Christopher Morgan
CS478 : Brother Christophe
Oct 23, 2013

Perform the following activities: (note that for your experience, I am encouraging you to do this homework using the R software, as it is one of the richest and most versatile freeware tool for statistical analysis. If you prefer, you may also perform the assigned activities in Weka. Needed algorithms and visualizations are under the Cluster tab.)

---

1. **Download and install the latest version of R on your computer. See the link on our syllabus under Resources/Software.**
   Done
2. **Load the R Stats Package and the R Datasets Package by typing 'library(stats)' and 'library(datasets)' respectively at the R prompt. Alternatively, you may use the Package Installer and the Package Manager of the R GUI to load these packages. Details on the various functions implemented in each package, as well as examples of usage may be found in the Package Manager by selecting the package of interest.**
   Done
3. **Run the k-means algorithm (kmeans) on the iris dataset (iris was loaded above when you loaded the R Datasets Package). Of course, iris has a target attribute. It must be excluded from the clustering. The simplest way to do this is to create a copy of iris consisting of only the first 4 attributes. This can be accomplished with the command: 'iris_copy <- subset(iris, select=c(1:4))'. You can then run kmeans on iris_copy.**
   - **Run k-means for k=2,3,4,5,7,9,11.** (5 Hours) (Figuring R out, Writing Program, Resolving PowerPoint issue and Re-adjusting) 'This seemed horrible!'

   - **For each value of k, report the size of the clusters and the F-measure.**

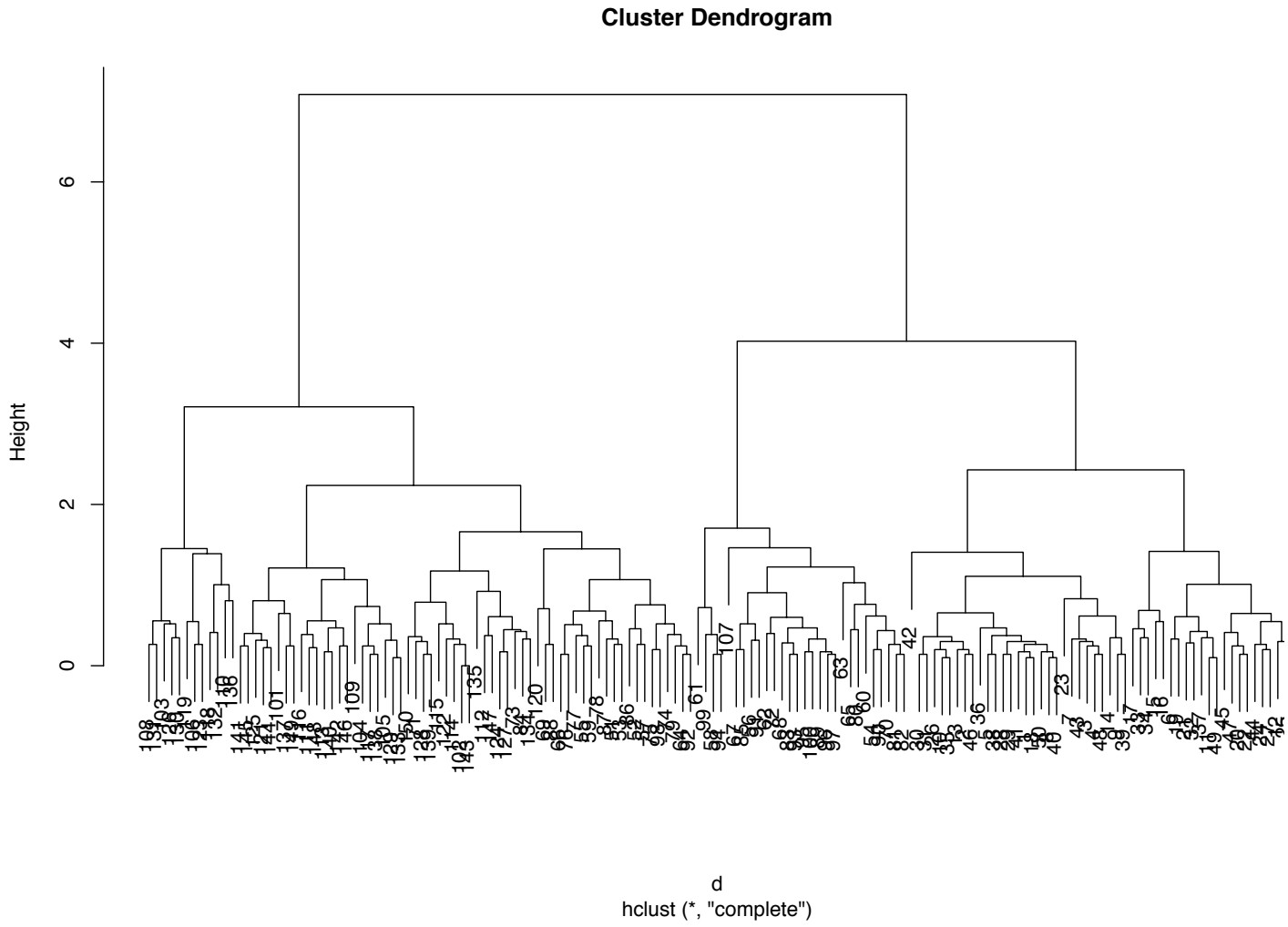
K	Size of Cluster	F-Measure
2	53,97	0.7321
3	50, 62, 38	0.8241
4	22, 62, 38, 28	0.7211
5	24, 12, 50, 39, 25	0.7215
7	22, 24, 20, 12, 28, 21, 23	0.5739
9	10, 12, 21, 36, 7, 16, 7, 17, 24	0.4788
11	7, 17, 20, 11, 6, 21, 5, 24, 23, 12, 4	0.4708

