**Assignment S2**

**Module 2: Introduction of Data Mining**

|  |  |
| --- | --- |
| 1. Discuss KDD.   Ans.KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process in data mining typically involves the following steps:   1. **Selection**: Select a relevant subset of the data for analysis. 2. **Pre-processing:** Clean and transform the data to make it ready for analysis. This may include tasks such as data normalization, missing value handling, and data integration. 3. **Transformation:**Transform the data into a format suitable for data mining, such as a matrix or a graph. 4. **Data Mining:** Apply data mining techniques and algorithms to the data to extract useful information and insights. This may include tasks such as clustering, classification, association rule mining, and anomaly detection. 5. **Interpretation**: Interpret the results and extract knowledge from the data. This may include tasks such as visualizing the results, evaluating the quality of the discovered patterns, and identifying relationships and associations among the data. 6. **Evaluation**: Evaluate the results to ensure that the extracted knowledge is useful, accurate, and meaningful. 7. **Deployment**: Use the discovered knowledge to solve the business problem and make decisions. |  |
| 1. What is data mining? What are its various techniques?   Ans. Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It can be used in a variety of ways, such as database marketing, credit risk management, [fraud detection](https://www.investopedia.com/financial-edge/0512/the-most-common-types-of-consumer-fraud.aspx), spam Email filtering, or even to discern the sentiment or opinion of users.  **Data Mining Techniques**  Data mining uses algorithms and various techniques to convert large collections of data into useful output. The most popular types of data mining techniques include:   * **Association rules**, also referred to as market basket analysis, searches for relationships between variables. This relationship in itself creates additional value within the data set as it strives to link pieces of data. For example, association rules would search a company's [sales](https://www.investopedia.com/terms/s/sale.asp)history to see which products are most commonly purchased together; with this information, stores can plan, promote, and forecast accordingly. * **Classification** uses predefined classes to assign to objects. These classes describe characteristics of items or represent what the data points have in common with each. This data mining technique allows the underlying data to be more neatly categorized and summarized across similar features or product lines. * **Clustering** is similar to classification. However, clustering identified similarities between objects, then groups those items based on what makes them different from other items. While classification may result in groups such as "shampoo", "conditioner", "soap", and "toothpaste", clustering may identify groups such as "hair care" and "dental health". * [**Decision trees**](https://www.investopedia.com/articles/financial-theory/11/decisions-trees-finance.asp) are used to classify or predict an outcome based on a set list of criteria or decisions. A decision tree is used to ask for input of a series of cascading questions that sort the dataset based on responses given. Sometimes depicted as a tree-like visual, a decision tree allows for specific direction and user input when drilling deeper into the data. * **K-Nearest Neighbor (KNN)** is an algorithm that classifies data based on its proximity to other data. The basis for KNN is rooted in the assumption that data points that are close to each are more similar to each other than other bits of data. This non-parametric, supervised technique is used to predict features of a group based on individual data points. * **Neural networks** process data through the use of nodes. These nodes is comprised of inputs, weights, and an output. Data is mapped through supervised learning (similar to how the human brain is interconnected). This model can be fit to give threshold values to determine a model's accuracy. * **Predictive analysis** strives to leverage historical information to build graphical or mathematical models to forecast future outcomes. Overlapping with [regression analysis](https://www.investopedia.com/terms/r/regression.asp), this data mining technique aims at supporting an unknown figure in the future based on current data on hand. |  |
| 1. “\_\_\_\_\_\_\_\_\_ is a common technique that separates data points into different classes”. Clarify it. | 9 |
| 1. A summarization of the general features of a target class of data is \_\_\_\_\_\_\_\_\_\_\_\_. Discuss. |  |
| 1. Differentiate between discrimination and characterization.   Ans. Data Characterization is all about slicing and dicing the data to understand what it is all about. Without understanding the contents in the data, practically the data is unusable, despite having lot of potential. Data mining builds on the results of characterization.  Data Discrimination is identifying splitting conditions to partition the data into independent bins. Although data discrimination is a part of characterization, the objectives are very different. In general, classifier tools could be used to perform the discrimination task. |  |
