CS 4980/6980: Predictive Data Analysis © Spring 2018

Lecture 7: Nearest Neighbor, k-Nearest Neighbors, and k-Means

Instructor: Daniel L. Pimentel-Alarcón Scribed by: Harnoor Singh and Jason Vaughn

7.1 Introduction

Nearest Neighbor is a method used for classification of data points through supervised learning. For an instance, given a labeled data set divided it to 2 or more classes, predict which class a new point X_{n+1} will belong to. k-Nearest Neighbors is used as a better alternative to this algorithm.

k-Nearest Neighbors is a non-parametric method used for classification and regression. As the labels are given to us, this method is called a Supervised Learning Model.
Goal: For an instance, given a labeled data-set divided into several classes, predict which category a new X point will lie.
This can be used for searching for semantically similar objects or even recommender systems.

k-Means Clustering is a clustering algorithm that tries to partition a set of points into K sets (clusters) such that the points in each cluster tend to be near each other. It is unsupervised because labels (classes/categories) are not given for the set of points.
This is useful for things such as clustering customers and segmenting them or trying to figure out where to open a new branch of a store with delivery.

7.2 Nearest Neighbor

In the Nearest Neighbor algorithm, the goal is to find the classifier of a new data point while being given a set S of points with known classifiers. The classifier of this new data point X_{n+1} is the classifier, K_{n+1}, of the nearest point to X_{n+1}. This is accomplished by finding the point i^* where:

$$i^* = \underset{i \in \{1,\dots,N\}}{\arg\min} ||X_{n+1} - X_i||_2^2 \tag{7.1}$$

The classifier that is given to i^* is now also given to X_{n+1}, meaning:

$$K_{n+1} = K_{i^*} \tag{7.2}$$

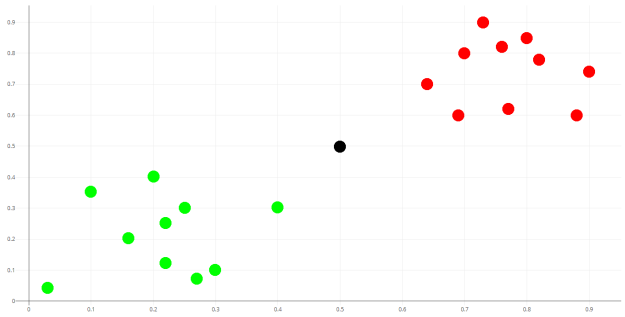

Figure 7.1: We only need to look at the closest neighbor to our new point.

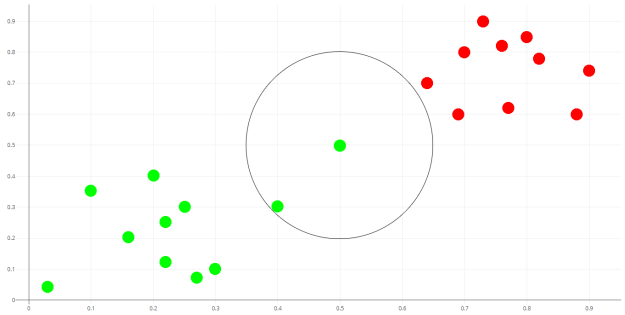

Figure 7.2: Looking at the nearest point only, the new point is classified as green.

7.3 k-Nearest Neighbors

k-Nearest Neighbors is just an extension of the Nearest Neighbor algorithm. The k in this algorithm is how many neighbors it is to consider. As such, Nearest Neighbor can be considered as 1-Nearest Neighbor. For k-Nearest Neighbors, we first find the k-closest points to a new point, X_{n+1}. From this, we take the argmax of the number of each classifier within our neighboring points. If two classifiers are tied for the most, we may randomly choose one of the tied classifiers for determining which our new point belongs to. k is usually chosen empirically.

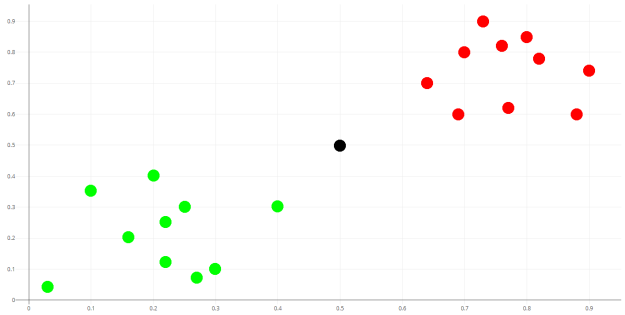

Figure 7.3: We need to look at the closest neighbors to find the new black point's classifier. We will use k = 3.

