MODULE 4: INTRODUCTION TO DATA MINING

WHAT IS DATA MINING?

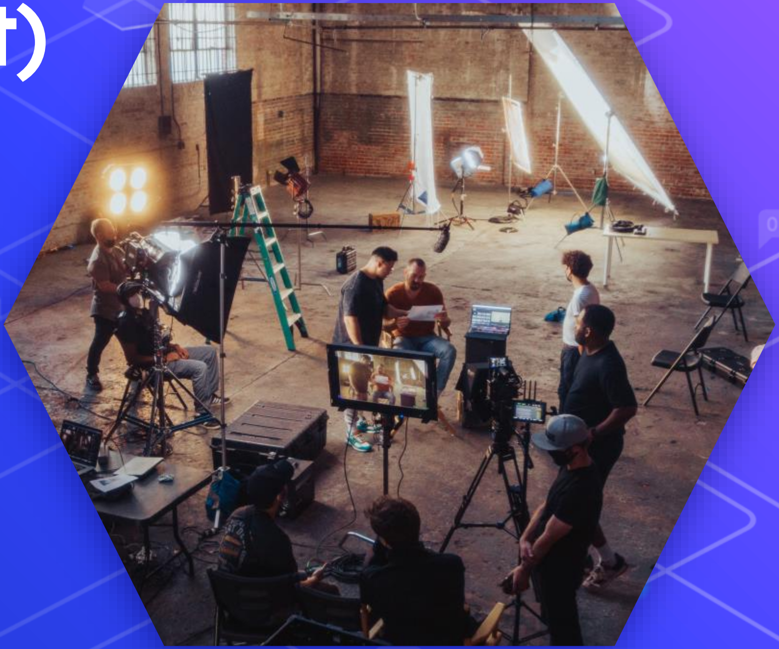
Data mining is the process of **extracting useful and actionable insights from large volumes of structured and unstructured data**. It involves using **various statistical and machine learning techniques** to **discover patterns, relationships, and hidden information** that can be used for decision-making and business strategy. Data mining helps uncover valuable knowledge and patterns that might not be readily apparent to humans, enabling organizations to gain a competitive advantage and make informed decisions.

STEPS IN DATA MINING

1. Data Collection: Gathering a large dataset from various sources.
2. Data Cleaning and Preprocessing: Removing errors and irrelevant information.
3. Data Exploration: Analyzing data to identify variables and initial patterns.
4. Data Mining Techniques: Applying statistical analysis, machine learning, and other techniques to discover patterns.
5. Pattern Evaluation: Assessing the significance and reliability of discovered patterns.
6. Knowledge Discovery: Interpreting and presenting patterns to make predictions and inform decision-making.

Why Mine data (Commercial Viewpoint)

- Data mining enables businesses to gain insights, drive strategic decision-making, and stay ahead in the competitive landscape.
- By mining customer data, businesses understand behavior, personalize experiences, and improve customer satisfaction.
- Data mining helps identify risks, detect fraud, and proactively mitigate potential financial losses.
- Analyzing operational data optimizes processes, resource allocation, and overall operational efficiency.
- Data mining aids in market analysis, competitor understanding, and product development for successful launches and innovation.
- Predictive analytics through data mining allows businesses to make accurate forecasts and proactive decisions in response to market dynamics.



“ DATA MINING – POTENTIAL APPLICATIONS

- Database analysis and decision support: Analyzing data stored in databases to support decision-making processes.
- Market analysis and management: Using data analysis to understand market trends, consumer behavior, and optimize marketing strategies.
- Risk analysis and management: Assessing and managing potential risks to minimize negative impacts on business operations.
- Fraud detection and management: Identifying fraudulent activities and implementing measures to prevent and manage fraud incidents.
- Text mining and web analysis: Extracting valuable insights from unstructured text data and analyzing web data for various purposes.
- Intelligent query answering: Developing systems that can understand and respond to user queries in a smart and efficient manner.
- Sports analytics: Applying data analysis to gain insights into player performance, team strategies, and game outcomes in sports.
- Astronomy: Analyzing astronomical data to study celestial objects, phenomena, and the universe.
- Internet web surf-aid: Assisting users in finding relevant information and navigating the web effectively.

1.
Data mining a KDD Process

Let’s start with the first set of slides

Data mining: Core of KDD

- Data cleaning: Identifying and correcting errors, inconsistencies, and inaccuracies in the data to ensure data quality and reliability.
- Data integration: Combining data from multiple sources and integrating them into a unified format for analysis and decision-making.
- Selection: Choosing and retrieving relevant data from the available dataset based on specific criteria and requirements.
- Task-relevant data: Identifying and selecting data that is most relevant and suitable for a specific analysis or task at hand.
- Data mining: Applying statistical and machine learning techniques to extract patterns, relationships, and insights from large datasets.
- Pattern evaluation: Assessing the discovered patterns and models for their significance, reliability, and usefulness in solving the problem or addressing the research question.
- Knowledge discovery: Interpreting and presenting the discovered patterns and insights in a meaningful way, leading to new knowledge and actionable insights.

