Final Project

Choose one of the options below for your final project. Undergraduates may work in groups of up to 3, while graduate students must work alone.

The audience for your report is someone who has taken this class, but doesn’t know any coding languages (must’ve been a boring class). In particular, your audience will not know how to interpret any code input or output, but they will be able to understand mathematical models, parameter estimation, diagnostic plots, details on training vs. testing, uncertainty quantification, etc. The project consists of a project proposal and the final project report.

Project proposal (due Friday March 22): The project proposal must include:

- A description of the dataset or scientific paper on which the project will be based (5 points)
- A carefully described scientific goal of the project (5 points)
- Anticipated methods that will be used in support of the project goal, including how project success/verification will be defined (5 points)
- A group of clearly identified students working together (5 points)

Project report (due Wednesday May 1): The grading rubric for the final report is:

- The report is typed (5 points)
- Adequately introduces the motivating dataset or paper under study (10 points)
- Clearly identifies goals for the project (i.e., the scientific problem of interest) (10 points)
- Adequate and correct use of plots and tables to describe and motivate modeling decisions and estimates (15 points)
- Proper use of statistical notation and comprehensive discussion of symbols as well as any relevant assumptions (20 points)
- Adequate model diagnoses, justification and verification (20 points)
- Relevant discussion that summarizes the report, and discusses unanswered questions and future directions students next year could take (5 points)
- Detail is sufficient that your professor could replicate your results without looking at your code (10 points)
- Code is contained in an appendix after the project description; no code (input or output) inline (5 points)
Option 1:

The file `SnowCoverData.RData` on the class website contains a set of remote sensing (i.e., satellite-based) data on snow cover percentage over a portion of the Sierra Nevada mountains. The data consist of over 4 million estimates of snow cover at a 30 m × 30 m resolution from the Landsat satellite constellation, held in the `landsat` variable. Snow cover is measured as a percentage in [0,100].

Available predictors include `day.of.year` (1 to 365), `elevation` (m), `slope` (degrees up to 90), `aspect` (degrees), `land.type` (factor with 16 types) and `modis` (snow cover in [0,100]). The latter of these are estimates that come from the MODIS satellite constellation, but are only available at a much coarser grid (≈ 500 m).

Generally the goals of this project are to understand the relationship between Landsat and all of the predictors. In particular, the rest of the predictors are available every day, but Landsat is only available about every two weeks, thus we need models to fill in the missing gaps.

Possible topics you may wish to examine for a project:

- Note that Landsat has many values of exactly 0 and exactly 100. Can you create a model that yields reasonable classifications for Landsat being 0, 100, or in between (3 classes)?

- Can you create a predictive model for values of Landsat strictly in (0,100), excluding the boundary cases of 0 and 100? A transformation would be useful here.

- Does the relationship between Landsat and the rest of the variables change depending on the day of the year?

- Is there evidence of nonlinear relationships?

- Is there evidence of interactions between any of the predictors?

Option 2:

The website [www.colorado.edu/fcq/fcq-results](http://www.colorado.edu/fcq/fcq-results) allows you to download FCQ historical data from CU’s Boulder, Colorado Springs and Denver campuses. In this project you might try to analyze:

- What are predictors for being a good instructor?

- How do ratings across campuses differ?

A possible approach to examining these might be to find a model that predicts whether a new faculty member will be a highly rated instructor (according to the `InstructorOverall` variable). Or, how do relationships between features and possible responses change across campuses? Is there anything that separates the Course Overall and Instructor Overall variables, or are they entirely colinear?
Option 3:
Read a research paper or analyze a dataset of your choice. Your project should include a summary of the manuscript as well as your attempt to implement the ideas of the paper, whether it be reproducing results of the paper or an application of the methodology to a new problem.

Possible topics you may include:

- How might a support vector machine be used in a probabilistic framework (Platt 1999)?
- Text classification using SVMs (Joachims 1998).
- Kernel principal components analysis (Scholkopf 1997).
- Elastic net regression (balance between ridge regression and lasso).
- Look into the optimization theory supporting SVMs.
- Introduce and compare a neural network to some of the models described in class.
- The website www.kaggle.com houses numerous possible datasets.
- Propose a project or data analysis that sounds interesting to you.