Stat 897 DAA 12

Penn State

December 3, 2017

In this assignment we will use the `NCI60` data found in the `ISLR` library.

(a) Run k-means clustering on the data using k = 3. Next, use the elbow method to choose an optimal number of clusters (based on total within sums of squares). Is there a clear choice? What is a potential way to choose the optimal k when the elbow is visually ambiguous? (Note: this is an open-ended question. I'm not looking for a specific answer, but for you to use your intuition.)

(b) Tabulate the clusters for k = 3 against the clusters using your optimal k. What do you observe?

(c) Now perform hierarchical clustering using both single and complete clustering. Plot the dendograms.

(d) Cut the trees to obtain the number of clusters you found optimal for kmeans. Tabulate the clusters for both single and complete versus the kmeans clusters. What do you observe? Based on the dendograms, does cutting the trees at this point make sense?

(e) Repeat parts (c) and (d) using a different distance measure (than euclidean). Give a reason for your choice. What differences (if any) do you see when you tabulate the results?

(f) Using PCA, pull out a number of principal components for the `NCI60` data. Explain your choice of number of PCs.

(g) Using these PCs, repeat kmeans clustering and hierarchical clustering (with your preferred distance function and both linkage methods). Compare the results to the results for the corresponding method on the original data.

(h) For which method(s) do you observe a substantial change in the clusters based on the original data versus the PC? Which method(s) don't change as much?