ER 190C: Statistical Learning for Energy and Environment

Duncan Callaway
dcal@berkeley.edu

Fall, 2018

Units: 4.0

Lecture Hours per week: 3.0
Lab Hours per week: 2.0

Course Description

This course will teach students to build, estimate and interpret models that describe phenomena in the broad area of energy and environmental decision-making. The effort will be divided between (i) learning a suite of data-driven modeling approaches, (ii) building the programming and computing tools to use those models and (iii) developing the expertise to formulate questions that are appropriate for available data and models. My goal is that students will leave the course as both critical *consumers* and responsible *producers* of data driven analysis.

We will work in Python in this course, and students must have taken Data 8 before enrolling. The course is designed to fit into Berkeley's emerging "data science" curriculum by providing students with a skill set similar to those developed in Data 100. However, in contrast to Data 100, here we will place a stronger emphasis on how to use prediction methods as decision-making tools in energy and environment contexts and less emphasis on web technologies, working with text, databases and statistical inference.

Materials

- You will need your own computer, but virtually any operating system will do (OSX, Windows, Linux, Chromebook).
- We will draw some material from Berkeley's Data 100 course book, freely available here: https://www.textbook.ds100.org
- Finally, we will draw material from the excellent text book, Introduction to Statistical Learning, available in both print and pdf form.

Prerequisites

Prerequisites:

- Foundations of Data Science (CS/ INFO/ STAT C8)
- Computing: An introductory programming course (CS61A or CS88).
- Math: Linear Algebra (Math 54, EE 16a, or Stat89a).

Course Structure

Class Structure

This is a four unit course, with three hours of lecture and two hours of lab section each week. Lectures will focus on theoretical and conceptual material but also introduce the programming structures required to use the material. Labs will be computer working sessions with a GSI and lab helpers available to work through weekly lab exercises.

Assessment

The course will have weekly homework assignments, a mid term and a final exam. Grading will be as follows:

- Homework: 20% (There will be ten, due most Fridays. We drop the lowest grade.)
- Lab assignments: 20% (There will be ten, due most Fridays. We drop the lowest grade.)
- Mid-term: 25% (October XX)
- Final Exam: 25% (December XX)
- Participation: 10% (Participation will be measured by answering questions in lecture with an online form.)

Schedule and weekly learning goals

The schedule is tentative and subject to change.

Week 01, 08/20 - 08/24: (First class on Thursday) Course overview. Prediction versus inference, examples of data-driven resource allocation problems. Ethics in data science.

Week 02, 08/27 - 08/31: The data science lifecycle. Review of basic pre-requisite concepts. Data manipulation, data frames, cleaning data.
Application: Patterns in global energy consumption

Week 03, 09/03 - 09/07: Exploratory data analysis and visualization.
Application: Global energy consumption patterns, continued.

Week 04, 09/10 - 09/14: Exploratory data analysis and visualization, continued.
Application: Global energy consumption patterns, continued.

Week 05, 09/17 - 09/21: What is a model? Uses and abuses of models. Overview of model estimation, loss functions and gradient descent. Revisit ethics in data science.
Application: Inferring behavior from electricity consumption data.

Week 06, 09/24 - 09/28: Supervised versus unsupervised learning; Clustering;
Application: Behavior and electricity consumption, continued.

Week 07, 10/01 - 10/05: Regression; Prediction versus inference revisited
Application: Behavior and electricity consumption, continued.

Week 08, 10/08 - 10/12: Regularization and the bias-variance tradeoff; Cross validation.
Application: Air quality and environmental justice. Data vs shared experiences and the spoken word.

Week 09, 10/15 - 10/19: Mid-term review. Mid-term on Thursday.

Week 10, 10/22 - 10/26: Model selection methods
Application: Air quality and environmental justice, continued.

Week 11, 10/29 - 11/02: Classification; regression trees.
Application: Air quality and environmental justice, continued.

Week 12, 11/05 - 11/09: Regression trees continued; Support vector machines
Application: Lead contamination in Los Angeles.

Week 13, 11/12 - 11/16: Support vector machines.
Application: Lead contamination in Los Angeles, continued.

Week 14, 11/19 - 11/23: Neural networks. (No Thursday lecture – Thanksgiving.)

Week 15, 11/26 - 11/30: TBD / slack.

Week 16, 12/03 - 12/07: (Reading week)

Week 17, 12/10 - 12/14: (Final Exam, date TBD)