Model Question Paper -1
DATA MINING

Instructions to Candidates: 1. Answer any Four questions from each part.
2. Answer All Parts

PART-A

I. Answer any Four questions, each carries Two marks. (4 x 2 = 8)

1. What do you mean by Data Mining?

Ans: Definition: Data Mining is defined as the procedure of extracting information from huge sets of data.

In other words, data mining is mining knowledge from data.

Terminologies involved in data mining: Knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis etc.

2. Define Prediction.

Ans: **PREDICTION:**

To find a numerical output, prediction is used. The training dataset contains the inputs and numerical output values. According to the training dataset, the algorithm generates a model or predictor. When fresh data is provided, the model should find a numerical output. This approach, unlike classification, does not have a class label. A continuous-valued function or ordered value is predicted by the model.

Example: 1. Predicting the worth of a home based on facts like the number of rooms, total area, and so on.

3. Define Regression.

Ans: **REGRESSION IN DATA MINING:**

Regression refers to a data mining technique that is used to predict the numeric values in a given data set. Regression involves the technique of fitting a straight line or a curve on numerous data points.

For example, regression might be used to predict the product or service cost or other variables. It is also used in various industries for business and marketing behavior, trend analysis, and financial forecast.

Regression is divided into five different types

1. Linear Regression
2. Logistic Regression
3. Lasso Regression

4. Ridge Regression
5. Polynomial Regression

4. What do you mean by outliers?

Ans: outliers are sample points with values much different from those of the remaining set of data. Outliers may represent errors in the data or could be correct data values that are simply much different from the remaining data. A person who is 2.5 meters tall is much taller than most people. In analysing the height of individuals; this value probably would be viewed as an outlier. Some clustering techniques do not perform well with the presence of outliers.

5. What is Decision Tree?

Ans: A decision tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. It is a tree-like structure where each internal node tests on attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction.

6. What do you mean by Distributed Algorithm?

Ans: The distribution of sample data values has to do with the shape which refers to how data values are distributed across the range of values in the sample. In simple terms, it means if the values are clustered around the average to show how they are symmetrically arranged around it or if there are more values to one side than the order.

Two ways to explore the distribution of the sample data are

1. Graphically

   through shape statistics

PART-B

II. Answer any Four questions, each carries Five marks. (4 x 4 = 20)

7) What are the difference between Data Mining and knowledge discovery in databases?

Ans: **DATA MINING VS KDD.**

Key Features	Data Mining	KDD
Basic Definition	Data mining is the process of identifying patterns and extracting details about big data sets using intelligent methods.	The KDD method is a complex and iterative approach to knowledge extraction from big data.
Goal	To extract patterns from datasets.	To discover knowledge from datasets.
Scope	In the KDD method, the fourth phase is called "data mining."	KDD is a broad method that includes data mining as one of its steps.
Used Techniques	Classification Clustering Decision Trees Dimensionality Reduction Neural Networks Regression	Data cleaning Data Integration Data selection Data transformation Data mining Pattern evaluation Knowledge Presentation
Example	Clustering groups of data elements based on how similar they are.	Data analysis to find patterns and links.

8) What are the various issues associated with the Data Mining?

Ans: **FACTORS THAT CREATE SOME ISSUES.**

1. Mining Methodology and User Interaction issues
2. Performance Issues
3. Diverse Data Types Issues

MINING METHODOLOGY AND USER INTERACTION ISSUES:

1. **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
2. **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
3. **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
4. **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
5. **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
6. **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
7. **Pattern evaluation** − The patterns discovered should be interesting or relevant.

PERFORMANCE ISSUES :

1. **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

2. **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

DIVERSE DATA TYPES ISSUES :

1. **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data(data related to a specific location on the Earth's surface)etc. It is not possible for one system to mine all these kind of data.
2. **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

9) Write short note on K-Nearest Neighbors algorithm and its applications.

Ans: K Nearest Neighbors:

One common classification scheme based on the use of distance measures is that of the K nearest neighbors (KNN). The KNN technique assumes that the entire training set includes not only the data in the set but also the desired classification for each item. In effect, the training data become the model. When a classification is to be made for a new item, its distance to each item in the training set must be determined. Only the K closest entries in the training set are considered further. The new item is then placed in the class that contains the most items from this set of K closest items.

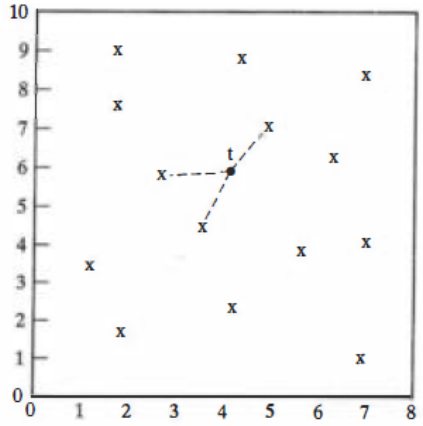


Fig: Classification using KNN

Here the points in the training set are shown and K = 3. The three closest items in the training set are shown; t will be placed in the class to which most of these are members. We use T to represent the training data. Since each tuple to be classified must be compared to each element in the training data, if there are q elements in the training set, this is O (q). Given n elements to be classified, this becomes an O (nq) problem. Given that the training data are of a constant size (although perhaps quite large), this can then be viewed as an O(n) problem.

```
Input:
   T      //Training data
   K      //Number of neighbors
   t      //Input tuple to classify
Output:
   c      //Class to which t is assigned
KNN algorithm:
          //Algorithm to classify tuple using KNN
   N = ∅;
             //Find set of neighbors, N, for t
   for each d ∈ T do
      if |N| ≤ K, then
         N = N ∪ {d};
      else
         if ∃ u ∈ N such that sim(t, u) ≤ sim(t, d), then
            begin
               N = N − {u};
               N = N ∪ {d};
            end
            //Find class for classification
   c = class to which the most u ∈ N are classified;
```

Applications

- Simplistic algorithm — uses only value of K (odd number) and the distance function (Euclidean, as mentioned today).
- Efficient method for small datasets.
- Utilises "Lazy Learning." In doing so, the training dataset is stored and is used only when making predictions therefore making it more quick than Support Vector Machines (SVMs) and Linear Regression.

