Supervised Learning

Supervised learning involves training a model on a labeled dataset, which means that each training example is paired with an output label. The goal is to learn a mapping from inputs to outputs that can be used to predict the labels of new, unseen examples.

Classification

Classification is a type of supervised learning where the goal is to assign inputs into one of several predefined categories.

- **Binary Classification**: The model distinguishes between two classes. Examples include spam detection (spam or not spam) and disease diagnosis (diseased or healthy).
- **Multiclass Classification**: The model distinguishes among three or more classes. Examples include handwriting recognition (recognizing each digit from 0 to 9) and image classification (identifying objects in images).

Common algorithms for classification include:

- **Logistic Regression**: Models the probability of a binary outcome using a logistic function.
- **Decision Trees**: Splits the data into subsets based on the value of input features.
- **Support Vector Machines (SVM)**: Finds the hyperplane that best separates the classes in the feature space.
- **k-Nearest Neighbors (k-NN)**: Classifies a data point based on the majority class among its k-nearest neighbors.

Regression

Regression is a type of supervised learning where the goal is to predict a continuous output.

- **Linear Regression**: Models the relationship between the input variables and the output by fitting a linear equation to the observed data.
- **Polynomial Regression**: Extends linear regression by considering polynomial relationships between the input variables and the output.
- **Ridge and Lasso Regression**: Variants of linear regression that include regularization terms to prevent overfitting.

Common metrics for regression include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

Ensemble Methods

Ensemble methods combine multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent models alone.

- **Random Forests**: An ensemble of decision trees, typically trained with the "bagging" method. Each tree in the forest is trained on a subset of the data, and the final prediction is made by averaging the predictions of all the trees.
- **Gradient Boosting**: Builds an ensemble of trees sequentially, where each new tree focuses on correcting the errors made by the previous trees. Examples include Gradient Boosting Machines (GBM) and XGBoost.

Support Vector Machines (SVM)

SVMs are supervised learning models that analyze data for classification and regression analysis. The core idea is to find a hyperplane in a high-dimensional space that distinctly classifies the data points.

- **Linear SVM**: Uses a linear hyperplane to separate the data.
- **Non-linear SVM**: Uses kernel functions (e.g., RBF, polynomial) to project data into higher dimensions where a linear separation is possible.

k-Nearest Neighbors (k-NN)

k-NN is a simple, instance-based learning algorithm where the class of a data point is determined by the majority class of its k-nearest neighbors in the feature space. It is computationally expensive during prediction since it involves calculating the distance to all training points.

Neural Networks

Neural networks are a set of algorithms inspired by the structure and function of the human brain, designed to recognize patterns.

- **Feedforward Neural Networks (FNN)**: The simplest type of neural network where information moves in one direction—from input to output.
- **Convolutional Neural Networks (CNNs)**: Primarily used for image data, CNNs use convolutional layers to automatically and adaptively learn spatial hierarchies of features.
- **Recurrent Neural Networks (RNNs)**: Suitable for sequential data, RNNs have connections that form directed cycles, allowing them to maintain a memory of previous inputs. Long Short-Term Memory (LSTM) networks are a type of RNN designed to handle long-term dependencies.

Unsupervised Learning

Unsupervised learning involves training a model on data without labeled responses. The goal is to infer the natural structure present within a set of data points.

Clustering

Clustering is a method of unsupervised learning that involves grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups.

- **k-Means Clustering**: Partitions the data into k clusters by minimizing the variance within each cluster.
- **Hierarchical Clustering**: Builds a hierarchy of clusters either through a bottom-up (agglomerative) or top-down (divisive) approach.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Groups together points that are close to each other based on a distance measurement, and marks points that are in low-density regions as outliers.

Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables.

- **Principal Component Analysis (PCA)**: Projects the data into a lower-dimensional space by maximizing the variance along the principal components.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**: Primarily used for visualizing high-dimensional data by reducing it to two or three dimensions.
- **Linear Discriminant Analysis (LDA)**: Finds the linear combinations of features that best separate two or more classes of objects or events.

Anomaly Detection

Anomaly detection aims to identify rare items, events, or observations that raise suspicions by differing significantly from the majority of the data.

- **Statistical Methods**: Assume a statistical distribution for the data and identify points that deviate significantly from this distribution.
- **Machine Learning Methods**: Include clustering-based methods (e.g., DBSCAN) and model-based approaches (e.g., autoencoders).

Association Rules

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.

- **Apriori Algorithm**: Identifies frequent itemsets and then derives association rules from these itemsets.
- **Eclat Algorithm**: Uses a depth-first search strategy to find frequent itemsets and is often faster than Apriori for large datasets.