Expanded Report on Key Topics from ML UNIT-II NOTES (1)\_removed.pdf

## Q: What is the K-Nearest Neighbor (KNN) algorithm, and what are its advantages and disadvantages?

The K-Nearest Neighbor (KNN) algorithm is one of the simplest Machine Learning algorithms based on a Supervised Learning technique. K-NN assumes similarity between new data and available data, classifying new data into the most similar existing category. The algorithm stores all available data and classifies new data points based on similarity. This means new data can be easily classified into an appropriate category using the K-NN algorithm. K-NN can be used for both Regression and Classification, but it's primarily used for Classification problems. It's a non-parametric algorithm, meaning it makes no assumptions about underlying data. It's also called a lazy learner algorithm because it doesn't learn from the training set immediately; instead, it stores the dataset and performs actions on it during classification. During the training phase, the KNN algorithm simply stores the dataset. When it receives new data, it classifies that data into the most similar category.

**Advantages of KNN Algorithm:**

* It is simple to implement.
* It is robust to noisy training data.
* It can be more effective if the training data is large.

**Disadvantages of KNN Algorithm:**

* It always requires determining the value of K, which can be complex.
* The computation cost is high due to calculating the distance between data points for all training samples.

## Q: How does a Decision Tree algorithm work, and what are its strengths and weaknesses?

A Decision Tree is a Supervised learning technique used for both classification and Regression problems, but it's mostly preferred for solving Classification problems. It's a tree-structured classifier where internal nodes represent features of a dataset, branches represent decision rules, and each leaf node represents an outcome. A Decision tree has two types of nodes: Decision Nodes and Leaf Nodes. Decision nodes are used for decision-making and have multiple branches, while Leaf nodes are the outputs of decisions and contain no further branches. Decisions or tests are performed based on features of the given dataset. It's a graphical representation for getting all possible solutions to a problem/decision based on given conditions. It's called a decision tree because, similar to a tree, it starts with a root node that expands into further branches and constructs a tree-like structure. To build a tree, the CART algorithm (Classification and Regression Tree algorithm) is used. A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees.

The algorithm works as follows:

* **Step-1:** Begin the tree with the root node (S), containing the complete dataset.
* **Step-2:** Find the best attribute in the dataset using Attribute Selection Measure (ASM).
* **Step-3:** Divide S into subsets containing possible values for the best attribute.
* **Step-4:** Generate the decision tree node containing the best attribute.
* **Step-5:** Recursively create new decision trees using the subsets from Step 3. Continue this until a stage is reached where nodes cannot be further classified; these final nodes are called leaf nodes.

**Advantages of the Decision Tree:**

* It's simple to understand, following the same process as humans use for real-life decision-making.
* It's useful for solving decision-related problems.
* It helps consider all possible outcomes for a problem.
* It requires less data cleaning compared to other algorithms.

**Disadvantages of the Decision Tree:**

* The decision tree can have many layers, making it complex.
* It may have an overfitting issue, resolvable using the Random Forest algorithm.
* For more class labels, computational complexity increases.

## Q: What is the Naïve 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. It's mainly used in text classification involving high-dimensional training datasets. It's a simple and effective classification algorithm that helps build fast machine learning models for quick predictions. It's a probabilistic classifier, meaning it predicts based on the probability of an object. Popular examples include spam filtration, sentimental analysis, and classifying articles.

**Naive:** It's called "Naïve" because it assumes the occurrence of a certain feature is independent of the occurrence of other features. For example, if a fruit is identified based on color, shape, and taste, a red, spherical, and sweet fruit is recognized as an apple. Each feature contributes individually to the identification, independent of the others.

**Bayes:** It's called "Bayes" because it depends on the principle of Bayes' Theorem. Bayes' theorem (Bayes' Rule or Bayes' law) determines the probability of a hypothesis with prior knowledge, depending on conditional probability.

**Advantages of Naïve Bayes Classifier:**

* It's a fast and easy ML algorithm for predicting class datasets.
* It can be used for Binary and Multi-class Classifications.
* It performs well in Multi-class predictions compared to other algorithms.
* It's a popular choice for text classification problems.

**Disadvantages of Naïve Bayes Classifier:**

* Naïve Bayes assumes features are independent or unrelated, so it can't learn relationships between features.

**Applications of Naïve Bayes Classifier:**

* Credit Scoring
* Medical data classification
* Real-time predictions (eager learner)
* Text classification (spam filtering and sentiment analysis)

## Q: What is the difference between Linear Regression and Logistic Regression?

**Linear Regression** is a popular and simple machine learning algorithm used for predictive analysis. It predicts continuous numerical values (like salary, age) and shows the linear relationship between dependent and independent variables. It shows how the dependent variable (y) changes according to the independent variable (x). It finds the best-fit line between variables; this line is called the regression line. The linear regression model provides a sloped straight line representing the relationship between variables. Linear regression can be simple (one independent variable) or multiple (more than one independent variable). A linear line showing the relationship between dependent and independent variables is called a regression line. This line can show a positive linear relationship (dependent variable increases as the independent variable increases) or a negative linear relationship (dependent variable decreases as the independent variable increases).

**Logistic Regression** is a popular Machine Learning algorithm under Supervised Learning. It predicts categorical dependent variables (using given independent variables). The outcome must be categorical or discrete (Yes/No, 0/1, True/False). Instead of giving exact values, it provides probabilistic values between 0 and 1. Logistic Regression is similar to Linear Regression but is used for classification problems instead of regression problems. In Logistic regression, an "S" shaped logistic function is fit, predicting two maximum values (0 or 1). The curve indicates the likelihood of something (e.g., whether cells are cancerous). Logistic Regression is significant because it can provide probabilities and classify new data using continuous and discrete datasets. It can classify observations using different data types and determine effective variables for classification.

## Q: What is a Support Vector Machine (SVM) algorithm, and what are some of its uses?

Support Vector Machine (SVM) is a popular Supervised Learning algorithm used for Classification and Regression problems, primarily classification. The goal is to create the best line (decision boundary) to separate n-dimensional space into classes, enabling easy categorization of new data points. This best decision boundary is called a hyperplane. SVM selects extreme points/vectors to create the hyperplane; these are called support vectors.

**Uses of SVM:**

* Face detection
* Image classification
* Text categorization

## Q: What are the different types of Logistic Regression?

There are three types of Logistic Regression:

* **Binomial:** Two possible types of dependent variables (0 or 1, Pass or Fail).
* **Multinomial:** Three or more possible unordered types of dependent variables ("cat", "dog", "sheep").
* **Ordinal:** Three or more possible ordered types of dependent variables ("low", "medium", "high").