10) Describe in detail one of the Decision Tree Algorithm give examples.

Ans: **Decision tree algorithm:**

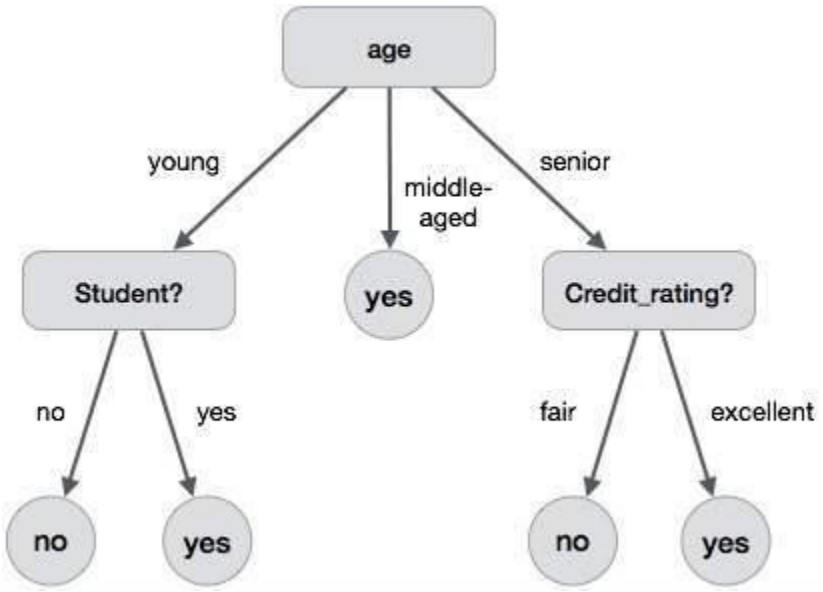
1. Begin with the entire dataset as the root node of the decision tree.
2. Determine the best attribute to split the dataset based on a given criterion,
3. Create a new internal node that corresponds to the best attribute and connects it to the root node.
4. Partition the dataset into subsets based on the values of the best attribute.
5. Recursively repeat steps 1-4 for each subset until all instances in a given subset belong to the same class or no further splitting is possible.
6. Assign a leaf node to each subset that contains instances that belong to the same class.

7. Make predictions based on the decision tree by traversing it from the root node to a leaf node that corresponds to the instance being classified.

The benefits of having a decision tree are

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

The following decision tree is for the concept to buy computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.



11) Explain Hierarchical clustering in detail.

Ans: **HIERARCHICAL ALGORITH MS**

As mentioned earlier, hierarchical clustering algorithms actually creates sets of clusters. Hierarchical algorithms differ in how the sets are created. A tree data structure, called a dendrogram, can be used to illustrate the hierarchical clustering technique and the sets of different clusters. The root in a dendrogram tree contains one cluster where all elements are together. The leaves in the dendrogram each consist of a single element cluster. Internal nodes in the

dendrogram represent new clusters formed by merging the clusters that appear as its children in the tree. Each level in the tree is associated with the distance measure that was used to merge the clusters. All clusters created at a particular level were combined because the children clusters had a distance between them less than the distance value associated with this level in the tree.

