset

Cluster	Customer ID	Age	Income k$	Spending score (1-100)
K1	1	19	15	39
K2	2	21	15	81
K1	3	20	16	6
K2	4	23	16	77
K1	5	31	17	40
K2	6	22	17	76.

K = no. of clusters which has to be known apriori

Let K = 2



Similarly is found b/w two clusters using Euclidean distance.

K1, ③ - $\sqrt{(20-19)^2+(16-15)^2+(6-39)^2} = 33.03$

[20,16,6] ⇒ K2, ③ - 75.013.

Place where dist. is min

K1	19	15	39
K2	21	15	81

⇒ Update centroid : Take average of $\frac{① + ③}{2}$ = 19.5, 15.2, 22.5

New :

K1	19	15	22
K2	21	15	81

[23, 16, 77]

K1, ⑤ - 55.154

K2, ⑤ - 4.582

∴ New centroid :

K1	19	15	22
K2	22	15	79.

[31, 17, 40]

K1, ⑤ → 21.7255

K2, ⑤ → 40.074

∴ New centroid

K1	25	16	31
K2	22	15	79.

[22, 17, 76]

K1, ⑥ → 45.11

K2, ⑥ → 3.605

K1	25	16	31
K2	22	16	77.

Algorithm :

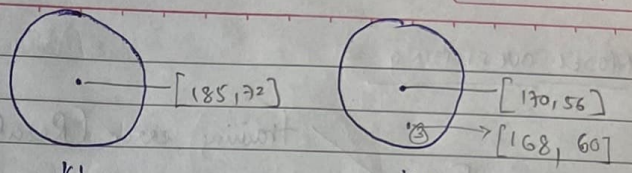
1. Start
2. Decide on the no. of cluster (k)
3. Find the centroids
4. Take each record & calculate the Euclidean distance from all centroids
5. Select the min. dist.
6. Update cluster centroid
7. ~~Upda~~ Update record with clust no.
8. Stop after executing all record.

Q]

	Height	Weight	Cluster
1	185	72	K1
2	170	56	K2
3	168	60	
4	179	68	
5	182	72	
6	188	77	
7	180	71	
8	180	70	
9	183	84	
10	180	88	
11	180	67	
12	177	70	

k=2.

Use K means



5. K1, 3 → 20.808
   K2, 3 → 4.472

online

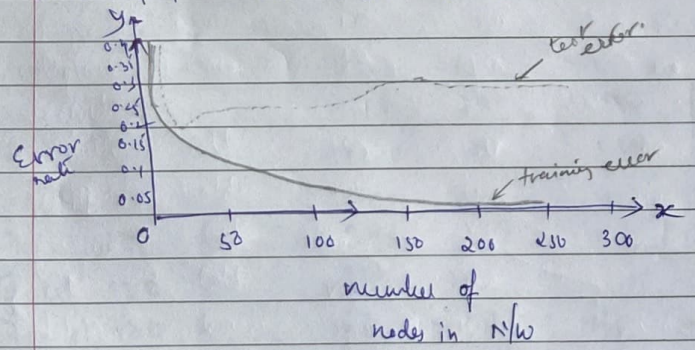
09/12/21 * MODEL OVERFITTING

Error < training error (Resubstitution error / apparent error)
Generalization errors.

Training errors: no. of misclassific^n errors committed on training records. (training records)
Generaliz^n error: expected error of model on previously unseen records (test data set)

Good classific^n model should have low training error as well as low generalization error.
Model that fits the data too well is known known as model overfitting.

Training & test error rates.



Model underfitting
Smaller the tree size, training & test errors are large.
This situation is called model underfitting.

When the tree becomes too large, its test error rate begins to increase even though the training error rate continues to decrease. This phenomenon is known as model overfitting.

