STATS 780/CSE 780, Winter Semester 2021

ASSIGNMENT 4

Submit to Crowdmark using the link that was emailed to you.

Due at 10.00am on Friday, March 5th.

Your assignment must conform to the Assignment Standards listed below.

Assignments submitted up to 24 hours late will incur a 30% penalty.
Assignments submitted more than 24 hours late will receive a zero grade.

Find a data set that is suitable for cluster analysis; you may not use a data set that was used for cluster analysis in class. Briefly describe your chosen data set and clearly explain where it was sourced. Carry out a thorough cluster analysis of your chosen data set using:

- agglomerative and/or divisive hierarchical clustering;
- k-means and/or k-medoids clustering; and
- model-based clustering **(use the GPCM family)**.

Provide a clear and concise description of the results. Your report must include a comparison of the clustering results obtained using these methods. Clearly state what conclusions can be drawn from your analysis; try to say something in the context of your chosen data set.

The marking scheme for the assignment is as follows:

- Description of data including source, details on variables, relationships between variables, descriptive analyses, etc. [5 marks]
- Description of the problem being addressed. [3 marks]
- Description of the techniques used. [6 marks]
- Descriptions of results, both in technical terms and in the context of the data. [6 marks]
- Conclusions, in the context of the data and the problem being addressed. [5 marks]

Assignment Standards

- LaTeX is strongly recommended but not strictly required.
- Eleven-point font (times or similar) must be used with 1.5 line spacing and margins of at least 1 inch all around.

- Your report may not exceed **three pages**, inclusive of tables, figures, and bibliography.
- Do not include a title page. The title and your name should be printed at the top of the first of the three pages.
- The writing and referencing should be appropriate to the graduate level.
- Various tools, including publicly available internet tools, may be used by the instructor to check the originality of submitted work.
