**Model Question Paper -1**

**DATA MINING**

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

2. Answer All Parts

**PART-A**

1. **Answer any Four questions, each carries Two marks. ( 4 x 2 = 8 )**
2. **What do you mean by Data Mining?**

**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.

1. **Define Prediction.**

To find a numerical output, prediction is used. The training dataset contains the inputs and numerical output values

1. **Define Regression.**

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.

1. **What do you mean by outliers?**

an Outlier is an observation in a given dataset that lies far from the rest of the observations. That means an outlier is vastly larger or smaller than the remaining values in the set

1. **What is Decision Tree?**

A Decision Tree is a popular machine learning algorithm used for both classification and regression tasks. It is a tree-like structure that represents a series of decisions and their possible outcomes.

1. **What do you mean by Distributed Algorithm?**

a distributed algorithm refers to an algorithmic approach that is designed to operate on a distributed computing environment. Distributed data mining involves the analysis of large datasets that are distributed across multiple nodes or computing device

**PART-B**

1. **Answer any Four questions, each carries Five marks. ( 4 x 4 = 20 )**
2. **What are the difference between Data Mining and knowledge discovery in databases?**

| **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 |

1. **What are the various issues associated with the Data Mining?**

Data mining, while a powerful and valuable tool for extracting patterns and insights from large datasets, is associated with several challenges and issues. Here are some of the key issues:

1. **Data Quality:**
   * **Incomplete Data:** Missing values or incomplete records in the dataset can affect the accuracy and reliability of mining results.
   * **Noisy Data:** Data with errors or outliers can lead to incorrect conclusions and models.
2. **Data Preprocessing:**
   * **Data Cleaning:** Cleaning and handling missing values, outliers, and inconsistencies require careful preprocessing.
   * **Data Transformation:** Converting data into suitable formats and scales is often necessary for effective mining.
3. **Dimensionality:**
   * **Curse of Dimensionality:** As the number of dimensions (features) increases, the data becomes sparse, and the computational complexity grows exponentially. This can lead to overfitting and difficulty in finding meaningful patterns.
4. **Computational Complexity:**
   * **Scalability:** Mining large datasets can be computationally intensive and may require scalable algorithms and distributed computing resources.
   * **Algorithmic Complexity:** Some data mining algorithms have high time and space complexity, making them impractical for large datasets.
5. **Overfitting and Generalization:**
   * **Overfitting:** Creating models that are too complex may capture noise in the data, leading to poor generalization on new, unseen data.
   * **Underfitting:** Overly simplistic models may fail to capture underlying patterns in the data.
6. **Privacy Concerns:**
   * **Data Sensitivity:** Mining sensitive or personal information can raise privacy concerns. Privacy-preserving techniques are necessary to address these issues.
   * **Anonymization:** Anonymizing data to protect privacy may lead to a loss of information or the possibility of re-identification.
7. **Bias and Fairness:**
   * **Algorithmic Bias:** Biases in training data or algorithms can result in unfair or discriminatory outcomes.
   * **Fairness Concerns:** Ensuring fairness in data mining results, especially in applications like hiring or lending decisions, is a critical ethical consideration.
8. **Interpretability and Explainability:**
   * **Black Box Models:** Complex models, such as deep learning models, can be challenging to interpret, making it difficult to understand how and why specific decisions are made.
   * **Model Explainability:** The lack of interpretability can be a barrier to gaining user trust and meeting regulatory requirements.
9. **Dynamic Nature of Data:**
   * **Concept Drift:** In dynamic environments, the relationships in the data may change over time. Models need to adapt to these changes to remain accurate.
10. **Knowledge Overload:**
    * **Information Overload:** Extracting meaningful patterns from vast amounts of data can result in information overload, making it challenging to identify truly valuable insights.
11. **Ethical Concerns:**
    * **Misuse of Results:** The potential misuse of data mining results for unethical purposes, such as surveillance or profiling, raises ethical concerns.
    * **Informed Consent:** Ensuring that individuals are informed about the use of their data for mining is an ethical consideration.
12. Write short note on K-Nearest Neighbors algorithm and its applications.

**( in unit 2 pg no 4)**

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

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.

**Example of a decision tree**

Suppose we want to build a decision tree to predict whether a person is likely to buy a new car based on their demographic and behavior data. The decision tree starts with the root node, which represents the entire dataset. The root node splits the dataset based on the “income” attribute. If the person’s income is less than or equal to $50,000, the decision tree follows the left branch, and if the income is greater than $50,000, the decision tree follows the right branch.

