2018 – dm – endsem

Sol(a) Major issues in Data Mining :-

(1.) Mining Methodology :-
- → Mining various and new kinds of knowledge
- → Mining knowledge in multi-dimensional space
- → Handling noise, uncertainty and incompleteness of data.
- → Pattern evaluation and pattern-or-constrain-guided mining

(2.) User Interaction :-
- → Interactive mining
- → Incorporation of background knowledge
- → Presentation and visualization of data mining results.

(3.) Efficiency and Scalability :-
- → Efficiency & scalability of data mining algorithms
- → Parallel, distributed, stream, and incremental mining methods.

(4.) Diversity of data types :-
- → Handling complex types of data
- → Mining dynamic, networked, and global data repositories.

(5.) Data mining and society :-
- → Social impacts of data mining
- → Privacy-preserving data mining
- → Invisible data mining

Sol (b) There are several major challenges associated with mining a huge amount of data as compared to mining a small amount of data :-

→ **Storage** – A huge amount of data requires a lot of storage space, which can be expensive and time-consuming to acquire and maintain.

→ **Processing** – A huge amount of data can take a long time to process, especially if data needs to be cleaned, transformed or analyzed in some way.

→ **Quality** – A huge amount of data is often more prone to errors, inconsistencies and missing values, which can impact the quality of the results obtained from mining the data.

→ **Interpretation** – It can be more challenging to interpret the results of mining a huge amount of data, as there may be many more variables to consider.

→ **Privacy** – A huge amount of data may contain sensitive or personal info. which requires careful handling and protection to ensure compliance with privacy regulations and ethical considerations.

sol 1(c) There are several factors that may need to be considered when performing data mining on web data vs text data. Some of these factors may include :–

(1.) **Data Source** – The source of data can affect the quality and relevance of the data for a specified task. For ex:– web data may come from social media, online news articles, e-comm websites, while text data may come from documents, books, or other written sources.

Aniket
Date
Page

(2) Data format — Web data is often unstructured and may be in the form of HTML or XML documents, while text data is typically structured and may be in the form of a text file or a document. This can affect the pre-processing steps required to extract and clean the data.

(3) Data Volume — Web data can be very large & may need to be sampled or aggregated to make it more manageable for data mining. Text data is often small in volume and may not require as much preprocessing.

(4) Data Quality — Web data can be noisy and may contain errors, duplicates, or irrelevant information. Text data is often more reliable, but may still contain errors or be incomplete.

(5) Data Privacy — Web data may contain personal information that needs to be protected in accordance with privacy laws & regulations. Text data may also contain sensitive info., but it is generally less of a concern compared to web data.

(6) Data relevance — The relevance of data for a specified task is an important consideration when performing data mining. Web data may be more relevant for tasks related to online trends and consumer behaviour, while ~~text~~ text data may be more relevant for tasks related to NLP and text Analysis.

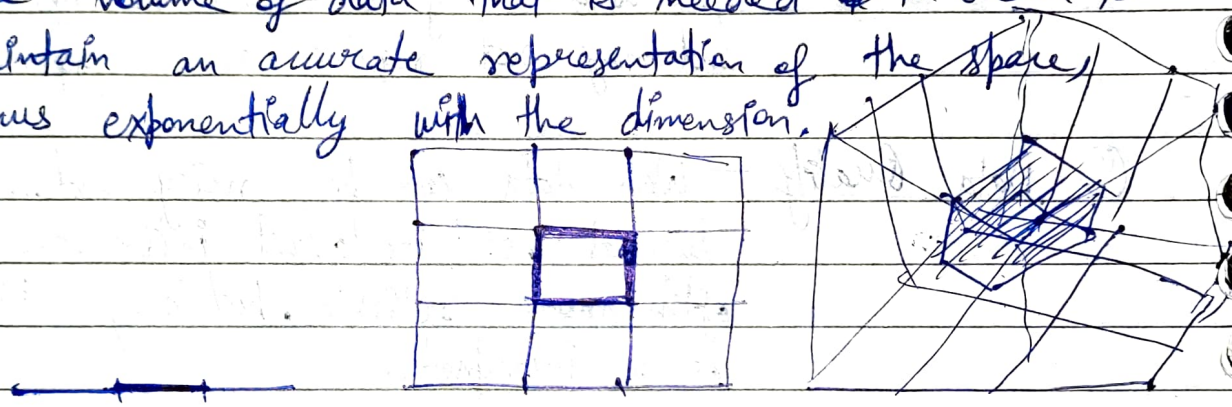
extra:- Types of Data Sets - { Record, Graph & network, ordered, Spatial, image, multimedia }

Aniket
Date
Page

sol 2(a) Curse of Dimensionality :-

→ means that the error increases with the increase in the number of features. It refers to the fact that the algorithms are harder to design in high dimensions and often have a running time exponential in the dimensions.

