Paper - 2018 DM

Q1 a) What are the different Issues in Data Mining

→ Mining Methodology :-
Mining various and new kinds of knowledge
Mining knowledge in multidimensional space
Data mining : An interdisciplinary effort
Boosting the power of discovery in a networked environment
Handling noise, uncertainty and incompleteness of data
Pattern evalution and pattern or contraint-guided mining

→ User Interaction - Interactive mining
- Incorporation of background knowledge
- Presentation and visualisation of data mining results

→ Efficiency and Scalability - of data mining algorithm
- Parallel, distributed, stream and incremental mining methods

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

→ Data mining and society - Social impacts of data mining
- Privacy-preserving data mining
- Invisible data mining

b) What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of dat (such as few hundred tuple data set) ?

1b:
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 the data needs to be cleaned, transformed, or analyzed in some way. This can require significant computing resources and can be cost-prohibitive.

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 and relationships to consider.

Privacy: A huge amount of data may contain sensitive or personal information, which requires careful handling and protection to ensure compliance with privacy regulations and ethical considerations.

Overall, mining a huge amount of data requires more resources and careful planning, and it can be more complex and time-consuming compared to mining a small amount of data.

Page No.:
Date : / /

Q(c) What are the factors that have to be considered when data mining is to be performed on the following data repository

(i) text data ⟵ Text mining

(ii) Web data — web mining

Text mining :- Subset of data mining, involves processing of data from various text documents. It is the process of transforming unstructured data (text) into structured format and interpreting these data to identify patterns. In Text Mining, various deep learning algorithms are used to evaluate the text and generate useful info effectively. Basic idea is to find patterns in large datasets that can be used for various purposes. Text Mining requires both sophisticated linguistic and ~~statistical~~ statistical techniques to analyze the unstructured ~~text~~ format data and provide valuable insights. Text mining consist variety of technologies :- 1) Keyword based Technologies :- depend on selecting keywords that input data contains and are then filtered as absence of character strings

2) Statistics Technologies :- refers to system that is completely based on machine learning. It uses certain text to model the data and, in turn uses the same model to manage and categorize text.

3) Linguistic - based Technologies :- Linguistic based system uses Natural language processing system. The NLP models read the input text and understand structure of text, grammar, logic and context of text.

Web mining :- process of extracting various useful information readily available on the internet (www). web mining is a subset of DM. It helps to analyze the user activities on diff web pages and track them over a period of time to understand customer's behaviour and surface patterns.

Web

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

Data source: The source of the data can affect the quality and relevance of the data for a specific task. For example, web data may come from a variety of sources such as social media, online news articles, and e-commerce websites, while text data may come from documents, books, or other written sources.

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 preprocessing steps required to extract and clean the data.

Data volume: Web data can be very large and may need to be sampled or aggregated to make it more manageable for data mining. Text data is often smaller in volume and may not require as much preprocessing.

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.

Data privacy: Web data may contain personal information that needs to be protected in accordance with privacy laws and regulations. Text data may also contain sensitive information, but it is generally less of a concern compared to web data.

Data relevance: The relevance of the data for a specific task is an important consideration when performing data mining. Web data may be more relevant for tasks related to online trends and consumer behavior, while text data may be more relevant for tasks related to natural language processing and text analysis.

Page No.:
Date : / /

Q2 What are the various dimensionality reduction techniques used in datamining. Explain any one technique.

Sol No. of input features, variables or columns present in a given dataset is known as dimensionality, and the process of reducing these features is called dim red.

A dataset contains a huge no. of input features, which makes predictive modeling task more complicated. Because it is very difficult to visualize or make predictions for the training dataset with high no. of features, for such cases, dim red techniques are used.

Curse of dimensionality :- Handling high-dimensional data increases complexity.

advantages of dim Red:- i) space ↓ (ii) time ↓ (iii) visualize quickly (iv) reduces redundant features

disadv. (i) data loss (ii) In PCA dim red technique, principal components req. to consider are unknown.

2 approaches of dim Red.

a) Feature Selection:- process of selecting subset of relavant features and leaving out the irrelavant features present in a dataset. to build a model of high accuracy. In other words it is a way of selecting optimal features from input dataset. → 3 methods:-

1) Filters method :- data set is filtered, subset is taken
   -) correlation -) Chi-square test -) ANOVA
   -) Info gain

2) Wrappers method :- same as filter method but takes ML model for evaluation. performance for a subset decide whether to add those features or remove to increase accuracy of model (more accurate than filter method)

3) forward Selection 2) Backward Selection 3) Bi-directional Elimination

3) Embedded methods: check diff training iterations of ML model and evaluate importance of each feature
- L2 Asso
- Elastic Net
- Ridge Regression.