Steps in KDD

- Learning the application domain
- Creating a target data set
- Data cleaning and preprocessing
- Data reduction and transformation
- Choosing functions of data mining
- Choosing mining algorithm
- Data mining
- Pattern evaluation and knowledge presentation
- Use of discovered knowledge

Data mining and Business Intelligence

Making decisions

Data presentation

Data mining

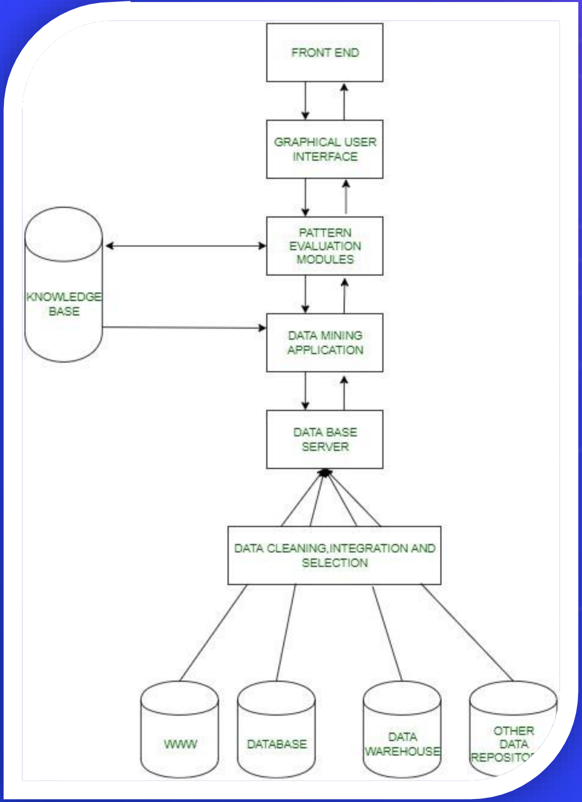
Data exploration

Data warehouses/ Data marts

Data sources

Architecture of a typical data mining system

The architecture of a typical data mining system consists of data sources, preprocessing, a data warehouse, a data mining engine, pattern evaluation, knowledge representation, decision-making, and deployment/maintenance.



Data mining on what kind of data

Relational databases

Data warehouses

Transactional databases

Advanced DB and information repositories

- Object-oriented and object relational databases
- Spatial databases
- Time series data and temporal data
- Text databases and multimedia databases

Data Mining Functionalities

Our process is easy

2

Classification and prediction

Classification assigns data instances to predefined categories, while prediction estimates numerical values. Both concepts rely on data mining algorithms to analyze and learn from data, enabling data-driven decision-making and forecasting.

1

Concept description: characterization and discrimination

Characterization and discrimination are data mining concepts used to provide a summary description of a dataset and identify significant differences or distinguishing factors between subsets or groups within the data, respectively.

3

Association (correlation and causality)

Correlation measures the statistical relationship between variables, while causality explores cause-and-effect relationships, requiring rigorous experimentation and control of confounding factors.

Our process is easy

1

Outlier analysis

Outlier analysis is a data mining technique used to identify and analyze data points that deviate significantly from the normal patterns or expected behavior, providing insights into unusual or anomalous observations in the data.

2

Cluster analysis

Cluster analysis is a data mining technique that groups similar data points together based on their similarities. It helps discover natural structures or relationships in the data without predefined categories. This analysis is used in various applications such as customer segmentation, image classification, and anomaly detection, providing insights into data patterns and supporting decision-making.

3

Trend and evolution analysis

Trend analysis examines data patterns over time, while evolution analysis studies the changes and development of data patterns or entities over time.

Are all the "Discovered" Patterns interesting?

Suggested approach:
Human-centered,
Query-based

Focused mining

Interestingness measures
Interestingnesserestingness measures are metrics used to assess the significance or importance of patterns or rules discovered through data mining. They help in identifying and prioritizing the most interesting and valuable patterns that are relevant to the problem at hand

Objective measures

Objective measures, also known as quantitative measures, are metrics used in data mining to evaluate and assess the quality or characteristics of patterns or models without relying on subjective judgments. These measures provide a systematic and unbiased way to compare and select patterns or models based on specific criteria

Subjective Measures

Subjective measures, also known as qualitative measures, are metrics used in data mining that involve subjective judgments or assessments based on human interpretation and domain expertise. These measures provide a qualitative evaluation of patterns, models, or insights that may not be easily quantifiable.

