DWM CAT2 ANSWERS

a) Enlist various types of data in cluster analysis.

Types Of Data Used In Cluster Analysis Are:

Interval-Scaled variables
Binary variables
Nominal, Ordinal, and Ratio variables
Variables of mixed types

Interval-Scaled Variables

- Interval-scaled variables are continuous measurements of a roughly linear scale.
- Typical examples include weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature.
- The measurement unit used can affect the clustering analysis. For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to a very different clustering structure.
- In general, expressing a variable in smaller units will lead to a larger range for that variable, and thus a larger effect on the resulting clustering structure.
- To help avoid dependence on the choice of measurement units, the data should be standardized. Standardizing measurements attempts to give all variables an equal weight.
- This is especially useful when given no prior knowledge of the data. However, in some applications, users may intentionally want to give more weight to a certain set of variables than to others.
- For example, when clustering basketball player candidates, we may prefer to give more weight to the variable height.

Binary Variables

- A binary variable is a variable that can take only 2 values.
- For example, generally, gender variables can take 2 variables male and female.

Contingency Table For Binary Data

Let us consider binary values 0 and 1

	1	0	*sum*
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

Let p=a+b+c+d

Simple matching coefficient (invariant, if the binary variable is symmetric):

$$d(i, j) = \frac{b + c}{a + b + c + d}$$

Jaccard coefficient (noninvariant if the binary variable is asymmetric):

$$d(i, j) = \frac{b + c}{a + b + c}$$

- **Nominal or Categorical Variables**

A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green.

Method 1: Simple matching

The dissimilarity between two objects i and j can be computed based on the simple matching.

m: Let m be no of matches (i.e., the number of variables for which i and j are in the same state).

p: Let p be total no of variables.

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

Method 2: use a large number of binary variables

Creating a new binary variable for each of the M nominal states.

- **Ordinal Variables**

- An ordinal variable can be discrete or continuous.
- In this order is important, e.g., rank.
- It can be treated like interval-scaled
- By replacing xif by their rank,
- $r_{if} \in \{1, \ldots, M_f\}$

  By mapping the range of each variable onto [0, 1] by replacing the i-th object in the f-th variable by,

  $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  Then compute the dissimilarity using methods for interval-scaled variables.

Ratio-Scaled Intervals

- **Ratio-scaled variable**: It is a positive measurement on a nonlinear scale, approximately at an exponential scale, such as Ae^Bt or A^e-Bt.
  Methods:
- First, treat them like interval-scaled variables — not a good choice! (why?)
- Then apply logarithmic transformation i.e.y = log(x)
- Finally, treat them as continuous ordinal data treat their rank as interval-scaled.

Variables Of Mixed Type

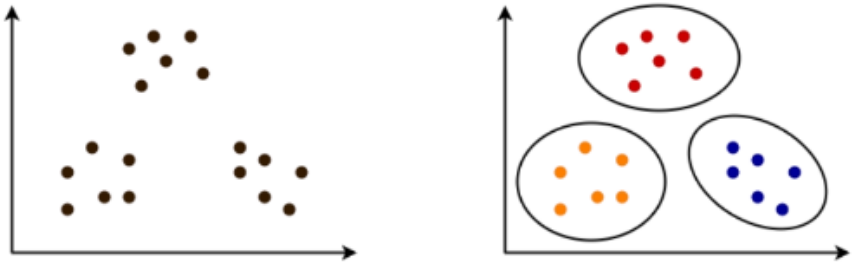
- A database may contain all the six types of variables
  symmetric binary, asymmetric binary, nominal, ordinal, interval, and ratio.
- And those combinedly called as mixed-type variables.

b) Explain k-means algorithm.

- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.
- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

K-Means Clustering-

K-Means clustering is an unsupervised iterative clustering technique.
It partitions the given data set into k predefined distinct clusters.
A cluster is defined as a collection of data points exhibiting certain similarities.



Before K-Means

After K-Means

It partitions the data set such that-

Each data point belongs to a cluster with the nearest mean.
Data points belonging to one cluster have high degree of similarity.
Data points belonging to different clusters have high degree of dissimilarity.

-

(Ques.18) Explain K-means algorithm.

→ ① K-means clustering is an unsupervised learning algo., which groups the unlabeled dataset into different clusters.

② It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

③ The K-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data points to its closest K-center, Those data points which are near to the particular K-center, create a cluster.

④ working :→

Step 1: Select the number K to decide the no. of clusters.

Step 2: Select random K points or centroids.

Step 3: Assign each data point to their ~~closest~~ closest centroid, which will form the predefined K clusters.

Step 4: calculate the variance and place a new centroid of each cluster.

Step 5: Repeat the third step, which means reassign each datapoint to the new closest centroid of each cluster.

Step 6: If any reassignment occurs, then go to step 4 else go to FINISH

Step 7: The model is ready.

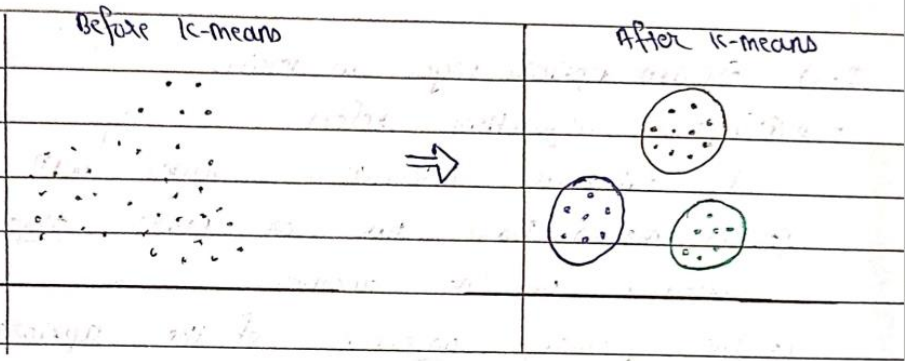


Fig: clustering by K-means algo.

a)Write a detailed note on split algorithm based on gini index.

b)What do you mean by hierarchical clustering approach? Explain agglomerative and divisive hierarchical clustering.

Hierarchical Method: In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

Agglomerative Approach: The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.

Divisive Approach: The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

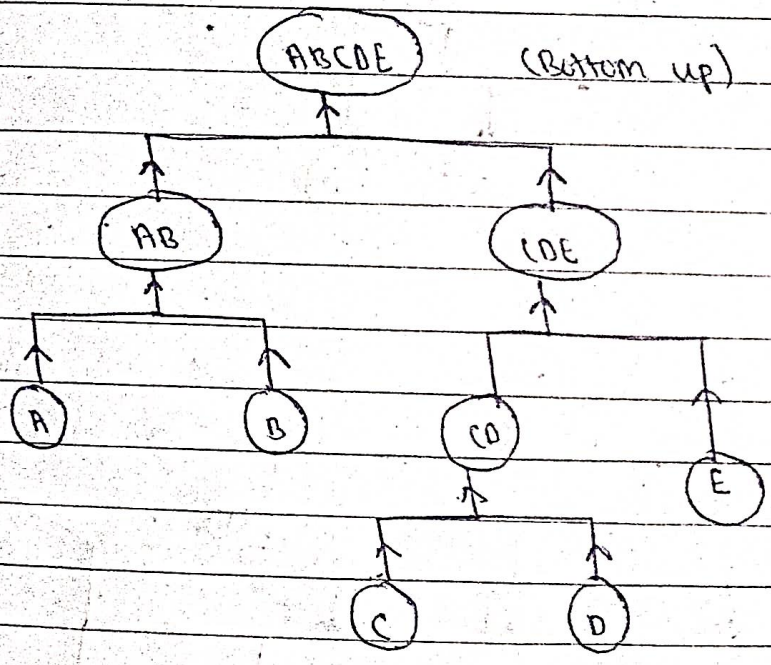
One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.