4) Feature Extraction:- process of transforming space containing many dimensions into space with fewer dimensions.
This approach is useful when we want to keep whole info but use fewer resources while processing info.
- PCA (Principal component Analysis)
- Linear Discriminant Analysis
- Kernel PCA
- Quadratic Discriminant Analysis.

→ Principal component Analysis is a statistical process that converts observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These transformed features are called principal components. Used for predictive modeling. PCA works by considering variance of each attribute because high attribute shows good split b/w classes, hence reduce dimensionality.

→ Backward feature elimination :- used while developing linear regression model.

n feature → check performance → n times repeat for n-1 features removing 1 feature. and check performance for each iteration

remove feature which made smallest change ← repeat for n-1 feature

↳ repeat until no feature can be dropped.

→ Forward feature selection:- we find best feature that can produce highest increase in performance.
Start with single feature → added each feature at a time
train models for each feature seperately.
↳ feature with best performance is selected
↳ repeat until no significant increase in performance

Missing Value Ratio :- If features have missing values drop them., Set a threshold level.
If #missing values > threshold ⇒ drop it.

Low variance filter :- Calculate variance for each variable.
variance < threshold ⇒ drop.

High correlation filter :- when 2 variable carry approx similar info. performace degraded, correlation > threshold ⇒ drop one of them.

Random Forest :- algo contains in-built feature importance — package
we dont need to program it seperately.
we need to generate large set of trees against the target variable, and with help of usage statistics of each attribut, we need to find subset of features. It takes only numerical variables, so we need to convert input data into numeric data using dot encoding

Factor analysis :- each variable is kept with a group accordily to correlation with other variables. eg income and spend.

Auto-encoders :- type of ANN (artificial neural network).
main aim is to copy the inputs to their outputs.
input is compressed into latent-space representation.
and output is occurred using this representation
it has 2 parts :- 1) Encoder :- compress input to form latent-space representation.
2) Decoder :- recreate output from latent-space representation.

(b) 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

a) mean = sum of values / no. of values = 809/27 = 29.96
median (for odd) = $\frac{n+1}{2} = \frac{28}{2}$ = 14th value = 25

b) 16, 20, 22, 33 occur twice
13, 15, 19, 21, 30, 36, 40, 45, 46, 52, 70 occurs once
25, 35 occur 4 times ⇒ data set is bimodal with 2 modes as 25 and 35

d) Q_1 (first quartile) = 25th percentile

$= \frac{25}{100} \times 27 = 6.75 = 7^{th}$ value $= 20$

Q_3 (third quartile) = 75th percentile

$= \frac{75}{100} \times 27 = 20^{th}$ value $= 35$

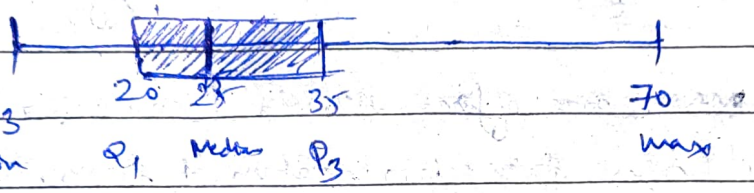
e) Five number summary :- (Q_0) min = 13, $Q_1 = 20$, median = 25, $Q_3 = 35$, Max = 70 (Q_4)

c) mid range = Average of largest and smallest

$= \frac{70+13}{2} = 41.5$

box plot

(f)



Q2(c) X

2c:
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 and 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 and the Tukey 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 include the DBSCAN algorithm and the Local Outlier Factor (LOF) algorithm.

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

Cluster-based methods: These methods identify observations that do not belong to any of the clusters formed by a clustering algorithm.

Outlier detection can be applied in a variety of applications, including:

Fraud detection: Outliers can indicate unusual or suspicious activity, such as fraudulent transactions or account activity.

Quality control: Outliers can indicate defects or errors in manufacturing processes or other types of data collection.

Medical diagnosis: Outliers can indicate unusual or rare medical conditions that need to be investigated further.

Environmental monitoring: Outliers can indicate unusual or extreme environmental conditions that need to be monitored and potentially mitigated.

Q3 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
Cookies – C
Diaper – D
Bread – E
Butter – F

a) Total no. of rules :-
There are 6 items
⇒ Total rules =
$3^6 - 2^6 + 1$
= 602

b) max size of freq itemsets (min sup > 0)
∵ longest transaction contain 4 items, ⇒ 4

not in paper
max no. of size 3 itemsets that can be derived
$= {}^6C_3 = \frac{6!}{3!\,3!} = \frac{6\times5\times4\times3\times2\times1}{3\times2\times1\times3\times2\times1}$
= 20

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

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

⇒

	Sup
AB	1
AC	1
AD	4
AE	3
AF	3
BC	2
BD	3
BE	0
BF	0
CD	2
CE	1
CF	1
DE	3
DF	3
EF	5

(d) Find a pair of items a and b such that rule {a} → {b} and {b} → {a} have same confidence

Freq(a,b) / Freq(a)

⇒ if freq(a) = freq(b)
then confidence(a→b) = confidence(b→a)

4 (B,C)	5 (E,F)
5 (A,E)	5 (A,F)

ans.