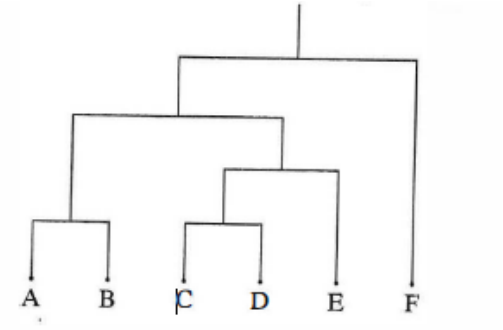


Fig: Dendrogram

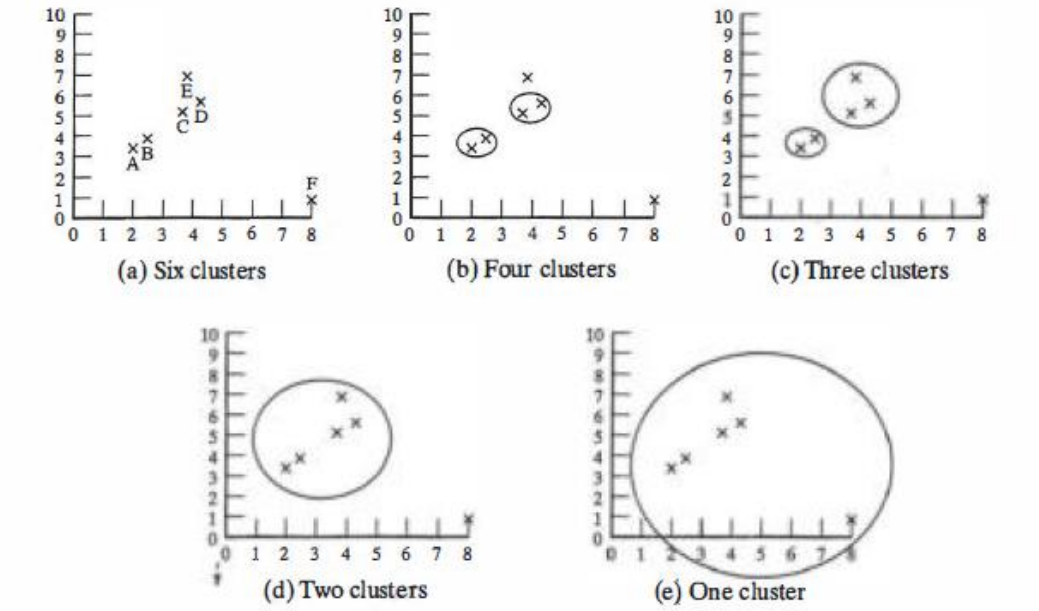


Fig: Five Levels of Clustering

shows six elements, {A, B, C, D, E, F}, to be clustered. Parts (a) to (e) of the figure show five different sets of clusters. In part (a) each cluster is viewed to consist of The space complexity for hierarchical algorithms is O (n2) because this is the space required for the adjacency matrix. The space required for the dendrogram is O (kn), which is much less than O (n2) . The time complexity for hierarchical algorithms is 0 (kn^2) because there is one iteration for each level in the dendrogram. Depending on the specific algorithm, however, this could actually be O (maxd n2) where maxd is the maximum distance between points. Different algorithms may actually merge the closest clusters from the next lowest level or simply create new clusters at each level with progressively larger distances. Hierarchical techniques are well suited for

many clustering applications that naturally exhibit a nesting relationship between clusters.

12) Write short note on Data Parallelism.

Ans:

Data Parallelisms

Data Parallelism means concurrent execution of the same task on each multiple computing core.

Let's take an example, summing the contents of an array of size N. For a single-core system, one thread would simply sum the elements [0] . . . [N − 1]. For a dual-core system, however, thread A, running on core 0, could sum the elements [0] . . . [N/2 − 1] and while thread B, running on core 1, could sum the elements [N/2] . . . [N − 1]. So the Two threads would be running in parallel on separate computing cores.

1. Same task are performed on different subsets of same data.
2. Synchronous computation is performed.
3. As there is only one execution thread operating on all sets of data, so the speedup is more.
4. Amount of parallelization is proportional to the input size.
5. It is designed for optimum load balance on multiprocessor system.

PART C

III. Answer any Four questions, each carries Five marks. (4 x 8 = 32)

13) How can you describe Data mining from the perspective of database?

Ans: **Data Mining from a Database Perspective.**

A data mining system can be classified according to the kinds of databases on which the data mining is performed. For example, a system is a relational data miner if it discovers knowledge from relational data, or an object-oriented one if it mines knowledge from object-oriented databases.

Database technology has b een successfully used in traditional busi-
ness data pro cessing. Companies have b een gathering a large amount

of data, using a DBMS system to manage it. Therefore, it is desirable that we have an easy and painless use of database technology within other areas, such as data mining.

DBMS technology o ers many features that make it valuable when implementing data mining applications. For example, it is p os-sible to work with data sets that are considerably larger than main memory, since the database itself is resp onsible for handling informa-tion, paging and swapping when necessary. Besides, a simpli ed data management and a closer integration to other systems are available (e.g. data may b e up dated or managed as a part of a larger op er-ational pro cess). Moreover, as emerging ob ject-relational databases are providing the ability to handle image, video and voice, there is a p otential area to exploit mining of complex data typ es. Finally, after rules are discovered, we can use ad-ho c and OLAP queries to validate discovered patterns in an easy way. We must not forget that information used during mining pro cessing is often con dential. Thus, DBMSs can also b e used as a means of providing data security, which is widely implemented in commercial databases, avoiding the need of using encryption algorithms to pro cess information

14) Write a short note on Scalable DT techniques.

Ans: refer notes

15) Explain how K-Means Clustering algorithm is working give examples.

Ans: **K Means Clustering:**

K- means is an iterative clustering algorithm in which items are moved among sets of clus-. ters until the desired set is reached. As such, it may be viewed as a type of squared error algorithm, although the convergence criteria need not be defined based on the squared error. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously

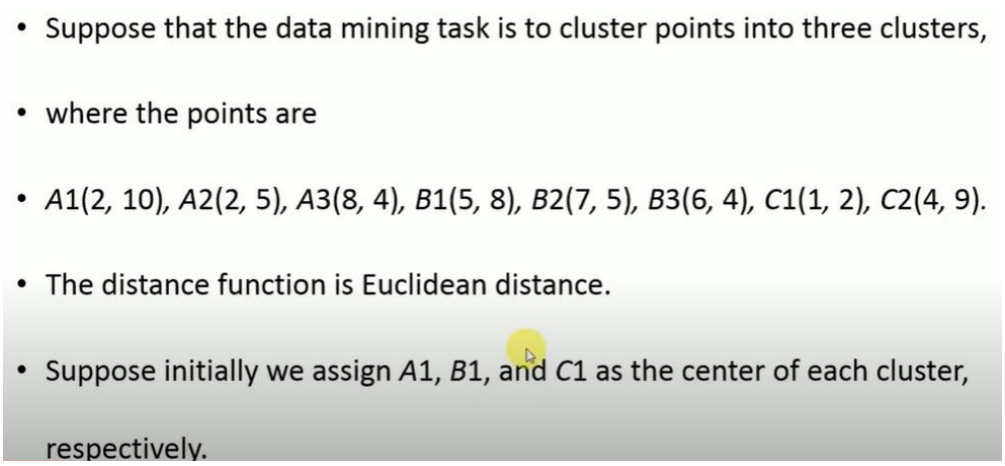
The *cluster mean* of $K_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$ is defined as