As the data space moves from 1D to 2D to 3D, the given data fills less and less of the data space. The volume of data that is needed in order to maintain an accurate representation of the space grows exponentially with the dimension.



diff^n :- b/w the space acquired against total space
(1/3, 1/9, 1/27)

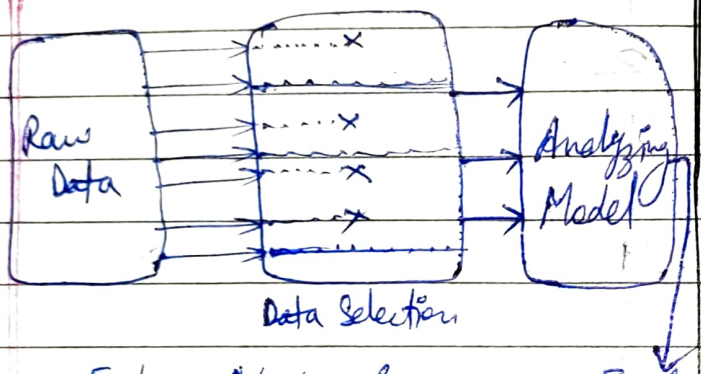
Dimentionality Reduction Techniques :-

→ Dimensionality Reduction is done based by either feature selection or feature extraction.

Feature Selection Techniques

→ It is based on omitting those features from the available measurements which do not contribute to class separability. In other words, redundant and irrelevant features are ignored.

→



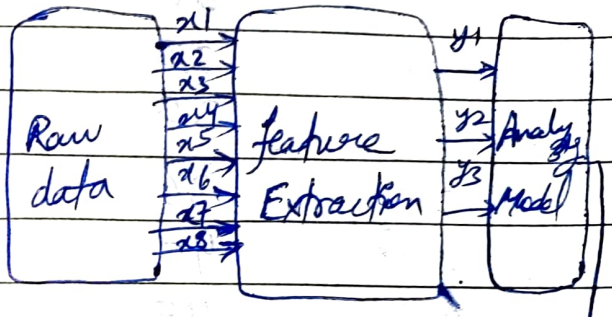
Feature Selection Process → Final Output

→ As a stand-alone task, feature selection can be unsupervised (eg. Variance Thresholds) or supervised (eg. Genetic Algorithms). You can also combine multiple methods if needed.

→
(1.) Variance Thresholds
(2.) Correlation Thresholds
(3.) Genetic Algorithms
(4.) Stepwise Regression

Feature Extraction Techniques

It considers the whole info. content and maps the useful info. content into a lower dimensional feature space.



Feature Extraction Process → Final output

Feature extraction is for creating a new, smaller set of features that still captures most of the useful information. This can come as supervised (eg. LDA) and unsupervised (eg. PCA) methods.

(1.) Linear Discriminant Analysis (LDA)
(2.) Principal Component Analysis (PCA)
(3.) t-distributed Stochastic Neighbor Embedding (t-SNE)
(4.) Autoencoders

Aniket
Date
Page

Sol 2(b)

age value	freq.
13	1
15	1
16	2
19	1
20	2
21	1
22	2
25	4
30	1
33	2
35	4
36	1
40	1
45	1
46	1
52	1
70	1
total freq	27

(a.) $\text{mean} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{809}{27} = 29.96$

$\text{median} = \left(\frac{n+1}{2}\right)^{th}\text{term} = \left(\frac{28}{2}\right)^{th}\text{term} = 14^{th}\text{term} = 25$

(n == odd) as 27 is odd

(b.) (25, 35) occurs 4 times. These are its modes.

⇒ data set is bi-modal with 25 & 35 are its modes.

(c) mid-range = average of largest & smallest of data

$$= \frac{70+13}{2}.$$

(d.) Q_1 (1^{st} quartile) = 25^{th} percentile

$$= \left(\frac{25}{100} * 27\right)^{th} = (6.75)^{th} \simeq 7^{th} \text{ term}$$

$$= 20$$

Q_3 (3^{rd} quartile) = 75^{th} percentile

$$= \left(\frac{75}{100} * 27\right)^{th} = (20.25)^{th} \text{ term.}$$

$$\simeq 20^{th} \text{ term}$$

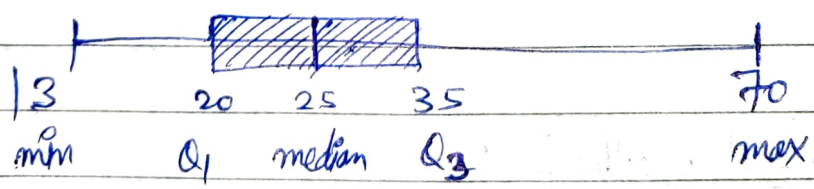
$$= 35$$

(e.) Five number summary of the data :-

min = 13 (Q_0), Q_1 = 20, median = 25 (Q_2), Q_3 = 35, max = 70 (Q_4)

~~(f) Boxplot of the~~

(f) Box-Plot of the data :-



Sol 2(c) Outlier detection is the process of identifying observations in a dataset that are significantly different from the majority of the data. Outliers can be caused by errors in data collection or processing, or they may represent legitimate but unusual observations. Identifying outliers is important in many applications because they can distort statistical analyses & influence the results of machine learning models.

There are several methods that can be used to detect outliers in a dataset :-

→ Statistical methods :-

These methods use statistical tests to identify observations that are significantly different from the majority of the data. Examples include the Z-score method & the Turkey method.

→ Distance-based methods :-

These methods identify observations that are far from the majority of the data using a distance measure such as Euclidean distance. Examples :- DBSCAN, etc.

→ Density-Based methods :-

These methods identify observations that are in low-density regions of data. Examples :- Isolation Forest algorithm and One-class Support Vector Machine (SVM).

