K-Means Clustering

Isaac G Callison

Middle Tennessee State University

Computer Science 5350: Introduction to Artificial Intelligence

Fall Semester 2019

**Introduction**

The purpose of this report is to discuss the concepts and practical implementation of a K-Means Clustering algorithm to classify simple data-sets. The K-Means algorithm is a relatively simple unsupervised machine learning algorithm used to classify data points. For purposes of this paper two data-sets will be used, the Iris and Cancer data-sets. As stated, K-Means is unsupervised, that means that the algorithm does not have the benefit classification labels when classifying data into different groups. The basic idea is grasped more readily if we think of a two-dimensional space where data points are graphed. If all the data points in the data-set are graphed in an X and Y plane, then we can imagine a two-dimensional space covered in dots. Similar data-points, which may likely represent the same classification, are likely to be graphed close to each other. So assuming that data-points with the same classification labels are likely to be clustered around each other, how do we identify and take advantage of this clustering? Also, if we are dealing with more attributes, we are dealing more dimensions than could be represented in a two-dimensional graph. For the Iris data-set, there are four attributes and thus four-dimensions.

The first step to clustering data for say, the Iris data-set, is to choose how many clusters we want.

**Program Basics**

**Program Implementation**

**Results**