Visualization and Learning Structure: Problem Session 7/14/16

Sticking with our famous olive oil data set used in clustering and classification. We have eight different chemical composition measurements on 572 Italian olive oils. Additionally, there are two sets of labels, the Region and the Area. There are three different Regions (Southern Italy, Sardinia, and Northern Italy).
Each region is comprised of a group of area: Southern Italy = North Apulia, Calabria, South Apulia, and Sicily; Sardinia = Inland Sardinia, Costal Sardinia; Northern Italy = Umbrian, East Liguria, and West Liguria.

Our current goal is to look for cluster structure in the olive oils and see how well it matches the two sets of labels.

1. Download the olive oil data from our website and read into R.

   `olive<-read.table("olive",header=T)`

   Use `attach(olive)` to create a variable for each column.

   Use `chem.meas<-olive[,3:10]` to create a matrix of variables to cluster.

2. Use single linkage hierarchical clustering to build a dendrogram for the eight chemical measurement variables. You can find the distance matrix using `dist()` (default is Euclidean). `hc.olive<-hclust(dist(chem.meas),method="single")`

3. We might first be interested in comparing our cluster structure to the Region labels. Plot your dendrogram using `plot(hc.olive,labels=Region)`.

   Then cut your tree to extract three clusters: `hc3<-cutree(hc.olive,k=3)`

   Compare your cluster labels to the Region labels using: `table(Region, hc3)`.

   If you want the misclassification rate, you can also use `classError(Region, hc3)` (you'll need to download/install the `mclust` library first).

   How did we do?

4. Now repeat (3) using complete and average linkage. Did we improve our misclassification rate? Are some Regions easier/harder to cluster than others?

5. We could also experiment with different distances when clustering.
   Try using `dist(chem.meas,method="manhattan")` or
   `dist(chem.meas,method="minkowski",p=3)`. See `help(dist)` for more.

   Did the change in distance measure change anything in your results?

6. Now we're interested in seeing how well we recover the Area labels. Experiment with different linkage types and distances to find your best nine-cluster solution.

   How did you do? Are some Areas easier/harder to find?

7. So far we've been choosing three or nine clusters because they are the suggested number of clusters based on the Region/Area labels. However, there might a better number of clusters that show more stable, robust separation.

   Looking through your dendrograms, how many clusters would you suggest?

8. Switching to k-means and partitioning our data into spherical clusters, find a three-cluster k-means solution for our chemical measurements. Compare the cluster labels to the Region labels.

   ```
   km3<-kmeans(chem.meas,centers=3)
   table(Region, km3$cl)
   classError(Region, km3$cl)
   ```

   How did we do?

9. Similarly, find a nine-cluster k-means solution for comparison to the Area labels.

   Better or worse than using hierarchical clustering?

10. Just as with hierarchical clustering, we need to choose K, the number of clusters. We'll use an *elbow graph* to visualize the improvement in the clustering solution as a function of increasing K.

    For an elbow graph, we plot the total within-sum-of-squares (the criterion we're minimizing) against K. We're looking for the point where increasing K doesn't necessarily improve the clustering solution enough to justify the use of more clusters, i.e. the elbow point.

    ```
    wcc<-NULL
    for(i in 2:15){
    wcc<-c(wcc,sum(kmeans(chem.meas,i)$withinss))
    }
    plot(seq(2,15),wcc,type="b",pch=16,xlab="Number of Clusters",
    ylab="Total Within Sum-of-Squares")
    ```

    Which K would you choose?