One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into microclusters, macro clustering is performed on the microcluster.

* Hierarchical clustering :→
  - It is groups of data into tree of clusters, which is represented by dendrogram (merging & splitting).
  - Dendrogram :
    It has sequence of all merges and splits
    ex. Normally in a company you will have
    leadership team
    ↓
    manager
    ↓
    Team leader
    ↓
    Employee

- There are two types of hierarchical clustering method.
  1. Agglomerative
  2. Divisive

1. Agglomerative :—→
   merging data items
   Bottom up approach.
- Calculate similarity of one cluster with respect to all other clusters.
- Consider every data point as individual item.
- merge the cluster with highest similarity.
- Recalculate similarity for each cluster.
- Repeat Step 3 and 4 until single cluster is obtain.



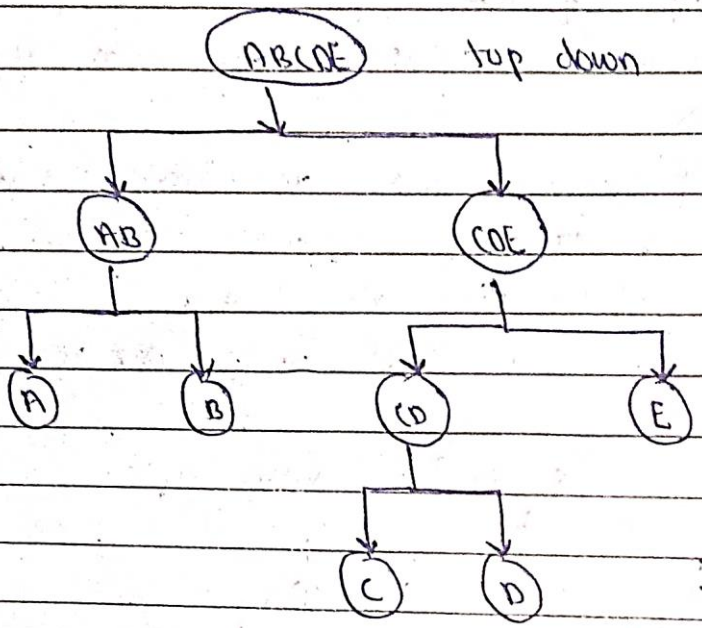
In agglomerative there are three methods:
1. Single linkage: Select min^m similarities.
2. complete linkage: Select max^m similarities.
3. Avg. linkage: Select Avg. similarities.

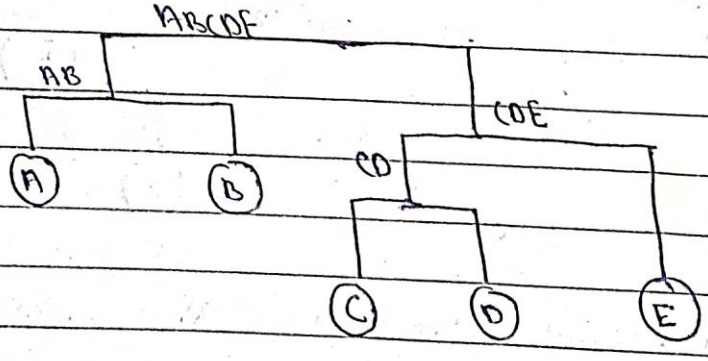
(2) Divisive :——→

top down approach

splitting of data items.

- Select all data items into single cluster and in iteration split the data till we get all n clusters





a)Explain Bayesian classification with suitable example.

* Bayesian classification :

Bayes Theorem :

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

↓

probability of A such that probability or evidence of B is true

Bayesian classifier are statistical classifier they are predict class membership probability such as the probability that the given tuple belongs to a particular class, Bayesian classification is based on Bayes theorem.

Bayes theorem :

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$P(A|B) \cdot P(B) = P(A \cap B) \quad \text{---} (1)$$

$$P(B|A) \cdot P(A) = P(B \cap A) \quad \text{---} (2)$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$\therefore P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

A is hypothesis

B is evidance / data

$P(A|B)$ → posterior probability.

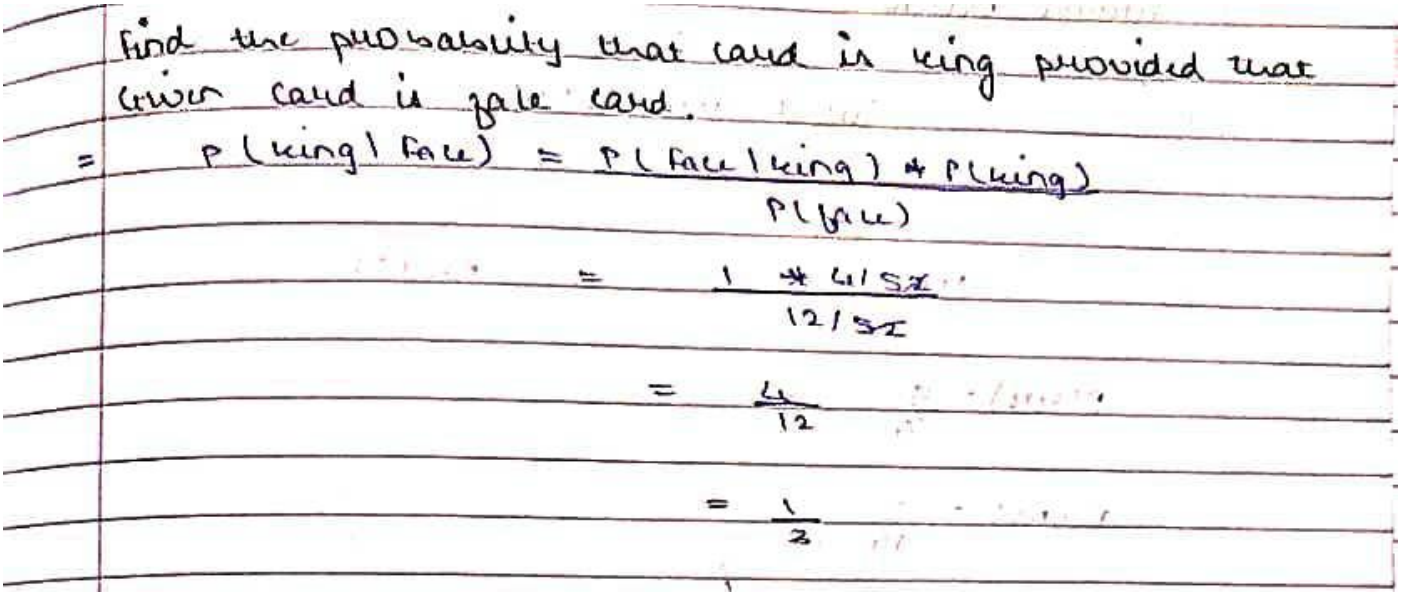
$P(B|A)$ → likelihood

$P(A)$ → prior probability

$P(B)$ → marginal probability.

we have to findout the hypothesis A given that we have observed the evidance (given data).