Bread, Butter

max

q3 (b) Given simple transactional database, find FP tree for this database in support threshold is 3

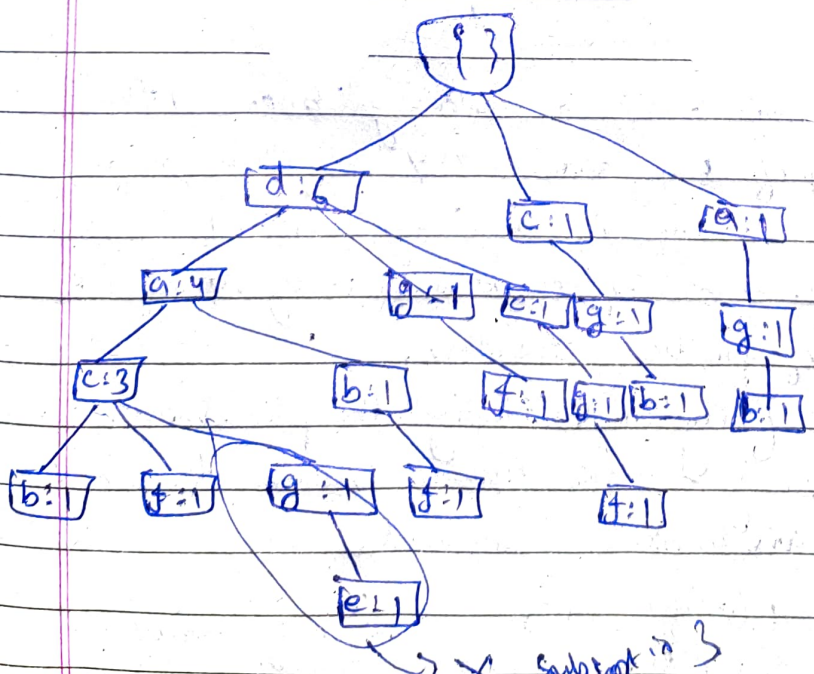
TID	Items
1	a, b, c, d
2	a, c, d, f
3	c d e g a
4	a d f b
5	b. c g
6	d f g
7	a b g
8	c d f g

dacb
dacf
dacge
dafb
cgb
dgf
agb
dcgf

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

⇒

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



→ ✗ support is 3

Give a short eg to show that items in a strong association rule may actually be negatively correlated.

3(c)(i) Let game refer to transactions containing computer games and video refer to those containing videos. Of 10,000 transactions analyzed, data show 6000 costumer transactions included computer games, and 7500 included videos, and 4000 included both games and video.

Suppose data mining program for discovering association rule is run on this data, using min support (30%) and a min confidence (60%) following rule is discovered:—

buys(X, "games") ⇒ buys(X, video) [support = 40%, confidence = 66%]

This is a strong association rule.

$$\text{Support} = \frac{4000}{10000} = 40\% \quad \text{and confidence} = \frac{4000}{6000} = 66\%.$$

Satisfy min sup and min conf thresholds.

However, this rule is misleading because, the probability of buying video = 75%. which is > 66%

In fact, games and vides are negatively correlated as purchase of one of these items actually decreases the likelihood of purchasing the other.

Without fully understanding the phenomenon, we could easily make unwise business decisions based on the association rule we found.

$$\text{if lift}(A \to B) = \frac{P(A \cup B)}{P(A)P(B)} < 1 \text{ then, negatively correlated}$$

$$= \frac{\text{conf}(A \to B)}{\text{sup}(B)} > 1 \text{ then, positively correlated}$$

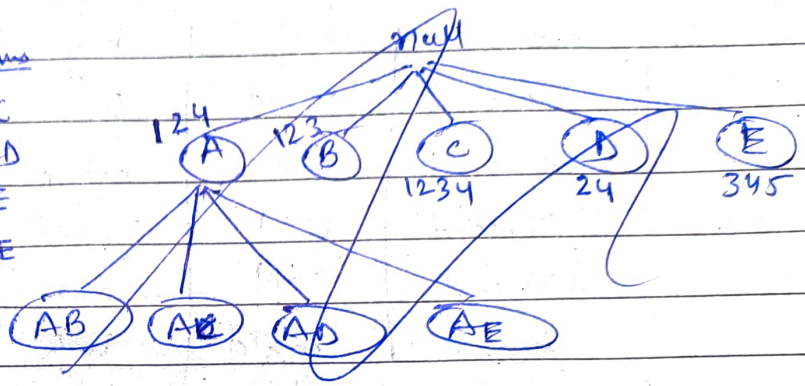
(vi) What is closed and max freq mining? Using an eg. show that how freq and closed items are mined.