→ Cluster-based methods :-

These methods identify observations that do not belong to any of the clusters formed by a clustering algorithm

Aniket
Date
Page

Outlier detection applications :-

1. Fraud detection — Outliers can indicate unusual or suspicious activity.
2. Quality control — Outliers can detect defects or errors in data collection process.
3. Medical diagnosis — Outliers can indicate unusual conditions to diagnose diseases.
4. Environmental monitoring — (similar)

Sol 3(a.)

Tid	Items bought
1	A, B, D
2	E, F, A
3	A, D, C
4	E, F, C
5	B, C, D
6	A, D, E, F
7	E, F, D
8	B, D
9	A, D, E, F
10	B, C

A – milk
B – beer
C – Cookies
D – Diaper
E – Bread
F – Butter

(a) Total no. of rules = max^m no. of association rules
(that can be extracted from this data (including rules having 0 support.)

$= 3^m - 2^m + 1$, $m = 6$ items
$= 3^6 - 2^6 + 1$
$= 602$

(b) max^m size of frequent items that can be extracted (assuming minsup > 0)
(largest transaction contain 4 items)
$= 4$
(corresponds to Tid 6 & 9)

Aniket
Date
Page

~~(extra question) max^m size of frequent items~~

(extra question) max^m no. of size-3 itemsets that can be derived from this data set

(no. of distinct 3-itemsets) $= {}^6C_3 = \frac{6!}{3!\,3!} = 20.$

(c) find an itemset (of size 2 or larger) that has the largest support.

Item	Sup
A	5
B	4
C	4
D	7
E	5
F	5

⇒

Items	Sup.
AB	1
AC	1
AD	4
AE	3
AF	3
BC	2
BD	
...	
EF	5 → max

Ignoring the 1-itemset and ϕ, the itemset with the largest support is {bread, butter} {E, F}

(d) find a pair of items a and b such that the rules {a} → {b} & {b} → {a} have same confidence.

confidence $= \frac{\text{freq}(a,b)}{\text{freq}(a)} = \frac{\text{freq}(b,a)}{\text{freq}(b)}$

∴, for equal confidence, freq.(a) = freq.(b)

pairs can be EF (sup = 5, c = 5/5 = 1), AE (sup = 5, c = 3/5), AF (sup = 4, c = 2/4 = 1/2), BC

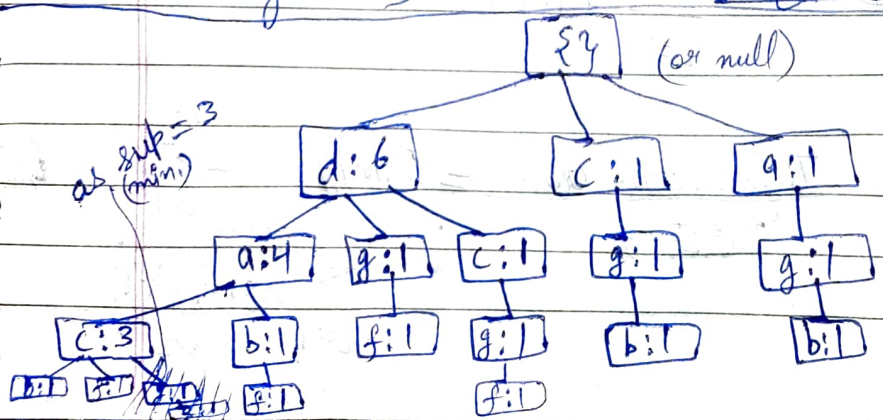
sol 3(b) Given simple transactional database, find FP tree for this database is support threshold is 3.

Tid	items	
1	a,b,d	d a b
2	a,c,d,f	d a c f
3	c,d,e,g,a	d a c g e
4	a,d,f,b	d a b f
5	b,c,g	c g b
6	d,f,g	d g f
7	a,b,g	a g b
8	c,d,f,g	d c g f

Item	freq
a	5
b	4
c	5
d	6
e	1
f	4
g	5

sorted items	freq
d	6
a	5
c	5
g	5
b	4
f	4
~~e~~	~~1~~

as sup = 3 (min.)



A misleading "strong" association rule. Suppose we are interested in analyzing transactions at *AllElectronics* with respect to the purchase of computer games and videos. Let *game* refer to the transactions containing computer games, and *video* refer to those containing videos. Of the 10,000 transactions analyzed, the data show that 6,000 of the customer transactions included computer games, while 7,500 included videos, and 4,000 included both computer games and videos. Suppose that a data mining program for discovering association rules is run on the data, using a minimum support of, say, 30% and a minimum confidence of 60%. The following association rule is discovered:

$$buys(X, \text{“}computer\ games\text{”}) \Rightarrow buys(X, \text{“}videos\text{”})\ [support = 40\%, confidence = 66\%] \quad (6.6)$$

Rule (6.6) is a strong association rule and would therefore be reported, since its support value of $\frac{4,000}{10,000} = 40\%$ and confidence value of $\frac{4,000}{6,000} = 66\%$ satisfy the minimum support and minimum confidence thresholds, respectively. However, Rule (6.6) is misleading because the probability of purchasing videos is 75%, which is even larger than 66%. In fact, computer games and videos are negatively associated because the purchase of one of these items actually decreases the likelihood of purchasing the other. Without fully understanding this phenomenon, we could easily make unwise business decisions based on Rule (6.6).