The left branch leads to a node that represents the “age” attribute. If the person’s age is less than or equal to 30, the decision tree follows the left branch, and if the age is greater than 30, the decision tree follows the right branch. The right branch leads to a leaf node that predicts that the person is unlikely to buy a new car.

The left branch leads to another node that represents the “education” attribute. If the person’s education level is less than or equal to high school, the decision tree follows the left branch, and if the education level is greater than high school, the decision tree follows the right branch. The left branch leads to a leaf node that predicts that the person is unlikely to buy a new car. The right branch leads to another node that represents the “credit score” attribute. If the person’s credit score is less than or equal to 650, the decision tree follows the left branch, and if the credit score is greater than 650, the decision tree follows the right branch. The left branch leads to a leaf node that predicts that the person is unlikely to buy a new car. The right branch leads to a leaf node that predicts that the person is likely to buy a new car.

In summary, a decision tree is a graphical representation of all the possible outcomes of a decision based on the input data. It is a powerful tool for modeling and predicting outcomes in a wide range of domains, including business, finance, healthcare, and more

**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.

.

**Explain Hierarchical clustering in detail.**

**Hierarchical Clustering Methods**

the given set of an object of data is created into a kind of hierarchical decomposition. The formation of hierarchical decomposition will decide the purposes of classification.

There are two types of approaches for the creation of hierarchical decomposition:

**1. Divisive Approach**

Another name for the Divisive approach is a top-down approach. At the beginning of this method, all the data objects are kept in the same cluster. Smaller clusters are created by splitting the group by using the continuous iteration. The constant iteration method will keep on going until the condition of termination is met. One cannot undo after the group is split or merged, and that is why this method is not so flexible.

**2. Agglomerative Approach**

Another name for this approach is the bottom-up approach. All the groups are separated in the beginning. Then it keeps on merging until all the groups are merged, or condition of termination is met.

**3. Density-Based Clustering Method**

In this method of clustering in Data Mining, density is the main focus. The notion of mass is used as the basis for this clustering method. In this clustering method, the cluster will keep on growing continuously. At least one number of points should be there in the radius of the group for each point of data.

**4. Grid-Based Clustering Method**

In this type of Grid-Based Clustering Method, a grid is formed using the object together. A Grid Structure is formed by quantifying the object space into a finite number of cells.

**5. Model-Based Clustering Methods**

In this type of clustering method, every cluster is hypothesized so that it can find the data which is best suited for the model. The density function is clustered to locate the group in this method.

**6. Constraint-Based Clustering Method**

Application or user-oriented constraints are incorporated to perform the clustering. The expectation of the user is referred to as the constraint. In this process of grouping, communication is very interactive, which is provided by the restrictions.

.

| **Write short note on Data Parallelism** |
| --- |
| 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**

1. **Answer any Four questions, each carries Five marks. ( 4 x 8 = 32 )**
2. **How can you describe Data mining from the perspective of database?**

**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.

**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.

**Types of Statistical Analysis**

1. [**Descriptive statistical analysis**](https://www.simplilearn.com/what-is-descriptive-statistics-article)involves collecting, interpreting, analyzing, and summarizing data to present them in the form of charts, graphs, and tables. Rather than drawing conclusions, it simply makes the complex data easy to read and understand.
2. **The**[**inferential statistical analysis**](https://www.simplilearn.com/inferential-statistics-article) focuses on drawing meaningful conclusions on the basis of the data analyzed. It studies the relationship between different variables or makes predictions for the whole population.
3. **Predictive statistical analysis** is a type of statistical analysis that analyzes data to derive past trends and predict future events on the basis of them. It uses [machine learning](https://www.simplilearn.com/tutorials/machine-learning-tutorial/statistics-for-machine-learning) algorithms, [data mining](https://www.simplilearn.com/what-is-data-mining-article), [data modelling](https://www.simplilearn.com/what-is-data-modeling-article), and [artificial intelligence](https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/what-is-artificial-intelligence) to conduct the statistical analysis of data.
4. **The prescriptive analysis** conducts the analysis of data and prescribes the best course of action based on the results. It is a type of statistical analysis that helps you make an informed decision.
5. [**Exploratory analysis**](https://www.simplilearn.com/exploratory-data-analysis-article) is similar to inferential analysis, but the difference is that it involves exploring the unknown data associations. It analyzes the potential relationships within the data.
6. **The causal statistical analysis** focuses on determining the cause and effect relationship between different variables within the raw data. In simple words, it determines why something happens and its effect on other variables. This methodology can be used by businesses to determine the reason for failure.