$$m_i = \frac{1}{m} \sum_{j=1}^{m} t_{ij}$$

```
   k      //Number of desired clusters
Output:
   K      //Set of clusters
K-means algorithm:
   assign initial values for means m1, m2, ..., mk;
   repeat
      assign each item ti to the cluster which has the closest mean;
      calculate new mean for each cluster;
   until convergence criteria is met;
```

The time complexity of K-means is O(tkn) where t is the number of iteratio ns.

K-means finds a local optimum and may actually miss the global optimum. K-⍰eans does not work on categorical data because the mean must be defined on the attnbute type.



16) Write a short note on hierarchical clustering.
Ans: repeated

17) What do you mean by Large item-sets explain in detail.
Ans: refer notes

18) What is Data Parallelism explain in detail?
Ans: repeated

Data Parallelisms

Data Parallelism means concurrent execution of the same task on each multiple computing core.

Let's take an example, summing the contents of an array of size N. For a single-core system, one thread would simply sum the elements [0] . . . [N − 1]. For a dual-core system, however, thread A, running on core 0, could sum the elements [0] . . . [N/2 − 1] and while thread B, running on core 1, could sum the elements [N/2] . . . [N − 1]. So the Two threads would be running in parallel on separate computing cores.

1. Same task are performed on different subsets of same data.
2. Synchronous computation is performed.
3. As there is only one execution thread operating on all sets of data, so the speedup is more.
4. Amount of parallelization is proportional to the input size.
5. It is designed for optimum load balance on multiprocessor system.

And refer notes;

Model Question Paper-2

DATA MINING

Instructions to Candidates: 1. Answer any Four questions from each part.
2. Answer All Parts

PART-A

I. Answer any Four questions, each carries Two marks. (4 x 2 = 8)

1. What do you mean by ETL process?

Ans: **ETL Tools** are applications/platforms that enable users to execute ETL processes. In simple terms, these tools help businesses move data from one or many desparate data sources to a destination. These help in making the data both digestible and accessible (and in turn analysis-ready) in the desired location – often a data warehouse.

ETL tools are the first essential step in the data warehousing process that eventually make more informed decisions in less time.

2. Define Regression and its types.

- Ans: **Regression**

Regression is a statistical tool that helps determine the cause and effect relationship between the variables. It determines the relationship between a dependent and an independent variable. It is generally used to predict future trends and events.

Regression is divided into five different types

1. Linear Regression
2. Logistic Regression
3. Lasso Regression
4. Ridge Regression
5. Polynomial Regression

3. How will you solve Classification problem?

Ans: The decision tree approach is most useful in classification problems. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in a classification for that tuple. There are two basic steps in the technique: building the tree and applying the tree to the database

4. What do you mean by outliers?
   Ans: repeated
5. What is CART classification?

Ans: CART is a predictive algorithm used in Machine learning and it explains how the target variable's values can be predicted based on other matters. It is a decision tree where each fork is split into a predictor variable and each node has a prediction for the target variable at the end.

6. What do you mean by Distributed Algorithm?

Ans: The distribution of sample data values has to do with the shape which refers to how data values are distributed across the range of values in the sample. In simple terms, it means if the values are clustered around the average to show how they are symmetrically arranged around it or if there are more values to one side than the order.

Two ways to explore the distribution of the sample data are

2. Graphically
3. through shape statistics.

PART-B

II. Answer any Four questions, each carries Five marks. (4 x 4 = 20)

7) What are the difference between Data Mining and knowledge discovery in databases?

Ans: **DATA MINING VS KDD.**

Key Features	**Data Mining**	**KDD**
Basic Definition	Data mining is the process of identifying patterns and extracting details about big data sets using intelligent methods.	The KDD method is a complex and iterative approach to knowledge extraction from big data.

Key Features	Data Mining	KDD
Goal	To extract patterns from datasets.	To discover knowledge from datasets.
Scope	In the KDD method, the fourth phase is called "data mining."	KDD is a broad method that includes data mining as one of its steps.
Used Techniques	Classification Clustering Decision Trees Dimensionality Reduction Neural Networks Regression	Data cleaning Data Integration Data selection Data transformation Data mining Pattern evaluation Knowledge Presentation
Example	Clustering groups of data elements based on how similar they are.	Data analysis to find patterns and links.

8) Explain Naive Bayesian method.

Ans: **Bayesian classification:**

Bayesian classification uses Bayes theorem to predict the occurrence of any event. Bayesian classifiers are the statistical classifiers with the Bayesian probability understandings. The theory expresses how a level of belief, expressed as a probability.

Bayes theorem came into existence after Thomas Bayes, who first utilized conditional probability to provide an algorithm that uses evidence to calculate limits on an unknown parameter.

