Fall 2021 — STAT 2221: Advanced Applied Multivariate Analysis

Homework 4. Due by Thursday, October 14, 2021 at 9:00 a.m. EST

Topic: Cluster analysis

Name: HERE

Additional instructions. Add `\newpage` immediately before each problem so that each has its own page. Add `\begin{proof}[Solution.] ... \end{proof}` below each problem for providing your solution. You are welcome to add additional packages to the preamble, but do not modify the existing commands and formatting.

Problem 4.1. Consider the `data1`, `data2`, `data3` synthetic data sets, available on Canvas. The file `alloc_vecs`, also on Canvas, consists of two column vectors of ground-truth cluster labels: "alloc1" for `data1` and "alloc2" for `data2`.

1. Conduct an exploratory cluster analysis of each data set, using k-means, Gaussian mixture modeling, and hierarchical clustering. For each approach, discuss its pros and cons. Be sure to also provide justification for your method-specific choices (e.g., any assumptions you make, etc.).

2. Investigate and write brief overviews of the *Jaccard index*, *Rand index*, *adjusted Rand index*, and *normalized mutual information* clustering criteria (see class sides for references). Then, using these different indices, compare your cluster outputs in step 1 with each other and with the ground truth provided in the file `alloc_vecs`.