K-Nearest Neighbor Algorithm

Mahmoud Fathy

1 Introduction

The Machine Learning systems which are categorized as instance-based learning are the systems that learn the training examples by heart and then generalizes to new instances based on some similarity measure. It is called instance-based because it builds the hypotheses from the training instances. It is also known as memory-based learning or lazy-learning (because they delay processing until a new instance must be classified). The time complexity of this algorithm depends upon the size of training data. Each time whenever a new query is encountered, its previously stores data is examined. And assign to a target function value for the new instance. And an example of instance based learning algorithms is the K-Nearest Neighbor (KNN) algorithm.

2 How it works

The K-Nearest Neighbors (KNN) algorithm operates on the principle of similarity where it predicts the label or value of a new data point by considering the labels or values of its K nearest neighbors in the training dataset.

Selecting the value of K: K is the number of nearest neighbors that needs to be considered.

Calculating the distance: To measure the similarity between target and training data points Euclidean distance is widely used. Distance is calculated between data points in the dataset and target point.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{1}$$

Finding Nearest Neighbors: The k data points with the smallest distances to the target point are nearest neighbors.

Voting for classification: When you want to classify a data point into a category, the KNN algorithm looks at the K closest points in the dataset. These closest points are called neighbors. The algorithm then looks at which category the neighbors belong to and picks the one that appears the most. This is called majority voting.

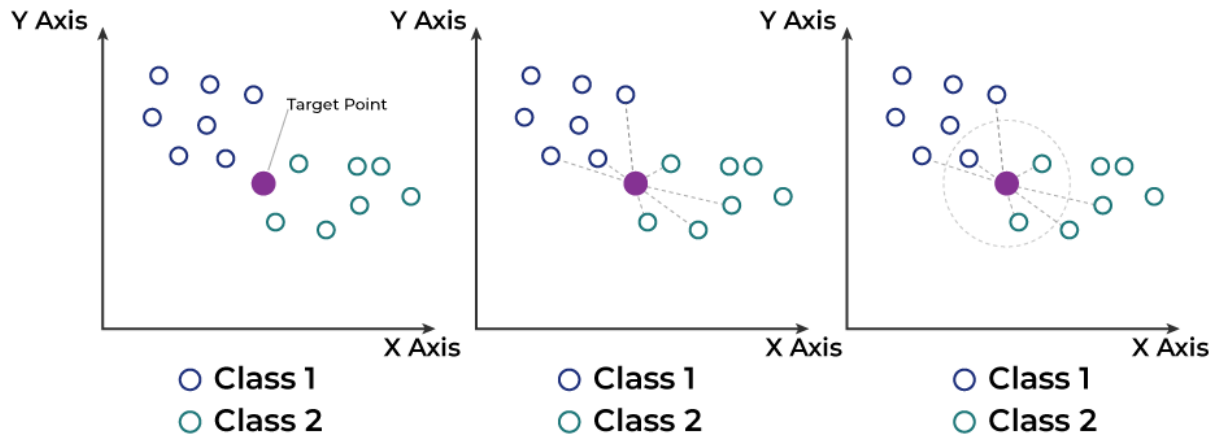

Figure 1: KNN

3 Pros and Cons

Advantages of KNN

Simple to use: Easy to understand and implement.
No training step: No need to train as it just stores the data and uses it during prediction.
Few parameters: Only needs to set the number of neighbors (k) and a distance method.
Versatile: Works for both classification and regression problems.

Disadvantages of KNN

Slow with large data: Needs to compare every point during prediction.
Struggles with many features: Accuracy drops when data has too many features.
Can Overfit: It can overfit especially when the data is high-dimensional or not clean.