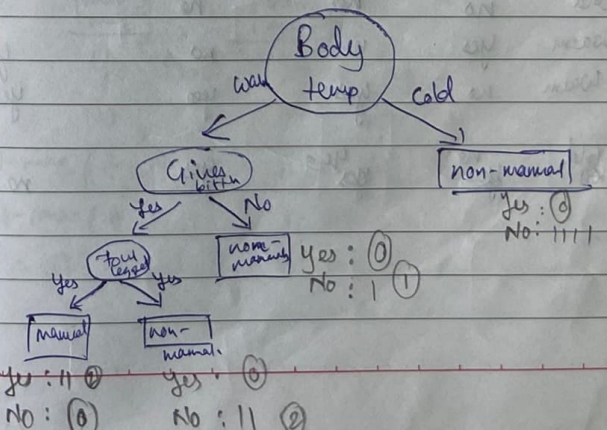
Reasons for overfitting

(1) Overfitting due to the presence of noise.

Eg: Manual classific^n problem.

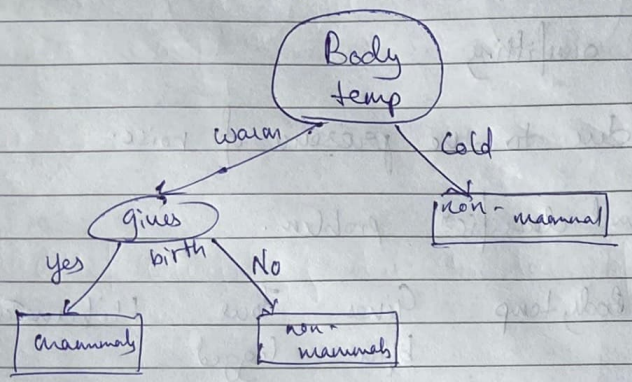
Name	Body temp	Gives birth	Four legged	Hibernates	Class label.
porcupine	warm	Yes	Yes	Yes	Yes
cat	"	Yes	Yes	No	Yes.
bat	"	Yes	No	Yes	No *
whale	"	Yes	No	No	No. *
salamander	cold	no	yes	yes	No.
Komodo dragon	cold	no	yes	no	no.
python	"	no	no	yes	no.
Salmon	"	no	no	no	no.
eagle	warm	no	no	no	no.
guppy	cold	yes	no	no	no

Decision tree & model m/c:



model

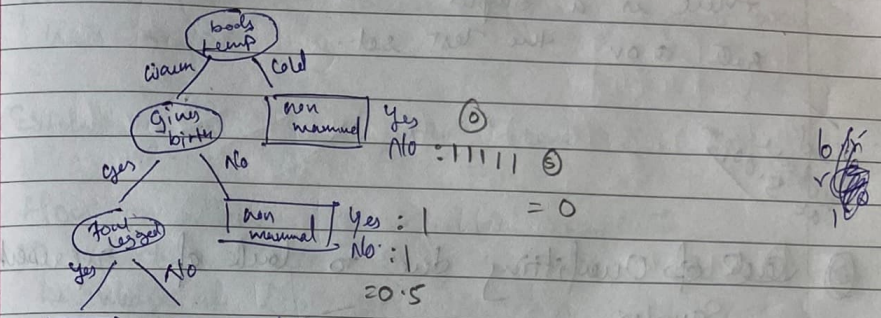
Decision tree of m2 :



Eg :- Test set for classifying mammals.

name	body temp	gives birth	four legged	Hibernats	Class Label
Human	warm	yes	no	no	yes
pigeon	warm	no	no	no	no
elephant	warm	yes	yes	no	yes
leopard shark	cold	yes	no	no	no
turtle	cold	no	yes	no	no
penguin	cold	no	no	no	no
eel	cold	no	no	no	no
dolphin	warm	yes	no	no	yes
spiny anteater	warm	no	yes	yes	yes
gila monster	cold	no	yes	yes	no

