IS590 DT

Assignment 5

Release Date: 02/27

Due Date: 03/06

Concepts:

1. What's X matrix and what's D matrix? Try defining both and explain why we need both.
2. Give your definition on K-Means and K-Medoids.
3. What're the factors of KMeans computation complexity (given X)? How about the KMedoid (given D)?

Case study:

4. If you are a biologist and is right now learning the genome sequence. How would you define the distance of the following sequences? Try at least 3 ways to calculate the distance of the two genome sequences. (Hint: you may define distance on the level of hint, condon, string, etc. You could also define your own distance. You can assume having any information that is useful for your definition.
    a. A - 'ATCTCGGGCATC'
    b. B - 'GTCACGGGCGTC'
5. (K-Means practice). Given dataset A(1,1), B(2,0), C(3,0), D(5,5), E(7,5), F(4,6), and the initial seed X(0,0), Y(3,2). Answering the following questions:
    a. In round 1: which points are clustered in Class X? and which in Y?
    b. What's the new centroids? (keep two decimal places, if needed.)
    c. In round 2: which points are clustered in class X? and which in Y?
    d. What's the new centroids? (keep two decimal places, if needed.)
    e. (bonus) Can you find a bad seed, which generates the false clusters?

Experiment:

6. R application for KMeans and KMedoids.
    a. Download the datafile (sportsranks.txt) and the R code (hw5_clustering.R) from github repository (https://github.com/yingjun2/IS590dt-19SP/tree/master/wk7_clustering).
    b. Complete the code to implement both KMeans and KMedoids, compare the results, and summarize your findings.
    c. Learn a statistical parameter, prediction strength [1], Gap$_k$ [2][3] or others. Introduce it in your report. You should also be able to implement it in R (for d).
    d. Introduce elbow point [4] in your own words in report. Use R to implement the elbow point methods utilizing the KMeans in b and the stat parameter you select in c. Try using clusters from 1 to 10, and discover your optimized cluster number. (Complete the code when in need.)

Submission requirement:

a. Learn to use R markdown, knit a pdf file [5][6] and have it submitted.
b. Submit your pdf file and your .rmd file.

Reference:

1. https://www.rdocumentation.org/packages/fpc/versions/2.1-11.1/topics/prediction.strength
2. https://datasciencelab.wordpress.com/2013/12/27/finding-the-k-in-k-means-clustering/
3. https://statweb.stanford.edu/~gwalther/gap
4. https://www.r-bloggers.com/finding-optimal-number-of-clusters/
5. https://www.earthdatascience.org/courses/earth-analytics/document-your-science/knit-rmarkdown-document-to-pdf/
6. https://rmarkdown.rstudio.com/authoring_quick_tour.html