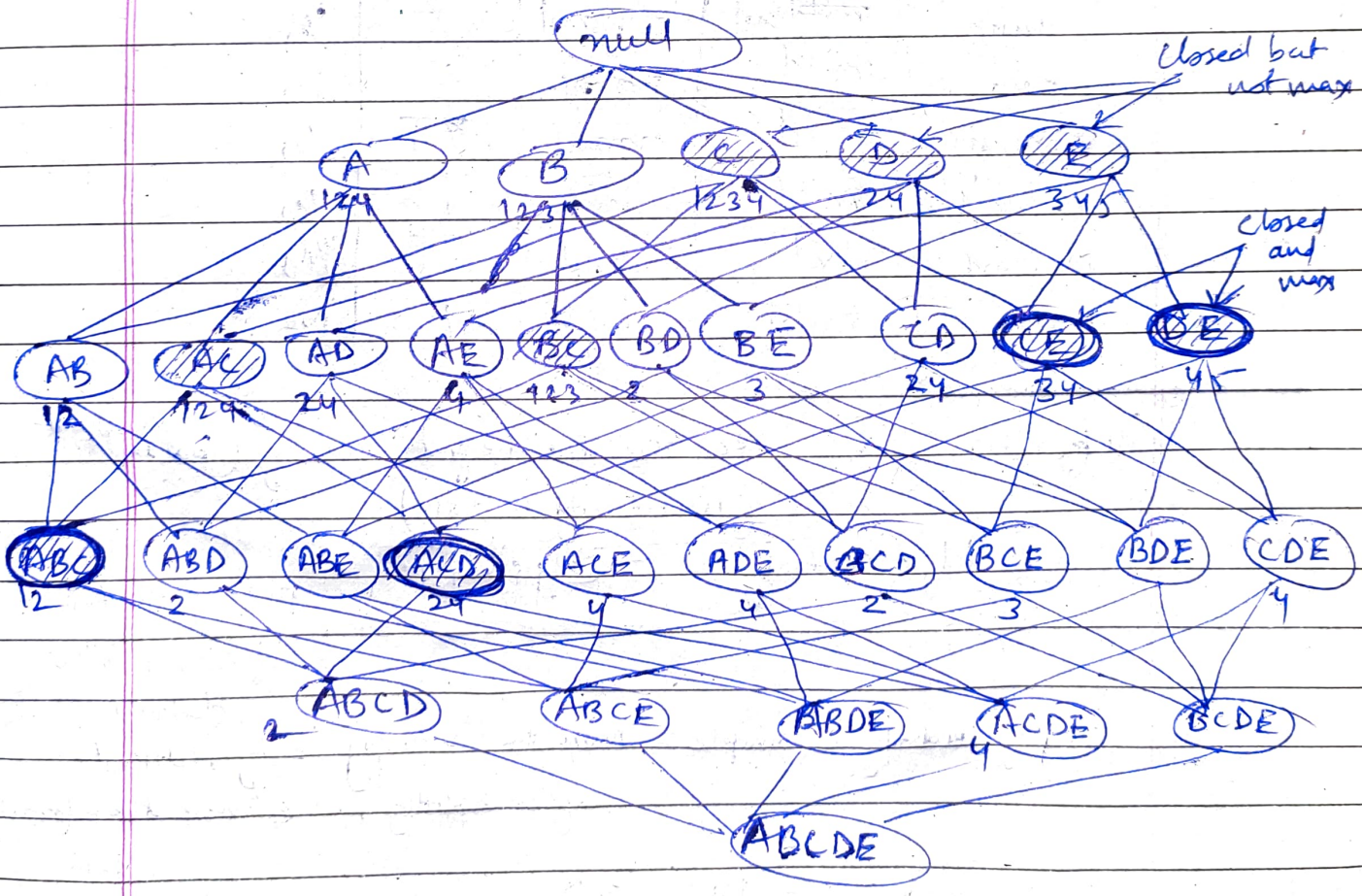
Page No.:
Date : / /

- 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.

