K-Nearest Neighbors (KNN) Algorithm

1 Introduction

The K-Nearest Neighbors (KNN) algorithm is a non-parametric, instance-based learning technique. It is commonly used for classification and regression tasks. The primary mechanism of KNN involves predicting the class label of a data point based on the majority class among its k nearest neighbors in the feature space.

2 Algorithm Description

Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector and y_i is the corresponding class label, the KNN algorithm proceeds as follows:

1. **Choose the number of neighbors**: Select an integer k, the number of nearest neighbors to consider.

2. **Compute distances**: For a test point $\mathbf{x}$, compute the distance between $\mathbf{x}$ and all points in the training set. The distance metric is often the Euclidean distance:

$$\text{distance}(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{d} (x_j - x_{ij})^2} \tag{1}$$

   where $\mathbf{x} = [x_1, x_2, \ldots, x_d]$ and $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$.

3. **Identify nearest neighbors**: Select the k points in the training set that are closest to $\mathbf{x}$.

4. **Majority vote (for classification)**: Determine the majority class among the k nearest neighbors. The predicted class $\hat{y}$ is given by:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{i=1}^{k} \mathbf{1}(y_i = c) \tag{2}$$

   where $\mathcal{C}$ is the set of all possible classes, and $\mathbf{1}(\cdot)$ is the indicator function.

3 Advantages and Disadvantages

KNN has several advantages:

- Simple to implement and understand.
- No assumptions about the data distribution.
- Can work well with a large number of features.

However, it also has some drawbacks:

- Computationally intensive as it requires distance calculation for each test point against all training points.
- The choice of k can significantly affect the algorithm's performance.
- Sensitive to the scale of the data and irrelevant features.

4 Conclusion

The KNN algorithm is a powerful tool for classification and regression tasks. Its simplicity and effectiveness make it a popular choice for many applications, despite its computational challenges.