Model M1



Yes ⓪
No : ||||| ⑤
= 0

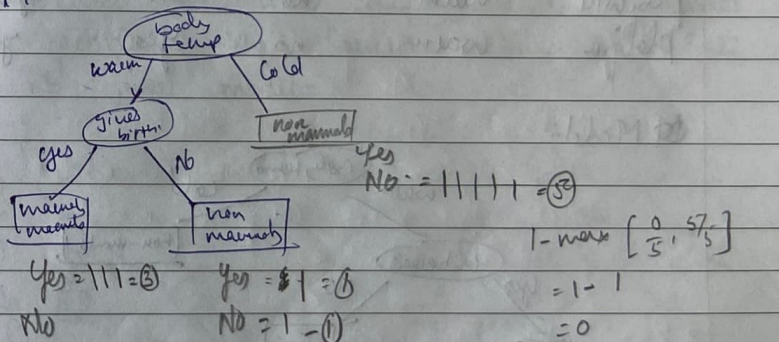
Yes : 1
No : 1
= 0.5

mamm
Yes : 1
No :
1 - max (1/1, 0/1)
= 0

non mamm
Yes : || ②
No : 1
1 - max (2/2, 0/2)
= 0

Test error rate of all leaf nodes
= 1/10 [0] + 2/10 [0] + 2/10 [0.5] + 5/10 (0)
= 0.1

Model M2



Yes ||| ③
No

1 - max (3/3, 0/3)
1 - max (1
= 0

Yes = $ | ①
No = 1 - ①

1 - max (1/2, 1/2)
1 - 0.5
= 0.5

Yes ||||| = ⑤
No

1 - max [0/5, 5/5]
= 1 - 1
= 0

∴ Test error rate of all leaf nodes
= 1/10 [0] + 2/10 [0.5] + 5/10 [0]
= 0.1 = 10%

Training error rate = 20%
Test error rate = 10%

M1 has overfitted the training data because there is a simpler model with lower error rate on the test set.

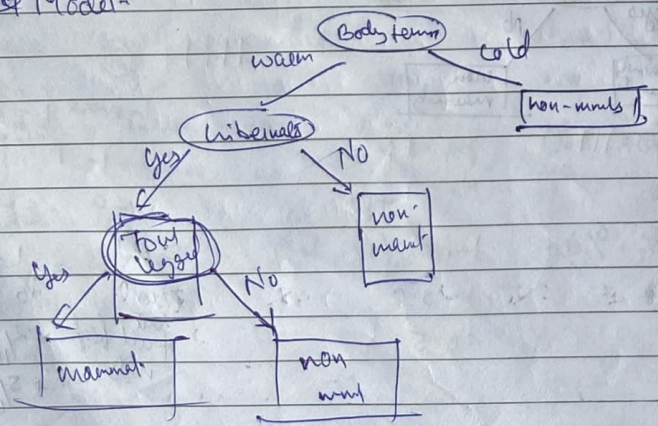
Next lecture at 3.00pm

② Lack of Overfitting due to lack of representative Samples

⇒ Eg:-

name	body temp	gives birth	four legged	Hibern	Class label
Salamander	cold	no	yes	yes	no
guppy	cold	yes	no	no	no
eagle	warm	no	no	no	no
poorwill	warm	no	no	yes	no
platypus	warm	no	yes	yes	yes

Model:-



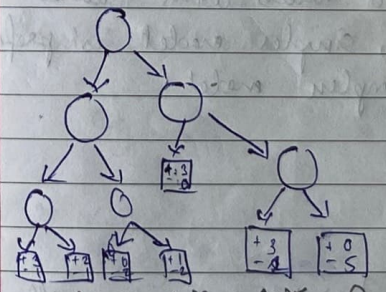
Error rate on test set, error rate is 30%. from test data set, humans, elephants, dolphins were misclassified

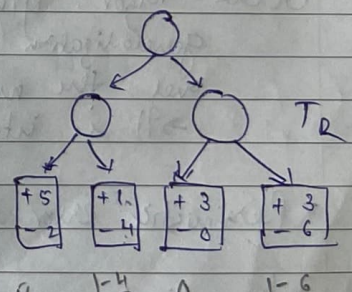
* Estimation of generalization errors.

⇒ How to determine right model complexity.
→ Model has access only to training dataset, & no knowledge of test set.
→ Hence, the need to estimate generalization error of induced tree

① Using re-substitution estimates / Optimistic approach.

→ Assumes that training set is a good rep. of overall data. & hence training error or resubstitution error can provide an optimistic estimate of the generalization error.
→ Decision tree algo simply selects the model that produces the lowest training error rate as is final model.





$\Rightarrow \frac{7}{24}(0.285) + \frac{5}{24}(0.2) + \frac{9}{24}(0.33)$

$= 0.25$

Training error rate for T_L.

