Assignment Week 2

Due: 1/29/2024

Shimon Greengart

The knn or k-nearest neighbors algorithm is based on the idea that similar things tend to be similar. So, to categorize a new datapoint, simply check which of the data whose classification you know are closest to your unknown point.

Knn is a supervised learning algorithm. The way it works is that you have a set of labeled test data. When classifying an unknown vector, you compare its distance with all the vectors in your test data, finding the k closest vectors. In our homework for Professor Rosenfeld, we used linear, euclidean, and hamming distance, but there are others. You then check those vectors' tags, returning either the mode tag (for classification) or the mean tag (for numeric tags).

The advantages of knn are that it is effective with good performance. It is very simple to implement. It is also easily understandable by humans. Its downsides are that it is memory intensive, since you need to store the entire test dataset wherever you are running the algorithm. It is also slow to run: O(n*m), where n is the size of the dataset and m is the length of the vectors. Other machine learning algorithms only need to check the dataset when training, but knn effectively trains itself every time it runs. Finally, the choice of k, how many neighboring vectors to take into account, is somewhat arbitrary, so you need to use trial and error to get the best performance.