"Group Wine"

Bryce Rowland

Taylor Lagler

Paloma Hauser

Seoyoon Cho

Michael Nodzenski

"Quant Psych"

William Belzak

Alexis Georgeson

Cara Arizmendi

"maRkov Chain Gang"

Hillary Heiling

Ethan Alt

Nathan Bean

Brady Nifong

Sarah Reifeis

"Tired Puppies"

Nikki Freeman

Tamy Harumy Moraes Tsujimoto

Yue Pan

John Sperger

Jie Chen

"Umbrella Academy"

Yutong Liu

Mochuan Liu

Yu Gu

Hannan Yang

Beilin Jia

~~~

Instructions:

Sources for data:

* <https://archive.ics.uci.edu/ml/index.php>
* <https://ai.google/tools/datasets/>
* <https://www.kaggle.com/datasets>

Each group of 3-5 should

1. Identifying a question in their selected dataset.
2. Identify a model to evaluate the question.
3. Code up a statistical model from scratch using knowledge from Module 2.
4. Implement a procedure to fit the model in an R package that uses knowledge from Module 1 (see details below).
5. Summarize results in Rmarkdown document, submitted separately from the R package. Aim for Introduction to the dataset and the question, a Methods section with any important pre-processing etc., three sections of Results (which may also contain some method details), and Discussion about interpretation or possible future directions. This will be due 5 pm, Thursday, April 25th.
6. Present the results in one of two sessions in the last week of class April 22nd and 24th (aim for 20 minute presentation).

~~~

Notes:

R package and Rmd details:

R package and Rmd file should be runnable by instructors (unless data has privacy restrictions, in which case we can arrange exception). Analyses are contained in Rmd file itself, which will not reside in the R package but will be turned in separately. Rmd should start by loading your package. You can point out where to download data, you don’t need to contain in a data/ directory in package. The R package should have man pages for any functions that are exported. The R package should pass build/check as shown in class (ask if you have questions about this).

What to implement from scratch and what not to

Don’t need to implement RF and SVM from scratch, but each project should use a likelihood-based model coded from scratch to compare results/performance.

* Parametric analogue for SVM —> logistic regression or penalized logistic regression.
* Parametric analogue for RF —> linear model/glm or their penalized variants.
* Parametric analogue for K-means —> multivariate gaussian mixture model, mixture models in general.

Bayesian methods may be used but not required as the department’s Bayesian course is not a 735 prerequisite.