Data Mining: Confluence of multiple disciplines

Database technology

Machine learning

Information science

Statistics

Visualization

Other disciplines

Data Mining: Classification schemes

General Functionality:

- Descriptive Data Mining: Focuses on summarizing and exploring the characteristics and patterns in the data.
- Predictive Data Mining: Emphasizes building models to make predictions or classify future data based on historical patterns.

Different Views, Different Classifications:

- Subject-Oriented Data Mining: Analyzes data from a specific subject or domain, such as sales, marketing, healthcare, or finance.
- Integrated Data Mining: Combines data from multiple sources or domains to gain insights from a holistic perspective.

Data Mining: Classification schemes

Kinds of Databases to be Mined:

- Relational Database Mining: Analyzes data stored in relational databases, such as SQL-based systems.
- Data Warehouse Mining: Extracts knowledge from large-scale, integrated data repositories known as data warehouses.
- Stream Data Mining: Deals with data streams that arrive continuously and in real-time, requiring online analysis and processing.

Data Mining: Classification schemes

Kinds of Knowledge to be Discovered:
Concept Mining: Focuses on identifying and understanding underlying concepts and relationships in the data.
Association Mining: Explores patterns of co-occurrence or association among items in the data.
Sequential Pattern Mining: Discovers patterns that occur in a specific sequence or order.
Text Mining: Extracts meaningful information and patterns from unstructured textual data.

Data Mining: Classification schemes

Kinds of Techniques Utilized:

Clustering: Groups similar data objects together based on their intrinsic similarities.

Classification: Builds models to predict or classify data into predefined categories or classes.

Regression: Analyzes the relationship between variables to make predictions or estimate values.

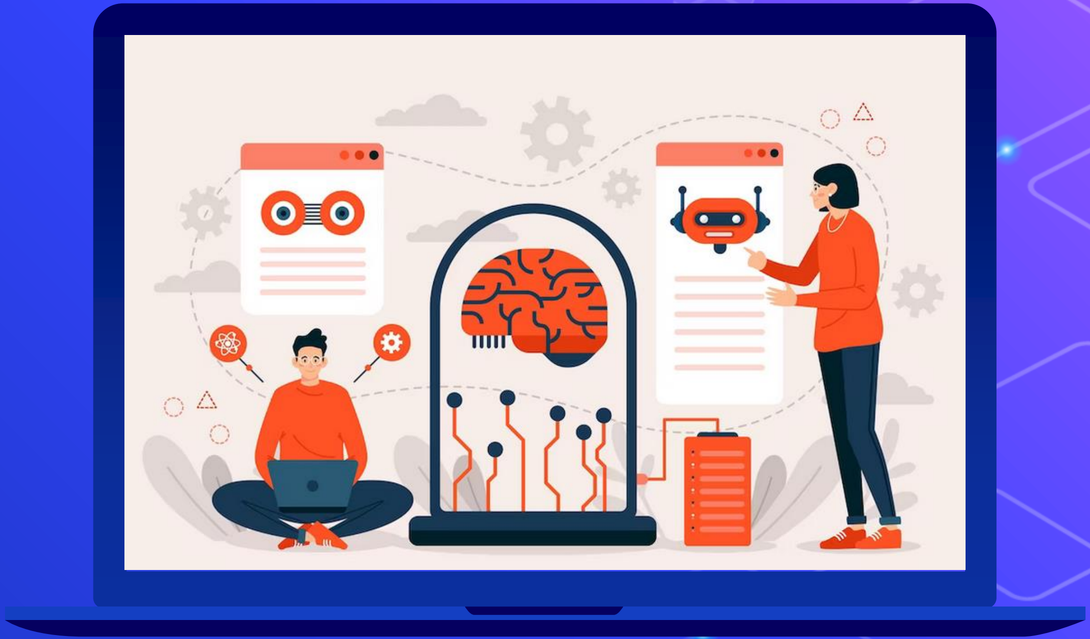
Anomaly Detection: Identifies rare or abnormal patterns or outliers in the data.

Data Mining: Classification schemes

Kinds of Applications Adapted:
Customer Relationship Management (CRM): Analyzes customer data to improve customer satisfaction, loyalty, and marketing strategies.
Fraud Detection: Identifies fraudulent activities or patterns in financial transactions or insurance claims.
Healthcare Analytics: Applies data mining techniques to healthcare data for disease prediction, diagnosis, and treatment planning.
Market Basket Analysis: Discovers associations and patterns in customer purchasing behavior for cross-selling and upselling opportunities.