Example:



Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorem

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities –

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

where X is data tuple and H is some hypothesis.

According to Bayes' Theorem,

$$P(H/X)= P(X/H)P(H) / P(X)$$

b)Explain Naive Bays algorithm with an example.

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- It is mainly used in *text classification* that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
- Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

- **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

Problem: If the weather is sunny, then the Player should play or not?

Solution: To solve this, first consider the below dataset:

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No

9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

- **Frequency table for the Weather Conditions:**

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

- **Likelihood table weather condition:**

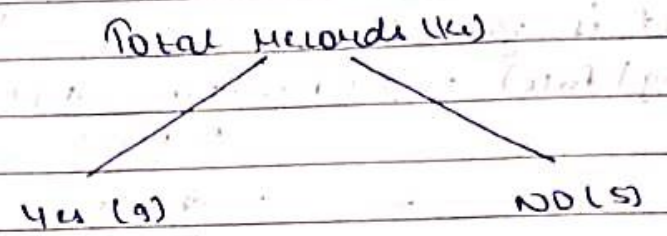
Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	

- **Applying Bayes'theorem:**
- **P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)**
  P(Sunny|Yes)= 3/10= 0.3
  P(Sunny)= 0.35
  P(Yes)=0.71
  So P(Yes|Sunny) = 0.3*0.71/0.35= **0.60**
  P(No|Sunny)= P(Sunny|No)*P(No)/P(Sunny)
  P(Sunny|NO)= 2/4=0.5
  P(No)= 0.29
  P(Sunny)= 0.35
  So P(No|Sunny)= 0.5*0.29/0.35 = **0.41**
  So as we can see from the above calculation that **P(Yes|Sunny)>P(No|Sunny)**
- **Hence on a Sunny day, Player can play the game.**

Day	Outlook	Temp	Humidity	Wind	Play-Cricket
1	Sunny	Hot	High	Weak	NO
2	Sunny	Hot	High	Strong	NO
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	NO
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	Normal	Weak	NO
9	Sunny	Cold	High	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	NO

Find the probability to play cricket on a day when condition are temp = cold humidity = high wind = strong outlook = sunny.

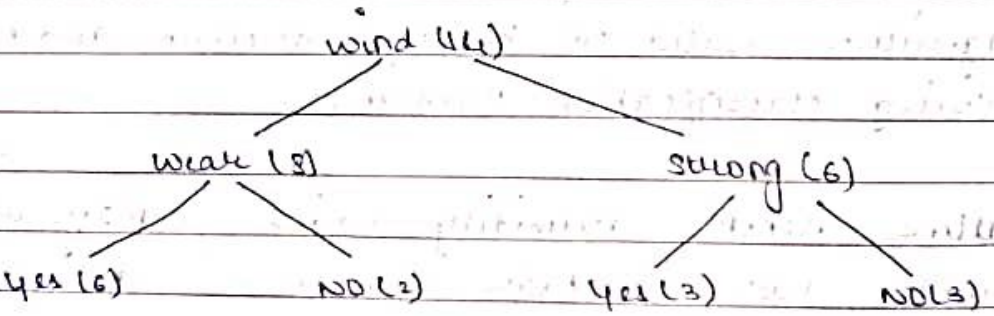
= Step 1 :