Bayes's theorem is expressed mathematically by the following equation that is given below.

$$P(X/Y) = \frac{P(Y/X)P(X)}{P(Y)}$$

Where X and Y are the events and P (Y) ≠ 0

P(X/Y) is a **conditional probability** that describes the occurrence of event **X** is given that **Y** is true.

P(Y/X) is a **conditional probability** that describes the occurrence of event **Y** is given that **X** is true.

P(X) and P(Y) are the probabilities of observing X and Y independently of each other. This is known as the **marginal probability**.

Bayesian interpretation:

In the Bayesian interpretation, probability determines a "**degree of belief**." Bayes theorem connects the degree of belief in a hypothesis before and after accounting for evidence. For example, Lets us consider an example of the coin. If we toss a coin, then we get either heads or tails, and the percent of occurrence of either heads and tails is 50%. If the coin is flipped numbers of times, and the outcomes are observed, the degree of belief may rise, fall, or remain the same depending on the outcomes.

For proposition X and evidence Y,

- P(X), the prior, is the primary degree of belief in X
- P(X/Y), the posterior is the degree of belief having accounted for Y.
- The quotient $\frac{P(Y/X)}{P(Y)}$ represents the supports Y provides for X.

Bayes theorem can be derived from the conditional probability:

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}, \text{ if } P(Y) \neq 0$$

$$P(Y/X) = \frac{P(Y \cap X)}{P(X)}, \text{ if } P(X) \neq 0$$

Where P (X∩Y) is the **joint probability** of both X and Y being true, because

$$P(Y \cap X) = P(X \cap Y)$$

$$\text{or, } P(X \cap Y) = P(X/Y)P(Y) = P(Y/X)P(X)$$

$$\text{or, } P(X/Y) = \frac{P(Y/X)P(X)}{P(Y)}, \text{ if } P(Y) \neq 0$$

Although the naive Bayes approach is straightforward to use, it does not always yield satisfactory results. First, the attributes usually are not independent. We could use a subset of the attributes by ignoring any that are dependent on others. The technique does not handle continuous data.

9) Write a short note on Data Mining tasks.

Ans: 1. Classification:

This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

Data mining techniques can be classified by different criteria

2. Clustering:

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. In other words, we can say that Clustering analysis is a data mining technique to identify similar data. This technique helps to recognize the differences and similarities between the data. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

3. Regression:

Regression analysis is the data mining process ,used to identify and analyze the relationship between variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modeling. For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand, and competition. Primarily it gives the exact relationship between two or more variables in the given data set.

4. Association Rules:

This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set.

Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases., For example, a list of grocery items that you have been buying for the last six months. It calculates a percentage of items being purchased together.

These are three major measurements technique: Lift,Support,Confidence.

5. Outlier detection:

This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behavior. This technique may be used in various domains like intrusion, detection, fraud detection, etc. It is also known as Outlier Analysis or Outlier mining. The outlier is a data point that diverges too much from the rest of the dataset. The majority of the real-world datasets have an outlier. Outlier detection plays a significant role in the data mining field. Outlier detection is valuable in numerous fields like network interruption identification, credit or debit card fraud detection, detecting outlying in wireless sensor network data, etc.

6. Sequential Patterns:

The sequential pattern is a data mining technique specialized for evaluating sequential data to discover sequential patterns. It comprises of finding interesting subsequences in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc.

In other words, this technique of data mining helps to discover or recognize similar patterns in transaction data over some time.

7. Prediction:

Prediction uses a combination of other data mining techniques such as trends, clustering, classification, etc. It analyzes past events or instances in the right sequence to predict a future event.

10) Describe in detail one of the Decision Tree Algorithm give examples.

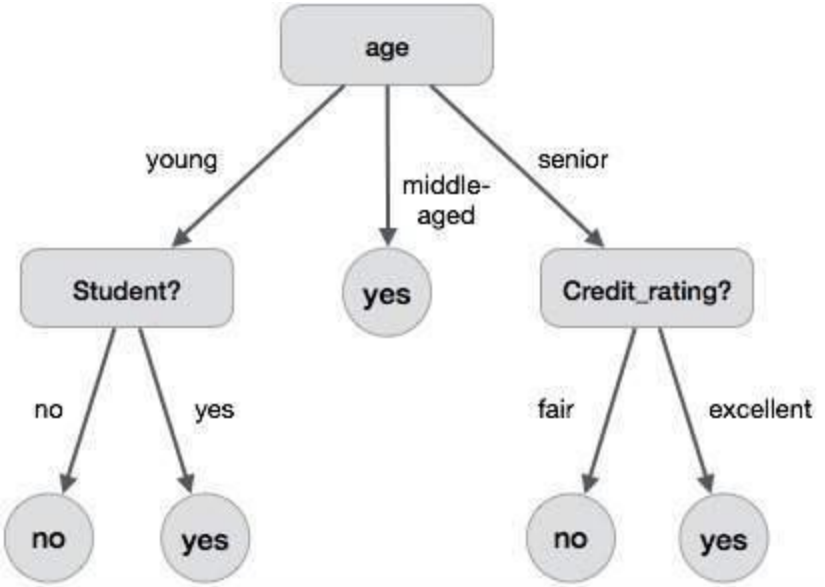
Ans: **Decision tree algorithm:**

8. Begin with the entire dataset as the root node of the decision tree.
9. Determine the best attribute to split the dataset based on a given criterion,
10. Create a new internal node that corresponds to the best attribute and connects it to the root node.
11. Partition the dataset into subsets based on the values of the best attribute.
12. Recursively repeat steps 1-4 for each subset until all instances in a given subset belong to the same class or no further splitting is possible.
13. Assign a leaf node to each subset that contains instances that belong to the same class.
14. Make predictions based on the decision tree by traversing it from the root node to a leaf node that corresponds to the instance being classified.

