Lecture 1: Introduction and Regression

Text References

Shalizi Chapter 1 (James et al. Chapter 2)

Overview

Statistics is the science of designing and analyzing methods for drawing *reliable inferences from imperfect data*.

One of the things that people most often want to know about the world is how different variables are related to each other, and one of the central tools statistics has for learning about relationships is *regression*. In your linear regression class, you learned about how it could be used in data analysis, and learned about its properties. In this course, we will build on that foundation, extending beyond basic linear regression in different directions, to answer many questions about how variables are related to each other.

But before we go beyond linear regression, we will first look at *prediction*, and how to predict one variable from nothing at all. Then we will look at predictive relationships between variables, and see how linear regression is just one member of a big family of smoothing methods.

Questions

1. What do we usually mean when we say that we want to predict Y from X? What are we estimating?
2. Why go beyond simple linear regression? What's the trade-off?

Definitions and Notations

1. Prediction and inference
2. Supervised versus unsupervised learning
3. The regression function
4. Linear smoothers
5. Training and test errors

Topics

1. Course overview and logistics.
2. The least-squares optimal prediction is the expectation value; the conditional expectation function is the *regression function*.
3. Ordinary least-squares revisited as a smoothing method.
4. Other linear smoothers: nearest-neighbor averaging, kernel-weighted averaging.
5. How to quantify "overfitting" versus "underfitting".

What's Next?

Lecture 2: The truth about linear regression (Shalizi Chapter 2)
Lecture 3: Error and Validation (Shalizi Chapters 2-3 + supplementary material)

Course Overview

Welcome

Welcome to 36-402 Advanced Methods for Data Analysis.

Who should take this course?

This course is intended for those who already know basic linear regression and who would like to learn about more advanced methods for analyzing data. See syllabus

In 36-401, you learned about how linear regression can be used for prediction and for inference (i.e. for learning about relationships between different variables; e.g. to understand how the response Y changes as a function of the predictors $X_1, \ldots, X_p$). In this course, we will build on that foundation, extending beyond basic linear regression to explore

- richer model classes
- more complex setups and data sets
- basic statistical learning theory
- supervised learning (e.g. regression; building a statistical model for predicting or estimating an output based on one or more inputs) *and* unsupervised learning (e.g. density estimation, dimension reduction, clustering; there are inputs but no supervising output; nevertheless we can learn relationships and structures from such data).

Course Mechanics

- **Lectures** Tuesdays and Thursdays. Print out and bring lecture outline; fill them out during class. Since this is a data analysis course, I will sometimes work through examples using R. If the R code is posted on Canvas beforehand, feel free to bring your laptop and actively follow along.
- **Required text** Shalizi, *Advanced Data Analysis from an Elementary Point of View* (September 2018)
- **Other references** Tector, *R Cookbook*; James, Witten, Hastie & Tibshirani, *An Introduction to Statistical Learning*; Wasserman, *All of Statistics*
- **Topics** (see syllabus and course schedule)
- **HW**. Weekly assignments due on **Fridays by 3:00 pm**; submit through Canvas (First HW due Jan 25.)
- **Computing**. Recommended to use R. There is RStudio, Unix command lines (the executable program is R), or the R app from *The Comprehensive R Archive Network*.
- **Office hours** We will schedule office hours starting next week.
- **Exams**. Two midterms (see syllabus for dates; one take-home, one in-class) and one take-home cumulative end-of-term assignment. Exam problems will be based on HW assignments, lecture examples and related concepts.