A multi-Dimensional view of data mining classification

- Databased to be mined
- Knowledge to be mined
- Techniques utilized
- Application adapted



OLAP Mining: Integration of data mining and data warehousing

- Coupling of data mining systems with DBMS and data warehouse systems: Integration and interaction levels can vary from no coupling to tight coupling, enabling seamless data transfer and real-time analysis.
- Online Analytical Mining (OLAM): Integration of data mining techniques with OLAP technologies for interactive and multidimensional data analysis.
- Interactive mining of multi-level knowledge: Exploration and analysis of data at different levels of abstraction using techniques like drilling, rolling, pivoting, and slicing and dicing.
- Integration of multiple mining functions: Combination of classification, clustering, and association mining to address complex analytical tasks and gain comprehensive insights.
- Characterized classification: Sequential process involving clustering to group similar instances and subsequent classification within each cluster for refined results.

OLAM ARCHITECTURE

OLAM (Online Analytical Mining) architecture refers to the framework and components involved in integrating data mining techniques with online analytical processing (OLAP) technologies.

Data Sources: These are the various data repositories or sources from which data is collected for analysis. They can include databases, data warehouses, data marts, and external data sources.

Data Preprocessing: This component involves cleaning, transforming, and integrating the data from different sources to create a unified and consistent dataset for analysis. It may include tasks such as data cleaning, data integration, data normalization, and data aggregation.

OLAP Engine: The OLAP engine is responsible for providing multidimensional analysis capabilities on the integrated and preprocessed data. It allows users to navigate and explore the data from different dimensions and hierarchies, perform aggregations and calculations, and create interactive visualizations.

Mining Engine: The mining engine incorporates various data mining algorithms and techniques to analyze the data and discover patterns, relationships, and insights. It applies classification, clustering, association, regression, or other mining methods based on the analysis requirements.

Integration Layer: The integration layer facilitates the seamless integration and communication between the OLAP engine and the mining engine. It ensures the smooth flow of data between the two components, allowing for interactive analysis and mining on the OLAP cubes or data cubes.

User Interface: The user interface provides a user-friendly and intuitive interface for interacting with the OLAM system. It allows users to define analysis tasks, specify mining queries, explore data dimensions, visualize results, and interpret the discovered patterns.

Challenges In Data Mining

- **Mining methodology and user interaction**
- Developing effective techniques for user-friendly data mining and interaction with the system.

  - Mining different kinds of knowledge in databases: Addressing the challenge of extracting various types of knowledge from diverse databases.
  - Interactive mining of knowledge at multiple levels of abstraction: Enabling users to explore data at different levels of detail for interactive analysis.
  - Incorporation of background knowledge: Integrating prior information or domain knowledge into the mining process for improved results.
  - Data mining query languages and ad-hoc data mining: Creating query languages and tools for expressing ad-hoc mining queries.
  - Expression and visualization of data mining results: Designing effective methods for visualizing and presenting mining results
  - Handling noise and incomplete data: Dealing with inaccuracies and missing values in the data during the mining process.
  - Pattern evaluation: the interestingness problem: Assessing the significance and relevance of discovered patterns.

Challenges In Data Mining

Performance and scalability

- Efficiency and scalability of data mining algorithms: Developing algorithms that can handle large datasets and complex tasks efficiently.
- Parallel, distributed, and incremental mining methods: Utilizing parallel and distributed computing to speed up the mining process and handle incremental updates.

Challenges In Data Mining

Issues relating to the diversity of data types:

Addressing the challenges of mining data with different types, such as text, images, videos, and sensor data.

Handling relational and complex types of data: Dealing with data stored in relational databases and complex data structures, such as XML or JSON.

Mining information from heterogeneous databases and global information systems: Extracting knowledge from diverse databases and systems with varying formats and structures.

Challenges In Data Mining

Issues related to applications and social impacts:

Considering the implications and ethical concerns of applying the discovered knowledge in real-world applications

Application of discovered knowledge: Ensuring that the mined knowledge is effectively applied to solve practical problems and make informed decisions.

Domain-specific data mining tools: Developing specialized tools and techniques tailored to specific domains or industries.

Intelligent query answering: Enabling intelligent and context-aware query answering to provide meaningful insights to users.

Process control and decision-making: Incorporating data mining results into process control systems and decision-making processes for improved efficiency and effectiveness.

Integration of the discovered knowledge with existing knowledge: Addressing the challenge of integrating newly discovered knowledge with existing knowledge to create a comprehensive knowledge base.

Protection of data security, integrity, and privacy: Ensuring the security, integrity, and privacy of sensitive data during the mining process and when applying the results.

Common data mining tasks

Classification

- Classification is a data mining task that involves categorizing or labeling data into predefined classes or categories.
- It uses historical data with known class labels to build a model that can classify new, unseen data.
- Classification is widely used in various applications, such as spam detection, customer segmentation, and disease diagnosis.
- It helps in making predictions and decisions based on the identified patterns and relationships in the data.

CLASSIFICATION METHODS

Decision Tree

- Decision tree learning is a popular classification method that uses a tree-like model to make decisions based on feature values.
- It recursively partitions the data based on the selected features, creating a tree structure where each node represents a feature and each leaf node represents a class label.
- Decision tree learning is intuitive, interpretable, and can handle both categorical and numerical data.
- It is widely used in various domains, such as medical diagnosis, customer segmentation, and credit scoring.

