Advanced Methods for Data Analysis: Spring 2014

Statistics 36-402/36-608

Instructor: Ryan Tibshirani, Dept. of Statistics, Baker Hall 229B, ryantibs@cmu.edu

Teaching assistants:
Jisu Kim, jisuk1@andrew.cmu.edu
Robert Lunde, rlunde@andrew.cmu.edu
Sonia Todorova sktodoro@andrew.cmu.edu

Lectures: Tuesdays and Thursdays 10:30-11:50am, Wean 7500

Overview and objectives

You've learned to use linear regression as your main tool in data analysis. So what's next?

This course aims to train you to use advanced statistical methods for analyzing data. We start with the linear model and build on the theory and applications that you've already seen in this setting in order to explore richer model classes, more kinds of data, and more complex setups. All the while, our intention is to develop an intuitive understanding of the methods and their limitations, a formal understanding of the same concepts, and the practical/programmatic skills to apply these methods in real problems.

Upon completing this course, you should be able to tackle new applied statistics problems, by: (1) selecting the appropriate techniques and justifying your choices; (2) implementing these techniques programmatically (using, say, the R programming language) and evaluating your results; (3) explaining your results to a researcher outside of statistics.

Outline of material

Here is a rough outline of the course topics.

- The statistical regression model
- The truth about linear regression
- Error and validation
- Kernel regression
- The bootstrap
- Degrees of freedom

- Smoothing splines
- Additive models
- Inference with linear smoothers
- Logistic regression
- Generalized linear models
- Generalized additive models
- Density estimation
- Testing goodness-of-fit
- Principal components analysis
- Nonlinear dimension reduction
- Clustering
- Regression shrinkage and selection
- High-dimensional statistics
- Time series

Logistics

Prerequisities: The formal prerequisite is 36-401: Modern Regression. Alternately, students can enter the course having taken an equivalent regression course, and with consent of the instructor (but this is exceptional).

In addition, we will assume that you are comfortable with basic probability, statistics, linear algebra, and R programming. Specifically, here is a list of topics that you should be more or less familiar with. (Note: in this course, we will cover R programming in depth for those who are feeling rusty/underexposed; if you don't know R, then you should at least be comfortable with programming in *some* language.)

- *Probability.* Event, random variable, indicator variable; probability mass function, probability density function, cumulative distribution function; joint and marginal distributions; conditional probability, Bayes's rule; independence; expectation, variance; binomial, Poisson, Gaussian distributions.
- *Statistics.* Sampling from a population; mean, variance, standard deviation, median, covariance, correlation, and their sample versions; histogram; likelihood, maximum likelihood estimation; point estimates, standard errors, confidence intervals, p-values; linear regression, response and predictor variables, coefficients, residuals.

- *Linear algebra.* Vectors and scalars; components of a vector, geometry of vectors; vector arithmetic: adding vectors, multiplying vectors by scalars, dot product of vectors; coordinate basis, change of basis; matrices, matrix arithmetic: matrix addition, matrix multiplication, matrix inversion, multiplication of matrices and vectors; eigenvalues and eigenvectors of a matrix.

- *R programming.* R arithmetic (scalar, vector, and matrix operations); writing functions; reading in data sets, using and manipulating data structures; installing, loading, and using packages; plotting.

Class website: The class website is `http://www.stat.cmu.edu/~ryantibs/advmethods/`. The class schedule, lecture notes, homeworks, etc., will be posted there.

Attendance: Attendance at lectures is highly encouraged. We think you will learn more by coming to lectures, paying attention, and asking questions. Plus, it will be more fun.

Office hours: The weekly schedule for office hours is given below.
RT: Wednesdays 3-4pm, Baker 229B
RL: Thursdays 3:30-4:30pm, Porter 117
ST: Mondays 3:30-4:30pm, Porter A20

Textbook: This class draws heavily on Professor Cosma Shalizi's previous offerings of the same course, and we will use his draft textbook as our class text. The title is *Advanced Data Analysis from an Elementary Point of View*, and relevant chapters will be posted on the course website.

An optional class textbook is *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani. This is also an excellent resource, and covers much of the same material that we will.

Evaluation: There will be 9 homework assignments, approximately one 1.5 weeks. The assignments will be a combination of written exercises and programming exercises. The assignments will be posted on the course website, and your homeworks will be submitted through Blackboard by midnight on the day it is due. Your lowest homework score will be dropped. This serves as your allowance for tardiness, and late homework will not be accepted (except only under highly exceptional circumstances, in which case you should email the instructor).

There will be two midterm exams, one take-home, and one in-class. There will be one take-home final exam. The grading breakdown is as follows:

Homeworks	60%
Midterm exam 1	10%
Midterm exam 2	10%
Final exam	20%

Plagiarism: You are encouraged to discuss homework assignments with each other. But you must submit your own original work, both written work and computer code. Explicitly sharing your written work or code with someone else is not allowed. See the student handbook's section on "Cheating and Plagiarism" (`http://www.cmu.edu/policies/documents/Cheating.html`). In the case that you do collaborate with other students, you must indicate on each homework your collaborators. And if you are unclear for any reason about the rules, come talk to the instructor—it is much, much better to be clear from the start what is considered collaboration and what is not.