

**KNN**

The k-NearestNeighbors (KNN) algorithm is a simple, non-parametric, supervised machine learning method used for both classification and regression tasks. It operates on the principle that similar data points tend to exist in close proximity, making predictions for a new data point based on the labels or values of its "k" nearest neighbors in the training dataset.

How the KNN Algorithm Works

The process for using KNN involves several steps:

1. Select the number of neighbors (k): A positive integer value for is chosen, which determines how many surrounding data points will be considered. It is common to choose an odd number for to avoid ties in classification problems.
2. Calculatedistances: The distance between the new (query) data point and all existing training data points is calculated using a distance metric, most commonly Euclidean distance for continuous variables. Other metrics include Manhattan, Minkowski, and Hamming distances.
3. Find the knearestneighbors: The data points are sorted by their distance from the query point, and the topclosest points are selected. **k**
4. Assign aclassorvalue: **k**
  - ForClassification: The new data point is assigned the class label that is most frequent among its nearest neighbors (a majority vote).
  - ForRegression: The algorithm predicts a continuous value by averaging the values of the nearest neighbors.

AdvantagesandDisadvantages

Advantages

Disadvantages

Simple and easyto implement

**k**

Computationallyexpensivefor large datasets, as it requires calculating distances for every prediction

No training phase ("lazy learner"), meaning new data can be added easily without retraining

**k**

requiring careful feature selection and data scaling (normalization or standardization)

assumptions about the underlying data distribution

**k**

Versatile, applicable to both classification and regression tasks

Performance suffers in high-dimensional data (the "curse of dimensionality")

Real-World Applications

KNN is used across various fields:

Recommendation Systems: Suggesting products or movies to users based on the preferences of similar users.

Image Recognition: Classifying images or handwritten digits based on pixel similarity.

Medical Diagnosis: Assisting in predicting disease likelihood by comparing patient data with similar historical cases.

Credit Scoring: Assessing the risk of loan applicants by comparing them to individuals with similar financial profiles.

Text Classification: Categorizing documents, such as for spam detection or sentiment analysis.

