ISE 5103 Intelligent Data Analytics
Homework #6 OPTIONAL

Instructor: Charles Nicholson

Due: December 16, 2015 by 2:00pm

Learning objective: Perform cluster analysis and interpret the results.

Submission notes:

1. You will submit exactly two files: a written document and your complete R script.
   - *Provide comments* on what your code is doing. Keep it clean and clear!
   - Include `library` commands to load *all* packages that are used in the completion of the assignment at the beginning of your code.
2. Do not zip your files for submission. Name the files "LastName-HW1" with the appropriate file extension (that is, .R, .pdf, or .docx) – no HTML files.

1 Cluster Analysis

Perform a complete cluster analysis of a data set.

(a) Identify a data set for cluster analysis. The data set must have a *minimum* of 8 numeric variables and 500 observations. You may use *any* data set meeting this requirement that we have *not* used in class, homeworks, or projects. Options include (but are not limited to): R data frames included in packages, UCI Machine Learning datasets, Kaggle data sets, etc. I highly recommend that you choose a data set with variables that you understand or can explain conceptually (see part f).

(b) **Briefly describe the data set (1-2 paragraphs)**.

(c) Perform a clustering analysis of this data set using:
   - partitional clustering (use statistical and/or rational reasoning to determine k)
   - hierarchical clustering (use statistical and/or rational reasoning to determine cut height)
   - density-based clustering (use statistical and/or rational reasoning to determine ϵ and *min points*)

   Note: For each of the above provide *no more* than 2 pages of data, results, and/or code. **Choose the *right* information to include in your document.**

(d) Provide a concise, statistical description of the final cluster results for each method (e.g. number of clusters, size of clusters, centroids, etc.) and compare/contrast the results (1 page)

(e) Identify the cluster solution you think is best and provide a rationale for your choice. (1-2 paragraphs)

(f) Provide an *interpretation* of the clusters belonging to your preferred solution. (2 page max) This interpretation will require a subjective analysis of the clusters.