Expanded Report on Key Machine Learning Algorithms

## Q: What is the k-Nearest Neighbors (k-NN) algorithm?

The k-Nearest Neighbors (k-NN) algorithm is a fundamental supervised learning algorithm renowned for its simplicity. It operates by classifying a new data point based on the proximity to existing data points within a training dataset. The algorithm leverages the concept of similarity: a new data point is assigned the category that is prevalent among its nearest neighbors. k-NN's versatility extends beyond classification; it can also be applied to regression problems. However, it finds particularly strong application in classification tasks. It is a non-parametric algorithm, meaning it doesn't make assumptions underlying the data distribution. Considered a lazy learner, it doesn't explicitly learn from the training data; instead, it stores the entire dataset and uses it later for classification. A key aspect of the k-NN algorithm requires determining the optimal value of 'k', representing the number of nearest neighbors to consider. While classification is straightforward to implement, it can be computationally intensive due to distance calculations between the new data point and all training points, limiting its efficiency, especially with large datasets.

## Q: How does a Decision Tree algorithm work for classification?

A decision tree is a powerful supervised learning algorithm frequently employed for both classification and regression problems, but it finds great utility in classification. It offers a visually intuitive representation of the decision-making process through its tree-like structure. Internal nodes represent features, branches represent decision rules based on those features, and leaf nodes represent the class label predicting the outcome. The construction of a decision tree involves recursively partitioning the data based on the most informative feature, determined by measures like information gain or Gini impurity. Algorithms like CART (Classification and Regression Trees) build this tree structure. Decision trees are relatively easy to understand and interpret, making them attractive for explanatory purposes. However, they are prone to overfitting, particularly with deep trees. Techniques like pruning (removing irrelevant branches) help mitigate this issue.

## Q: What is the Naive Bayes classifier, and what is its key assumption?

The Naive Bayes algorithm is a probabilistic supervised learning algorithm widely used in text classification applications involving high-dimensional data. It uses Bayes' theorem to calculate the probability of an event occurring given prior knowledge of related events. The "naive" aspect refers to its key assumption: feature independence. This means that the presence or absence of one feature is considered unrelated to the presence or absence of any other feature. This simplification considerably reduces calculation, allowing for efficient classification. The algorithm's strength lies in its simplicity, speed, and effectiveness in scenarios requiring quick predictions, making it ideal for applications such as spam filtering, sentiment analysis, and article categorization. It's important to note that the independence assumption is rarely perfectly met in real-world data, which can affect its accuracy in certain cases.

## Q: What is linear regression, and what are its different forms?

Linear regression is a fundamental statistical machine learning algorithm used in predictive analysis involving continuous numerical variables. It models a linear relationship between a dependent variable (the variable we aim to predict) and one or more independent variables (predictor variables). The goal is to find a well-fitting line that minimizes the difference between predicted and actual values of the dependent variable. Simple linear regression involves a single independent variable, while multiple linear regression uses multiple independent variables. The model is represented by an equation with coefficients that define the slope and intercept of the line. The accuracy of a linear regression model is evaluated using metrics like mean squared error.

## Q: What is logistic regression, and how does it differ from linear regression?

Logistic regression is a supervised learning algorithm specifically designed for classification problems where the dependent variable is categorical (e.g., yes/no, true/false). Unlike linear regression, which predicts continuous values, logistic regression predicts the probability of a particular outcome. It uses a sigmoid function to map a linear combination of independent variables to a probability. This probability then classifies the data point into a category. Logistic regression is a popular choice for binary classification (two categories), but it can be extended to handle multiple categories using techniques like multinomial logistic regression. It provides classification probability estimates for each category, offering valuable insight into the confidence of the prediction.

## Q: What are Support Vector Machines (SVMs), and what is their core idea?

Support Vector Machines (SVMs) are powerful supervised learning algorithms used for both classification and regression, though predominantly for classification. The core idea behind SVMs is to find the optimal hyperplane that best separates data points of different classes in a high-dimensional space. This hyperplane maximizes the margin between classes, making the classification robust. Points close to the hyperplane are called support vectors; they are crucial in defining the hyperplane. SVMs can handle linearly separable data directly. For non-linearly separable data, kernel functions map the data into a higher-dimensional space where linear separation becomes possible. SVMs are known for their effectiveness in high-dimensional space applications, including image classification and text categorization.