Cogs 109: Modeling and Data Analysis

Final project guidelines, Spring 2021

- Work in teams of at least 2 and no more than 3 students. Every student in the group will be expected to contribute substantially to the final product(s), and all students should be able to understand and explain all aspects of the project when you present your work in the final symposium.
- Your project should.
    1. Identify a **real problem**, challenge or scientific question which could benefit from data analysis and modeling. Your final report must explain why the question is interesting or important.
    2. Identify a **relevant data set**. You should learn about how the data was collected and be able to explain key features of the data, for example: How many observations are there (n)? What are the noise sources or other factors affecting the data quality? What are the relevant predictors?
    3. Identify at least one **relevant data analysis approach**, choosing from the methods covered in the course (linear or nonlinear regression, classification, clustering, PCA, etc.). Explain why this analysis approach is appropriate for addressing your question.
    4. Identify and explain **one or more hypotheses** or initial expectations that you will test using the data.
    5. **Model selection:** You should compare and contrast multiple different models (at least 2, but usually more). Your comparison should make use of cross-validation, bootstrap sampling, regularization, and/or other relevant techniques. For example, you might compare K-Nearest Neighbors classification for a range of k values (k=1,2,…,50), and select the k value that provides the lowest test set (cross-validation) error.
    6. **Model estimation**: Implement your data analysis and present the results using a combination of data visualizations (box plots, scatter plots), statistical analyses and models.
    7. Present your conclusions and outlook for next steps/future directions.
- The final product will be a written report, 5-10 pages in length, as well as a short ~3 minute presentation. We will provide more information about the final paper and presentation in a few weeks.

Schedule:

- HW5 due May 7 (Friday)
    1. On HW5 you will be asked to name your final project group members, title, and 2-4 sentence “abstract”
- We will provide comments on your proposals in Week 7.
- On HW7 you will submit a revised and updated abstract
- Final report will be due Thursday June 3
- Project presentations will be scheduled Friday June 4

Written report:

Your final report must include the following sections (use these headings).

a. **Introduction.**
   i. Define the real problem and explain its motivation
   ii. Identify the dataset you will use and explain its key characteristics.
   iii. Explain at least one hypothesis that you will test.
b. **Methods.** Identify the data analysis approach you will use and explain the rationale/motivation for your choice of this approach.
c. **Results**
   i. **Model selection.** You MUST compare at least 2 models, using cross-validation, regularization, and/or other relevant techniques.
   ii. **Model estimation.** What are the final parameter estimates? What is the final accuracy of the model's predictions?
d. **Conclusions and discussion.** What can you conclude about your hypothesis? (Note that negative or ambiguous results are perfectly acceptable, you just need to explain what you found.) What are some potential implications/next steps for researchers interested in this topic?

Suggested datasets/resources (these will be updated):

- https://github.com/awesomedata/awesome-public-datasets
- Neuroscience
  - https://crcns.org/data-sets
  - https://neurodata.io/ocp/
  - Allen Institute: https://alleninstitute.org/bigneuron/data/
  - EEG dataset: https://github.com/meagmohit/EEG-Datasets
  - EMG dataset: https://archive.ics.uci.edu/ml/datasets/EMG+data+for+gestures
- Government/economics/politics
- Sports
  - https://www.kaggle.com/c/march-machine-learning-mania-2015/data
- Climate
- Kitti: http://www.cvlibs.net/datasets/kitti/
- Coco imaging dataset: http://cocodataset.org/#home
- Health data: https://healthdata.gov/
- Smart City Dataset: https://smartcities.data.gov.in/