Statistics and Data Science 365 / 565

Data Mining and Machine Learning

February 1

Yale

Outline

- Overview of Course
- Syllabus and Logistics
- Some Terminology and Concepts

What is Machine Learning?

What is Machine Learning?

Many people have many different views.

What is Machine Learning?

Many people have many different views.

Machine Learning is Statistics with a focus on computation, scalability, prediction, representation, and complex problems

What is Machine Learning?

Many people have many different views.

Machine Learning is Statistics with a focus on computation, scalability, prediction, representation, and complex problems

- Speech recognition
- Machine translation
- Object recognition and scene classification
- Autonomous driving

Subproblems of these and other complex problems are concrete, statistical estimation and inference problems that can be studied in isolation.

What is Machine Learning?

Many people have many different views.

Machine Learning is Statistics with a focus on computation, scalability, prediction, representation, and complex problems

- Speech recognition
- Machine translation
- Object recognition and scene classification
- Autonomous driving

Subproblems of these and other complex problems are concrete, statistical estimation and inference problems that can be studied in isolation.

Recognized as increasingly important to many areas of science, commerce, medicine, business

Machine Learning and AI

AI used to be a core component of the Computer Science curriculum

Machine Learning and AI

AI used to be a core component of the Computer Science curriculum

- Logic
- Search
- Games

Courses as many universities kept swapping out classical material for more machine learning and statistics. Now much more “core” ML material than can be fit in a one semester or quarter course.

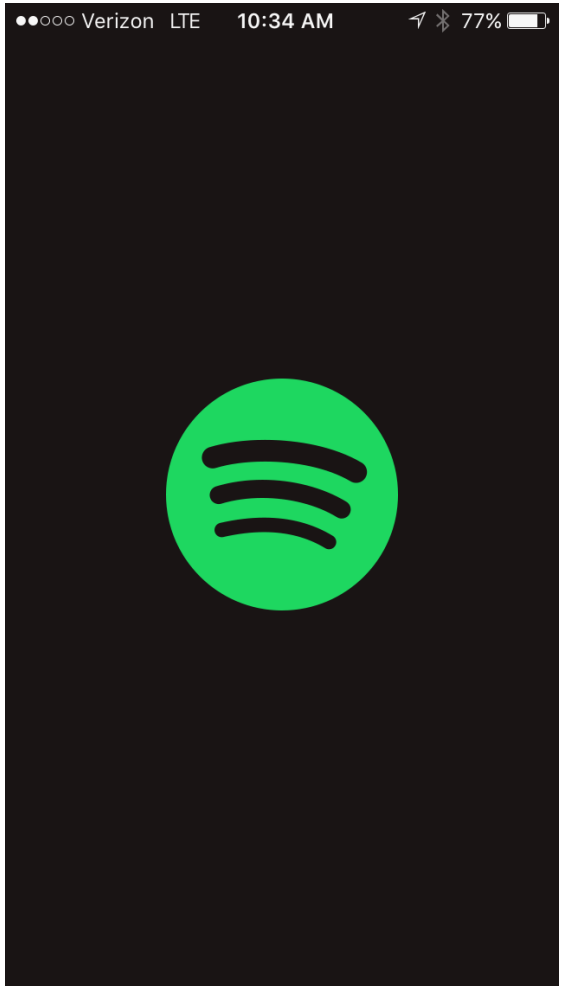
Confluence of CS and Stat activities $\Rightarrow$ Data Science

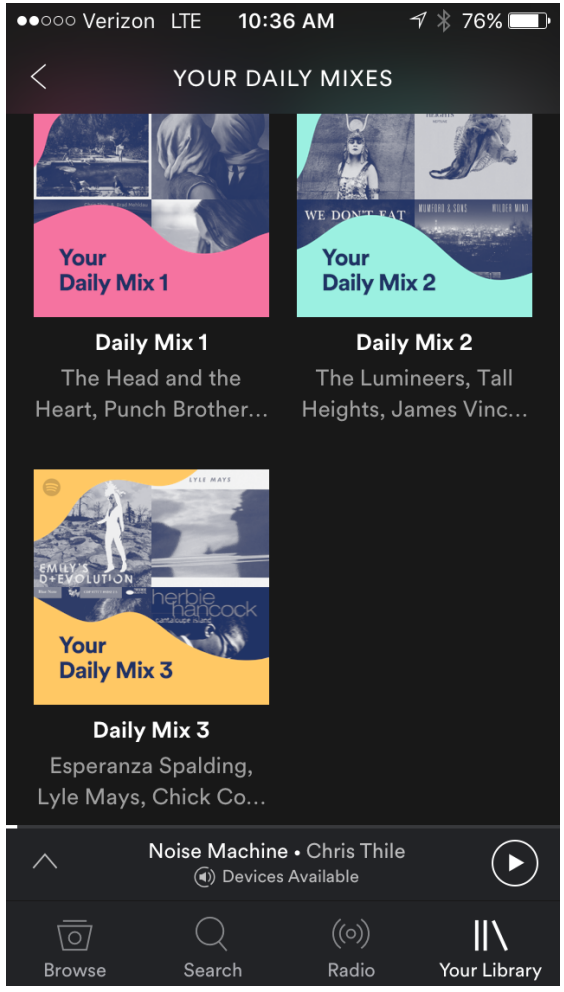
Statistical Machine Learning

This course emphasizes the statistical view of machine learning

Methods and theory grounded in probability and statistical principles

Examples of ML in daily life?