**Statistical Analysis Methods**

**Mean:**[Mean or average mean](https://www.simplilearn.com/tutorials/data-analytics-tutorial/measures-of-central-tendency) is one of the most popular methods of statistical analysis. Mean determines the overall trend of the data and is very simple to calculate.

**Standard Deviation:-**Standard deviation is another very widely used statistical tool or method. It analyzes the deviation of different data points from the mean of the entire data set

**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

1. **Write a short note on Scalable DT techniques**.

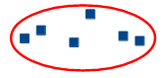
Scalability is a crucial aspect in data mining, especially when dealing with large datasets. Decision Trees (DT) are popular in data mining due to their simplicity and interpretability. Here are some scalable techniques and strategies for Decision Trees in the context of data mining:

1. **Parallelization and Distributed Computing:**
   * Use parallel and distributed computing frameworks to build decision trees. Algorithms like Hadoop and Spark provide the infrastructure to distribute the computation across multiple nodes, making it feasible to handle large datasets.
2. **Random Forests:**
   * Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions. This approach often leads to more accurate and robust models. Random Forest can be easily parallelized, and building multiple trees in parallel is a common strategy for scalability.
3. **Incremental Learning:**
   * Instead of building a decision tree on the entire dataset at once, consider an incremental learning approach. This involves updating the decision tree as new data arrives or in batches. Techniques like Online Random Forests adapt well to incremental learning.
4. **Feature Selection and Dimensionality Reduction:**
   * Large datasets often come with a high dimensionality. Use techniques like feature selection or dimensionality reduction (e.g., Principal Component Analysis) to reduce the number of features and focus on the most relevant ones. This can significantly speed up the tree-building process.
5. **Optimized Splitting Criteria:**
   * Choose splitting criteria that are computationally efficient. While Gini impurity and information gain are common measures, there may be alternatives that are more efficient for large datasets. Consider using approximate algorithms or techniques that require less computational resources.
6. **Tree Pruning and Depth Limiting:**
   * Prune decision trees to avoid overfitting and reduce their size. Limiting the depth of the trees can also improve scalability. Smaller trees are quicker to build and evaluate.
7. **Sampling Techniques:**
   * Instead of using the entire dataset, work with a representative sample. This is particularly useful when the dataset is too large to handle in memory. Techniques like bagging and boosting leverage sampling for improved scalability.
8. **Optimized Data Structures:**
   * Utilize efficient data structures for storing and manipulating decision trees. In-memory structures that allow for quick traversal and modification are crucial for scalability. Consider using sparse representations if the data is sparse.
9. **GPU Acceleration:**
   * Leverage Graphics Processing Units (GPUs) to accelerate the computation of decision trees. Some libraries and frameworks support GPU acceleration, providing a significant speedup for certain operations.
10. **Caching and Memoization:**
    * Cache intermediate results during tree construction to avoid redundant computations. Memoization techniques can be employed to store and reuse intermediate results, especially during the evaluation of splitting criteria.

**Explain how K-Means Clustering algorithm is working give examples** :

**First Property of K-Means Clustering Algorithm**

**All the data points in a cluster should be similar to each other.**

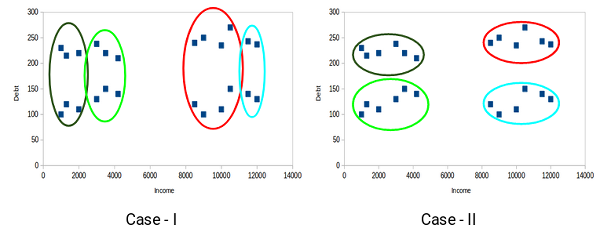
Example:[](https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-from-2019-08-08-14-49-11.png)

If the customers in a particular cluster are not similar to each other, then their requirements might vary, right? If the bank gives them the same offer, they might not like it, and their interest in the bank might reduce. Not ideal.