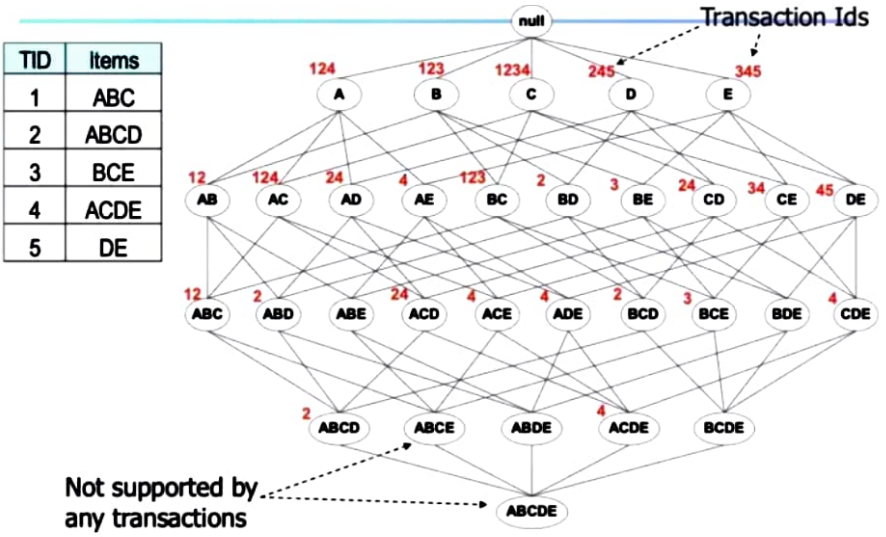
eg :-

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



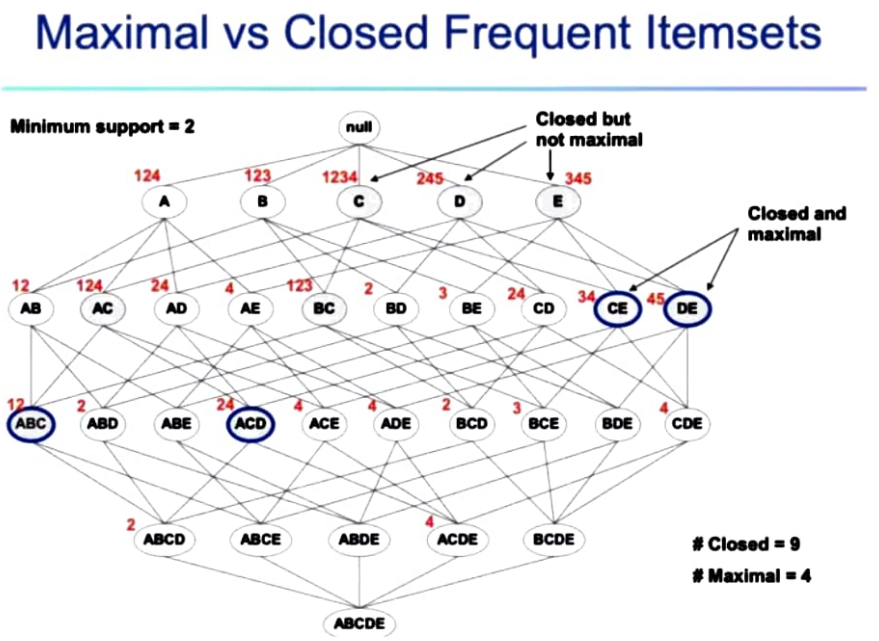


min support is 2 ⇒ if support = 1 ⇒ non-frequent



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.



Q4a) What is the difference b/w Information Gain, Gain Ratio and Gini Index? Using decision tree classifier find the root of the decision tree for the following training data. (Play = output)

Outlook	Temperature	humidity	windy	play
Sunny	hot	high	F	No
Sunny	hot	high	T	No
overcast	hot	high	F	Yes
rain	mild	high	F	Yes
rain	cool	normal	F	Yes
rain	cool	normal	T	No
overcast	cool	normal	T	Yes
Sunny	mild	high	F	No
sunny	cool	normal	F	Yes
rain	mild	normal	F	Yes
Sunny	mild	normal	T	Yes
overcast	mild	high	T	Yes
overcast	hot	normal	F	Yes
rain	mild	high	T	No

Information Gain:- For this first we find entropy (measurement of uncertainty in data)

$$H_j \text{ Entropy} = -\sum p_i \log_2 p_i$$

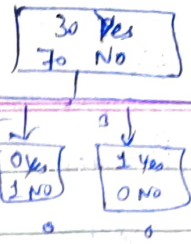
p_i is proportion of class i in the dataset

idea of entropy is simple:- more the entropy being reduced after splitting, the more the information gain

$$IG_{split} = H - \left(\sum \frac{|D_j|}{|D|} \times H_j\right)$$

IG has an undesired characteristic, which is to favor the predictor variable with a large no. of values. Those highly branching predictors are likely to split the data into subsets

with low entropy values eg :- extreme case



disadvantages of these splits
- model becomes more prone to overfitting
- no of nodes in tree may be very large

Gain Ratio :- this attempts to lessen the bias of IG on highly branched predictors by introducing normalizing term called Intrinsic information. $II = -\sum \left(\frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} \right)$

$$\text{Gain ratio} = \frac{IG}{II}$$

Gini index :- $\text{Gini} = 1 - \left(\sum p_i^2 \right)$

$$\text{Gini}_{split} = \sum \frac{|D_j|}{|D|} \text{Gini}_j$$

developed independently with intention of assessing income dispersion of countries but adapted to work as heuristic for splitting optimization.