Naïve Bayes

- The Naive Bayes method is a simple yet effective probabilistic classification technique based on Bayes' theorem.
- It assumes that all features are independent of each other given the class label, which is why it is called "naive."
- Naive Bayes calculates the probability of each class label for a given set of feature values and assigns the label with the highest probability.
- It is computationally efficient, particularly for large datasets, and is commonly used in email spam filtering, sentiment analysis, and text categorization.

K- Means Clustering

- The K Nearest Neighbor classifier is a non-parametric method that classifies data based on their proximity to other data points.
- KNN assigns a class label to a new data point by considering the class labels of its k nearest neighbors in the feature space.
- The value of k determines the number of neighbors to consider, and the class label is determined by majority voting.
- KNN is versatile, as it can handle both classification and regression tasks, and it is used in image recognition, recommender systems, and anomaly detection.

TYPES OF DECISION TREE

ID3 (Iterative Dichotomiser 3)

an algorithm for building decision trees. It selects the best attribute at each step based on information gain to split the data and construct the tree.

S

4

C4.5

C4.5 is an extension of ID3 that addresses its limitations. It uses information gain ratio instead of information gain, handles missing values, and allows for continuous attributes. C4.5 produces more accurate and robust decision trees

C

C

a decision tree algorithm that handles categorical target variables. It uses chi-squared tests to determine the best attribute for splitting and produces multiway splits in the tree.

CHAID

(Classification and Regression Trees) is a decision tree algorithm that can be used for both classification and regression tasks. It uses the Gini index to measure impurity and makes binary splits in the tree.

CART

Advantages of Decision Trees

- Interpretability: Decision trees are easy to understand and interpret, as the tree structure represents a sequence of if-then rules that explain the classification or prediction process.
- Handling of both categorical and numerical data: Decision trees can handle both categorical and numerical features without requiring extensive preprocessing.
- Non-parametric: Decision trees make minimal assumptions about the underlying data distribution, making them flexible and applicable to a wide range of problems.
- Feature importance: Decision trees can provide insights into the importance of different features in the classification or prediction task.
- Handling of missing values: Decision trees have built-in mechanisms to handle missing values by considering surrogate splits and assigning appropriate probabilities.

Limitations of Decision Trees

- Overfitting: Decision trees tend to overfit the training data, capturing noise and irrelevant details, which can lead to poor generalization on unseen data.
- Lack of smoothness: Decision trees create piecewise constant models, which may not capture complex relationships or subtle variations in the data.
- Sensitivity to small changes: Decision trees are sensitive to small changes in the training data, which can result in different tree structures and potentially different predictions.
- Bias towards features with more levels: Decision trees tend to favor features with a larger number of levels or categories, potentially neglecting important but less represented features.
- Difficulty in handling continuous variables: Traditional decision tree algorithms may struggle with continuous variables, as they require binning or discretization to convert them into categorical features.

Bayesian Classification

Bayesian classification is a statistical classification technique based on Bayes' theorem and probability theory. It calculates the posterior probability of a class label given a set of feature values by combining prior knowledge and observed evidence.

Probabilistic Prediction:
Probabilistic prediction refers to making predictions or estimating the likelihood of an event occurring based on probabilistic models and statistical analysis. It involves quantifying uncertainty and providing prediction outcomes with associated probabilities.

Standard Explanation:
In the context of data mining and machine learning, "standard explanation" refers to providing human-understandable explanations for the models and predictions generated by algorithms. It aims to make complex models and predictions more transparent and interpretable to users or stakeholders.

MAIN DIFFERENCES

Probabilistic Prediction

- Builds models based on probability theory and statistical analysis.
- Quantifies uncertainty and provides prediction outcomes with associated probabilities.
- Uses probabilistic models to capture the underlying probability distribution of the data.
- Generates prediction outcomes as probability distributions, acknowledging inherent uncertainty.
- Enables decision-making under uncertainty by considering predicted probabilities and associated risks.

Standard Explanation

- Provides human-understandable explanations for models and predictions.
- Emphasizes interpretable models like decision trees or linear regression.
- Highlights feature importance to understand influential factors in the model.
- Extracts simplified rules or explanations from complex models for better interpretability.
- Focuses on local explanations to comprehend model decision-making on specific instances.

K- Nearest Neighbour

Euclidean Distance

The most commonly used distance metric for k-NN is the Euclidean distance, which measures the straight-line distance between two points in Euclidean space.