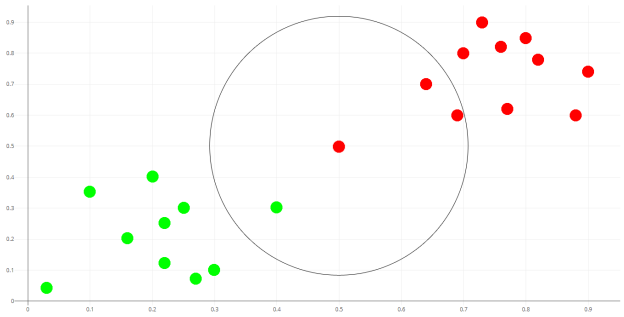

Figure 7.4: Looking at the 3 nearest points, the majority classifier is red, so the new point is classified as red.

7.4 k-Means

Goal: Find some centers ranging from $\mu_1, \mu_2, ..., \mu_k$ that minimizes the following equation:

$$\sum_{k=1}^{k} \sum_{x \in S_k} ||X_i - \mu_k||_2^2$$

Here S_k indicates points that correspond to the k^{th} cluster

We will be using **Lloyd's Algorithm** to solve the problem of k-Means by using the following steps:

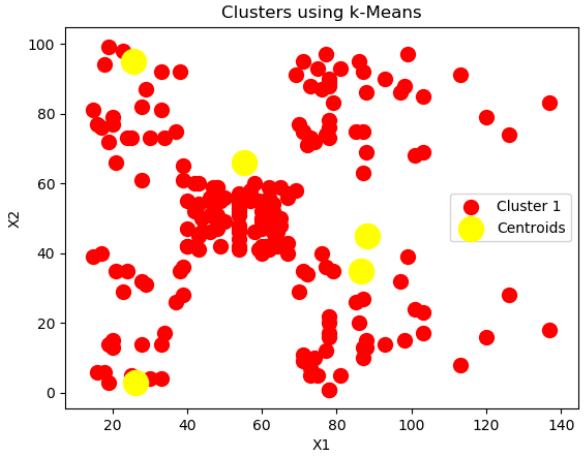

Figure 7.5: Initial Starting point with all points.

- Step 0: Initialize $\mu_1, \mu_2, ..., \mu_k$
- Step 1: Assign points $X_1, X_2, ..., \mathbf{X}_N$ to their nearest nearest neighbours center. This will produce the sets as $S_1, S_2, ..., S_N$

- Step 2: Compute new centers. The mathematical form is as follows

$$\mu_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} X_i$$

Repeat steps 1 and 2 until convergence. We know that the the centers will remain the same when the point of convergence arrives. Refer to figure 7.6 for the final graph.

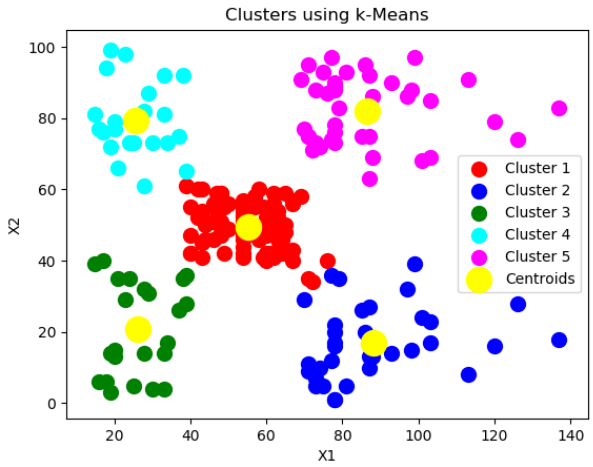

Figure 7.6: Final graph with total 5 clusters.

7.5 Conclusion

This lecture shows logic behind Nearest Neighbor, k-Nearest Neighbors and k-Means Clustering. While Nearest Neighbor and k-Nearest Neighbor predict the class in which a new data point belongs, k-Means classifies an unlabeled data set without any estimate for number of classes into different clusters. Both are completely different methods and they have different applications.