IG is biased towards high branching features
GR as a result of II, prefers splits with some partitions being much smaller than others.
GI is balanced around 0.5, while entropy penalizes small proportions more than large ones.

$$H = -\left(P(\text{Yes}) \log P_{\text{Yes}} + P_{\text{No}} \log P_{\text{No}} \right)$$
$$= -\left(\frac{9}{14} \log \frac{9}{14} + \frac{5}{14} \log \frac{5}{14} \right)$$
$$= 0.283$$

Outlook :-

Outlook	Yes	No
sunny	2	3
overcast	4	0
rainy	3	2

$$H_{\text{sunny}} = -\left(\frac{2}{5} \log \frac{2}{5} + \frac{3}{5} \log \frac{3}{5} \right) = 0.29$$

$$H_{\text{overcast}} = -\left(\frac{4}{4} \log \frac{4}{4} + \frac{0}{4} \log \frac{0}{4} \right) = 0$$

$$H_{\text{rainy}} = -\left(\frac{3}{5} \log \frac{3}{5} + \frac{2}{5} \log \frac{2}{5} \right) = 0.29$$

$$IG_{\text{outlook}} = 0.283 - \left(\frac{5}{14} \times 0.29 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.29 \right) = 0.0758$$

Temperature :-

Temp	Yes	No
hot	2	2
mild	4	2
cool	3	1

$H_{hot} = -\left(\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}\right) = 0.301$

$H_{mild} = -\left(\frac{4}{6}\log\frac{4}{6} + \frac{2}{6}\log\frac{2}{6}\right) = 0.276$

$H_{cool} = -\left(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right) = 0.244$

$IG_{Temp} = 0.283 - \left(\frac{4}{14} \times 0.301 + \frac{6}{14} \times 0.276 + \frac{4}{14} \times 0.244\right)$

$= 0.009$

Humidity :-

Humidity	Yes	No
high	3	4
normal	6	1

$H_{high} = -\left(\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}\right) = 0.296$

$H_{normal} = -\left(\frac{6}{7}\log\frac{6}{7} + \frac{1}{7}\log\frac{1}{7}\right) = 0.178$

$IG_{Humidity} = 0.283 - \left(\frac{7}{14} \times 0.296 + \frac{7}{14} \times 0.178\right) = 0.046$

windy

Windy	Yes	No
T	3	3
F	6	2

$H_T = -\left(\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6}\right) = 0.301$

$H_F = -\left(\frac{6}{8}\log\frac{6}{8} + \frac{2}{8}\log\frac{2}{8}\right) = 0.244$

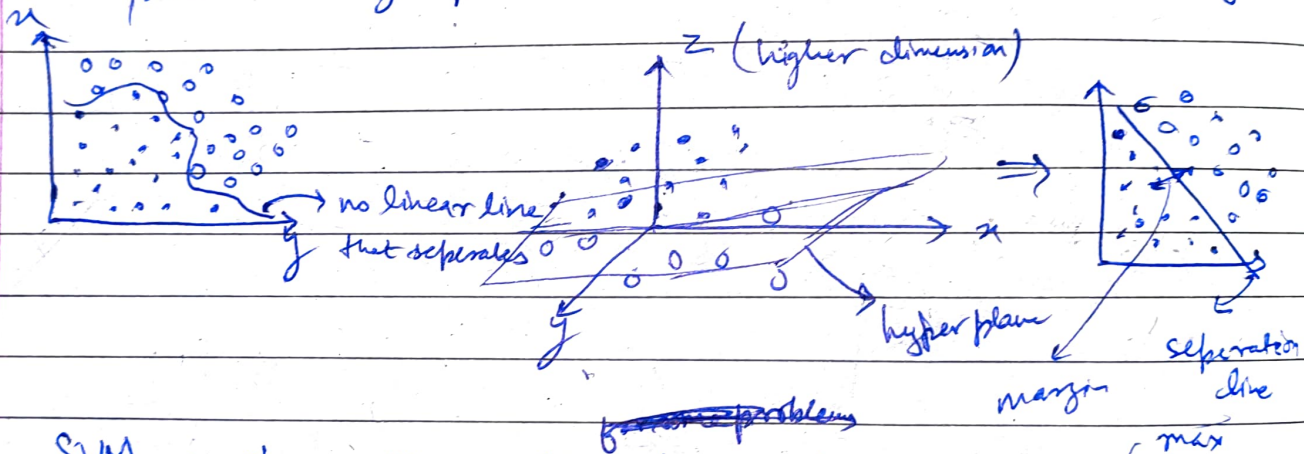
$IG_{windy} = 0.283 - \left(\frac{6}{14} \times 0.301 + \frac{8}{14} \times 0.244\right) = 0.01457$

⇒ IG for outlook is greater than every other split

⇒ root node is (Outlook)

Q4(b) Explain the working principle of Support vector machine. Why SVM is sometimes considered superior to other classification algorithm on some problems.

