Clustering in R - STA 521

Abbas Zaidi

1 Agenda

1. CH Index
2. The clusterCrit Package
3. paste() Function
4. Assigned Reading

2 Lab Tasks

1. Read in the file titled Wimbledon.csv into the variable myTennisDataD. In your read.csv() remember to set the stringsAsFactors() parameter to FALSE. Remove the first two columns of your dataset, paste the names together, and replace the row numberings with these match names.

2. There are two columns in your dataset that contain only NA values. Remove these. For all other columns that contain missing values, replace the missing values with the median values for that column. Store this cleaned dataset into the variable iFinalTennisData

3. Create a log-Euclidean distance matrix for your data and perform single linkage and complete linkage clustering on the data and store these in the variables singeLinkage and completeLinkage respectively

4. Using the command intCriteria() in the clusterCrit package, write a function that takes as its inputs the number of clusters to be tested (say 10), a clustered object (e.g SingleLinkage from above) and the original data frame (e.g. iFinalTennisData), and computes the CH Index for each assumed number of clusters and creates a plot of the resulting CH indices for each cluster count with a dotted vertical line indicating the maximum value. Also return the maximum CH Index computed. Test this function for up to 10 clusters with both single and complete linkage on iFinalTennisData, and include the maximum CH Index calculated and the related plots. What should the CH Index value be when the number of clusters is 1 and why?

5. For each clustering (based on linkage) what is the optimal number of clusters?

6. Create a dendogram for each of your two clustering assignments. Based on the data type, what is the most appropriate type of clustering and why?

3 Directions

In general for Labs, at the top of any file you are asked to submit, please list the following:

1. First Name Last Name

2. Lab Date

3. Team Member(s)

With respect to any item for which you are asked to generate any output, please provide the actual R output as a part of your solution and any explanation needed as well. For any functions/ computations that you will write, please list the following as comments before the step in R:

1. Task number and descriptions.

2. Input(s) with descriptions.

3. Outputs(s) with descriptions.

4. Function/ output summary (along with intermediate step comments).

For Lab 5, please provide the following deliverable items:

1. Please provide your solutions using Markdown as a .pdf with the following naming convention: LastName_FirstName_Solutions_Lab5.pdf.

2. Provide your .Rmd file (this **MUST** compile) for the lab using the following naming convention: LastName_FirstName_Solutions_Lab5.Rmd