K-Nearest Neighbors and K-Means

Somani, Neel - Bao, Jason

April 17, 2018

1 K-Nearest Neighbors

Today we explore a particularly simple classification model: *k-nearest neighbors*. As the name suggests, classification is performed by looking at the k "nearest" neigbors to a data point, and classifying the data point as the same value as the majority of its neighbors. k is some natural number between 1 and N, where N is the number of points.

If k is large, we risk underfitting. If k is small, we risk overfitting.

Properties:

- Very flexible, can represent a large set of functions
- Number of parameters is variable
- Requires storing all data
- Training time is constant
- Poor behavior in high dimensions. Specifically, adding dimensions makes points farther away.
- Takes a long time to evaluate
- Easy to interpret

What are some improvements that we can make? For one, we can pick better (fewer) features, to make points closer together. We can also get more data.

2 K-Means

K-means is an unsupervised learning algorithm, means that it operates on data that is not labeled. We still want to find some structure in the data, though.

At a high level, we want to divide the dataset into K clusters (groups), where

each cluster has a centroid (the mean of the points in that cluster). We want to minimize the following objective function:

$$argmin_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{x \in C_k} (||x - c_k||^2)$$

where c_k represents the centroid of a cluster, C_i represents a group of points from X, our set of data points, and $\bigcup_{K}^{i=1} C_i = X$. How should we solve this problem?

That actually turns out to be a very hard problem. In fact, it's NP-hard. We have some heuristic algorithms, though. For example:

1. Initialize the clusters randomly. While C_k has not converged:
2. Update each C_k by assigning each x_i to the closest cluster.
3. Update c_k for each cluster.

Some shortcomings?

- Not guaranteed to work
- Hard to pick K
- Each feature is treated equally, which means that clusters tend to be spherical
- Elements are either in or not in clusters

Some other fancier techniques include soft k-means, Gaussian mixture models, Dirichlet Process GMM (DP-GMM), and spectral clustering.