The benefits of having a decision tree are

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

The following decision tree is for the concept to buy computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.



11) Explain Hierarchical clustering in detail.

Ans: **HIERARCHICAL ALGORITH MS**

As mentioned earlier, hierarchical clustering algorithms actually creates sets of clusters. Hierarchical algorithms differ in how the sets are created. A tree data structure, called a dendrogram, can be used to illustrate the hierarchical clustering technique and the sets of different clusters. The root in a dendrogram tree contains one cluster where all elements are together. The leaves in the dendrogram each consist of a single element cluster. Internal nodes in the dendrogram represent new clusters formed by merging the clusters that appear as its children in the tree. Each level in the tree is associated with the distance measure that was used to merge the clusters. All clusters created at a particular level were combined because the children clusters had a distance between them less than the distance value associated with this level in the tree.

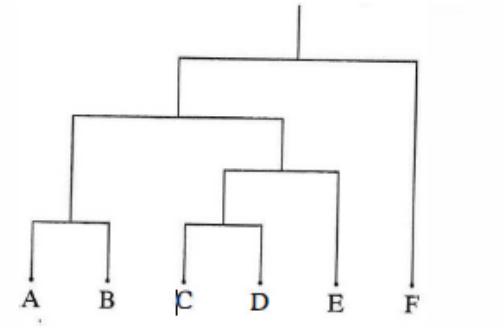


Fig: Dendrogram

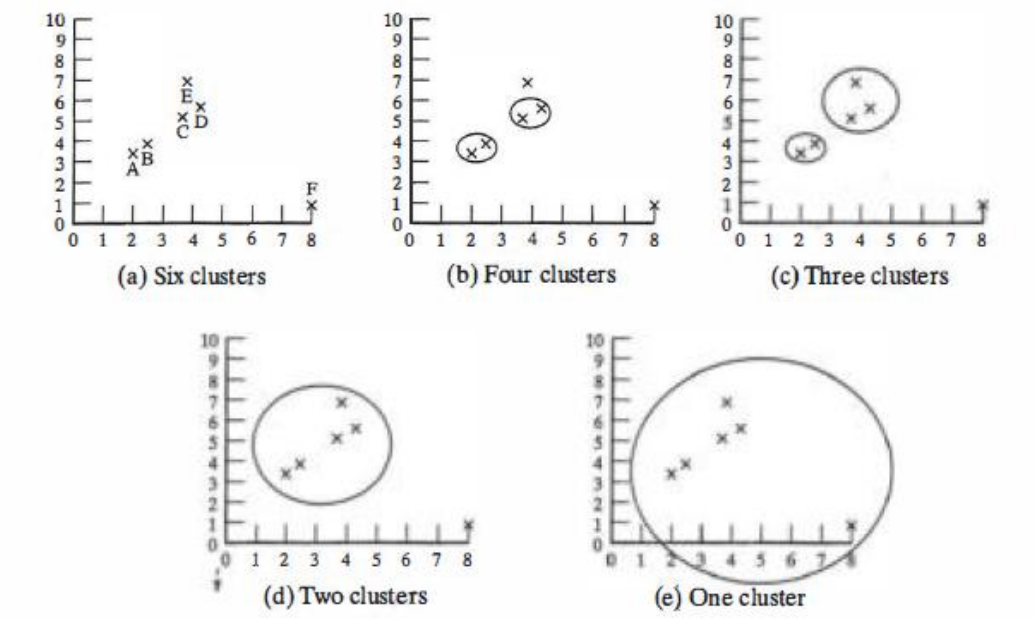


Fig: Five Levels of Clustering

shows six elements, {A, B, C, D, E, F}, to be clustered. Parts (a) to (e) of the figure show five different sets of clusters. In part (a) each cluster is viewed to consist of The space complexity for hierarchical algorithms is O (n2) because this is the space required for the adjacency matrix. The space required for the dendrogram is O (kn), which is much less than O (n2) . The time complexity for hierarchical algorithms is 0 (kn^2) because there is one iteration for each level in the dendrogram. Depending on the specific algorithm, however, this could actually be O (maxd n2) where maxd is the maximum distance between points. Different algorithms may actually merge the closest clusters from the next lowest level or simply create new clusters at each level with progressively larger distances. Hierarchical techniques are well suited for many clustering applications that naturally exhibit a nesting relationship between clusters.

12) Write a short note on Data warehouse.

Ans: A **data warehouse**, or enterprise data warehouse (EDW), is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, artificial intelligence (AI), and machine learning.

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Using Data Warehouse Information

There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be used in any of the following domains

Tuning Production Strategies − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.

Customer Analysis − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.

Operations Analysis − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

FUNCTIONS OF DATA WAREHOUSE TOOLS AND UTILITIES:

- **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning** − Involves finding and correcting the errors in data.
- **Data Transformation** − Involves converting the data from legacy format to warehouse format.
- **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- **Refreshing** − Involves updating from data sources to warehouse.

PART C

III. Answer any Four questions, each carries Five marks. (4 x 8 = 32)

13) How can you describe Data mining from the perspective of a database?

Ans: **Data Mining from a Database Perspective.**

A data mining system can be classified according to the kinds of databases on which the data mining is performed. For example, a system is a relational data miner if it discovers knowledge from relational data, or an object-oriented one if it mines knowledge from object-oriented databases.