Error rate $\left[\frac{3}{4}, \frac{1}{4}\right]$.

$$\frac{4}{24}(0.25) + \frac{3}{24}(0.33) + \frac{3}{24}(0.33) + \frac{3}{24}(0.25) + \frac{4}{24}(0.25)$$

$\Rightarrow 0.1666.$

* lower rate results error better it is

$T_R = 0.25$

$\therefore T_L$ is left.

* Incorporate model complexity

→ Chance for model overfitting increases as model becomes more complex

→ Prefer simpler models

↓

→ Occam's Razor or principle of parsimony.

Occam's Razor: Given two models with same generalization error, simpler model is preferred over the more complex model.

→ It is intuitive

* Pessimistic error estimate

Generalization error = training error + penalty term for model complexity.

→ Let $n(t)$ be no. of training records classified by node t & $e(t)$ be no. of misclassified records

∴ Pessimistic error estimate of a decision tree T, $E_g(T)$ will be given as:

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

$$= \frac{\sum_{i=1}^{k} e(t_i) + \Omega(T)}{N_t}$$

where $k \rightarrow$ no. of leaf nodes

$e(T) \rightarrow$ overall training error of the decision tree

$N_t \rightarrow$ no. of training records.

$\Omega(t_i) \rightarrow$ penalty term associated with each node t_i

eg → for the binary decision trees, T_L & T_R if penalty term = 0.5

Pessimistic error estimate for T_L

$$e_g(T_L) = 0.167 + \frac{7 \times 0.5}{24} = 0.3128$$

" " " T_R

$$e_g(T_R) = 0.25 + \frac{4 \times 0.5}{24} = 0.3333.$$

[Lower the pessimistic error, better it is]

② Min. description length principle or MDL.

→ Overall cost of decision tree would be cost of encoding the model + cost of encoding misclassified records.

→ cost (model, data) = cost (model) + cost (data | model)

According to MDL, we should seek a model that minimizes the overall cost func^n.

* How overfitting is handled in decision tree induction.

① Prepruning (Early stopping rule)

→ Tree-growing algo. is halted before generating a fully grown tree that perfectly fits an entire training data.

Eg: Stop expanding leaf nodes when gain in impurity measure falls below certain threshold.

→ Advantage : Avoids generating overly complex sub-tree that overfit the training data.

→ Disadvantage : Choosing the right threshold value

② Post pruning

→ The decision tree is initially grown to its maximum size

→ Tree-pruning step is applied to the fully grown tree in a bottom up fashion.

⇒ Pruning is done by replacing a subtree with

① A new leaf node whose class label is determined from majority class of records affiliated with subtree.

② Most freq. used branch of the subtree

10/12/2021

* Rule based classifier

→ Direct method : RIPPER algo

2 class prob : choose one as the pos and other as -ve

→ Learn rules for positive class

→ Negative class will be default class.

Eg: ~~Classes are~~ :

~~y_1 & y_2~~

Multiclass prob :

→ Order classes acc. to increasing class prevalence
→ $y_1, y_2, \ldots y_c$; y_1 is the least freq. class

→ Learn rule set for smallest class first, treat rest as negative

→ Repeat with next smallest class as the class

→ Growing a rule:

→ Start from empty rule.

→ Add conjuncts as long as they improve FOIL's info. gain

→ Stop when rule no longer covers negative egs.

→ Prune the rule immediately using incremental reduced error pruning.

→ Measure of pruning ⇒ $v = \frac{p-n}{p+n}$.

P: no. of +ve egs. covered by rule in validation set.

n: no. of -ve egs " " " "

→ Pruning method: delete any final sequence of conditions that maximize v.

→ End till advantages

Nearest-neighbour classifier:

Eager learners	Lazy learner.
→ Generalized model from training data set is constructed → Using the model, the class of test dataset is predicted Eg: Decision trees, rule based classifiers	→ Training dataset is stored → On querying, similarity b/w test data & training set records is calculated to predict the class of test data. Eg: K nearest neighbours

Algorithm:

Step 1: Find dist. from test example to each of training records.

