APSTA-GE.2011 Professor Marc Scott
Project #s 1 & 2 January 2019

Project #1 is due 8am January 18, 2019 (a Friday) and is a slightly guided cluster analysis.

Project #2 is due in three parts: i) PROJECT STRUCTURE MUST BE APPROVED BY ME BY TUESDAY January 22; ii) A partial draft must be completed by Friday January 25, 8am, so that we may discuss it in class; iii) The full draft is due Thursday January 31st, 2018, 8am.

The project is YOUR CHOICE BETWEEN: i) a cluster analysis organized in a manner similar to Project #1, but using your own data (or data we agree on, such as the Stop & Frisk data); (ii) a classification analysis using your own data (or data we agree on, such as the Stop & Frisk data); (iii) a simulation study comparing and contrasting methods to determine their sensitivity to modeling choices (e.g., choice of distance metric, clustering method).

Group work: Project #1 is to be done individually; Project 2 may be done in groups up to size 3, but they must be approved by me (the main criterion I use for approval is a balance of skills and experience in each group). Group projects must necessarily tackle more and/or harder problems. This is why the approval process is so important.

Project #1 DESCRIPTION

The objective of a cluster analysis is knowledge discovery – somehow, by identifying groups in the data, you learn something interesting about the substantive area being explored.

You will look for potential clusters in the Australian Leptograpsus Crabs data. As you know from the handouts, 200 crab specimens were collected at Fremantle, Western Australia in the mid-1970s (Campbell and Mahon, 1974). Each specimen has measurements on: frontal lip (FL), rear width (RW), length of midline of the carapace (CL), maximum width of carapace (CW), and body depth (BD), all in millimeters.

You also know the sex and species of these crabs – these are the *demographics you will explore **after** clustering.*

First, explore the five features using bivariate plots. You should explore the need to *transform* or *rescale* the measurements. **Make a recommendation based on those bivariate plots [CLARIFICATION: I want you to state whether you recommend transforming (e.g., take logs) or rescale (e.g., standardize) based on some data exploration. You will not actually change data yet]**. However, in order to save time, we have decided (for you) that you should standardize the measurements (the usual z-score transform). The simplest way to standardize is to make a NEW crabs dataframe as follows:

```
crabs.stdz <- crabs;
crabs.stdz[,4:8] <- scale(crabs[,4:8])
```

from this point forward, you would use crabs.stdz in your analysis (not crabs).

NOTE THAT when examining YOUR OWN data, e.g., in Project 2, you should always consider whether you need to transform (e.g., take logs) and/or standardize; you explore transforms in part by looking at univariate densities, and then replotting the bivariate plots with new measures.

You should also examine bivariate plots using principal components on the standardized version of the data, as these might reveal the clusters better. **Having already standardized the data, you can just use the function `princomp` to generate the principle components, but remember that only columns 4:8 contain features.** *Do the actual clustering on the raw (standardized) measures, not the principal components.*

As a first approach, if you see fairly well separated clusters, particularly if they are "stringy," you can use **single linkage** hierarchical clustering; otherwise, use **centroid linkage** [justify your choice in your writeup, *but only choose ONE method*]. We will assume that Euclidean distance (L2 norm – *not squared*) is appropriate for these data.

First, choose the number of clusters (you think provide good separation between groups and homogeneity within) by examining the dendogram and evaluating several alternative "cut points" for the number of clusters.

Next, determine the optimal number of clusters based on a criterion: compute the ratio $C(g)=(\sum \text{msb})/(\sum \text{msw})$ and choose the g such that C(g) is maximized. **There is another package, NbClust, which will compute C(g) for you – it is index 'ch' in the output. I recommend trying the program NbClust in R's cluster library, as you will save time over computing C(g) manually (students who use STATA will find it somewhat easier to compute C(g) with STATA).**

As a comparison approach, redo the analysis using **k-means** clustering (make sure you use enough random starts to have a consistent result), and **SEARCH FOR optimal number of clusters for this method,** again determined by C(g). **NbClust is probably the easiest way – there is a kmeans option**.
Here is some code to get you started:

```
NbClust(crabs.stdz[,4:8],method='centroid',index='ch')
NbClust(crabs.stdz[,4:8],method='kmeans',index='ch')
```

I leave it to you to find a way to report and use the information you get from this very convenient function.

Compare the results from these last two methods, e.g., optimal using C(g) and centroid or single linkage depending on your prior choice and the **optimal k-means** result. **Use a crosstab comparison.** State the maximal agreement between methods (and justify using the crosstab). Evaluate the distribution of the known demographics (sex, species) for the k-means cluster solution (**you can use a crosstab here as well**). Do the clusters seem to divide in a manner consistent with demographic differences? Justify your answer by comparing the frequency distribution of demographics within each cluster.

PROJECT #2: DESCRIPTION OF "PROTOTYPICAL" CLUSTERING

Using Project #1 as a guideline, choose your own data that contains at least one demographic (as well as more than two features) and perform a cluster analysis. You are welcome to choose methods other than single linkage, centroid, or k-means. As examples (you can be more creative here), you could compare Ward's with Model-Based Clustering (if you use R), or you could compare k-means to Nagin Clusters (if you use Stata). Choose methods that relate to the type of data you are using and justify the choice.

Here is a description of how you should write up the project:

At a minimum, your analysis should include:

1. Brief (1 paragraph) description of the data source and measurement scales associated with your feature set.
2. Any additional variables that have been measured, not part of the clustering, to which you wish to make comparisons. Good examples are demographics, such as gender, race, ses, region.
3. A brief (1 paragraph) outline of your analysis plan, which should include:
    a. Any pre-processing transformations you will perform, and/or whether you will do the clustering on the original scale, transformed/ standardized scale, or use (a full set of) principle components.
    b. Choice of clustering method (and rationale)
    c. How you will determine the number of clusters (C(g) is only one approach, there is 1-Wilk's Lambda, BIC, e.g.)
4. The write up should go into more detail about the choice of the number of clusters, and then provide:
    a. Graphical display of the clustering (e.g., dendogram, if you used a hierarchical approach) and "best" cluster solution (decide and justify)
    b. Comparison to alternative choices for clustering (if any were explored)
    c. Description of the clusters with respect to the feature set and the additional variables (demographics)
    d. Any substantive conclusions you may be able to draw from the analysis (briefly)