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**](http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Syllabus-F12.html)**under Resources/Software.**

Done

1. **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

1. 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.
   * For each value of k, report the size of the clusters and the F-measure (see [this](https://learningsuite.byu.edu/Slides/CS%20478%20-%20Clustering%20Evaluation.pdf) for details). Both size and cluster assignments are available in variables computed during k-means (see the documentation in the R Stats Package under k-means). You will need the target values from the original iris dataset to compute the F-score. You may write a small program in R to do this or export the data and compute elsewhere.

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

* + Report the value of k that produces the highest F-score.
  + Comment on anything interesting about your experiment.

1. Run the hierarchical clustering algorithm (hclust) on the iris dataset using complete link for the distance. Be mindful that hclust requires a distance matrix rather than a set of points as input. You can easily transform a set of points into its equivalent distance matrix using the dist() function. From iris\_copy, you could thus construct iris\_dist with the command: 'iris\_dist <- dist(iris\_copy)'. You can then run hclust on iris\_dist.
   * Display and include in your report the result of hierarchical clustering.

Macintosh HD:Users:clm:junk.eps

* + 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. (In principle, we would do this by computing some quality measure during the clustering process, but for simplicity, we are just eyeballing here).
  + How does the corresponding number of clusters compare with that obtained with k-means above?

1. 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.