Step 2: Find rank

Step 3: Find nearest neighbour

Step 4: Assign class of nearest neighbour to the test record.

Eg:

Sepal length	Sepal width	Species	dist
5.3	3.7	Setosa	0.608
5.1	3.8	Setosa	0.704
7.2	3.0	virginica.	2.002
5.4	3.4	Setosa	0.36
5.1	3.3	Setosa	0.22
5.4	3.9	"	0.82
7.4	2.8	Virginica	2.22
6.1	2.8	versicolor	0.94
7.3	2.9	virginica.	2.1
6.0	2.7	versicolor	0.89
5.8	2.8	virginica.	0.67
6.3	2.3	versicolor	1.36
5.1	2.5	"	0.60
6.3	2.5	"	1.25
5.5	2.4	"	0.75

Query

Sepallength sepal width species.

5.2 3.1 ?

k = 1 , 2 ... 5

Step 1 : Distance from test eg. to each of train eg

⇒ Euclidean distance. : $\sqrt{(x-a)^2+(y-b)^2}$

∴ Dist (sepal length, sepal width) = $\sqrt{(5.2-5.3)^2+(3.1-3.9)^2}$

= 0.608

→ Calculate for all records

k = 1, test eg. would be assigned the species setosa.

k = 2, " " " " " " " "

k = 3, " " " " " " " "

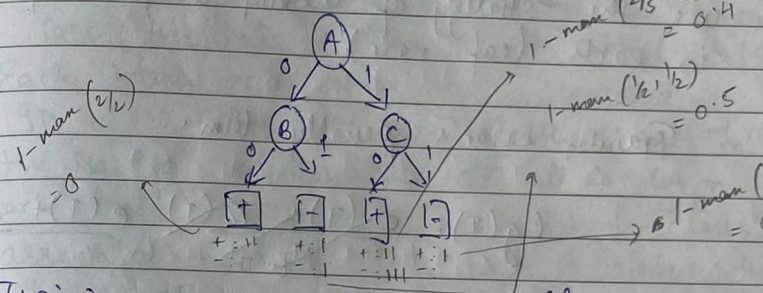
⋮

k = 5, majority voting scheme is used & test eg. is assigned to the class setosa.

16/12/21

(1) Consider decision tree given below,

(a) Compute generalization error rate of tree using using optimistic approach.



Training Instance	A	B	C	Class
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	−
5	1	0	0	+
6	1	0	0	+
7	1	1	0	−
8	1	0	1	+
9	1	1	0	−
10	1	1	0	−

Formula:

$$\text{Classification error}(t) = 1 - \max_i [P(i/t)]$$

→ $\frac{2}{10}(0) + \frac{2}{10}(0.5) + \frac{5}{10}(0.4) + \frac{1}{10}(0)$

⇒ 0.3 or 30%.

(b) Compute the generalization error rate of the tree using pessimistic approach (for simplicity, use the strategy of adding a factor of 0.5 to each leaf node).

Training error + penalty term.

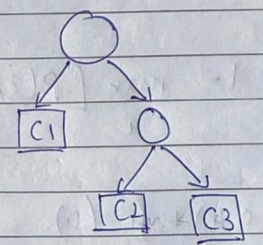
$$e_g(T) = \frac{e(T) + \Omega(T)}{N_t} \quad \frac{e(T) + \Omega(T)}{N_t}$$

$$e_g(T) = \frac{0.3 + 4 \times 0.5}{10}$$

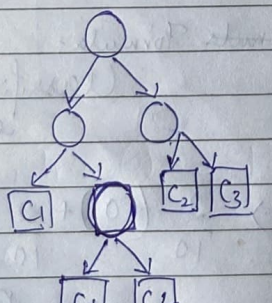
$$= 0.5$$

6] Consider the decision shown below, assume they are generated from a dataset that contains 16 binary attributes & 3 classes.

16 binary attributes & $C_1, C_2 \& C_3$.



(a) Decision tree with 7 errors



(b) Decision tree with 4 errors.