Euclidean Distance = √(∑(xi - yi)^2)

K-Nearest Neighbor Classification

In k-NN classification, the algorithm determines the class label of a new data point by considering the class labels of its k nearest neighbors in the feature space

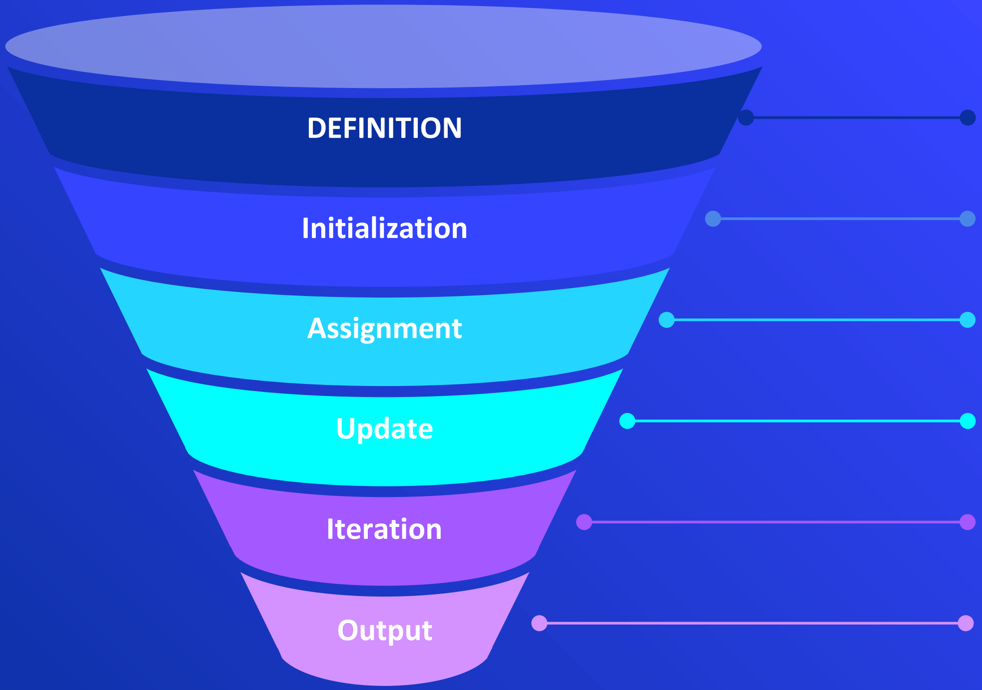
Class Label = mode(neighbors)

K-Nearest Neighbor Regression

In k-NN regression, the algorithm predicts a continuous value for a new data point by averaging the values of its k nearest neighbors

Predicted Value = mean(neighbors) or median(neighbors)

K- MEANS



K-means is a clustering algorithm that aims to partition a dataset into K distinct non-overlapping clusters, where each data point belongs to the cluster with the nearest mean (centroid). The algorithm iteratively assigns data points to the closest cluster centroid and updates the centroids based on the mean of the assigned data points

Choose the number of clusters K and randomly initialize K cluster centroids.

Assign each data point to the nearest cluster centroid based on a distance metric, commonly the Euclidean distance

Update the cluster centroids by calculating the mean of all data points assigned to each cluster.

Repeat steps 2 and 3 until convergence, where convergence is achieved when the assignment of data points to clusters no longer changes or the predefined number of iterations is reached.

The algorithm outputs the final cluster centroids and the assignment of each data point to a specific cluster.

CENTROID

- A centroid is a central point or a representative point within a cluster in the context of clustering algorithms. It is typically defined as the mean or average of all the data points that belong to a specific cluster.
- In the case of the K-means clustering algorithm, a centroid represents the center of a cluster. During the algorithm's execution, the centroids are updated iteratively by computing the mean of the data points assigned to each cluster. The updated centroids serve as reference points that help define the clusters' positions and boundaries.

HIERARCHICAL CLUSTERING

- Hierarchical clustering is a clustering algorithm that aims to build a hierarchy of clusters. It does not require a predefined number of clusters like the K-means algorithm. Instead, it starts with each data point as a separate cluster and merges or divides clusters based on their similarity or dissimilarity.

Agglomerative Hierarchical Clustering

- Agglomerative hierarchical clustering, also known as bottom-up clustering, is a common approach in hierarchical clustering. It starts by considering each data point as an individual cluster and then iteratively merges the most similar clusters based on a chosen distance metric. This process continues until all data points are grouped into a single cluster or until a stopping criterion is met.

Divisive Hierarchical Clustering

- Divisive hierarchical clustering, also known as top-down clustering, takes the opposite approach. It starts with all data points belonging to a single cluster and then divides the clusters recursively into smaller clusters based on dissimilarity. This process continues until each data point is assigned to its own individual cluster or until a stopping criterion is met.

Regression

- Regression is a data mining task used to predict numerical values or estimate continuous variables based on input features.
- It identifies the relationship between the input variables and the target variable to make predictions.
- Regression is useful in applications like sales forecasting, price prediction, and demand analysis.
- It helps in understanding how different factors influence the target variable and enables informed decision-making.

