K-Nearest Neighbors (KNN)

Introduction

K-Nearest Neighbors (KNN) is a simple, non-parametric, and supervised learning algorithm used
for both classification and regression tasks. It works by finding the 'K' closest data points (neighbors) to a query point and making predictions based on majority voting (classification) or averaging (regression).

How KNN Works

1. Choose the number of neighbors (K).
2. Calculate the distance between the query point and all training points.
3. Select the K nearest neighbors based on distance.
4. For classification: assign
the majority class among neighbors. For regression: take the average of neighbor values.

Distance Metrics

Common distance metrics include: - Euclidean Distance - Manhattan Distance - Minkowski Distance - Hamming Distance (for categorical variables)

Choosing K

The value of K greatly influences performance: - Small K → more flexible but sensitive to noise (high variance). - Large K → more stable but can oversmooth decision boundaries (high bias).

Advantages

- Simple and intuitive. - No training phase (lazy learner). - Works well with smaller datasets.

Disadvantages

- Computationally expensive for large datasets. - Sensitive to irrelevant or scaled features. - Performance drops in high-dimensional data (curse of dimensionality).

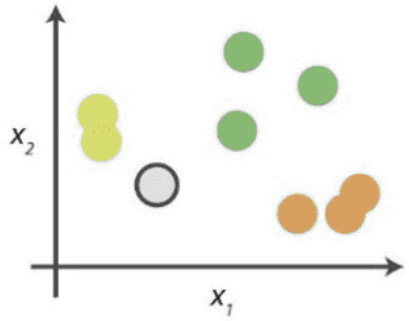
Applications

- Recommender systems - Image recognition - Medical diagnosis - Anomaly detection

Conclusion

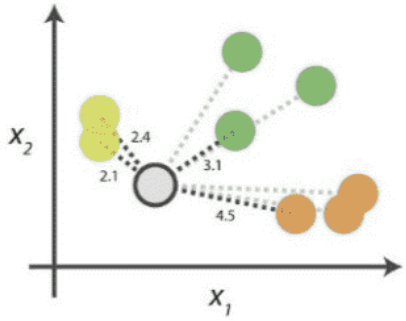
KNN is a powerful yet simple algorithm suitable for various applications. Its effectiveness depends on the choice of K, distance metric, and proper feature scaling.

0. Look at the data



Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances

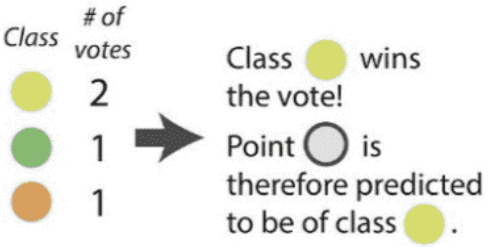


Start by calculating the distances between the grey point and all other points.

2. Find neighbours

Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels



Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.