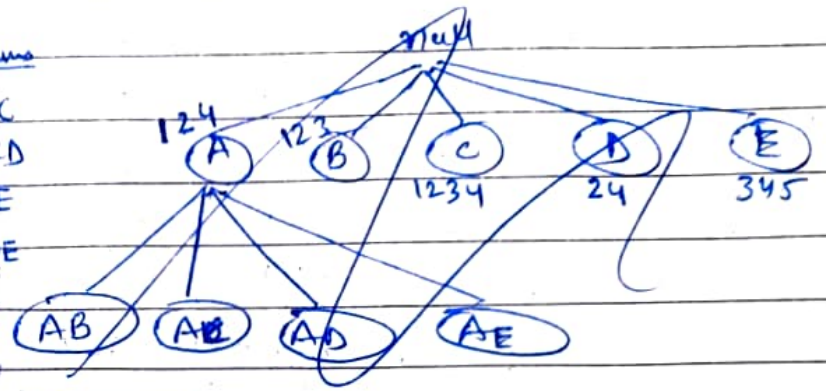
(ii) What is closed and max freq mining? Using an eg. show that how freq and closed items are mined.

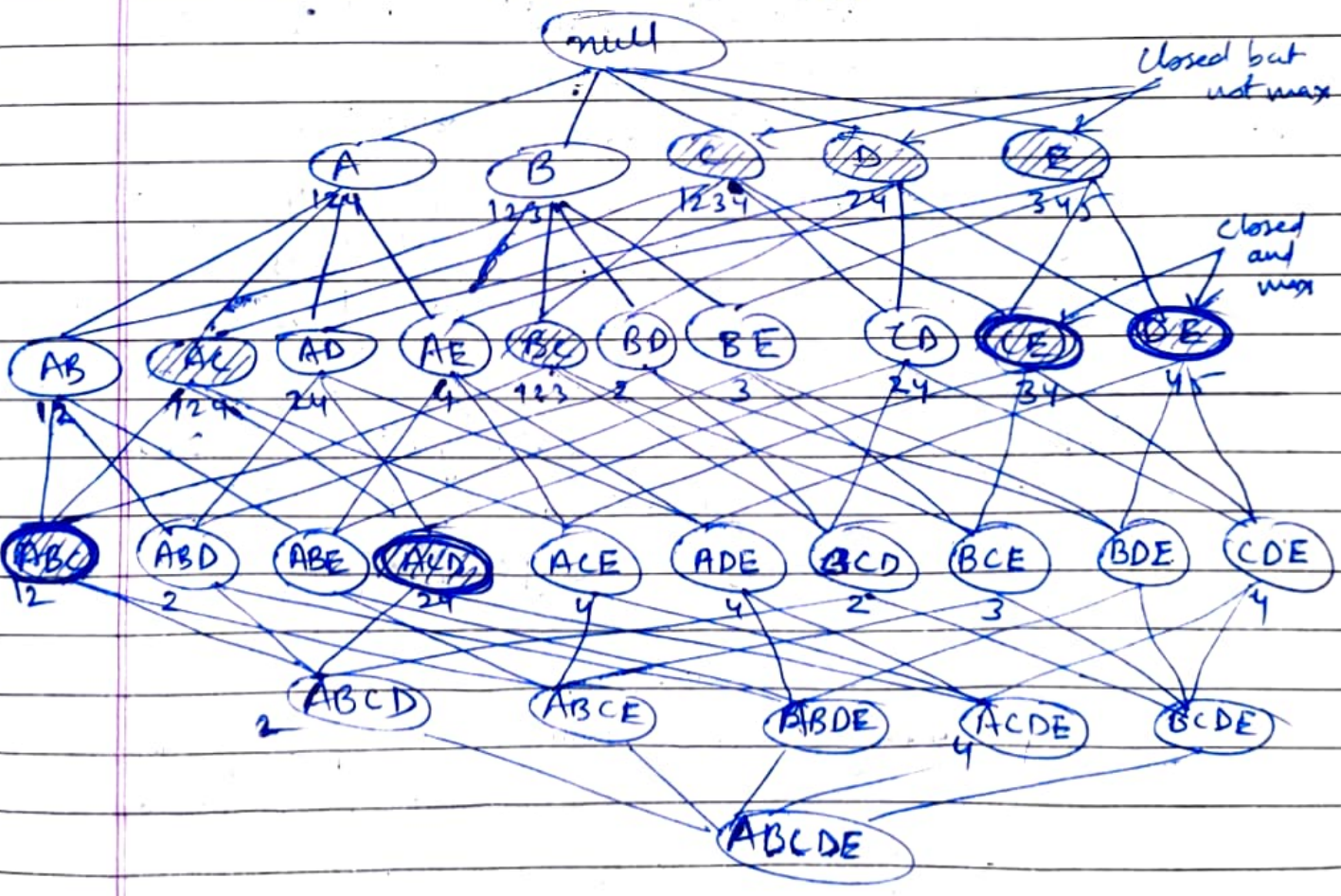
∴ An itemset is closed if none of its immediate supersets has the same support as the itemset.

An itemset is max frequent if none of its immediate supersets is frequent.

