Introduction to Data Mining
Individual Project III: Clustering

Due: 11/23/2015 - 10:30am

What to Deliver

You should submit:

- A *Readme.txt* file, explaining how to run your script. (If there is any ambiguity in working with your functions)
- A report clearly showing the requested items.
- Three .R scripts, for each part. By running these scripts, we should be able to get your reported results.

Submit your files as a .zip file: **FirstnameLastname_UFID.zip**.

Part 1

For the first dataset (*dataset1.csv*) apply one clustering technique of your choice from the following categories:

- Distance-based clustering
- Density-based clustering
- Graph-based clustering

For each cluster assign a color, and plot (in 3D)

- Actual data
- Labels from your distance-based clustering
- Labels from your density-based clustering
- Labels from your graph-based clustering

In your report, you should show these plots and clearly reason why you would expect these outcomes from these methods. It is not required to use R for plotting.

Part 2

For the same dataset (*dataset1.csv*) write your own distance-based clustering method such that assigns weight to x attribute 4 times, and y attribute 2 times (z attribute is treated normally, with weight 1). Compare your method's result with *KMeans* clustering method. Clearly state why your method improved the outcome (or why it could not).

Part 3

For the second dataset (*dataset2.csv*) you should provide an analytical report, covering:

- How many clusters you think is the best to seek.
- Which methods can be applied for this case, and which one you think works best.

Based on your reasoning, choose your clustering method and apply it on the dataset. Report the accuracy of the method, as well as your challenges and solutions.

Please also note that, this dataset is too large to apply different clustering methods and check their results. The decision you make for the clustering method is the important aspect of this part.

Final Note: Cluster labels are provided to check the performance of your methods. You should exclude that (for both datasets) and row index (for *dataset2.csv*) before applying any clustering technique.