(a) Compute the total description length of each decision tree according to min. description length principle. The total description length of a tree is given by

Cost (tree, data) = Cost (tree) + Cost (data/tree)

Each internal node of the tree is encoded by the ID of splitting attribute. If there are 'm' attributes, the cost of encoding each attribute is $\log_2 m$ bits

Each leaf is encoded using Id of class it is associated with. If there are k classes, the cost of encoding a class is $\log_2 k$ bits

- Cost (tree) is the cost of encoding all the nodes in the tree. To simplify, the computation you can assume that the total cost of the tree is obtained by adding up the cost of encoding each internal node & each leaf node
- Cost (data/tree) is encoded using classification errors the tree commits on the training set. Each error is encoded by $\log_2 n$ bits where n is the total no. of training instances

Which decision tree is better according to the MDL principle.

→ Cost (tree, data) = Cost (tree) + Cost (data/tree)

m attributes → $\log_2 m$ bits

16 attributes → $\log_2 16 = 4$ bits

k classes → $\lceil \log_2 k \rceil$ bits

3 classes → $\lceil \log_2 3 \rceil = 2$ bits

The cost of each misclassification error is $\log_2 n$ bits.

Internal nodes = O (circles)
leaf nodes = □

∴ Overall cost of decision tree (a) is

$2 \times 4 + 3 \times 2 + 7 \times \log_2 n$

$14 + 7 \log_2 n.$

∴ Overall cost of decision tree (b) is

$4 \times 4 + 5 \times 2 + 4 \times \log_2 n.$

$\Rightarrow 16 + 10 + 4 \times \log_2 n.$

$\Rightarrow 26 + 4 \log_2 n.$

n = 1

(a) $14 + 7 \log_2(1) = 14$

(b) $26 + 4 \log_2(1) = 26$

n = 2

(a) $14 + 7 \log_2(2) = 21$

(b) $26 + 4 \log_2(2) = 30$

→ According to MDL principle, tree (a) is better than (b) if $n < 16$ & worse than (b) if $n > 16$

a] Implement k-nearest neighbour classifier on dataset

Height (cm)	Weight (kg)	class
167	51	underweight
182	62	normal
176	69	"
173	64	"
172	65	"
174	56	underweight
169	55	normal
173	57	"
170	55	"
Test set: 170	57	?

k = 5

Predict class of above record.

15/12/21

Unit 4: Association analysis

→ Helps find hidden relationship among data.

eg: Customer Purchase Data

Market Basket Data:

TID	Items
1	{Bread, milk}
2	{Bread, diapers, beer, eggs}
3	{Milk, diapers, beer, cola}
4	{Bread, milk, diapers, beer}
5	{Bread, milk, diapers, cola}

→ transaction

* Association rules:

Eg: {Diapers} → {Beer}.

Let $I = \{i_1, i_2, i_3 \ldots\ldots i_d\}$ be the set of all items in market basket data.

I = {bread, milk, beer, diapers, cola, eggs}.

→ A collection of zero or more items in a transaction is called an itemset.

If there are k-items in item set, it is called k-item set.

eg: {milk, bread, eggs} → 3 item set

{beer, eggs} → 2 item set

Support count: This refers to a no. of transaction that contain a particular item set.

Mathematically, represented using:

$$\sigma(X) = \left| \{ t_i \mid X \subseteq t_i, t_i \in T \} \right|$$

Support count {Beer, diapers, milk} = 2.

Association Rule: Implication expression of the form $X \to Y$ where X & Y are disjoint itemsets i.e $(X \cap Y = \phi)$

Strength of association rules:

Measured using two metrics: support & confidence.

Support: Determine how often a rule is applicable to a given dataset.

$$\text{Support} : s(X \to Y) = \frac{\sigma(X \cup Y)}{N}$$

Confidence: Determines how frequently items in Y appear in transactions that contain X.

$$\text{Confidence } c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

eg: Consider the rule:

{milk, diapers} → {Beer}.

s = 2/5

c = 2/3

* Formulation of Association rule mining problem:

→ Given a set of transactions T, find all the rules having support ≥ minimum support and confidence ≥ min. confidence where minimum support & minimum confidence are corresponding support & confidence thresholds.