$P(Yes) = \frac{9}{14}$

$P(NO) = \frac{5}{14}$

Step 2 : Expansion Step

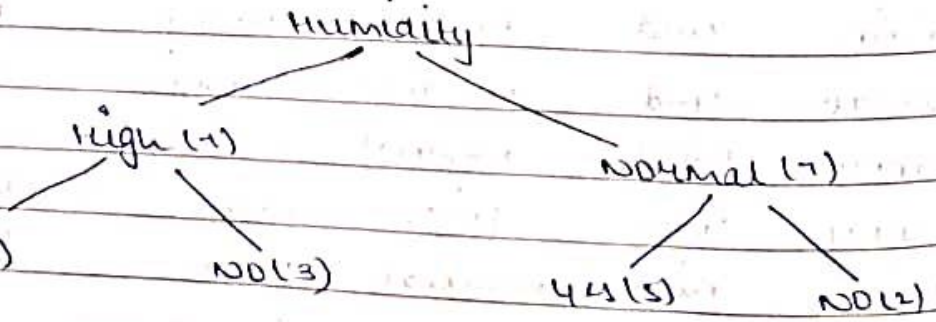


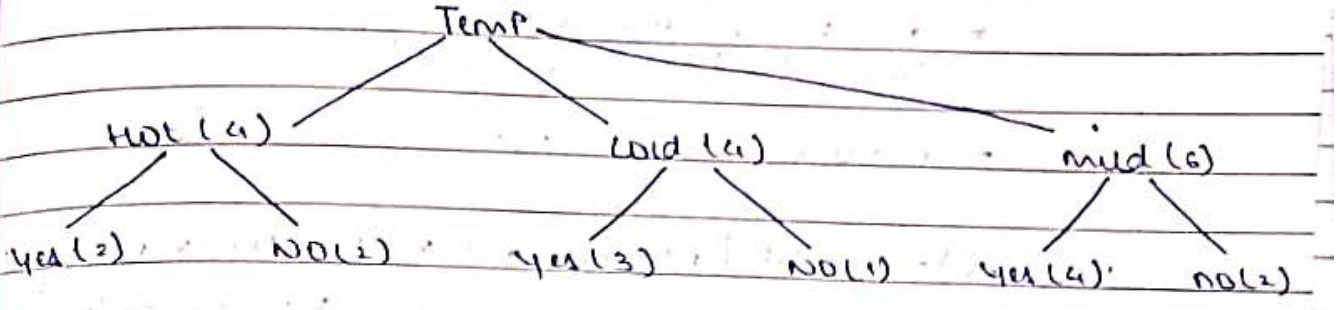
P(weak / Yes) = 6/9

P(weak / NO) = 2/5

P(strong / Yes) = 3/9

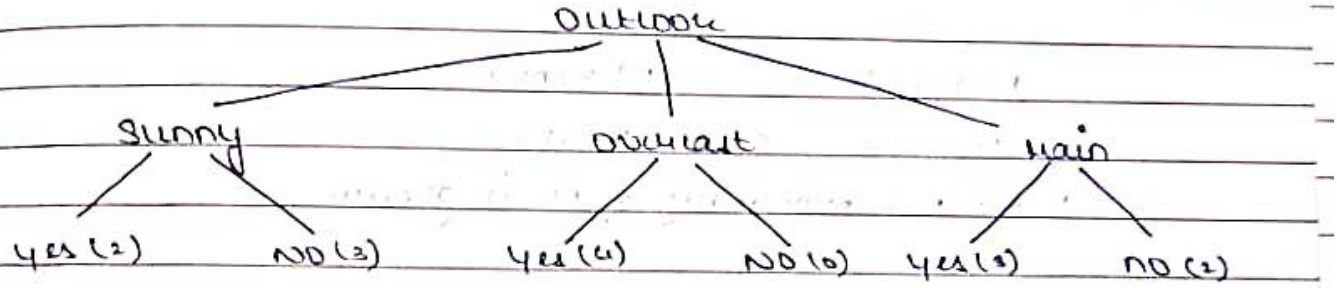
P(strong / NO) = 3/5





P(Hot/Yes) = 2/9 P(Cold/Yes) = 3/9 P(Mild/Yes) = 4/9

P(Hot/No) = 2/5 P(Cold/No) = 1/5 P(Mild/No) = 2/5



P(Sunny/Yes) = 2/9 P(Overcast/Yes) = 4/9 P(Rain/Yes) = 3/9

P(Sunny/No) = 3/5 P(Overcast/No) = 0/5 P(Rain/No) = 2/5

Outlook	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temp	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cold	3/9	1/5

Humidity	Yes	No
High	4/9	3/5
Low	5/9	2/5

Wind	Yes	No
Weak	3/9	3/5
Strong	6/9	2/5

$x = [\text{Sunny}, \text{Cold}, \text{High}, \text{Strong}]$

$P(x/Yes) = P(Yes) * [P(\text{Sunny}/Yes) * P(\text{Cold}/Yes) * P(\text{High}/Yes) * P(\text{Strong}/Yes)]$

$$= \frac{9}{14} * \left[\frac{2}{9} * \frac{3}{9} * \frac{4}{9} * \frac{3}{9}\right]$$

$$= 0.0070$$

$P(x/no) = P(no) * [P(sunny/no) * P(cold/no) * P(high/no) * P(strong/no)]$

$$= \frac{5}{14} * \left[\frac{3}{5} * \frac{1}{5} * \frac{3}{5} * \frac{3}{5}\right]$$

$$= 0.015$$

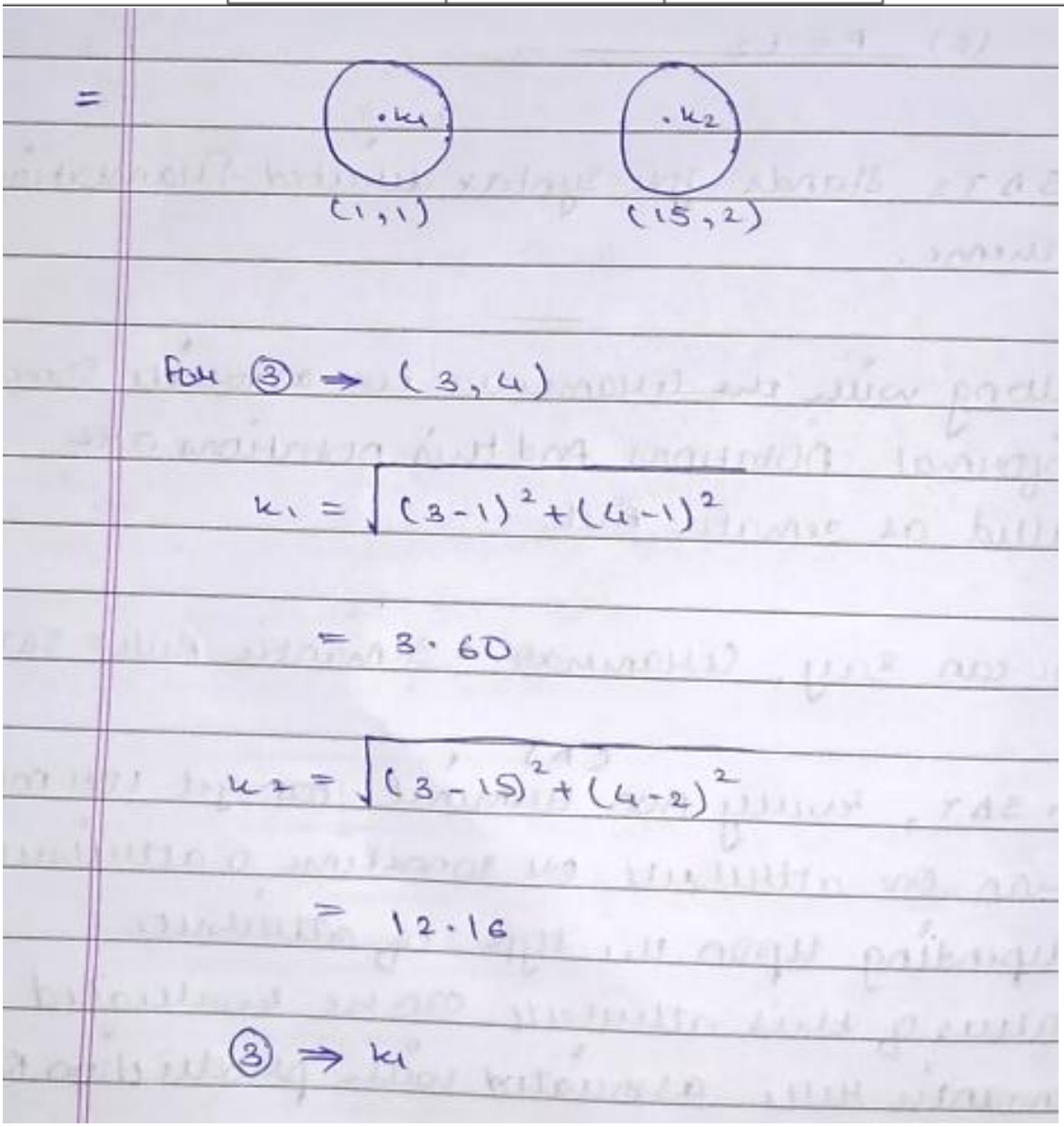
$P(x/yes)$		$P(x/no)$
0.0070	<	0.015

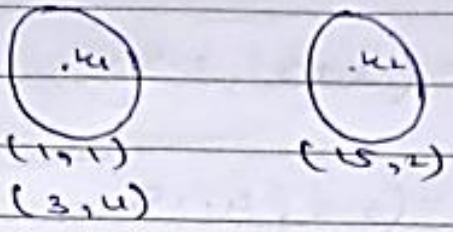
Hence the probability of no is greater.

a) Consider the following dataset consisting of the scores of two variables on each of seven subjects. Design K Means clustering for the data set.