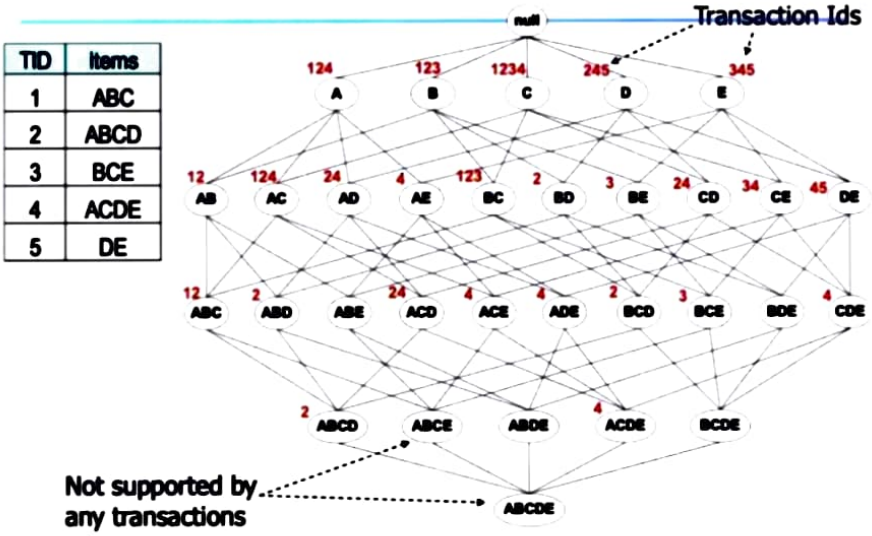
eg :-

TID	items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



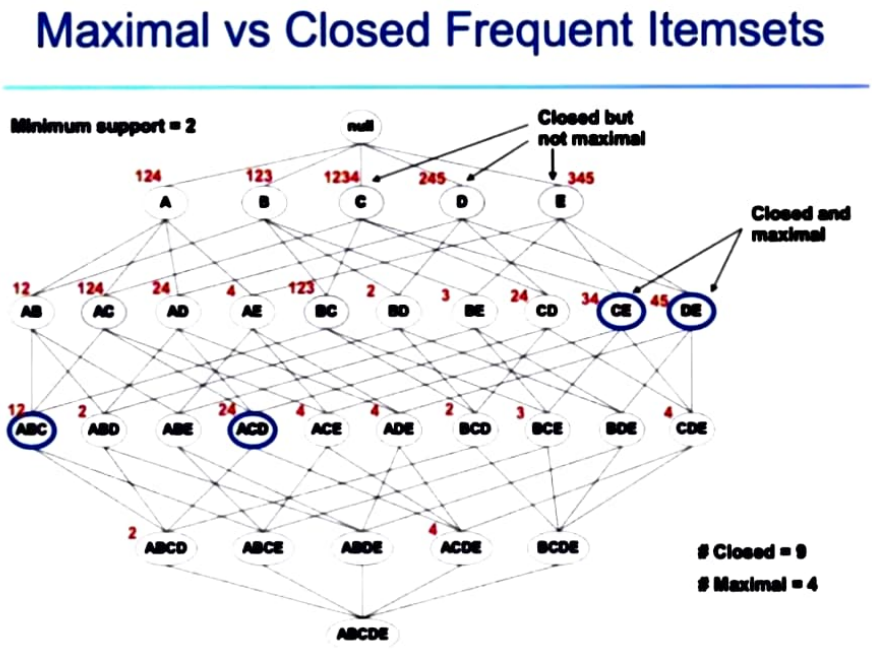


min support is 2 ⇒ if support = 1 ⇒ non-frequent



Example demonstration.

If we set the minsup to be 2, any itemsets that appear more than twice will be frequent itemsets. And among those frequent itemsets, we can find closed and maximal frequent itemsets by comparing their support(frequency of occurrence) to their supersets.



Aniket
Date
Page

sol 4(a.) Comparing Attribute Selection Measures :-

The three measures, in general, return good results but

→ Information gain :-
  → biased towards multivalued attributes.

→ Gain ratio :-
  → tends to prefer unbalanced splits in which one partition is much smaller than the others.

→ Gini index :-
  → biased to multivalued attributes
  → has difficulty when # of classes is large
  → tends to favor tests that results in equal-sized partitions and purity in both partitions.

$$\text{Entropy} = \sum_{i=1}^{n} -P(c_i) \log_2 P(c_i)$$

where $P(c_i)$ is the probability of class c_i in a node.

