Model-based clustering in R-Markdown: parametric and nonparametric methods

- A REU project supervised by Dr. Jiancheng Jiang

Department of Mathematics and Statistics, UNC Charlotte

E-mail: jjiang1@charlotte.edu

In this project, students learn how to use resampling techniques to make statistical inference for model-based clustering. Given a sample, various resampling procedures, such as bootstrap, jackknife, cross-validation, and randomization techniques, will be employed to estimate the model and to obtain accuracy of clustering. Software R with parallel computation is expected to be used. The methodology used in this project will be reported and implemented through the R-Markdown on R studio. The completed project will be used for classroom teaching and (or) for publication if appropriate.

Keywords: Clustering, MAP, Mixture Distribution, Parallel Computation, Resampling, Smoothing.

- Prerequisites:
  - Software: one-year experience with R
  - Course: a course in multivariate analysis

- Reference:
  - Lecture Notes (lec_13_Resampling.pdf, Lec_15_Parallel Computing.pdf)
  - R codes: rmarkdown.Rmd

- Tentative task:

1. To sample from a mixture of parametric or nonparametric models，e.g. $X \sim \pi_1 f_1(x) + \pi_2 f_2(x) + \cdots + \pi_K f_K(x)$, where $\sum_{k=1}^{K} \pi_k$=1 and $f_k(x)$ are pdfs of d-dimensional r.v. representing possible different populations from which X may come. When $K = 2$ and f_k are normal, it is so-called the mixed normal distribution. The sample we have is $X_1, \ldots, X_n$.
2. To visualize the sample data.
3. To estimate the model
4. To cluster the data and to make comparisons with K-means or others.
5. To resample from the original sample (random weighting or bootstrapping)
6. To estimate the bias, variance, or distribution of the estimator (bootstrap or jackknife); Parallel computation is to be made.

7. To assess accuracy of the resulting cluster.
8. To visualize/summarize the accuracy by varying the sample.
9. To cluster a real dataset.
10. To write a report on the project in an R-Markdown file. It is necessary to report the results in tables and graphs (or videos).
11. To run your R-Markdown codes and to generate a html or ppt document for presentation.