   - **Report the value of k that produces the highest F-score.**
     K == 3 Produced the highest F score.

   - **Comment on anything interesting about your experiment.**
     It appears that for Iris the F-Measure is higher for sizes of K such that K is less than or equal to the number of actual classes in our data set. This might be because after that point we begin to over fit the data. It is also interesting that

when K equals the number of actual classes we have the highest F-Measure. This might be because this is the point at which we have the exact number of classes as we do cluster points, meaning this is the best fit for the data.

4. **Run the hierarchical clustering algorithm (hclust) on the iris dataset using complete link for the distance.**
   - **Display and include in your report the result of hierarchical clustering.**



- **By looking at the display or using the values of clustering heights, select a threshold at which you feel the clustering would be optimal and justify your choice.**

  Given what little we have talked about this, I would suggest stopping at height 3. It appears that from the graph at this level there are three different clusters and there are three different classes in our data set. Therefore, to me this sees like a good place to stop.

It also appears that at this point the number of converging clusters has begun to settle down, this would suggest to me that all of the close points have now been group together.

- **How does the corresponding number of clusters compare with that obtained with k-means above?**

  They seem to line up and support one another. K-means seems to favor 3 clusters, where hclust also seems to propose about 3 clusters.

5. **Consider the swiss dataset. Use clustering, either k-means or hierarchical clustering (whichever seems to make most sense), to produce a list of the Swiss cities predominantly protestant and those predominantly catholic. You may produce a graph or simply a list.**

   Copy_swiss<- subset(swiss, c(1:4,6))
   Kmeans(copy_swiss, 2)

   Produced the following list for predominantly catholic:

   ['Delemont', 'Franches-Mnt', 'Moutier', 'Neuveville', 'Porrentruy', 'Broye', 'Glane', 'Gruyere', 'Sarine', 'Veveyse', 'Aigle', 'Aubonne', 'Avenches', 'Cossonay', 'Echallens', 'Lavaux', 'Morges', 'Moudon', 'Nyone', 'Orbe', 'Oron', 'Payerne', "Paysd'enhaut", 'Rolle', 'Yverdon', 'Conthey', 'Entremont', 'Herens', 'Martigwy', 'Monthey', 'StMaurice', 'Sierre', 'Sion', 'ValdeRuz']

   Produced the following list for predominantly protestant:

   ['Courtelary', 'Grandson', 'Lausanne', 'LaVallee', 'Vevey', 'Boudry', 'LaChauxdfnd', 'LeLocle', 'Neuchatel', 'ValdeTravers', 'V.DeGeneve', 'RiveDroite', 'RiveGauche']