Stats 503 - Project, 2018

Instructor: Long Nguyen

Description. This project is designed to provide an opportunity to apply multivariate and categorical data analysis techniques to analyze a data set of your choice. Given a data set, you are asked to perform the following:

1. do as much data exploration as possible by visualization, summary statistics and a dimension reduction technique (if applicable)

2. apply <u>some</u> of the methods learned in the class (clustering, classification/regression, mixture modeling, hiearchical and graphical modeling)... You should discuss your implementation, data analysis results, and the implication of various choices for parameters selection. You are also welcome to consider more advanced methods not learned in the class, if the data suggests so.

3. where to get the data? A place to start looking is the UCI Machine Learning Repository: `http://archive.ics.uci.edu/ml/`
   However, it is *better* to consider data sets other than those those from the link above.

4. This is a team project consisting of three people. Projects will be evaluated in terms of <u>the amount</u> and <u>the quality</u> of work performed on the data analysis and the exposition contained in the project report.

Desiderata. The choice of data sets will be subject to the instructor's approval. In general, I would like to see teams taking on new and challenging data sets, and formulating novel inferential questions. The project can be open-ended. What to *avoid*:

- older data sets that have already been tortured to death

- data sets that have little information about the data attributes/variables and inferential questions that can be asked of them

- a project that simply runs through a zoo of classification techniques without any substantiave data analysis

Relevant dates.

1. *Proposal.* Each team submits to *both* the instructor and the GSI via email a project proposal, including information about team members, the data set, and the methods they intend to apply with a brief justification. <u>Please include [503 project] in the subject line.</u> The proposal should be approximately no more than one page. The deadline for the proposal is **Thursday, March 8**.

2. *Final Project report.* Each team submits a final project report by 12pm **Monday, April 16**. Here are important information

If you torture the data enough, it may confess
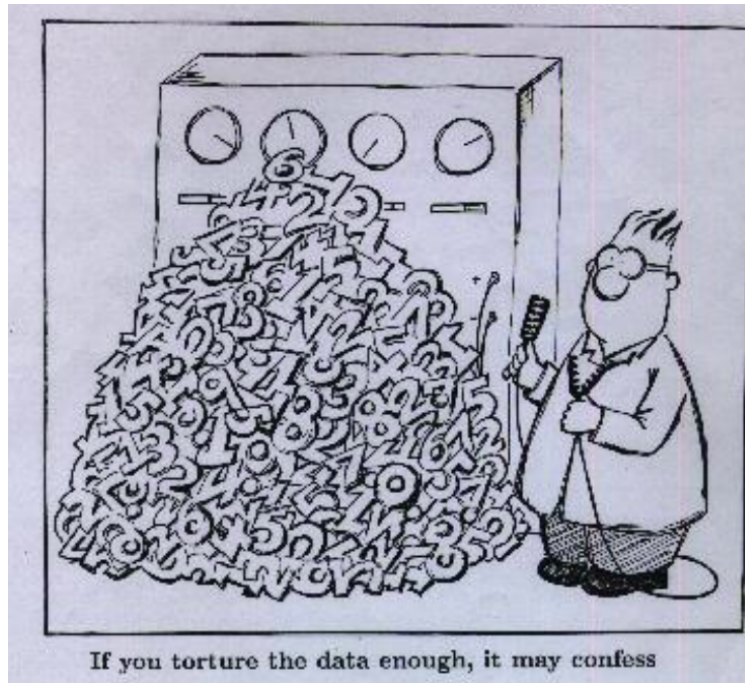
- Please limit your report to no more than 15 pages, including figures and tables, and references.
- The first page of the report contains only project title, project members. Moreover, please clearly state **the contribution made to the project by each member** in this page.
- All members will receive the same grade for the project.
- Please submit two versions of the final report, a full version and another version without the identification information (in the first page of the full version).
- Technical description of the project starts at page 2.
- You are not required to submit codes in your report – if you would like to, submit it in a separate file.

3. *Poster session.* We plan to organize a poster session in the week of April 16. Time and location TBA.

4. *Project evaluation.* Each project report will be evaluated by the instructor and the GSIs. In addition, it will be assigned in an anonymous and random fashion to three other students, who will rate the report (according to a guideline to be announced). The reviewers will also provide a brief comments a project's strength and weaknesses. You may notice that we are adopting a reviewing system similar to that of a top-tier conference in machine learning such as NIPS or ICML. This is also an opportunity for each to learn about other teams' work, and to see how data analysis is done well (or not so well). All review reports are due by 12pm **Friday, April 20**.