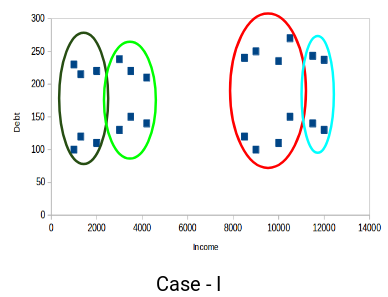
Having similar data points within the same cluster helps the bank to use targeted marketing. You can think of similar examples from your everyday life and consider how clustering will (or already does) impact the business strategy.

**Second Property of K-Means Clustering Algorithm**

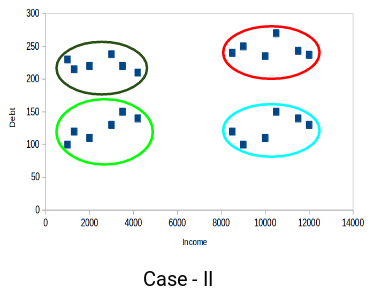
**The data points from different clusters should be as different as possible.** This will intuitively make sense if you’ve grasped the above property. Let’s again take the same example to understand this property:

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-from-2019-08-08-14-51-31.png)

Which of these cases do you think will give us the better clusters? If you look at case I:

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-from-2019-08-08-14-52-26.png)

Customers in the red and blue clusters are quite similar to each other. The top four points in the red cluster share similar properties to those of the blue cluster’s top two customers. They have high incomes and high debt values. Here, we have clustered them differently. Whereas, if you look at case II:

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2019/08/Screenshot-from-2019-08-08-14-52-58.png)

.

1. **Write a short note on hierarchical clustering.**

**Hierarchical Clustering in Data Mining:**

Hierarchical clustering is a versatile technique in data mining that facilitates the exploration of inherent structures within a dataset. This method groups similar data points into clusters, creating a hierarchical representation of relationships. Here are key aspects of hierarchical clustering in the context of data mining:

**1. Hierarchical Structure:**

* One of the distinctive features of hierarchical clustering is its ability to organize data in a hierarchical tree-like structure, known as a dendrogram. The dendrogram visually represents the merging (agglomerative) or splitting (divisive) of clusters at different similarity levels.

**2. Agglomerative and Divisive Approaches:**

* In the agglomerative approach, each data point begins as a separate cluster, and pairs of clusters are iteratively merged based on their similarity. The divisive approach starts with all data points in a single cluster and recursively splits them into smaller clusters until each data point is in its own cluster.

**3. Similarity Measures:**

* The choice of a similarity metric or distance measure is crucial in hierarchical clustering. Common metrics include Euclidean distance, Manhattan distance, and correlation coefficient. The selection depends on the nature of the data and the specific goals of the analysis.

**4. Linkage Methods:**

* Linkage methods determine how the distance between clusters is calculated and play a significant role in hierarchical clustering. Common linkage methods include:
  + **Single Linkage:** Minimizes the distance between the closest members of two clusters.
  + **Complete Linkage:** Maximizes the distance between the farthest members of two clusters.
  + **Average Linkage:** Averages the distances between all pairs of members from two clusters.
  + **Ward's Method:** Minimizes the variance within clusters.

**5. Interpretability and Visualization:**

* Hierarchical clustering provides an interpretable output in the form of a dendrogram, which allows analysts to explore the structure of the data at different levels of granularity. Visualization aids in understanding the relationships between clusters and identifying natural groupings.

**6. Dynamic Cluster Exploration:**

* One advantage of hierarchical clustering is its ability to dynamically explore clusters at various granularity levels. Analysts can cut the dendrogram at different heights to obtain different numbers of clusters, providing flexibility in the analysis.

**7. Applications in Data Mining:**

* Hierarchical clustering finds applications in various data mining tasks, including:
  + **Customer Segmentation:** Identifying groups of customers with similar behaviors or preferences.
  + **Document Clustering:** Organizing documents based on content similarities.
  + **Biology and Genomics:** Analyzing gene expression patterns and taxonomy.
  + **Image Analysis:** Grouping similar images based on visual features.

**8. Challenges and Scalability:**

* While hierarchical clustering is powerful for smaller datasets, it may face scalability challenges with large datasets. Efficient algorithms and parallel computing techniques are often employed to address these challenges.

