|  |  |
| --- | --- |
| **UNIT- 3**  **Machine Learning**  Machine learning is a branch of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data. The goal of machine learning is to create systems that can automatically learn and improve from experience without being explicitly programmed for every task  **Machine Learning Types:**   1. Supervised Learning  * Housing Price Prediction * Medical Imaging  1. Unsupervised Learning  * Customer Segemntation * Market Basket Analysis  1. Semi-Supervised Learning  * Text Classification * Lane-finding on GPS data  1. Reinforcement Learning  * Optimized Marketing * Driverless Cars | **Regression:**  • Simple Linear Regression Algorithm  • Multivariate Regression Algorithm  • Decision Tree Algorithm  • Lasso Regression  **Classification:**  • Naïve Bayes Classifier Algorithm  • Random Forest Algorithm  • Decision Tree Algorithm  • Logistic Regression Algorithm  • Support Vector Machine Algorithm  **Clustering**  •K-Means Clustering algorithm  •Mean-shift algorithm  •DBSCAN Algorithm  •Principal Component Analysis  •Independent Component Analysis |
| **Supervised Learning:**  Supervised learning is a type of machine learning where the algorithm learns from labeled training data. This means each input data point is paired with a corresponding correct output or target variable. The objective is to train the algorithm to learn a mapping function from inputs to outputs, which can then be used to predict or classify new, unseen data accurately. Common tasks in supervised learning include regression, where the output is a continuous value (e.g., predicting house prices), and classification, where the output is a discrete label (e.g., classifying emails as spam or not spam). By iterating through the training data and adjusting its parameters, the algorithm minimizes the error between its predictions and the actual outcomes, effectively learning from the examples provided.  **Example:** Suppose you have a dataset of housing prices with features like size, location, number of bedrooms, and the corresponding sale prices. In supervised learning, you would use this labeled data to train a model to predict the price of a house given its features.  **Types of Supervised Learning Algorithms:**  **Classification:** In classification tasks, the algorithm predicts a discrete label or category. For example, predicting whether an email is spam or not spam based on its content.  **Regression:** Regression tasks involve predicting a continuous value. For instance, predicting the price of a house based on its features. | **Unsupervised Learning:** Unsupervised learning deals with unlabeled data, where the algorithm attempts to identify patterns, structures, or relationships within the data without explicit guidance or pre-defined categories. The primary goal is to explore the data and extract meaningful insights or groupings that were not previously known. Common tasks in unsupervised learning include clustering, where the data points are grouped into clusters based on similarity (e.g., customer segmentation in marketing), and dimensionality reduction, where the data is simplified while preserving its essential features (e.g., reducing the number of variables in a dataset while retaining significant information). Unlike supervised learning, unsupervised learning does not rely on labeled training data, making it particularly useful for exploratory data analysis and discovering hidden structures within large datasets.  **Example:** Consider a dataset containing customer purchase histories without any labels. In unsupervised learning, you might use clustering algorithms to group similar customers together based on their purchasing behavior.  **Types of Unsupervised Learning Algorithms:**  **Clustering:** Clustering algorithms group similar data points together into clusters. K-means clustering is a popular technique used for this purpose.  **Dimensionality Reduction:** These algorithms aim to reduce the number of features in a dataset while preserving important information. Principal Component Analysis (PCA) is a widely used dimensionality reduction technique. |

|  |  |
| --- | --- |
| **Linear Regression**  Linear regression is a statistical method used in machine learning and statistics to model the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the independent variables and the dependent variable, represented by a straight line. The basic concept of linear regression involves fitting a line to a set of data points to predict the value of the dependent variable based on the values of the independent variables. The equation of a simple linear regression model can be written as:  **y=mx+b**  Where:   * 𝑦 is the dependent variable (target). * 𝑥 is the independent variable (feature). * 𝑚 is the slope of the line, representing the relationship between 𝑥*x* and 𝑦*y*. * 𝑏 is the y-intercept, indicating the value of 𝑦*y* when 𝑥*x* is 0.   **Types of Linear Regression:**  **Simple Linear Regression:** Involves one independent variable (feature) and one dependent variable (target).  **Multiple Linear Regression:** Deals with multiple independent variables to predict a single dependent variable. | **Classification**  In classification, the goal is to build a model that can accurately assign class labels to instances based on their features. The model learns the patterns and relationships in the training data and uses this knowledge to make predictions on new data.  **Types of Classification Algorithms:**  **Logistic Regression:** Although named "regression," logistic regression is a classification algorithm used for binary classification tasks. It predicts the probability that an instance belongs to a particular class.  **Decision Trees:** Decision trees partition the feature space into regions and assign a class label to each region based on the majority class of training instances within that region  **Support Vector Machines (SVM):** SVMs find the optimal hyperplane that separates classes in the feature space, maximizing the margin between classes  **Random Forest:** A random forest is an ensemble technique that combines multiple decision trees to improve classification accuracy and robustness.  **Naive Bayes:** Naive Bayes classifiers are based on Bayes' theorem and assume that features are independent given the class. They are particularly effective for text classification tasks.  **Neural Networks:** Deep learning models such as neural networks can also be used for classification tasks, especially when dealing with complex, high-dimensional data. |
