Computational Methods for American Politics
Autumn 2019

Problem Set #1: Clustering

*Remember to submit a **single** rendered PDF or HTML file (either from .Rmd or a Jupyter Notebook) via GitHub **by Friday at 5 pm**: https://github.com/macss-cmap19*

1. Load the state legislative professionalism data from the folder. See the codebook for reference in the same folder and combine with our discussion of these data and the concept of state legislative professionalism from class for relevant background information.

2. Munge the data:
    a. select only the continuous features that should capture a state legislature's level of "professionalism" (session length (total and regular), salary, and expenditures);
    b. restrict the data to only include the 2009/10 legislative session for consistency;
    c. omit all missing values;
    d. standardize the input features;
    e. and anything else you think necessary to get this subset of data into workable form (hint: consider storing the state names as a separate object to be used in plotting later)

3. Diagnose clusterability in any way you'd prefer (e.g., sparse sampling, ODI, etc.); display the results and discuss the likelihood that natural, non-random structure exist in these data.

4. Fit a simple agglomerative hierarchical clustering algorithm using any linkage method you prefer, to these data and present the results. Give a quick, high level summary of the output and general patterns.

5. Fit a k-means algorithm to these data and present the results. Give a quick, high level summary of the output and general patterns. Initialize the algorithm at k=2, and then check this assumption in the validation questions below.

6. Fit a Gaussian mixture model via the EM algorithm to these data and present the results. Give a quick, high level summary of the output and general patterns. Initialize the algorithm at k=2, and then check this assumption in the validation questions below.

7. Compare output of all in **visually** useful, simple ways (e.g., present the dendrogram, plot by state cluster assignment across two features like salary and expenditures, etc.). There should be several plots of comparison and output.

8. Select a single validation strategy (e.g., compactness via min(WSS), average silhouette width, etc.), and calculate for all three algorithms. Display and compare your results for all three algorithms you fit (hierarchical, k-means, GMM).

9. Discuss the validation output.
    a. What can you take away from the fit?
    b. Which approach is optimal? And optimal at what value of k?
    c. What are reasons you could imagine selecting a technically "sub-optimal" partitioning method, regardless of the validation statistics?