Subject	A	B
1	1	1
2	15	2
3	3	4
4	5	7
5	3.5	5
6	4.4	5
7	3.5	4.5

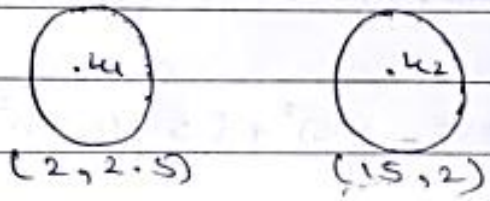
=





Mean of $k_1 = \left(\frac{1+3}{2}, \frac{1+4}{2}\right)$

$= (2, 2.5)$



For ④ $\Rightarrow (5, 7)$

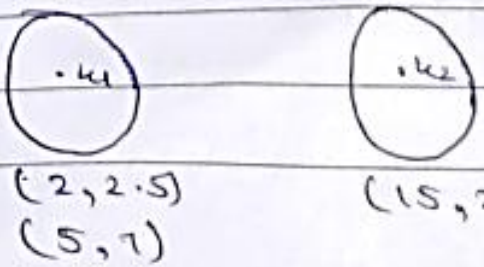
$k_1 = \sqrt{(5-2)^2 + (7-2.5)^2}$

$= 5.40$

$k_2 = \sqrt{(5-15)^2 + (07-2)^2}$

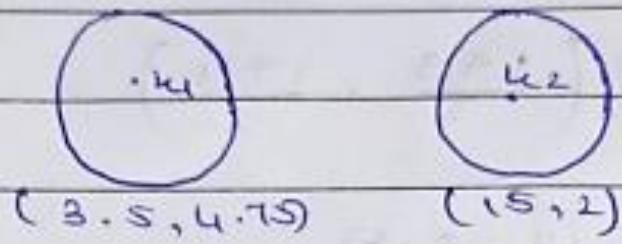
$= 11.18$

④ $\Rightarrow$ k1



Mean of $k_1 = \left(\frac{2+5}{2}, \frac{2.5+7}{2}\right)$

$= (3.5, 4.75)$



FOR (5) $\Rightarrow (3.5, 5)$

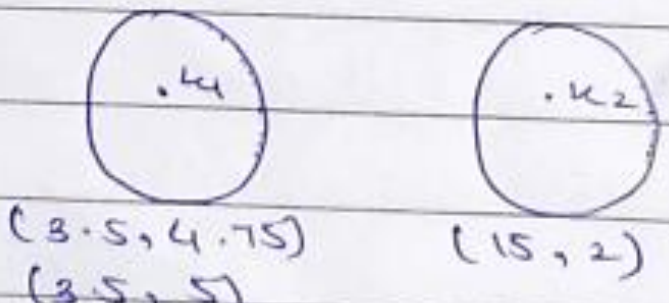
$$k_1 = \sqrt{(3.5 - 3.5)^2 + (5 - 4.75)^2}$$

$$= 0.25$$

$$k_2 = \sqrt{(3.5 - 15)^2 + (5 - 2)^2}$$

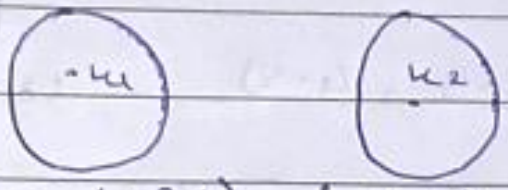
$$= 11.88$$

(5) $\Rightarrow k_1$



Mean of $u = \left(\frac{3.5+3.5}{2}, \frac{4.75+5}{2}\right)$

$= (3.5, 4.87)$



(3.5, 4.87) (15, 2)

FOH ⑥ $\Rightarrow$ (4.4, 5)

$u_1 = \sqrt{(4.4-3.5)^2 + [illegible](5-4.87)^2}$

$= 0.92$

$u_2 = \sqrt{(4.4-15)^2 + (\text{illegible}\ 5-2)^2}$

$= 11.01$

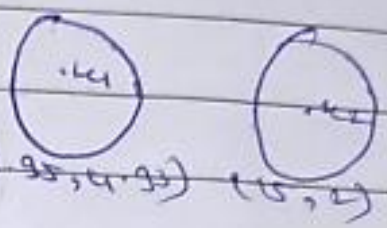
⑥ $\Rightarrow u$

(3.5, 4.87) (15, 2)

(4.4, 5)

Mean of $k_1 = \left(\frac{3.5 + 4.4}{2}, \frac{4.87 + 5}{2}\right)$

$= (3.95, 4.93)$



For ① $\Rightarrow$ (3.5, 4.5)

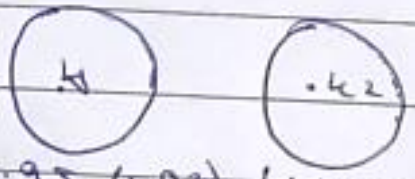
$$k_1 = \sqrt{(3.5 - 3.95)^2 + (4.5 - 4.93)^2}$$

$$= 0.62$$

$$k_2 = \sqrt{(3.5 - 15)^2 + (4.5 - 2)^2}$$

$$= 11.76$$

① $\Rightarrow k_1$



(3.5, 4.5)

Mean of $k_1 = \left(\frac{3.95 + 3.5}{2}, \frac{4.93 + 4.5}{2}\right)$

$= (3.725, 4.715)$

$k_1 = \{1, 3, 4, 5, 6, 7\}$

$k_2 = \{2\}$

Subject	A	B	k-mean
1	1	1	1
2	15	2	2
3	3	4	1
4	5	7	1
5	3.5	5	1
6	4.4	5	1
7	3.5	45	1

b) What is Frequent pattern mining and Association Rules? What is the use of both? Explain.

Association Rule:

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

The association rule learning is one of the very important concepts of machine learning, and it is employed in **Market Basket analysis, Web usage mining, continuous production, etc.** Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby.

Association rule learning can be divided into three types of algorithms:

1. **Apriori**
2. **Eclat**
3. **F-P Growth Algorithm**

To measure the associations between thousands of data items, there are several metrics. These metrics are given below:

- **Support**
- **Confidence**
- **Lift**

Let's understand each of them:

Support

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

$$\text{Supp(X)} = \frac{Freq(X)}{T}$$

Confidence

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

$$\text{Confidence} = \frac{Freq(X,Y)}{Freq(X)}$$

Lift

It is the strength of any rule, which can be defined as below formula:

$$\text{Lift} = \frac{Supp(X,Y)}{Supp(X) \times Supp(Y)}$$

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

- If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1**: It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1**: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

Uses of Association rules:

- **Market Basket Analysis:** It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
- **Medical Diagnosis:** With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
- **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
- It is also used for the **Catalog Design** and **Loss-leader Analysis** and many more other applications.

Frequent Pattern Mining:

- Frequent pattern mining in data mining is the process of identifying patterns or associations within a dataset that occur frequently. This is typically done by analyzing large datasets to find items or sets of items that appear together frequently.
- Frequent pattern mining is a major concern it plays a major role in associations and correlations and disclose an intrinsic and important property of dataset.
- Frequent patterns are patterns(such as items, subsequences, or substructures) that appear frequently in the database. It is an analytical process that finds frequent patterns, associations, or causal structures from databases in various databases. This process aims to find the frequently occurring item in a transaction.
- By frequent patterns, we can identify strongly correlated items together and we can identify similar characteristics and associations among them. By doing frequent data mining we can go further for clustering and association.
- Frequent data mining can be done by using association rules with particular algorithms eclat and apriori algorithms.
- Frequent pattern mining searches for recurring relationships in a data set.
- It also helps to find the inheritance regularities. to make fast processing software with a user interface and used for a long time without any error.

