Version 2.0
12/04/2017 Subject to Change.

Introduction to Data Science - 6.s077 -
Spring 2018 - 12 units (4-0-8)

Tuesday, Thursday 1-2.30pm (32-155)

**** This course has a maximum enrollment of 100 ****

A new course, "Introduction to Data Science" will be offered in the Spring of 2018. This message is to provide some advance information, so that students can plan ahead their courses.

Style and objectives

These days, data are plentiful, as is software for its processing. The challenge however is to be able to discern what method can be applied where, and to understand the underlying assumptions and the possible pitfalls.

The objective of this class is to introduce basic **statistical concepts** through a rich collection of applications and hands-on experience with real data. It is **not** a theoretical class; it is **not** a collection of recipes; and it is **not** about databases or programming.

Positioning

6.s077 counts as an EECS Foundation class. However, students cannot get EECS credit for both 6.008 and 6.s077, for either the Foundation or other departmental EECS requirements. These two classes have different emphasis and are offered in alternate semesters.

While not a prerequisite of 6.036, 6.s077 will provide useful foundations for topics covered in that class.

Catalog description

Introduction to the methodological foundations of data science, emphasizing basic concepts, but also modern methodologies. Learning of distributions and their parameters. Testing of multiple hypotheses. Linear and nonlinear regression and prediction. Classification. Learning of dynamical models. Uncertainty quantification. Model validation. Causal inference. Applications and case studies drawn from electrical engineering, computer science, the life sciences, finance, and social networks.

Version 2.0
12/04/2017 Subject to Change.

Instructors

Profs. Polyanskiy, Shah, and Tsitsiklis

Prerequisites

6.s077 assumes familiarity with basic probability, at the level covered in 6.041A (the first half of 6.041) or equivalent. Programming comfort at the level of 6.0001 is also expected.

The mechanics

There will be two weekly lectures, of 1.5 hours each. There will also be a weekly (1 hour) recitation to be delivered by TAs. Enrollment for the first offering is to be capped at 100 and we anticipate the need for 4 TAs.

There will be a weekly homework involving theoretical problems, and computational exercises. There will also be about 4 major, and somewhat more open-ended, computational mini-projects. There will be an evening, 2-hour midterm exam, and a final exam during finals week.

Homework (including computational projects):30%
Midterm: 30% (04/10/2018, 7.30-9.30pm in 32-123).
Final exam: 40%

Textbook

There will be assigned readings from the following required text, available online through the MT libraries, although we will not be always following the text closely. When deviating from the text, course notes will be provided.

An Introduction to Statistical Learning: with Applications in R,
by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 2013 (with corrected 2017 edition).

The text assumes background in probability, for which you may consult:

Introduction to Probability, by Dimitri Bertsekas and John Tsitsiklis, Athena Scientific, 2008
or the archived version of 6.041x.

Philosophy

The course is intended to provide a solid conceptual foundation to the field of data-driven modeling, prediction, and inference.

The main objective of the course is to introduce students to mature and informed approaches, including a discussion of the main and common statistical pitfalls that a modern-day data scientist should be aware of. Instead of covering a long list of popular statistical methods and algorithms, the course is to highlight the key conceptual ingredients of sound methods, developed in the context of select applications (drawn from EE, CS, social science, and the life sciences) that will serve for both motivation and practice.

More concretely, the course will assume a grounding in basic probability, including an exposure to conditional distributions and expectations, and variants of the Bayes' rule, at the level of 6.041A. It will cover key concepts of Bayesian and frequentist estimation and hypothesis testing, regression (from simple linear regression all the way to nonlinear regression and neural networks), important issues related to uncertainty quantification and validation (especially in the context of problems involving a large number of parameters and/or hypotheses), and an understanding of causality. The course will conclude wth a sample of high-end applications involving complex modern-day problems and approaches, e.g., for collaborative filtering.

Some of the conceptual issues and potential pitfalls to be emphasized will be related to:
a) misinterpretations of confidence intervals and p-values
b) issues that arise when testing large numbers of hypotheses (false discoveries)
c) the distinction between association, prediction, and causal inference

Tentative syllabus (25 lectures)

The lecture titles refer to methodologies. However, the methodologies will be introduced and discussed in the context of motivating applications.

Introduction (1)
Overview of data science and methodological issues.

Learning of distributions and their parameters (5)
1. Histograms, empirical distributions. Plug in estimators. Maximum likelihood.
2. Properties and assessment of parameter estimators: standard errors, confidence intervals, bootstrap.
3. Estimation of multiple parameters. Prediction and empirical risk minimization.
4. Bayesian methods.
5. Linear normal models.

Hypothesis testing (5)
1. Basics: t-test and z-test. p value.
2. Generalized likelihood ratio test. Testing for uniformity and independence.
3. Two-sample tests: parametric and Kolmogorov-Smirnov.
4. Multiple comparison problem. False discovery rate. Benjamini-Hochberg procedure.
5. Optimality properties of tests. Neyman-Pearson Lemma. ROC curve.

Regression and prediction (5)

1. Formulation of the regression problem. Simple linear regression. Test on regression coefficient and interpretation.
2. Multiple linear regression. Test for coefficients.
3. Cross-validation, Boot-strap for model evaluation and uncertainty quantification.
4-5. Overfitting, Model selection, Regularization and Optimization.

Classification (5)

1. Formulation of classification problem. Logistic regression.
2. Linear and Quadratic Discriminant Analysis.
3. K-nearest neighbors. Comparative study. Collaborative Filtering.
4-5. Perceptron, Feed-forward neural network, backpropagation and stochastic gradient descent.

Dynamical models (3)

1. Linear models (autoregressive and state space models)
2. Nonlinear models: maximum likelihood and empirical risk minimization
3. Difficulties with nonlinear models (e.g., unobserved variables, causality)

Conclusion(1)

Wrapping up lecture