Algorithm:

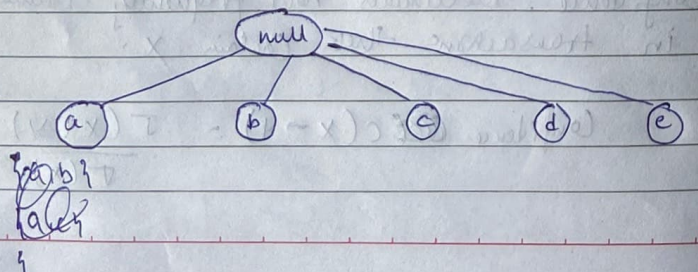
① Frequent item set generation:

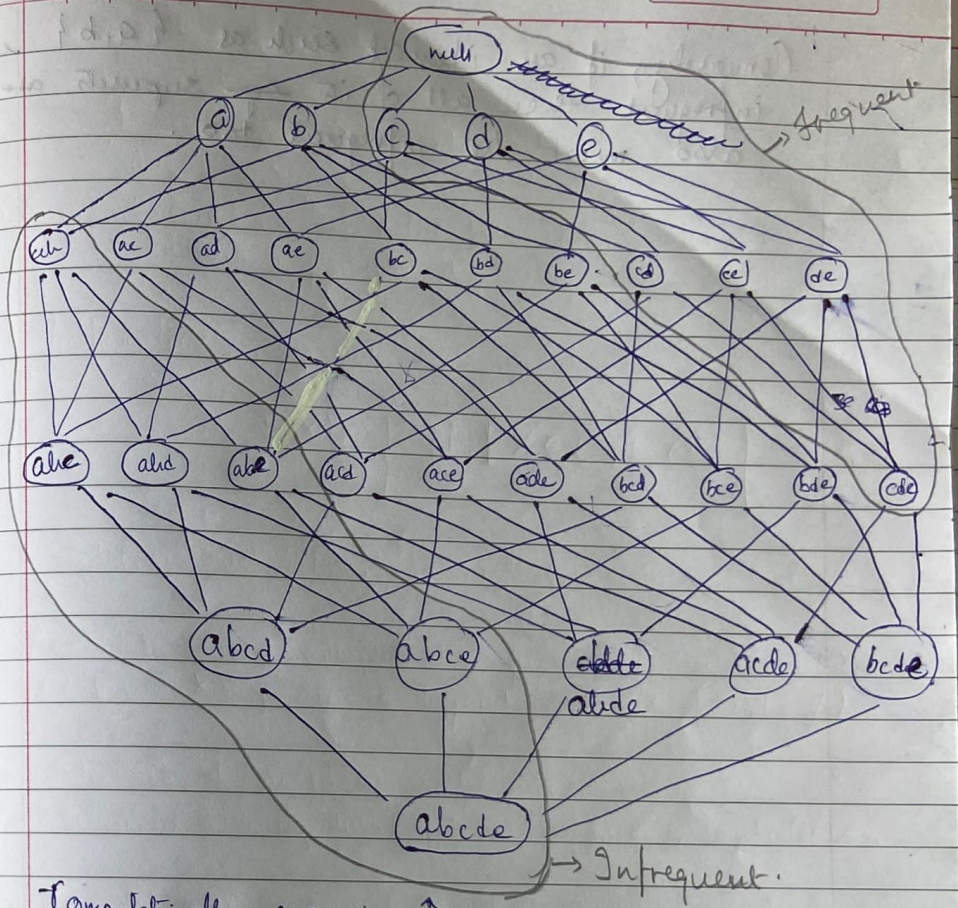
② Rule generation → All high confidence rules are extracted from frequent item set.

Lattice

eg: I = {a, b, c, d, e}.

Lattice structure:





Computationally expensive ↑

① Reduce the number of candidate itemsets

② Reduce " " " comparisons.

* Apriori principle

→ If an item set is frequent then all of its subsets must also be frequent.

→ Eg: {c, d, e} ⇒ {c, d}, {d, e}, {c, e}, {c}, {d}, {e}.

DOMS | Page No.
Date / /

Conversely, if an itemset such as {a, b} is infrequent then all of its ~~sup~~ supersets are also must be infrequent too.