Advantages:

It can find useful information which is not visible in simple data browsing

It can find interesting association and correlation among data items

Disadvantages:

It can generate a large number of patterns

With high dimensionality, the number of patterns can be very large, making it difficult to interpret the results.

Uses:

- Market Basket Analysis: This is the process of analyzing customer purchasing patterns in order to identify items that are frequently bought together. This information can be used to optimize product placement, create targeted marketing campaigns, and make other business decisions.
- Recommender Systems: Frequent pattern mining can be used to identify patterns in user behavior and preferences in order to make personalized recommendations.
- Fraud Detection: Frequent pattern mining can be used to identify abnormal patterns of behavior that may indicate fraudulent activity.
- Network Intrusion Detection: Network administrators can use frequent pattern mining to detect patterns of network activity that may indicate a security threat.
- Medical Analysis: Frequent pattern mining can be used to identify patterns in medical data that may indicate a particular disease or condition.
- Text Mining: Frequent pattern mining can be used to identify patterns in text data, such as keywords or phrases that appear frequently together in a document.
- Web usage mining: Frequent pattern mining can be used to analyze patterns of user behavior on a website, such as which pages are visited most frequently or which links are clicked on most often.
- Gene Expression: Frequent pattern mining can be used to analyze patterns of gene expression in order to identify potential biomarkers for different diseases.

c) Explain the technique for improving efficiency of FP growth algorithm.

A database has five transactions. Let min sup = 2.

TID	items bought
T1	{A, B, C, D, E,}
T2	{B, C, D}
T3	{B, C, D, E}
T4	{A, B, C, D, E}
T5	{B, C, D, E}

Find all frequent itemsets using FP-growth Algorithm.

Mini Sup = 2

TID	Items Bought
T1	{A, B, C, D, E}
T2	{B, C, D}
T3	{B, C, D, E}
T4	{A, B, C, D, E}
T5	{B, C, D, E}

Step 1: Frequency and Priority.

Item	Frequency	Priority
A	2	5
B	5	1
C	5	2
D	5	3
E	4	4

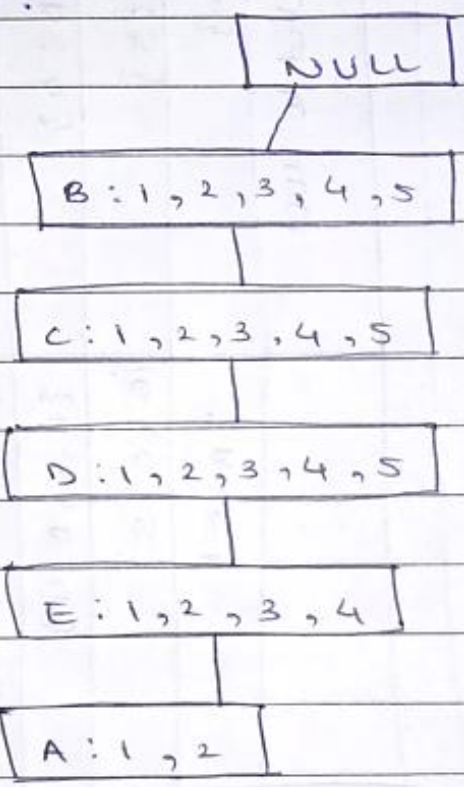
<1, 2, 3, 4, 5> the priorities

<B, C, D, E, A> Ordered.

Step 2: Ordered Itemset

TID	Items	Ordered Items
T_1	{A, B, C, D, E}	{B, C, D, E, A}
T_2	{B, C, D}	{B, C, D}
T_3	{B, C, D, E}	{B, C, D, E}
T_4	{A, B, C, D, E}	{B, C, D, E, A}
T_5	{B, C, D, E}	{B, C, D, E}

STEP 3 :



Step 4:

Item	conditional Path Base	condition Frequency Pattern Tree
A	{B, C, D, E : 2}	{B, C, D, E : 2}
E	{B, C, D : 4}	{B, C, D : 4}
D	{B, C : 5}	{B, C : 5}
C	{B : 5}	{B : 5}
B	connected to ROOT	

Frequent Pattern Generator

{<B, A : 2> <C, A : 2> <D, A : 2> <E, A : 2> <B, C, A : 2> <B, D, A : 2> <B, E, A : 2> <C, D, A : 2> <C, E, A : 2> <D, E, A : 2> <B, C, D, A : 2> <B, C, E, A : 2> <B, D, E, A : 2> <C, D, E, A : 2> <B, C, D, E, A : 2>}
{<B, E : 4> <C, E : 4> <D, E : 4> <B, C, E : 4> <C, D, E : 4> <B, D, E : 4> <B, C, D, E : 4>}
{<B, D : 5> <C, D : 5> <B, C, D : 5>}
{<B, C : 5>}

Consider Transactional data for an *AllElectronics* branch.

TID	*List of item IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2,

There are nine transactions in this database, that is, $|D| = 9$. Apply the Apriori algrithm for finding frequent itemsets in D. consider min support=2.

a)Explain Apriori Algorithm in detail.

Q.3) Explain Apriori Algo. in detail.

→ (1) Apriori algorithm refers to algo. which used to calculate association rule betn objects.

(2) It means how two or more objects are related to one another.

(3) The primary objective of the apriori algo. is in to create the association rule between different objects. generally

(4) Generally the apriori algo. is operated on a db that consist of huge no. of transaction.

(5) This algo. refers to an algo. that is used in mining frequent products sets and relevant association rules.

(6) It helps the customers to buy their products with ease and increases the sales performance of particular store.

(7) There are three component of apriori algo. they are:

- Support :→

It refers to the default popularity of any product. you find the support as a quotient of the division of the no. of transaction comprising that

that product by the total number of transaction

support = Transactions relating biscuit / total transaction (Biscuit)

= 400/4000 = 10 percent.

- Confidence :→

Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the no. of transactions that comprise both biscuit and chocolates by the total no. of transactions to get the confidence.

Confidence = Transactions relating both biscuit & chocolate / Total transactions involving biscuit

= 200/400

= 500 percent.

It means that 50 % of customers who bought biscuits, ~~bought~~ chocolates also.

- Lift :→

It refers to the increase in the ratio of the sale of chocolates when you sell biscuits. The mathematical equ:

Lift = (Confidence (Biscuit - chocolates)) / support (Biscuit)

= 50/10

= 5

It means that the probability of people buying both biscuit and chocolates together is five times more than that of purchasing the biscuit alone.

b)How FP growth algorithm works? Explain.

→ (1) The FP growth algo. is an alternative way to find frequent item sets without using candidate generations, thus improving performance.

(2) For so much, it uses a divide and conquer strategy.

(3) The core of this method is the usage of a special data structure named frequent-pattern tree, which retain the item set / association information.

* This algo. works as follows:

- First, it compresses the input database creating an FP tree instance to represent frequent item.
- After this, it divides the compressed db into a set of conditional db, each associated with one frequent pattern.
- Finally each such db is mined separately.

a) Explain Market Basket Analysis for mining frequent pattern set and association rules with suitable example.