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} P^2(c_i)$$

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Info}} = \frac{\text{Entropy(before)} - \sum_{i=1}^{m} \text{Entropy}(i, \text{after})}{\sum_{i=1}^{m} w_i \cdot \log_2 w_i}$$

Aniket
Date
Page

decision tree question :-

classifying attribute = "play"

out of 14, No = 5 and Yes = 9

Attribute: outlook

values (Outlook) = Sunny, Overcast, Rain

+ ⇒ Yes
− ⇒ No

S = [9+, 5−]

S_{sunny} ← [2+, 3−]

$S_{overcast}$ ← [4+, 0−]

S_{rain} ← [3+, 2−]

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

$$= 0.94$$

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

$$= 0.971$$

$$Entropy(S_{overcast}) = 0$$

$$Entropy(S_{rain}) = 0.971$$

$$I.\ Gain(S, Outlook) = Entropy(S) - \sum_{v \in \{Sunny, Overcast, Rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= Entropy(S) - \frac{5}{14} Entropy(S_{sunny}) - \frac{4}{14} Entropy(S_{overcast}) - \frac{5}{14} Entropy(S_{rain})$$

$$= 0.94 - \frac{5}{14}(0.971) - \frac{4}{14}(0) - \frac{5}{14}(0.971)$$

$$= 0.2464$$

Aniket
Date
Page

Attribute : Temp

Values (Temp) = Hot, Mild, Cool

$S = [9+, 5-]$ Entropy(S) $= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$

$S_{Hot} = [2+, 2-]$ Entropy(S_{Hot}) $= -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$

$S_{Mild} = [4+, 2-]$ Entropy(S_{Mild}) $= -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.918$

$S_{Cool} = [3+, 1-]$ Entropy(S_{Cool}) $= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$

I. Gain(S, Temp) $= 0.94 - \frac{4}{14}(1.0) - \frac{6}{14}(0.918) - \frac{4}{14}(0.811)$

$= 0.0289$

Attribute : Humidity

Values (Humidity) = High, Normal

$S = [9+, 5-]$ Entropy(S) $= 0.94$

$S_{High} \leftarrow [3+, 4-]$ Entropy(S_{High}) $= -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$

$S_{Normal} \leftarrow [6+, 1-]$ Entropy(S_{Normal}) $= -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.592$

Info. Gain (S, Humidity) $= 0.94 - \frac{7}{14}(0.985) - \frac{7}{14}(0.592)$

$= 0.1516$

Aniket
Date
Page

Attribute : Wind

Values (Wind) = Strong, Weak

$S = [9+, 5-]$ Entropy (S) = 0.94

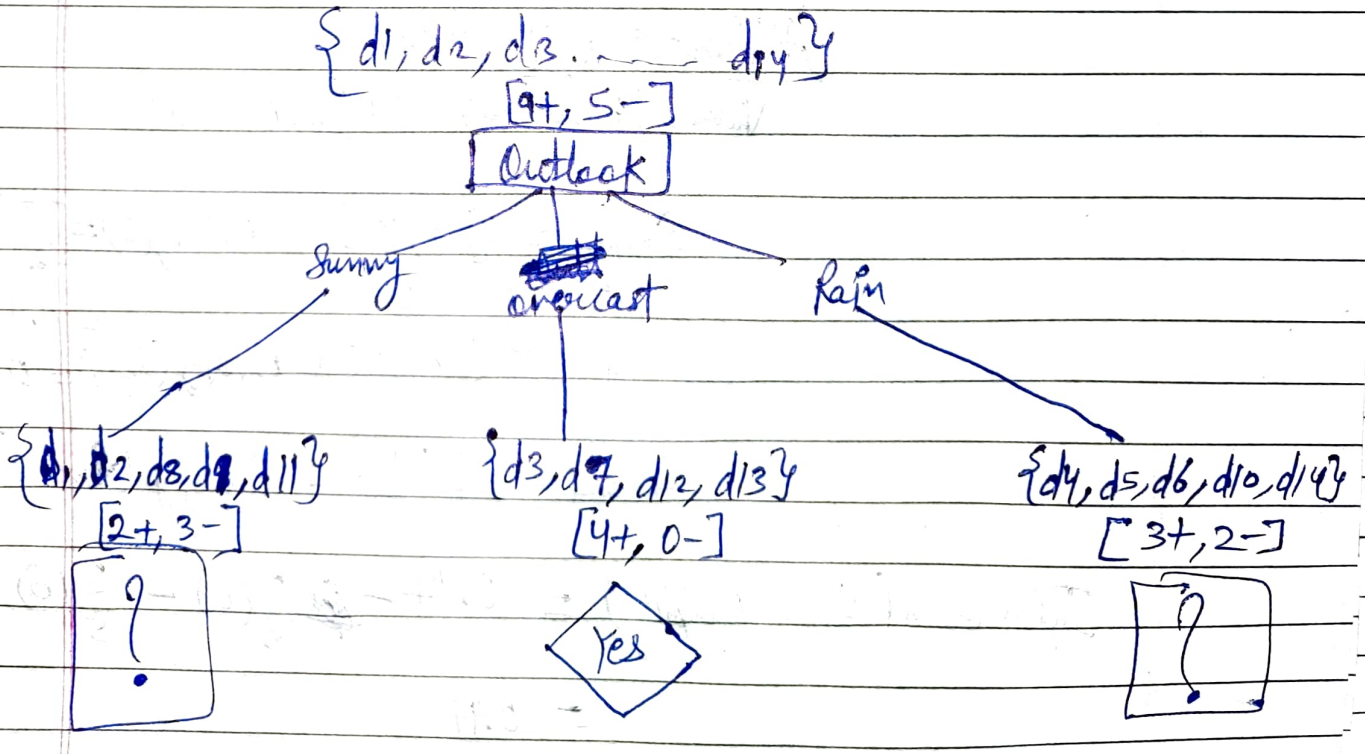
$S_{Strong} \leftarrow [3+, 3-]$ Entropy (S_{Strong}) = 1.0