| **Bayes' theorem**  Bayes' theorem is a fundamental concept in probability theory and statistics, named after the Reverend Thomas Bayes. It provides a way to update our beliefs about the probability of an event based on new evidence or information. The theorem is often used in machine learning, especially in Bayesian inference and Bayesian networks, where it plays a key role in making predictions and decisions under uncertainty.  Where:   * 𝑃(𝐴∣𝐵)*P*(*A*∣*B*) is the probability of event 𝐴*A* occurring given that event 𝐵*B* has occurred (posterior probability). * 𝑃(𝐵∣𝐴)*P*(*B*∣*A*) is the probability of event 𝐵*B* occurring given that event 𝐴*A* has occurred (likelihood). * 𝑃(𝐴)*P*(*A*) is the prior probability of event 𝐴*A* occurring. * 𝑃(𝐵)*P*(*B*) is the prior probability of event 𝐵*B* occurring. | **Naive Bayes Classification**  Naive Bayes is a popular classification algorithm in machine learning, particularly suited for text classification tasks such as spam detection, sentiment analysis, and document categorization. It's based on Bayes' theorem, assuming independence among features, hence the term "naive." The Naive Bayes classifier assumes that the presence of a particular feature in a class is independent of the presence of other features. Despite this simplifying assumption, Naive Bayes often performs well, especially with large datasets and when the independence assumption approximately holds.  **Example:**  Naive Bayes for email classification works like this: Imagine you have a few emails, some are spam, and some are not. You look at these emails to see how often certain words like "free" or "buy" appear and how long the emails are. Then, using this information, when a new email arrives, you check if it has words like "free" or "buy" and how long it is. Based on how often these features occur in spam or non-spam emails in your data, you make a guess about whether the new email is likely to be spam or not. It's like making an educated guess based on patterns you've seen before. |

|  |  |
| --- | --- |
| **Classification in Machine Learning**  • Types of predictive models in machine learning are:  **Binary Classification for Machine Learning:**  • The most popular algorithms which are used for binary classification are :  ‒ K-Nearest Neighbours (Also supports multiple labels)  ‒ Logistic Regression (Only for two labels)  ‒ Support Vector Machine (Only for two labels)  ‒ Decision Trees (Also supports multiple labels)  ‒ Naive Bayes (Also supports multiple labels)  **Multi-Class Classification**  • These types of classification problems have no fixed two labels but can have any number of labels.  **Multi-Label Classification for Machine Learning**  • Here, we refer to those specific classification tasks where we need to assign two or more specific class labels that could be predicted for each example.  • The main difference with multi-class is the ability to predict multiple labels and not just one  **Imbalanced Classification**  • An Imbalanced Classification refers to those tasks where the number of examples in each of the classes are unequally distributed.  Generally, imbalanced classification tasks are binary classification jobs where a major portion of the training dataset is of the normal class type and a minority of them belong to the abnormal class. | **Clustering**  Clustering is an unsupervised machine learning technique that groups similar data points into clusters based on their characteristics, aiming to uncover inherent patterns and structures within the data without predefined labels. This method is useful for exploratory data analysis, helping to identify natural groupings that can inform decision-making across various fields. Popular algorithms like k-means, hierarchical clustering, and DBSCAN analyze datasets to form these clusters by measuring the similarity between data points. Clustering has diverse applications, such as customer segmentation for targeted marketing, image segmentation in computer vision, and pattern recognition in bioinformatics, making it a versatile tool for revealing hidden insights and organizing complex datasets.  For example, in healthcare, clustering can be used to group patients with similar medical histories or symptoms, enabling healthcare providers to identify patterns in disease progression or response to treatments. This can lead to more personalized and effective treatment plans, improve patient outcomes, and assist in early detection of health issues. By analyzing patient data through clustering, medical researchers can also uncover new insights into the relationships between different health conditions, facilitating advancements in medical research and public health strategies. |
| **K-Means Clustering Algorithm:**  K-means is a clustering algorithm in machine learning that aims to partition a set of data points into K clusters. It works by iteratively assigning each data point to the nearest centroid (center) of a cluster and then recalculating the centroids based on the new assignments. The goal is to minimize the sum of squared distances between data points and their respective cluster centroids, creating clusters where data points within the same cluster are similar to each other and dissimilar to those in other clusters. K-means is an unsupervised learning algorithm commonly used for data exploration, pattern recognition, and segmentation tasks.  **K-Means Steps:**   1. Choose the number of clusters (K) and initialize centroids randomly or using K-means++. 2. Assign data points to the nearest centroid based on distance, updating assignments until convergence. 3. Update centroids by computing the mean of data points in each cluster, repeating until convergence. 4. Convergence occurs when assignments and centroids stabilize or change minimally. 5. Output final cluster assignments and centroids as groups of similar data points. | **Elbow Method:**  The Elbow method is a technique used to determine the optimal number of clusters 𝐾*K* for a dataset in K-means clustering. It helps find the "elbow point" in a plot of the within-cluster sum of squares (WCSS) versus the number of clusters 𝐾*K*. Here's how it works:   1. **WCSS Calculation:**   Run the K-means algorithm for a range of 𝐾*K* values, typically from 1 to a maximum number based on domain knowledge or trial and error.  For each 𝐾*K*, calculate the within-cluster sum of squares (WCSS), which measures the sum of squared distances of data points to their cluster centroids.   1. **Elbow Point Identification:**   Plot a graph of 𝐾*K* versus WCSS, where 𝐾*K* is on the x-axis and WCSS is on the y-axis.  Look for the "elbow point" on the graph, which is the point where the rate of decrease in WCSS slows down significantly (forming an elbow-like shape).   1. **Optimal 𝐾*K* Selection:**   The optimal number of clusters 𝐾*K* is typically chosen at the elbow point, as it represents a balance between minimizing WCSS (better clustering) and avoiding overfitting (too many clusters). |