1. **What do you mean by Large item-sets explain in detail.**

In data mining, large itemsets refer to sets of items that appear frequently in a given dataset. The concept is commonly associated with association rule mining, a technique used to discover interesting relationships or patterns in large datasets. Large itemsets play a crucial role in the first step of association rule mining, where the goal is to identify sets of items that occur together frequently.

Let's break down the key components of this concept:

1. **Itemsets:**
   * An itemset is a collection of items. In the context of data mining, items could represent products, services, or any other entities of interest in a dataset.
   * There are two types of itemsets:
     + **Frequent Itemsets:** These are sets of items that occur in the dataset with a frequency above a specified threshold.
     + **Infrequent Itemsets:** These are sets of items that occur less frequently and are typically filtered out in the early stages of the mining process.
2. **Support:**
   * The support of an itemset is a measure of how frequently the itemset appears in the dataset.
   * Mathematically, support is defined as the proportion of transactions in the dataset that contain the itemset.
   * Support is usually expressed as a percentage or a fraction.
3. **Large Itemsets:**
   * Large itemsets are frequent itemsets that meet a predefined support threshold. The support threshold is a user-defined parameter that determines what is considered "frequent."
   * The process of finding large itemsets involves scanning the dataset to identify sets of items that meet or exceed the specified support threshold.
4. **Association Rule Mining:**
   * Once large itemsets are identified, the next step is to generate association rules. An association rule is a relationship between sets of items in the dataset.
   * Association rules are typically represented in the form "if X, then Y," indicating that the occurrence of items in set X is associated with the occurrence of items in set Y.
   * The strength of association rules is often measured using metrics like confidence and lift.

In summary, large itemsets are a fundamental concept in association rule mining, helping to identify patterns and relationships within a dataset by focusing on sets of items that occur together frequently. The support threshold is a crucial parameter in this process, as it determines the minimum level of occurrence required for an itemset to be considered "large" or "frequent."

Top of Form

1. **What is Data Parallelism explain in detail?**

Data parallelism in the context of data mining refers to the parallelization of computational tasks by dividing the data into smaller chunks and processing these chunks concurrently across multiple processors or computing nodes. This parallel processing approach aims to improve the efficiency and speed of data mining algorithms, especially when dealing with large datasets. Let's delve into the details of data parallelism in data mining:

1. **Parallel Processing:**
   * In data mining, various algorithms involve processing large volumes of data to discover patterns, associations, or other valuable insights.
   * Parallel processing divides the workload into smaller, manageable tasks that can be executed simultaneously on multiple processors or computing units.
2. **Data Partitioning:**
   * Data parallelism involves partitioning the dataset into subsets, and each subset is processed independently by a separate computing unit.
   * The dataset can be divided based on different criteria, such as splitting it into equal-sized partitions or using a more sophisticated approach that considers the characteristics of the data.
3. **Parallel Algorithms:**
   * To leverage data parallelism, data mining algorithms need to be designed or modified to work in a parallelized environment.
   * Certain algorithms are naturally suited for parallelization, while others may require adaptation to exploit parallel processing capabilities efficiently.
4. **MapReduce Framework:**
   * The MapReduce programming model is often employed for implementing data parallelism in large-scale data processing, including data mining tasks.
   * MapReduce divides the computation into two phases: the "Map" phase, where data is processed in parallel across multiple nodes, and the "Reduce" phase, where the results are aggregated.
5. **Distributed Computing:**
   * Data parallelism often involves distributed computing, where computing nodes are spread across a cluster of machines.
   * Each node works independently on its assigned subset of data, and the results are later combined to obtain the final output.
6. **Advantages:**
   * **Scalability:** Data parallelism allows for the efficient scaling of computations as the size of the dataset grows.
   * **Speedup:** By processing data in parallel, the overall processing time can be significantly reduced compared to a sequential approach.
   * **Resource Utilization:** Utilizing multiple computing resources concurrently helps maximize the utilization of available hardware.
7. **Challenges:**
   * **Communication Overhead:** Coordinating the results from parallel processes can introduce communication overhead, impacting performance.
   * **Load Balancing:** Ensuring that the workload is evenly distributed among computing nodes is crucial to prevent some nodes from becoming bottlenecks.
8. **Applications in Data Mining:**
   * Data parallelism is applicable to various data mining tasks, such as clustering, classification, and association rule mining.
   * Algorithms like k-means clustering, decision trees, and frequent itemset mining can be parallelized using data parallelism.