$S_{Weak} \leftarrow [6+, 2-]$ Entropy (S_{Weak}) = 0.8113

$$\text{Information Gain}(S, \text{Wind}) = 0.94 - \frac{6}{14}(1.0) - \frac{8}{14}(0.8113)$$

$$= 0.0478$$

Now, I.G.(S, Outlook) = 0.2464 ✓ highest gain (root node)
I.G.(S, Temp) = 0.0289
I.G.(S, Humidity) = 0.1516
I.G.(S, Wind) = 0.0478



Sunny

Day	Temp	Humidity	Wind	Play Tennis
d_1	Hot	High	Weak	No
d_2	Hot	High	Strong	No
d_8	Mild	High	Weak	No
d_9	Cool	Normal	Weak	Yes
d_{11}	Mild	Normal	Strong	Yes

Attribute : Temp

Values (Temp) = Hot, Mild, Cool

$S_{sunny} = [2+, 3-]$, $Entropy(S_{sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$S_{hot} = [0+, 2-]$ $Entropy(S_{hot}) = 0$

$S_{Mild} = [1+, 1-]$ $Entropy(S_{Mild}) = 1.0$

$S_{cool} = [1+, 0-]$ $Entropy(S_{cool}) = 0$

$$Info.\ Gain(S_{sunny}, Temp) = Entropy(S) - \frac{2}{5}Entropy(S_{hot}) - \frac{2}{5}Entropy(S_{Mild}) - \frac{1}{5}Entropy(S_{cool})$$

$$= 0.97 - \frac{2}{5}(0) - \frac{2}{5}(1) - \frac{1}{5}(0)$$

$$= 0.57$$

Attribute : Humidity

Values (Humidity) = High, Normal

$S_{sunny} = [2+, 3-]$ $Entropy(S) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$S_{High} \leftarrow [0+, 3-]$ $Entropy(S_{High}) = 0$

$S_{Normal} \leftarrow [2+, 0-]$ $Entropy(S_{Normal}) = 0$

$$Info.\ Gain(S_{sunny}, Humidity) = 0.97 - \frac{3}{5}(0) - \frac{2}{5}(0)$$

$$= 0.97$$

Attribute : Wind

Values (Wind) = Strong, Weak

$S_{sunny} = [2+, 3-]$ Entropy (S) = 0.97

$S_{strong} \leftarrow [1+, 1-]$ Entropy(S_{strong}) = 1.0

$S_{weak} \leftarrow [1+, 2-]$ Entropy (S_{weak}) = 0.9183

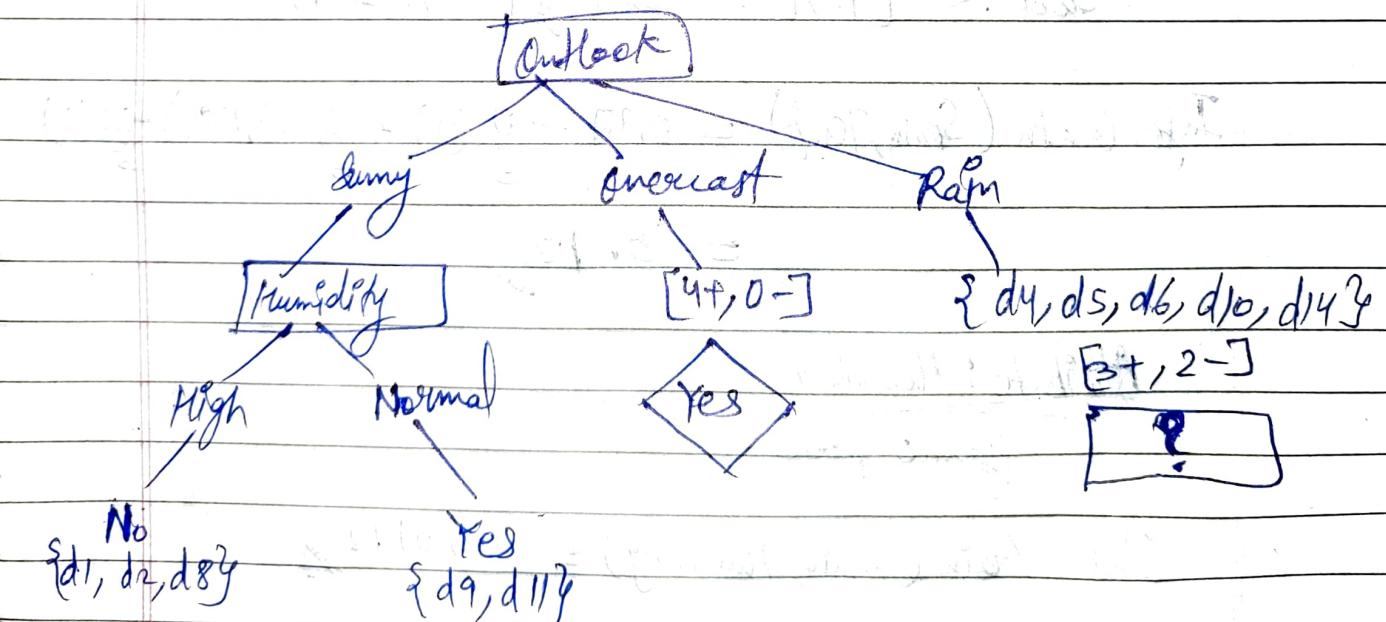
Info. Gain (S_{sunny}, Wind) = $0.97 - \frac{2}{5}(1) - \frac{3}{5}(0.9183)$