* Market Basket analysis :→

(1) Technique which identifies the strength of association betn pairs of products purchase together and identify patterns of co-occurance.

(2) It is data mining technique used by retailers to increase sales for better understanding customer purchasing patterns.

(3) It analyse purchase that commonly happened together.

(4) Identifies customer buying habits by finding association betn the different items that customer ~~peop~~ buys in their shopping basket.

(5) This kind of association will be helpful for retailers to develop marketing strategies by gaining inside into which items are frequently got together by customers

(6) For example: people who buy bread and peanut butter also buy jelly.

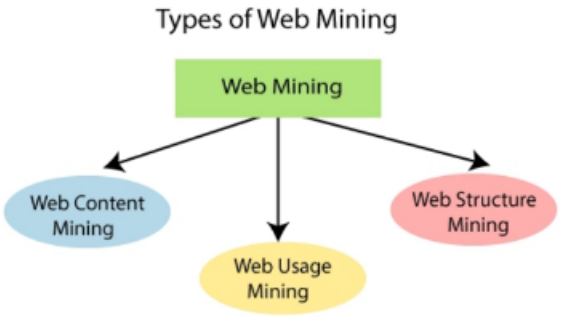
(7) market basket analysis creates if-then scenario rules that is for example if item A is purchased then item B is likely to be purchased.

(8) The rules are probabilistic, (they are derived from the frequencies of occurance in the observation).

Explain Web Mining in detail.

Web mining can widely be seen as the application of adapted data mining techniques to the web, whereas data mining is defined as the application of the algorithm to discover patterns on mostly structured data embedded into a **knowledge discovery process**. Web mining has a distinctive property to provide a set of various data types. The web has multiple aspects that yield different approaches for the mining process, such as web pages consist of text, web pages are linked via hyperlinks, and user activity can be monitored via web server logs. These three features lead to the differentiation between the three areas are web content mining, web structure mining, web usage mining.

There are three types of data mining:



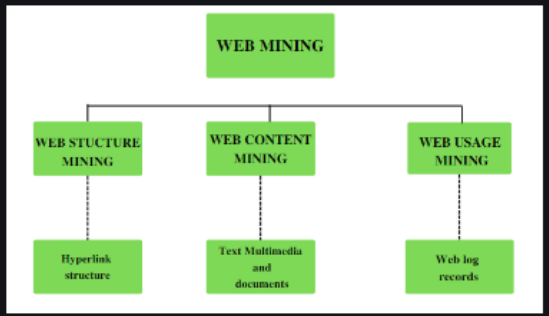
Applications of Web Mining:

- Web mining helps to improve the power of web search engines by classifying web documents and identifying web pages.
- it is used for Web Searching e.g., Google, Yahoo, etc, and Vertical Searching e.g., FatLens, Become, etc.
- Web mining is used to predict user behavior.
- Web mining is very useful for a particular Website and e-service e.g., landing page optimization.

Process of Web Mining:

Web Mining Process

Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.



Categories of Web Mining

Describe various Graph properties of Web.

How to access accuracy of text retrieval in text mining system?

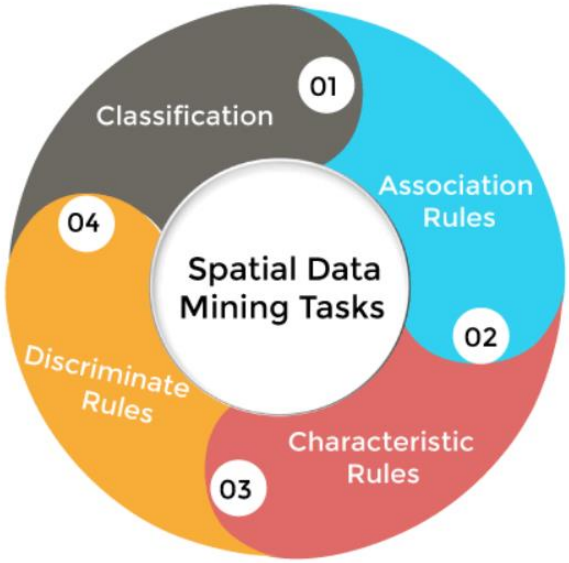
Explain Spatial Data Mining in detail.

1. Spatial Data Mining :

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. In spatial data mining analyst use geographical or spatial information to produce business intelligence or other results. Challenges involved in spatial data mining include identifying patterns or finding objects that are relevant to research project.

Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands the unification of data mining with spatial database technologies. It can be used for learning spatial records, discovering spatial relationships and relationships among spatial and nonspatial records, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries.

It is expected to have broad applications in geographic data systems, marketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are used.



Classification:

Classification determines a set of rules which find the class of the specified object as per its attributes.

Association rules:

Association rules determine rules from the data sets, and it describes patterns that are usually in the database.

Characteristic rules:

Characteristic rules describe some parts of the data set.

Discriminate rules:

As the name suggests, discriminate rules describe the differences between two parts of the database, such as calculating the difference between two cities as per employment rate.

What do you mean by web Data mining? Explain various steps involved in WDM.

Web Data Mining

- Web mining can widely be seen as the application of adapted data mining techniques to the web, whereas data mining is defined as the application of the algorithm to discover patterns on mostly structured data embedded into a knowledge discovery process.
- Web mining has a distinctive property to provide a set of various data types.
- The web has multiple aspects that yield different approaches for the mining process, such as web pages consist of text, web pages are linked via hyperlinks, and user activity can be monitored via web server logs.
- These three features lead to the differentiation between the three areas are web content mining, web structure mining, web usage mining.

How data mining is useful in customer relationship management in e-business world?

Explain the concept of visual web data mining in detail.

Differentiate temporal and spatial data mining in detail.

SNO.	Spatial data mining	Temporal data mining
1.	It requires space.	It requires time.
2.	Spatial mining is the extraction of knowledge/spatial relationship and interesting measures that are not explicitly stored in spatial database.	Temporal mining is the extraction of knowledge about occurrence of an event whether they follow Cyclic , Random ,Seasonal variations etc.
3.	It deals with spatial (location , Geo-referenced) data.	It deals with implicit or explicit Temporal content , from large quantities of data.
4.	Spatial databases reverses spatial objects derived by spatial data. types and spatial association among such objects.	Temporal data mining comprises the subject as well as its utilization in modification of fields.
5.	It includes finding characteristic rules, discriminant rules, association rules and evaluation rules etc.	It aims at mining new and unknown knowledge, which takes into account the temporal aspects of data.
6.	It is the method of identifying unusual and unexplored data but useful models from spatial databases.	It deals with useful knowledge from temporal data.
7.	Examples – Determining hotspots , Unusual locations.	Examples – An association rule which looks like – “Any Person who buys a car also buys steering lock”. By temporal aspect this rule would be – ” Any person who buys a car also buys a steering lock after that “.

Discuss the challenges that occurred during knowledge discovery on the web.

