K-Nearest Neighbors (KNN) Explained Using Set Theory

K-Nearest Neighbors (KNN) is a simple and intuitive algorithm used for classification and regression tasks. It can be explained using the concepts of set theory as follows:

Definitions

Let:

- $X = \{x_1, x_2, \ldots, x_n\}$ be the set of all data points in the feature space.
- $Y = \{y_1, y_2, \ldots, y_n\}$ be the set of corresponding labels (for classification) or values (for regression).
- $q \in X$ be the query point for which we want to predict the label or value.
- $N_k(q) \subseteq X$ be the subset of k nearest neighbors of q, determined by a distance metric $d(x, q)$ (e.g., Euclidean distance).

Algorithm

1. Compute the distance $d(x, q)$ for all $x \in X$. 2. Identify the subset $N_k(q) \subseteq X$ such that $|N_k(q)| = k$ and $\forall x_i \in N_k(q), \forall x_j \notin N_k(q) : d(x_i, q) \leq d(x_j, q)$. 3. For classification:

- Define a mapping $f : N_k(q) \to Y$ that assigns labels to the neighbors.
- Predict the label $\hat{y}$ for q as the mode (most frequent label) in the multiset $f(N_k(q))$.

4. For regression:

- Predict the value $\hat{y}$ for q as the mean of the values in $f(N_k(q))$.

Set Theory Representation

$$\begin{aligned} N_k(q) &= \{x \in X \mid \text{rank}(d(x, q)) \leq k\}, \\ \hat{y} &= \begin{cases} \text{mode}(f(N_k(q))) & \text{for classification,} \\ \text{mean}(f(N_k(q))) & \text{for regression.} \end{cases} \end{aligned}$$