= 0.0192

Now, Info. gain (S_{sunny}, Temp) = 0.57

" (S_{sunny}, Humidity) = 0.97 — selected

" (S_{sunny}, Wind) = 0.0192



Aniket
Date
Page

Rain

day	Temp	Humidity	Wind	Play
d4	Mild	High	Weak	Yes
d5	Cool	Normal	Weak	Yes
d6	Cool	Normal	Strong	No
d10	Mild	Normal	Weak	Yes
d14	Mild	High	Strong	No

Attribute : Temp

Values (Temp) = Hot, mild, cool

$S_{Rain} = [3+, 2-]$ Entropy (S_{Rain}) $= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$S_{Hot} \leftarrow [0+, 0-]$ $= 0.97$

Entropy (S_{Hot}) = 0

$S_{mild} \leftarrow [2+, 1-]$ Entropy (S_{mild}) = 0.9183

$S_{cool} \leftarrow [1+, 1-]$ Entropy (S_{cool}) = 1.0

Info. Gain (S_{Rain}, Temp) $= 0.97 - \frac{0}{5}(0) - \frac{3}{5}(0.9183) - \frac{2}{5}(1)$

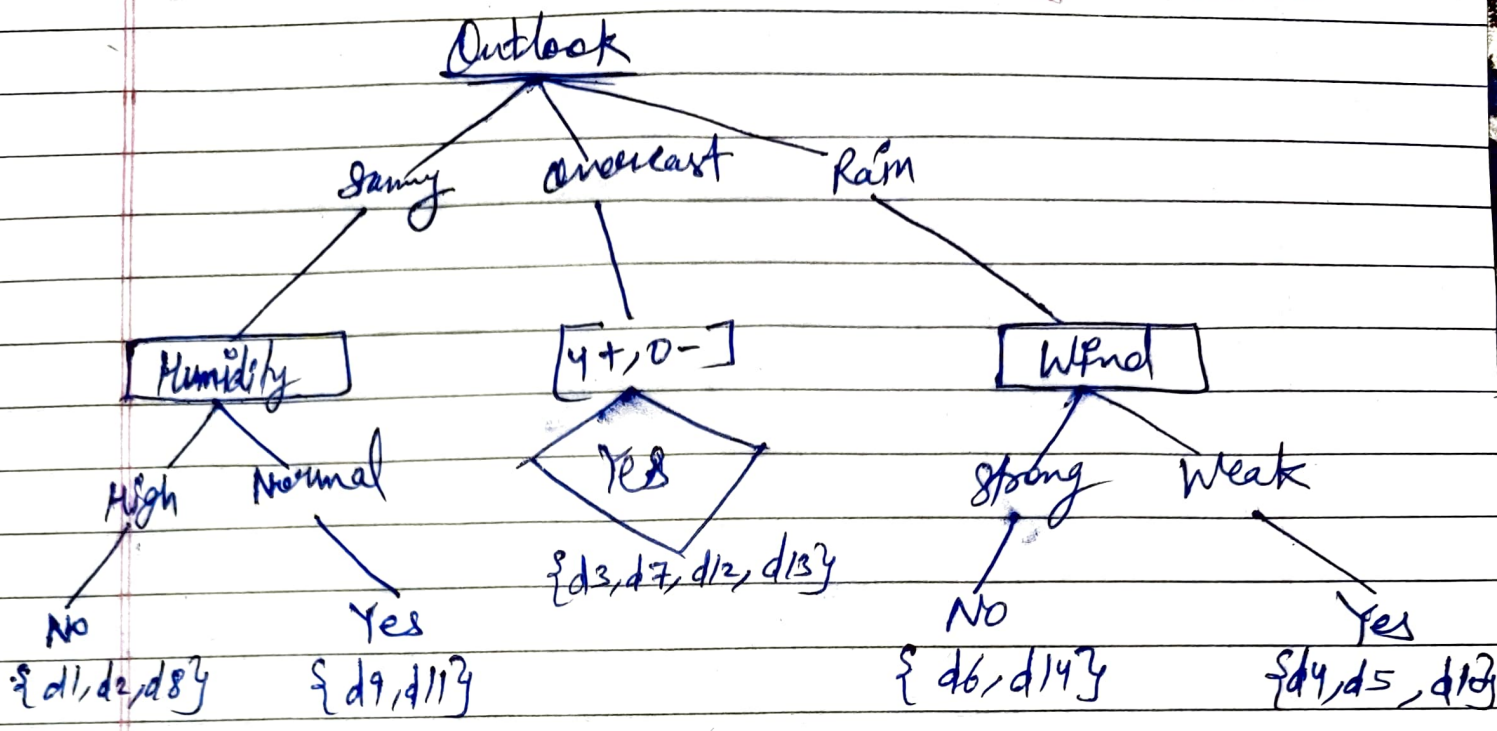
$= 0.0192$

Attribute : Humidity
(same pattern)

Info. Gain (S_{Rain}, Humidity) = 0.0192

Attribute : Wind
(same pattern)

Info. Gain (S_{Rain}, Wind) = 0.97 ← selected



Decision Tree