Sol:- SVM works by mapping data to a high dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A seperator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of ~~that~~ new data can be used to predict the group to which a new record should belong.



SVM works well sometimes because there is a clear margin of seperation bw classes for some problems.

SVM is more efficient in high dimensional spaces and is relatively memory efficient.

SVM is effective in cases where dimensions are greater than number of samples.

⇒ SVM performs well when number of features for each data points exceeds the no. of ~~sample~~ data samples.

4(c)

NAIVE BAYES CLASSIFIER

EXAMPLE - 3

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

$$p(Yes) = \frac{5}{10} = 0.5$$

$$p(No) = \frac{5}{10} = 0.5$$

Color	Yes	No
Red	3/5	2/5
Yellow	2/5	3/5

Type	Yes	No
Sports	4/5	2/5
SUV	1/5	3/5

Origin	Yes	No
Domestic	2/5	3/5
Imported	3/5	2/5

$New\ Instance = (Red, SUV, Domestic)$ No

$$P(Yes|New\ Instance) = p(Yes) \cdot P(Color = Red|Yes) \cdot P(Type = SUV|Yes) \cdot P(Origin = Domestic|Yes)$$

$$P(Yes|New\ Instance) = \frac{5}{10} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} = \frac{3}{125} = 0.024$$

$$P(No|New\ Instance) = p(No) \cdot P(Color = Red|No) \cdot P(Type = SUV|No) \cdot P(Origin = Domestic|No)$$

$$P(No|New\ Instance) = \frac{5}{10} \cdot \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} = \frac{9}{125} = 0.072$$

$$P(No|New\ Instance) > P(Yes|New\ Instance)$$

5a:
Numerical (interval-scaled) variables: There are several ways to compute the dissimilarity between objects described by numerical (interval-scaled) variables. Some common options include:
Euclidean distance: This is the most widely used distance measure for numerical variables. It calculates the dissimilarity between two objects as the square root of the sum of the squared differences between their corresponding variables.

Manhattan distance: This distance measure calculates the dissimilarity between two objects as the sum of the absolute differences between their corresponding variables.

Minkowski distance: This distance measure is a generalization of both Euclidean and Manhattan distance. It calculates the dissimilarity between two objects as the sum of the absolute differences between their corresponding variables, raised to a power (p) and then taking the pth root.

Asymmetric binary variables: To compute the dissimilarity between objects described by asymmetric binary variables, one option is to use the Jaccard coefficient. This measure calculates the dissimilarity between two objects as the number of variables that are different between the objects, divided by the number of variables that are different or are the same in both objects.

Categorical variables: To compute the dissimilarity between objects described by categorical variables, one option is to use the Hamming distance. This measure calculates the dissimilarity between two objects as the number of variables that are different between the objects.

Ratio-scaled variables: To compute the dissimilarity between objects described by ratio-scaled variables, one option is to use the Pearson correlation coefficient. This measure calculates the linear relationship between two variables and ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). A value of 0 indicates no correlation.

Nonmetric vector objects: To compute the dissimilarity between nonmetric vector objects, one option is to use the cosine similarity measure. This measure calculates the angle between two vectors and ranges from -1 (vectors are perfectly dissimilar) to 1 (vectors are perfectly similar). A value of 0 indicates no similarity.

5b:-
(a) k-medoids clustering:

Shapes of clusters that can be determined: k-medoids clustering can determine clusters of any shape, including non-convex shapes.

Input parameters that must be specified: The number of clusters (k) must be specified in advance. The distance metric to be used for measuring the similarity between data points must also be specified.

Limitations: k-medoids clustering is sensitive to the choice of initial medoids and can get stuck in local optima. It can also be computationally expensive for large datasets.

(b) AGNES (Agglomerative Nesting):

Shapes of clusters that can be determined: AGNES can determine clusters of any shape, including non-convex shapes.

Input parameters that must be specified: The distance metric to be used for measuring the similarity between data points must be specified. The linkage criterion (e.g., single, complete, average, etc.) must also be specified, which determines how the similarity between clusters is measured.

Limitations: AGNES is sensitive to the choice of distance metric and linkage criterion and can be computationally expensive for large datasets. It is also prone to producing elongated clusters.

(c) BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):

Shapes of clusters that can be determined: BIRCH can determine clusters of any shape, including non-convex shapes.

Input parameters that must be specified: The desired number of clusters (k) must be specified in advance. The threshold for the minimum number of points in a subcluster must also be specified.

Limitations: BIRCH can be sensitive to the choice of threshold and may not perform well on datasets with high dimensionality or outliers. It is also sensitive to the order in which the data points are processed.

