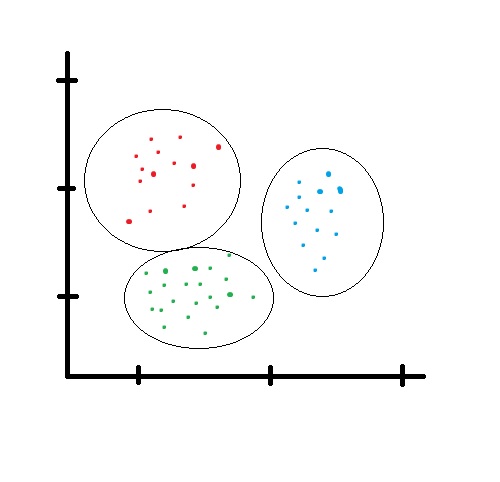
IE411 Term Project

Vlad Predovic

Monday, December 7th, 2015.

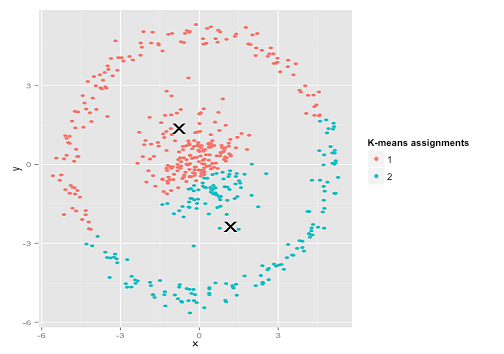
# Introduction

The Visual Basic program that goes along with this paper is designed to provide a quick and user friendly way to input data and perform cluster analysis. The program allows the user to pick between K-Means and K-Medoid, two different methods used for performing this type of analysis .

Cluster Analysis in a short statement is the grouping of data into separate categories based on a quantitative attribute. The grouping can be done through different methods but ultimately each data point becomes attributed to a distinct group. This results of this type of analysis are easier to understand graphically, but can be imagined to be somewhat like the picture on the right. A common use of this analysis is to try to differentiate data and see if it came from different sources.

*Fig. 1*

K-Means

http://www.purplemath.com/modules/xyplane/dist07b.gif K-Means is a common method employed to solve the clustering problem and a commonly associated mathematical implementation to solve it is labeled as Lloyd’s algorithm. It begins with defining an n amout of centroids within the range of your data set.There are many different ways to place these centroids but ideally you would want them spread throughout the data. After this you associate each data point to the centroid which it has the lowest distance from. Since this program only allows 2-dimensional objects, finding the distances can be done using the pythagorean theorem as in figure 2. Afterwards new centroids are deduced from all the objects that currently make up each cluster. The process loops, attributing each data point to a given cluster and deducing the center point until these stop changing altogether.

*Fig. 2*

The K-Means method has some drawbacks. The easiest to represent of these is the idea that it tries to create circular groupings of data as in figure 1. If you had an example with a different type of pattern, the K-means method would fail such as the one shown in figure 3 to the right where it is obvious that all the data in the middle is one cluster and the large ring on the outside is a separate set of data.

*Fig. 3*

K-Medoid

The K-medoid method is similar to the k-means method described above in that it attempts to attribute data data objects into clusters depending on the distance of these points to the designated cluster centroids. One of the more common realizations of this method is the Partititiong Around Medoids (PAM) method which begins by selecting random existing data points as your medoids. Each data point is then associated to the closest medoid. Each data point in a cluster is then taken as the medoid and the total distance cost of all data points to this point is calculated. The central point of the configuration with the lowest cost is selected and that point then becomes the medoid. The algorithm loops as all the data is then re-associated to the closest of the new medoids. Once the medoids do not change anymore the algorithm ends.

The main drawback of the K-medoid method is that it is usually much more expensive to run than the K-means method. In this case you have to calculate every distance from every point to all others each time you run an iteration. The latter method only requires that you calculate the distance to the given centroid points. However, the K-medoid method is considered more reliable since it is comparable to the median, and therefore less susceptibe to being displaced by outliers. The K-means method is more comparable to the mean, and therefore an aggressive outlying point can severely

# Implementation Details

This next section will go into over the advantages and drawbacks of the algorithms implemented in the visual basic solution. These are known as Lloyd’s algorithm and the Partitioning Around Mediods algorithm, used to solve the K-Means method and K-Medoids method respectively.

# User Guide

This section provides a comprehensive guide to running the visual basic program. It assumes the user is beginning on the opening page that appears when the solution is ran and details the many effects that different combinations of inputs will have on the solution. It also is aimed to walk the user through how to correctly set up and run the program, as well as how to interpret the results.

Section I: Guide Introduction

Upon starting up the program you should be taken to an initial form that introduces the program. The first page of the solution presents the user with the concept of data clustering. Afterwards, it gives a brief numbered explanation on how to run the program and how to manage your results.

Clicking next takes the user to a page which explains in a few sentences basics of the K-Means method. It then gives a quick listing of the pros and cons of this method meant to give the user a better idea of whether or not they should employ the K-Means method. Finally the page provides a sentence on what the data that program displays represents.

Clicking next a second time will take them to the K-Medoids page. The page is in the same format as the prior, with a quick introduction to the method and a listing of pros and cons, ending with an explanation of how the data is presented. In-depth descriptions of both these methods can be found in the Implementation Details section of this write-up.

Clicking next again will initiate the main part of the program.

Section II: The program

{{ INSERT PICTURE OF THE FINISHED PRGORAM HERE }}

Section III: Exported Data

How to export yo motha fukin data nergerrhhh