Clustering

- Clustering is a data mining task that groups similar data points together based on their intrinsic characteristics.
- It aims to discover hidden patterns or structures in the data without any predefined class labels.
- Clustering is used for customer segmentation, image recognition, and anomaly detection.
- It helps in understanding the natural grouping or distribution of data points and provides insights for targeted actions.

Association Rule Learning

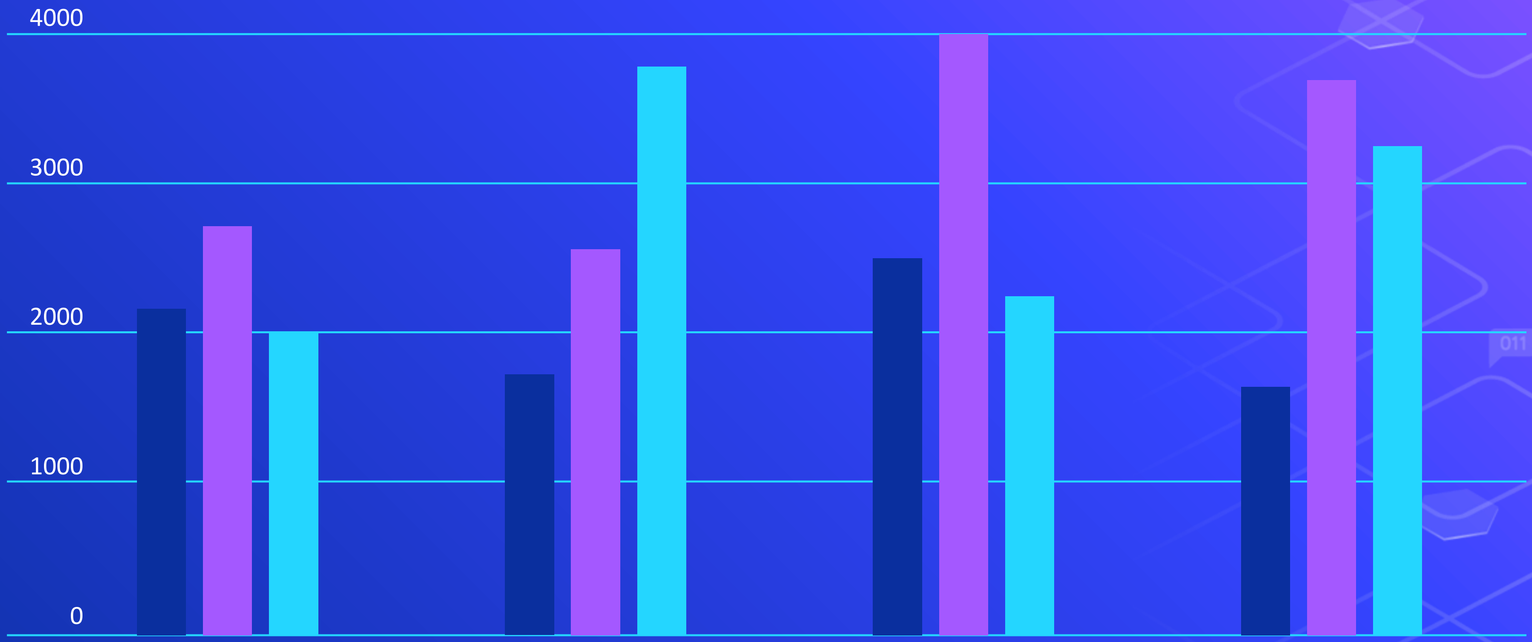
- Association rule learning is a data mining task that discovers interesting relationships or associations between items in a dataset.
- It identifies co-occurrence patterns and frequent itemsets to generate actionable rules.
- Association rule learning is widely used in market basket analysis, recommender systems, and customer behavior analysis.
- It helps in understanding purchasing patterns, cross-selling opportunities, and user preferences.

Anomaly Detection

- Anomaly detection is a data mining task that identifies abnormal or unusual patterns in the data.
- It focuses on detecting data points or events that deviate significantly from the expected or normal behavior.
- Anomaly detection is applied in fraud detection, network intrusion detection, and quality control.
- It helps in identifying outliers or anomalies that may indicate potential problems or anomalies in the system.

Summarization

- Summarization is a data mining task that aims to provide concise and meaningful summaries of large datasets.
- It extracts key features, trends, or patterns to provide a high-level overview of the data.
- Summarization is useful in data exploration, report generation, and decision support.
- It helps in understanding the main characteristics of the data and facilitates efficient data analysis.