Statistical Methods in Data Mining

Data mining refers to extracting or mining knowledge from large amounts of data. In other words, data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns. Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective, and accurate. Any situation can be analyzed in two ways in data mining:

1. **Non-statistical Analysis:** This analysis provides generalized information and includes sound, still images, and moving images.
2. **Statistical Analysis:** In statistics, data is collected, analyzed, explored, and presented to identify patterns and trends. Alternatively, it is referred to as quantitative analysis. It is the analysis of raw data using mathematical formulas, models, and techniques. Through the use of statistical methods, information is extracted from research data, and different ways are available to judge the robustness of research outputs. It is created for the effective handling of large amounts of data that are generally multidimensional and possibly of several complex types.

14) Write a short note on Scalable DT techniques.

Ans: Refer notes

15) Explain how K-Means Clustering algorithm is working and give examples.

Ans: **K Means Clustering:**

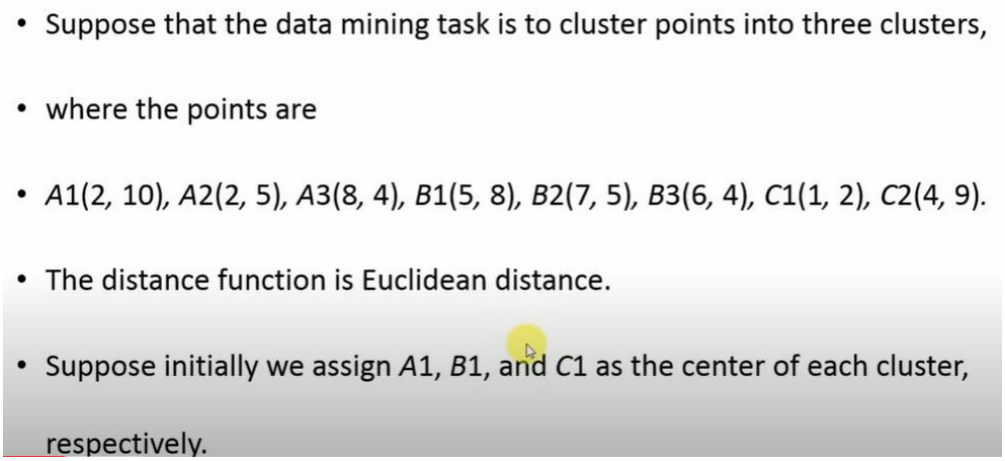
K- means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. As such, it may be viewed as a type of squared error algorithm, although the convergence criteria need not be defined based on the squared error. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously

The *cluster mean* of $K_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$ is defined as

$$m_i = \frac{1}{m} \sum_{j=1}^{m} t_{ij}$$

```
   k      //Number of desired clusters
Output:
   K      //Set of clusters
K-means algorithm:
   assign initial values for means m1, m2, ..., mk;
   repeat
      assign each item ti to the cluster which has the closest mean;
      calculate new mean for each cluster;
   until convergence criteria is met;
```

The time complexity of K-means is O(tkn) where t is the number of iteratio ns.
K-means finds a local optimum and may actually miss the global optimum. K-⍰eans
does not work on categorical data because the mean must be defined on the attnbute



type.

16) Write a short note on clustering techniques.

Ans: Clustering is similar to classification in that data are grouped. However, unlike classification,
the groups are not predefined. Instead, the grouping is accomplished by finding similarities between data according to characteristics found in the actual data. The groups are called clusters.

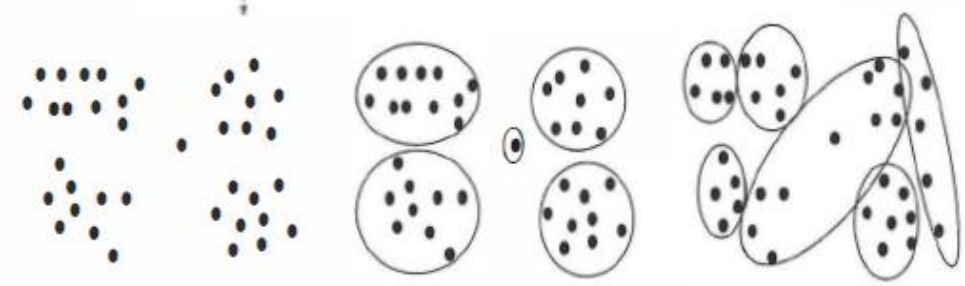
-Set of like elements. Elements from different clusters are not alike.

-The distance between points in a cluster is less than the distance between a point in the cluster and any point outside it.

A term similar to clustering is database segmentation, where like tuples (records) in a database are grouped together. This is done to partition or segment the database into components that then give the user a more general view of the data. This example illustrates the fact that determining how to do the clustering is not straightforward

TABLE 5.1: Sample Data for Example 5.1

Income	Age	Children	Marital Status	Education
$25,000	35	3	Single	High school
$15,000	25	1	Married	High school
$20,000	40	0	Single	High school
$30,000	20	0	Divorced	High school
$20,000	25	3	Divorced	College
$70,000	60	0	Married	College
$90,000	30	0	Married	Graduate school
$200,000	45	5	Married	Graduate school
$100,000	50	2	Divorced	College



(a) Group of homes (b) Geographic distance-based (c) Size-based

Clustering has been used in many application domains, including biology, medicine, anthropology, marketing, and economics. Clustering applications include plant and animal classification, disease classification, image processing, pattern recognition, and document retrieval. One of the first domains in which clustering was used was biological taxonomy.