| 1. Discuss support and confidence in terms of the transactional dataset.  Ans. Confidence In data mining, confidence is a measure of the reliability or support for a given association rule. It is defined as the proportion of cases in which the association rule holds true, or in other words, the percentage of times that the items in the antecedent (the “if” part of the rule) appear in the same transaction as the items in the consequent (the “then” part of the rule). Support In data mining, support refers to the relative frequency of an item set in a dataset. For example, if an itemset occurs in 5% of the transactions in a dataset, it has a support of 5%. Support is often used as a threshold for identifying frequent item sets in a dataset, which can be used to generate association rules. For example, if we set the support threshold to 5%, then any itemset that occurs in more than 5% of the transactions in the dataset will be considered a frequent itemset. |  |
| 1. Briefly describe the following advanced database systems and applications: object-relational databases, spatial databases, text databases, multimedia databases, and the World Wide Web. |  |
| 1. Discuss cluster analysis and how it can help to detect outliers.  Ans. Cluster-Based****Approaches for detecting Outliers:**** Clustering-based outlier detection methods assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters. Clustering-based approaches detect outliers by extracting the relationship between Objects and Cluster. An object is an outlier if   * Does the object belong to any cluster? If not, then it is identified as an outlier. * Is there a large distance between the object and the cluster to which it is closest? If yes, it is an outlier. * Is the object part of a small or sparse cluster? If yes, then all the objects in that cluster are outliers.  **Checking an outlier:**  * To check the objects that do not belong to any cluster we go with [DENSITY BASED CLUSTERING (DBSCAN)](https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/) * To check outlier detection using distance to the closest cluster we go with [K-MEANS CLUSTERING (K-Means)](https://www.geeksforgeeks.org/k-means-clustering-introduction/) |  |
| 1. Describe challenges to data mining regarding data mining methodology and user interaction issues. |  |
| 1. Discuss various Applications of data mining.   **Data Mining Applications**   1. Financial Data Analysis. 2. Retail Industry. 3. Telecommunication Industry. 4. Biological Data Analysis. 5. Other Scientific Applications. 6. Intrusion Detection. |  |
| 1. Explain various types of learning.  Supervised Learning**Overview:** Supervised learning is a type of machine learning that uses labeled data to train machine learning models. In labeled data, the output is already known. The model just needs to map the inputs to the respective outputs.  An example of supervised learning is to train a system that identifies the image of an animal. Unsupervised Learning**Overview:** Unsupervised learning is a type of machine learning that uses unlabeled data to train machines. Unlabeled data doesn’t have a fixed output variable. The model learns from the data, discovers the patterns and features in the data, and returns the output. Reinforcement Learning**Overview** Reinforcement Learning trains a machine to take suitable actions and maximize its rewards in a particular situation. It uses an agent and an environment to produce actions and rewards. The agent has a start and an end state. But, there might be different paths for reaching the end state, like a maze. In this learning technique, there is no predefined target variable. |  |
| 1. Write short note on: 2. Classification   Classification in data mining is a common technique that separates data points into different classes. It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones.  It primarily involves using algorithms that you can easily modify to improve the data quality.   1. Association Rule Mining   Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis.  Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.   1. Difference between regression and classification   Regression algorithms predict a continuous value based on the input variables. The main goal of regression problems is to estimate a mapping function based on the input and output variables. If your target variable is a quantity like income, scores, height or weight, or the probability of a binary category (like the probability of rain in particular regions), then you should use the regression model.  Classification is a predictive model that approximates a mapping function from input variables to identify discrete output variables, which can be labels or categories. The mapping function of classification algorithms is responsible for predicting the label or category of the given input variables. A classification algorithm can have both discrete and real-valued variables, but it requires that the examples be classified into one of two or more classes. |  |