Expanded Report on Key Topics from ML UNIT-II NOTES

## Q: What is the K-Nearest Neighbors (KNN) algorithm, and what are its main uses?

The K-Nearest Neighbors (KNN) algorithm is a simple machine learning algorithm based on supervised learning. It assumes similarity between new data and available data, classifying new data into the most similar existing category. KNN stores all available data and classifies new data points based on this similarity. This allows for easy classification of new data into appropriate categories. While usable for both regression and classification, KNN is primarily employed for classification problems. It's a non-parametric algorithm, meaning it makes no assumptions about the underlying data distribution. It's also termed a "lazy learner" because it doesn't learn from the training set immediately; instead, it stores the dataset and performs actions on it during classification. During training, KNN simply stores the dataset; classification involves finding the category most similar to the new data point.

## Q: What are the advantages and disadvantages of using the KNN algorithm?

**Advantages of KNN Algorithm:**

* Simple to implement.
* Robust to noisy training data.
* More effective with large training datasets.

**Disadvantages of KNN Algorithm:**

* Requires determining the optimal value of K, which can be complex.
* High computational cost due to the need to calculate distances between the data point and all training samples.

## Q: How does a decision tree algorithm work, and what are its key components?

A decision tree is a supervised learning technique used for both classification and regression problems, though it's preferred for classification. It's a tree-structured classifier where internal nodes represent dataset features, branches represent decision rules, and leaf nodes represent outcomes. Decision nodes make decisions and have multiple branches, while leaf nodes are the final outputs and have no further branches. Decisions are based on the features of the given dataset. Decision trees provide a graphical representation of possible solutions based on given conditions. The tree starts at the root node, expanding into branches to create a tree-like structure. The CART (Classification and Regression Tree) algorithm is used to build the tree. A decision tree works by asking questions based on features, splitting the tree into subtrees based on the answers (yes/no). The process involves beginning with a root node containing the complete dataset, finding the best attribute using Attribute Selection Measure (ASM), dividing the dataset into subsets based on the attribute's possible values, generating a decision tree node containing the best attribute, and recursively creating new decision trees for the subsets until reaching leaf nodes (points where no further classification is possible).

## Q: What are the advantages and disadvantages of using a decision tree?

**Advantages of the Decision Tree:**

* Simple to understand, mirroring real-life decision-making processes.
* Useful for solving decision-related problems.
* Helps consider all possible problem outcomes.
* Requires less data cleaning compared to other algorithms.

**Disadvantages of the Decision Tree:**

* Can become complex with many layers.
* Prone to overfitting (can be mitigated using the Random Forest algorithm).
* Computational complexity increases with more class labels.

## Q: What is the Naive Bayes classifier, and where is it commonly applied?

The Naïve Bayes algorithm is a supervised learning algorithm based on Bayes' theorem, used for solving classification problems, particularly text classification involving high-dimensional datasets. It's a simple and effective classification algorithm for building fast machine learning models that provide quick predictions. It's a probabilistic classifier, predicting based on object probabilities. Examples of its application include spam filtering, sentiment analysis, and article classification. The "naive" aspect refers to its assumption that feature occurrences are independent of each other. The "Bayes" aspect stems from its reliance on Bayes' theorem, a formula to determine the probability of a hypothesis given prior knowledge and conditional probability.

## Q: What is the difference between linear and logistic regression?

**Linear Regression:** A popular and simple machine learning algorithm for predictive analysis, predicting continuous numerical values (e.g., salary, age). It shows the linear relationship between dependent and independent variables, illustrating how the dependent variable changes according to the independent variable. It aims to find the best-fitting line representing this relationship, known as the regression line. Linear regression can be simple (one independent variable) or multiple (multiple independent variables). The relationship can be positive (dependent variable increases as the independent variable increases) or negative (dependent variable decreases as the independent variable increases).

**Logistic Regression:** A popular supervised learning algorithm for predicting categorical dependent variables (e.g., yes/no, true/false). Instead of a regression line, it uses an S-shaped logistic function to predict probabilities (between 0 and 1). It's similar to linear regression but solves classification problems instead of regression problems. The curve indicates the likelihood of an event. It's significant for its ability to provide probabilities and classify new data using continuous and discrete datasets. It can handle different data types and identify effective variables for classification. There are types of logistic regression: binomial (two categories), multinomial (three or more unordered categories), and ordinal (three or more ordered categories).

## Q: What is a Support Vector Machine (SVM) algorithm used for?

A Support Vector Machine (SVM) is a popular supervised learning algorithm for both classification and regression problems, mainly used for classification. Its goal is to create an optimal decision boundary (hyperplane) in an n-dimensional space to separate data points into classes, enabling easy categorization of new data points. It selects extreme points (support vectors) to define this hyperplane. SVMs are used in applications like face detection, image classification, and text categorization.