(d) DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

Shapes of clusters that can be determined: DBSCAN can determine clusters of any shape, including non-convex shapes and clusters with arbitrary densities.

Input parameters that must be specified: The minimum number of points required to form a cluster (minPts) and the maximum distance between points in a cluster (eps) must be specified.

Limitations: DBSCAN requires a good choice of the minPts and eps parameters in order to perform well. It is also sensitive to the scale of the data and can be computationally expensive for large datasets. It may also not perform well on datasets with large variations in density.

Part 2:
Density-based clustering methods and partition-based clustering methods are two main categories of clustering algorithms.

Density-based clustering methods identify clusters by finding areas of the data that are high in density, surrounded by areas of low density. These algorithms can identify clusters of any shape and can handle datasets with noise and outliers. Examples of density-based clustering methods include DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points To Identify the Clustering Structure).

Partition-based clustering methods, on the other hand, divide the data into a predefined number of clusters by minimizing the within-cluster sum of distances. These algorithms can identify clusters that are convex or spherical in shape, but may struggle with non-convex or unevenly sized clusters. Examples of partition-based clustering methods include k-means and k-medoids.

In summary, density-based clustering methods are able to identify clusters of any shape and can handle datasets with noise and outliers, while partition-based clustering methods are limited to identifying convex or spherical clusters and may be sensitive to the presence of noise and outliers.

Q5(c)

Data point	x value	y value	dissimilarity from C_1	dissimilarity from C_2	medoid
X_1	3	7	0+2 = 2	7+8 = 15	C_1
X_2	4	8	1+1 = 2	6+7 = 13	C_1
X_3	3	9	0+0 = 0	7+6 = 13	C_1
X_4	4	9	1+0 = 1	6+6 = 12	C_1
X_5	6	10	3+1 = 4	4+5 = 9	C_1
X_6 missing	5	12	2+3 = 5	5+3 = 8	C_1
X_8	7	11	4+2 = 6	3+4 = 7	C_1
X_9	8	12	5+3 = 8	2+3 = 5	C_2
X_{10}	11	16	8+7 = 15	1+1 = 2	C_2
X_{11}	10	15	7+6 = 13	0+0 = 0	C_2
X_{12}	9	10	6+1 = 7	1+5 = 6	C_2

$C_1 = X_3$ (3,9) and $C_2 = X_{11}$ (10,15).

dissimilarity bw $X_1 Y_1$ and $X_2 Y_2$ is calculated by using manhattan distance $= |X_1 - X_2| + |Y_1 - Y_2|$

a) cost for selecting these points as medoids :—

Cost for X_3 and X_{11} as C_1 and C_2 =

$= (2+2+1+4+5+6) + (5+2+6)$

$= 20 + 13 = 33$

b) cost for shifting the medoids to X_4 (4,9) and X_{10} (11,16) from current selection :—

$=$ distance(X_4, X_3) + distance(X_{10}, X_{11})

$= (|4-3| + |9-9|) + (|11-10| + |16-15|)$

$= (1+0) + (1+1) = 1+2 = 3$

(c) What will be the data points after first iteration in both clusters, after shifting medoids has taken place :—

Page No.:
Date : / /

Initially :- $C_1(X_1, X_2, X_4, X_5, X_7, X_8)$ C_1 is X_4

$C_2(X_9, X_{10}, X_{12})$ C_2 is X_{10}

Datapoint	X value	Y value	dissimilarity from $C_1 = X_4$	dissim from $C_2 = X_{10}$	medoid
X_1	3	7	1+2 = 3	8+8 = 17	C_1
X_2	4	8	0+1 = 1	7+8 = 15	C_1
X_3	3	9	1+0 = 1	8+7 = 15	C_1
X_4	4	9	0+0 = 0	7+7 = 14	C_1
X_5	6	10	2+1 = 3	5+6 = 11	C_1
X_7	5	12	1+3 = 4	6+4 = 10	C_1
X_8	7	11	3+2 = 5	4+5 = 9	C_1
X_9	8	12	4+3 = 7	3+4 = 7	C_1
X_{10}	11	16	7+7 = 14	0+0 = 0	C_2
X_{11}	10	15	6+6 = 12	1+1 = 2	C_2
X_{12}	9	10	5+1 = 6	2+6 = 8	C_1

Cost = shifting + ~~medoid~~ (dissimilarity)

= 3 + (3+1+1+3+4+5+7+6) + (2)

~~+ (1+2)~~

= 3 + 30 + 2

= 35

$\Rightarrow$ $cost_{new} > cost_{old}$

$\Rightarrow$ no change in clusters

$\Rightarrow$ $C_1(X_1, X_2, X_4, X_5, X_7, X_8)$ C_1 is X_3

and $C_2(X_9, X_{10}, X_{12})$ C_2 is X_{11}