You can insert graphs from Excel or Google Sheets

CLUSTERING APPLICATION

Customer Segmentation: Clustering can be used to group customers based on their similarities, allowing businesses to tailor marketing strategies, personalize offerings, and provide targeted customer support.

Image Segmentation: Clustering algorithms can be applied to segment images by grouping similar pixels together. This is useful in various fields such as computer vision, medical imaging, and object recognition

Document Clustering: Clustering can be used to organize large collections of documents into meaningful groups based on their content. This aids in information retrieval, text categorization, and topic modeling.

Anomaly Detection: Clustering techniques can help identify anomalies or outliers in a dataset by grouping normal data points and detecting deviations from these groups. This is valuable in fraud detection, network intrusion detection, and anomaly-based intrusion detection systems

Social Network Analysis: Clustering can be employed to discover communities or clusters within social networks based on patterns of connections between individuals. This aids in understanding network structures, identifying influencers, and targeted marketing

Genomics and Bioinformatics: Clustering is used to analyze gene expression data, DNA sequences, and protein structures. It helps identify patterns, classify genes, and understand biological processes.

1 2 3 4 5 6

Association Rule

Association rules in data mining are used to discover interesting relationships between items in a dataset. Support measures the frequency of an association rule in the dataset, while confidence measures the reliability of the association. The formula for calculating support is

(Number of transactions containing both items) / (Total number of transactions),

and the formula for confidence is

(Number of transactions containing both items) / (Number of transactions containing the first item). These measures help identify meaningful associations and determine their significance.

Application of Association Rule

1. Market Basket Analysis: Discover associations between products for cross-selling and targeted marketing in retail and e-commerce.

2. Fraud Detection: Identify anomalies and detect fraudulent activities in transactions or financial data.

3. Customer Behavior Analysis: Uncover patterns and preferences to personalize recommendations and improve customer experience.

4. Web Usage Mining: Analyze web log data to optimize websites, advertising, and user experience

5. Healthcare Analytics: Predict diseases, optimize treatment plans, and improve healthcare outcomes.

6. Supply Chain Management: Optimize inventory management, demand forecasting, and supplier selection

Data Mining Softwares

HIGH VALUE 2

Orange

Orange is an open-source data mining and machine learning tool that provides a visual programming interface for data analysis. It offers a wide range of data exploration, visualization, and predictive modeling techniques, making it suitable for both beginners and advanced users.

Weka

Weka is a widely used open-source data mining tool that provides a collection of machine learning algorithms and data preprocessing techniques. It offers a graphical user interface (GUI) and a command-line interface (CLI) for data analysis tasks such as classification, regression, clustering, and association rule mining.

LOW VALUE 1

HIGH VALUE 1

Rattle GUI

Rattle GUI is a free and open-source graphical interface for the R programming language. It provides a user-friendly environment for data mining and statistical analysis using R. Rattle GUI includes various data visualization, preprocessing, and modeling functionalities

Rapid Miner

RapidMiner is a commercial data mining tool that offers a visual workflow-based interface for building predictive models and performing data analytics. It supports a wide range of data mining tasks, including data preprocessing, feature engineering, classification, clustering, and text mining.

LOW VALUE 2

Application of data mining

Banking

Data mining is used in banking to detect fraudulent transactions, identify patterns of customer behavior, predict credit risk, and optimize marketing campaigns.

Targeted Marketing

Data mining techniques are used to analyze customer demographics, preferences, and purchase history to create targeted marketing campaigns and personalized recommendations

Fraud Detection

Data mining is employed to identify patterns and anomalies in data to detect fraudulent activities, such as credit card fraud, insurance fraud, and identity theft.

Manufacturing and Production

Data mining is used to optimize production processes, predict equipment failures, improve quality control, and analyze supply chain data for efficient inventory management.

Medicine

Data mining helps in analyzing patient records, identifying risk factors, predicting disease outcomes, optimizing treatment plans, and discovering patterns in drug effectiveness and adverse reactions

Molecular/ Pharmaceutical

Data mining techniques are applied in drug discovery and development, analyzing genomic data, predicting protein structures, and identifying potential drug targets.

Scientific Data Analysis

Data mining is used in scientific research to analyze large datasets, discover patterns, and make predictions in various fields such as astronomy, genomics, environmental science, and social sciences.

Customer Relationship Management (CRM)

Data mining helps businesses analyze customer data to improve customer satisfaction, identify cross-selling and upselling opportunities, and personalize marketing and customer service

Website/Store Design and Promotion

Data mining helps businesses analyze website traffic, user behavior, and customer preferences to optimize website design, improve user experience, and enhance promotional strategies.

Extra resources

SlidesCarnival icons are editable shapes.

This means that you can:

- Resize them without losing quality.
- Change fill color and opacity.
- Change line color, width and style.

Isn’t that nice? :)

Examples:

Find more icons at slidescarnival.com/extra-free-resources-icons-and-maps