BASIC ASSIGNMENT 7
CORE METHODS IN EDUCATIONAL DATA MINING
PROFESSOR BAKER
CLUSTERING
DUE NOON, MONDAY DECEMBER 1

In this assignment, you need to conduct clustering on data set asgn7-clustering.

This is not a real data set, but is simulated data, generated for the purpose of this assignment.

Please conduct this assignment in RapidMiner.

Question 1:

Conduct k-means clustering with K=2, using k-Means. (Not W-SimpleKMeans). Use factory settings for all other parameters. Drag arrows between both of the output ports of k-Means to the results at the right. Look at the Centroid Table within the Cluster Model tab. Which two attributes have the biggest difference between cluster_0 and cluster_1?

Question 2:

Now look at the Plot view in the Example set tab. Set the X axis to be the first answer from Question 1, and set the Y axis to be the second answer from Question 1. (i.e. if you chose G and H, set X to G, and Y to H). Now set the color column to cluster. There are six major groupings ("lumps") in this data. How many of them are red?

Question 3:

What did k-Means do here?

A) It split the lumps in the data approximately evenly in clusters
B) It did a median split on two key variables
C) It found the least central lump in the data and made it a cluster
D) It found the most central lump in the data and made it a cluster

Question 4:

Now re-run k-Means with k=6. Did k-Means find the 6 lumps in the data that you saw earlier?

A) Yes

B) No

Question 5:

Plot each of the other variables against each other (not including the variables in question 1). Is there meaningful structure in any of these variables?

A) Yes
B) No

Question 6:

Filter out all of the variables except the ones in question 1, and re-run k-Means using just these two variables. Are all six of the six data lumps now more or less incorporated into six reasonable clusters?

A) Yes
B) No

Question 7:

What happened?

A) It looks the same as when k= 2
B) One region of space without a lump got a cluster, and two lumps got a single cluster
C) Several clusters were devoted to regions of space without a lump
D) Two regions of space without lumps got clusters, and three lumps got a single cluster

Question 8:

Try again with k=7. Are the six data lumps now more or less incorporated into six reasonable clusters?

A) Yes
B) No

Question 9:

For fun, you might want to try playing with different values of k, and the other parameters within k-Means. When you're done, try running Expectation Maximization Clustering with k=7. Look at the cluster probabilities for each cluster in Plot View. Which clusters are focused on a single data lump? (As opposed to including lots of outliers?)

A) All of them but a single outlier cluster

B) All but cluster 2 and cluster 6
C) All but cluster 1 and cluster 5
D) All but cluster 3 and cluster 4

Question 10:

For fun, you might notice that outliers close to the center of the top-right cluster still got placed into an outlier cluster. This is the power of having centers and radii. When you're done looking, try running Agglomerative Clustering. Look at the Dendrogram. Nifty, huh?

A) Yes, that is nifty.
B) I dispute the value of this question as assessment.

Question 11:

OK, fine. Squint really hard and look at the top-right of the dendogram. You'll see at the very top fork, that a branch goes down the right side. How many nodes are in this branch? (e.g. how many data points end up in this branch). Note that an immediate branch to a small subset of the data indicates strong outliers.