When clustering is applied to a real-world database, many interesting problems occur:

- Outlier handling is difficult. Here the elements do not naturally fall into any cluster
- Dynamic data in the database implies that cluster membership may change over time
    - Interpreting the semantic meaning of each cluster may be difficult. With classification, the labeling of the classes is known ahead of time. However, with clustering, this may not be the case. Thus, when the clustering process finishes creating a set of clusters, the exact meaning of each cluster may not be obvious. Here There is no one correct answer to a clustering problem. In fact, many answers may be

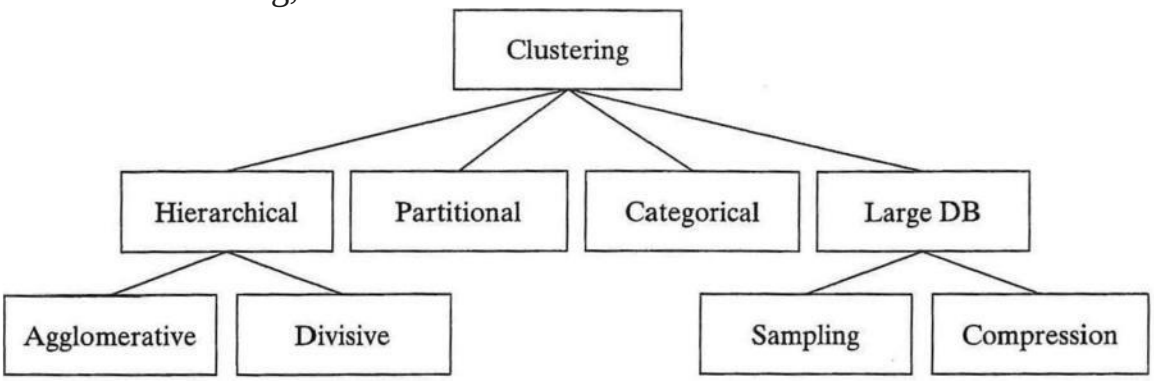
found. The exact number of clusters required is not easy to determine. Again, a domain expert may be required.

- Another related issue is what data should be used for clustering. Unlike learning

during a classification process, where there is some a priori knowledge concerning what the attributes of each classification should be, in clustering we have no supervised learning to aid the process. Indeed, clustering can be viewed as similar to unsupervised learning.

DEFINITION 5.1. Given a database $D = \{t_1, t_2, \ldots, t_n\}$ of tuples and an integer value k, the **clustering problem** is to define a mapping $f : D \rightarrow \{1, \ldots, k\}$ where each t_i is assigned to one cluster K_j, $1 \leq j \leq k$. A **cluster**, K_j, contains precisely those tuples mapped to it; that is, $K_j = \{t_i \mid f(t_i) = K_j, 1 \leq i \leq n, \text{ and } t_i \in D\}$.

A classification of the different types of clustering algorithms is shown Clustering algorithms themselves may be viewed as hierarchical or partitional. With hierarchical clustering, a nested set of clusters is created.



17) Explain apriori algorithm..

Ans: **Apriori Algorithm – Frequent Pattern Algorithms**

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It was later improved by R Agarwal and R Srikant and came to be known as Apriori. This algorithm uses two steps "join" and "prune" to reduce the search space. It is an iterative approach to discover the most frequent itemsets.

Apriori says:

The probability that item I is not frequent is if:

- P(I) < minimum support threshold, then I is not frequent.

- P (I+A) < minimum support threshold, then I+A is not frequent, where A also belongs to itemset.
- If an itemset set has value less than minimum support then all of its supersets will also fall below min support, and thus can be ignored. This property is called the Antimonotone property.

The steps followed in the Apriori Algorithm of data mining are:

1. **Join Step**: This step generates (K+1) itemset from K-itemsets by joining each item with itself.
2. **Prune Step**: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

Steps In Apriori

Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

#1) In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.

#2) Let there be some minimum support, min_sup (eg 2). The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min_sup, are taken ahead for the next iteration and the others are pruned.

#3) Next, 2-itemset frequent items with min_sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.

#4) The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.

#5) The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min_sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned.

#6) Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min_sup criteria. The algorithm is stopped when the most frequent itemset is achieved.

Advantages

1. Easy to understand algorithm
2. Join and Prune steps are easy to implement on large itemsets in large databases

Disadvantages

1. It requires high computation if the itemsets are very large and the minimum support is kept very low.
2. The entire database needs to be scanned.
3. FPM has many applications in the field of data analysis, software bugs, cross-marketing, sale campaign analysis, market basket analysis, etc.

18) What is Data Parallelism explain in detail?

Ans:

Data Parallelisms

Data Parallelism means concurrent execution of the same task on each multiple computing core.

Let's take an example, summing the contents of an array of size N. For a single-core system, one thread would simply sum the elements [0] . . . [N − 1]. For a dual-core system, however, thread A, running on core 0, could sum the elements [0] . . . [N/2 − 1] and while thread B, running on core 1, could sum the elements [N/2] . . . [N − 1]. So the Two threads would be running in parallel on separate computing cores.

1. Same task are performed on different subsets of same data.
2. Synchronous computation is performed.
3. As there is only one execution thread operating on all sets of data, so the speedup is more.
4. Amount of parallelization is proportional to the input size.
5. It is designed for optimum load balance on multiprocessor system.