Discuss
i) Web Content Mining

ii) Web Usage Mining

iii) Web Structure Mining

iv) Visual Web Data Mining

1. Web Content Mining:

Web content mining can be used to extract useful data, information, knowledge from the web page content. In web content mining, each web page is considered as an individual document. The individual can take advantage of the semi-structured nature of web pages, as HTML provides information that concerns not only the layout but also logical structure. The primary task of content mining is data extraction, where structured data is extracted from unstructured websites. The objective is to facilitate data aggregation over various web sites by using the extracted structured data. Web content mining can be utilized to distinguish topics on the web. For Example, if any user searches for a specific task on the search engine, then the user will get a list of suggestions.

What is Web Content Mining?

Web Content Mining can be used for the mining of useful data, information, and knowledge from web page content. Web content mining performs scanning and mining of the text, images, and group of web pages according to the content of the input by displaying the list in search engines.

It is also quite different from data mining because web data are mainly semi-structured or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structured nature of the web, while text mining focuses on unstructured texts. Thus, Web content mining requires creative applications of data mining and text mining techniques and its own unique approaches.

In the past few years, there has been a rapid expansion of activities in the web content mining area. This is not surprising because of the phenomenal growth of web content and the significant economic benefit of such mining. However, due to the heterogeneity and the lack of structure of web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems. Web content mining could be differentiated from two approaches, such as:

1. Agent-based Approach

This approach involves intelligent systems. It aims to improve information finding and filtering. It usually relies on autonomous agents that can identify relevant websites. And it could be placed into the following three categories, such as:

- **Intelligent Search Agents:** These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
- **Information Filtering or Categorization:** These agents use information retrieval techniques and characteristics of open hypertext Web documents to retrieve automatically, filter, and categorize them.
- **Personalized Web Agents:** These agents learn user preferences and discover Web information based on other users' preferences with similar interests.

2. Data based approach

Data based approach is used to organize semi-structured data present on the internet into structured data. It aims to model the web data into a more structured form to apply standard database querying mechanisms and data mining applications to analyze it.

Web usage mining is used to extract useful data, information, knowledge from the weblog records, and assists in recognizing the user access patterns for web pages. In Mining, the usage of web resources, the individual is thinking about records of requests of visitors of a website, that are often collected as web server logs. While the content and structure of the collection of web pages follow the intentions of the authors of the pages, the individual requests demonstrate how the consumers see these pages. Web usage mining may disclose relationships that were not proposed by the creator of the pages.

Some of the methods to identify and analyze the web usage patterns are given below:

I. Session and visitor analysis:

The analysis of preprocessed data can be accomplished in session analysis, which incorporates the guest records, days, time, sessions, etc. This data can be utilized to analyze the visitor's behavior.

The document is created after this analysis, which contains the details of repeatedly visited web pages, common entry, and exit.

II. OLAP (Online Analytical Processing):

OLAP accomplishes a multidimensional analysis of advanced data.

OLAP can be accomplished on various parts of log related data in a specific period.

OLAP tools can be used to infer important business intelligence metrics

What is Web Usage Mining?

Web Usage Mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining, discovering user navigation patterns from web data, trying to discover useful information from the secondary data derived from users' interactions while surfing the web. Web usage mining collects the data from Weblog records to discover user access patterns of web pages. Several available research projects and commercial tools analyze those patterns for different purposes. The insight knowledge could be utilized in personalization, system improvement, site modification, business intelligence, and usage characterization.

What is Web Structure Mining?

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis has increased. These efforts resulted in a newly emerging research area called **Link Mining**, which is located at the intersection of the work in link analysis, hypertext, web mining, relational learning, inductive logic programming, and graph mining.

Web structure mining uses graph theory to analyze a website's node and connection structure. According to the type of web structural data, web structure mining can be divided into two kinds:

- **Extracting patterns from hyperlinks in the web:** a hyperlink is a structural component that connects the web page to a different location.
- **Mining the document structure:** analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

The web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are in, out, and co-citation (two pages linked to by the same page). Attributes include HTML tags, word appearances, and anchor texts. Web structure mining includes the following terminology, such as:

- **Web graph:**directed graph representing web.
- **Node:** web page in the graph.
- **Edge:** hyperlinks.
- **In degree:** the number of links pointing to a particular node.
- **Out degree:** number of links generated from a particular node.

An example of a technique of web structure mining is the **PageRank** algorithm used by Google to rank search results. A page's rank is decided by the number and quality of links pointing to the target node.

Link mining had produced some agitation on some traditional data mining tasks. Below we summarize some of these possible tasks of link mining which are applicable in Web structure mining, such as:

1. **Link-based Classification:** The most recent upgrade of a classic data mining task to linked Domains. The task is to predict the category of a web page based on words that occur on the page, links between pages, anchor text, html tags, and other possible attributes found on the web page.
2. **Link-based Cluster Analysis:** The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Unlike the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.
3. **Link Type:** There is a wide range of tasks concerning predicting the existence of links, such as predicting the type of link between two entities or predicting the purpose of a link.
4. **Link Strength:** Links could be associated with weights.
5. **Link Cardinality:** The main task is to predict the number of links between objects. page categorization used to

    - Finding related pages.
    - Finding duplicated websites and finding out the similarity between them.

Visual web Data Mining:

Visual data mining is an idea that uses recent technology to apply some specific principles to how humans interpret data. Data mining is the process of detecting patterns in a certain chunk of information. It is a very general method, applied in detailed ways to get specific results, in areas like finance, medicine, public administration and government, transportation, and much more.

Much of technology behind data mining has only been invented relatively recently, with the microcomputer being the most relevant example. Computers can collect and display a huge amount of data that, with data mining, can be interpreted in different ways. This is the power that data mining brings to the human community, and the potential that its practitioners are looking at for improving modern methodologies.

In visual data mining, programmers build interfaces that allow for visual presentations to be a part of how users interpret the data. The data might be places into graphs or charts so that users can spot patterns or outliers that wouldn't otherwise be immediately obvious. Scientists and programmers are still looking at the possibilities of this method and suggesting some recommendations for best possible techniques.

i) web content mining :→

- web content mining is referred to as text mining.
- Content mining is the browsing and mining of text images and graphs of a web pages to decide the relevance of the content to the search query.
- It can be defined as the phase of extracting essential data from standard language text.
- Some data that it can generate via text messages, files, emails, documents are written in common language text.
- Text mining can draw beneficial insights or patterns from such data.

ii) web usage mining :→

- It is used to derive useful data, information, knowledge from the weblog data and helps in identifying the user access designs for web pages.
- The management of web resources the individual is thinking about data of request of visitors of a websites that are composed as web server logs.
- web usage mining can disclose relationship that were not suggested by the designer of the pages.

iii) web structure mining :→

- It is tool that can recognize the relationship bet^n web pages linked by data or direct link connection.
- This structured data is discoverable by the provision of web structure schema through db technique for web pages.
- web mining can widely be viewed as the application of adapted data mining method to the web, whereas data mining is represented as the application of algo. to find patterns on mostly structure data fixed into a knowledge discovery process.
- Structure mining uses minimize two problem that first problem is irrelevant to search outcome and second is the inability to index the large amount of data supported on the web.