|  |  |
| --- | --- |
| **UNIT - 4** Data Visualization  Data Visualization techniques involve the generation of graphical or pictorial representation of DATA, form which leads you to understand the insight of a given data set. This visualization technique aims to identify the Patterns, Trends, Correlations, and Outliers of data sets.  **Benefits of Data Visualization**  **Patterns in business operations:** Data visualization techniques help us to determine the patterns of business operations. By understanding the problem statement and identifying the solutions in terms of pattering and applied to eliminate one or more of the inherent problems.  **Identify business trends and relate to data:** These techniques help us identify market trends by collecting the data on Day-To-Day business activities and preparing trend reports, which helps track the business how influences the market. So that we could understand the competitors and customers. Certainly, this helps to long-term perspective.  **Storytelling and Decision making:** Knowledge of storytelling from available data is one of the niche skills for business communication, specifically for the Data Science domain which is playing a vital role. Using best visualization this role can be enhanced much better way and reaching the objectives of business problems. | **1) Line Chart**  A line chart is a type of graph that displays data points connected by straight line segments. It is commonly used to visualize trends over time, such as stock prices, temperature changes, or sales figures. The x-axis typically represents time or some other continuous variable, while the y-axis represents the value being measured. Line charts are useful for identifying patterns, trends, and fluctuations in data over time.  **2) Histogram**  A histogram is a graphical representation of the distribution of numerical data. It consists of a series of bars, where each bar represents the frequency or count of data points falling within a specific range, known as a bin. Histograms are particularly useful for visualizing the distribution of continuous data, such as exam scores, ages of people, or product prices. They help to understand the central tendency, variability, and shape of the data distribution.  **3) Pie Chart**  A pie chart is a circular graph divided into sectors, where each sector represents a proportion or percentage of the whole dataset. Pie charts are commonly used to show the composition or distribution of categorical data, such as market share, budget allocation, or survey responses. Each sector's size is proportional to the quantity it represents, making it easy to compare relative sizes of categories within the dataset. |
| **4) Scatter plots**  A scatter plot is a two-dimensional graph that uses dots to represent individual data points. It is used to visualize the relationship between two variables, typically one plotted on the x-axis and the other on the y-axis. Scatter plots are useful for identifying patterns, correlations, clusters, or outliers in the data. They are especially effective for exploring relationships between continuous variables, such as height versus weight or temperature versus humidity.  **5) Hexbins plots**  A hexbin plot is a variation of a scatter plot that uses hexagonal bins to represent the density of data points in a two-dimensional space. Each hexagon's color or intensity indicates the number of data points it contains, helping to visualize areas of high or low density. Hexbin plots are useful for handling large datasets and identifying patterns or clusters that may not be obvious in a traditional scatter plot.  **6) Heatmap**  A heatmap is a graphical representation of data where values in a matrix are represented as colors. It is commonly used to visualize correlations, distributions, or relationships in a dataset. Heatmaps are particularly effective for highlighting patterns, trends, or anomalies in large datasets. They are commonly used in fields such as biology (gene expression), finance (stock correlations), and social sciences (survey responses). | **7) Boxplot**  A boxplot, also known as a box-and-whisker plot, is a graphical representation of the distribution of numerical data through quartiles. It consists of a box that represents the interquartile range (IQR) of the data, with a line inside the box indicating the median. Whiskers extend from the box to show the range of the data, excluding outliers. Boxplots are useful for identifying central tendency, variability, and outliers in the data distribution.  **8) Pairplot**  A pairplot is a grid of scatter plots and histograms that visualizes pairwise relationships between variables in a dataset. It displays scatter plots of each pair of variables along the diagonal and histograms on the off-diagonal cells. Pairplots are useful for exploring correlations, distributions, and patterns between multiple variables simultaneously. They are commonly used in exploratory data analysis (EDA) to gain insights into data relationships.  **9) Bar Chart**  A bar chart is a graphical representation of data using rectangular bars of varying lengths. It is used to compare categories or show changes over time. The length of each bar represents the value it represents, making it easy to compare quantities across different categories. Bar charts are effective for visualizing categorical data, such as sales by product category, population by country, or survey responses by option. |