Nearest Neighbors

The k-nearest neighbors (KNN) algorithm is a simple, instance-based learning method used for classification and regression tasks. The core idea behind KNN is to predict the label of a new data point based on the labels of its k-nearest neighbors in the training dataset.

The model is the training data itself. As a result, KNN is a heavyweight algorithm. Using it is time and memory intensive, especially with large datasets. To make things more efficient, we could use some sort of graph solution or partition the data in some way.

Design Decisions

We can use a similarity metric or distance metric to find the nearest neighbors. Common choices include Euclidean distance and Manhattan distance.

Should we standardize our feature? It depends on the data. If the features are on different scales, standardizing them (e.g., using z-score normalization) can help ensure that no single feature dominates the distance calculations.

Overfitting

Overfitting is more likely to happen when k is small because the model becomes too sensitive to noise in the training data. Choosing a larger k value can help smooth out the decision boundary and reduce the risk of overfitting, but it may also lead to underfitting if k is too large.