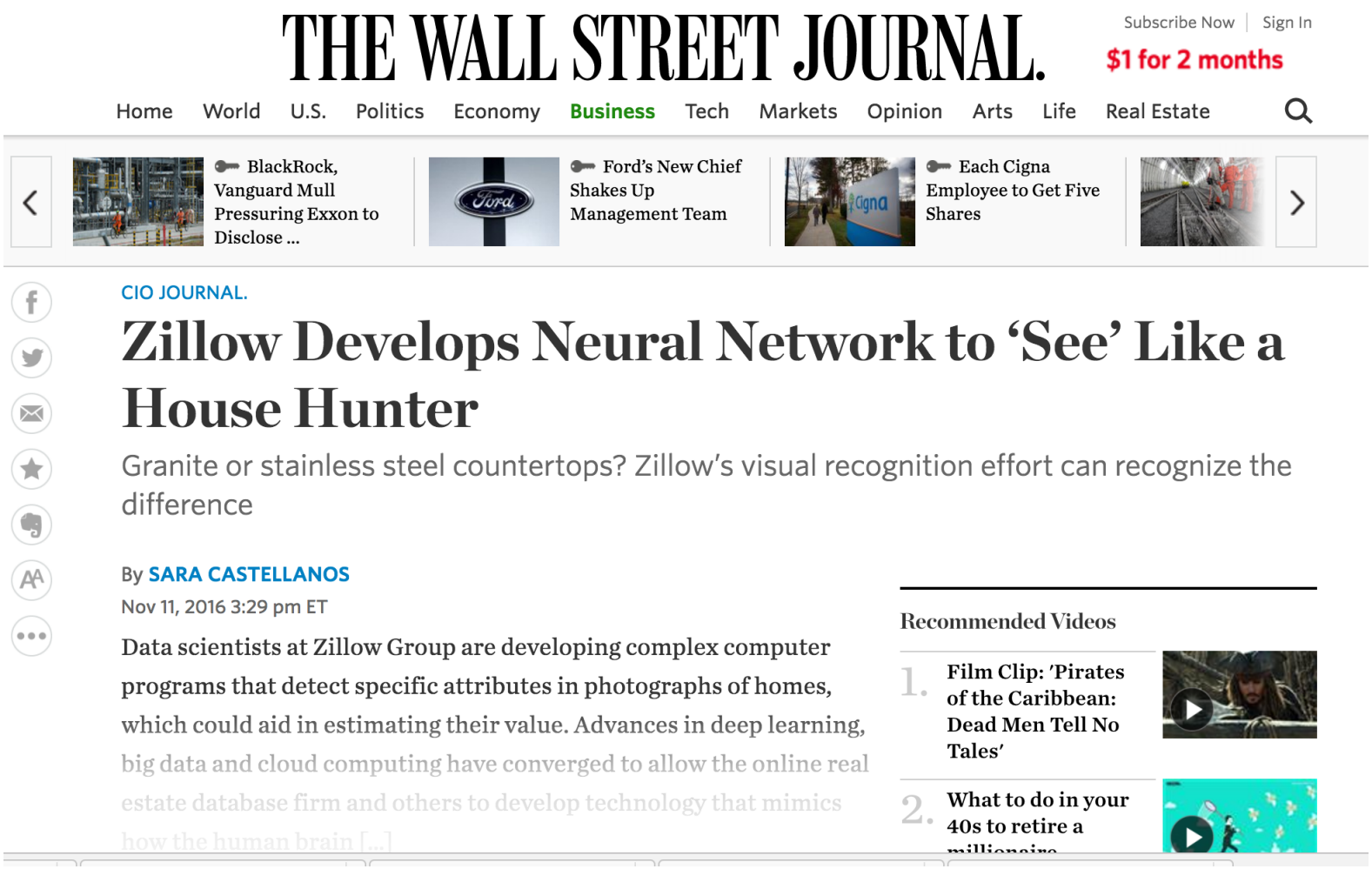




Podcast episode: `https://overcast.fm/+OcVctJBEA`

Examples of ML in daily life?

Predicting home values

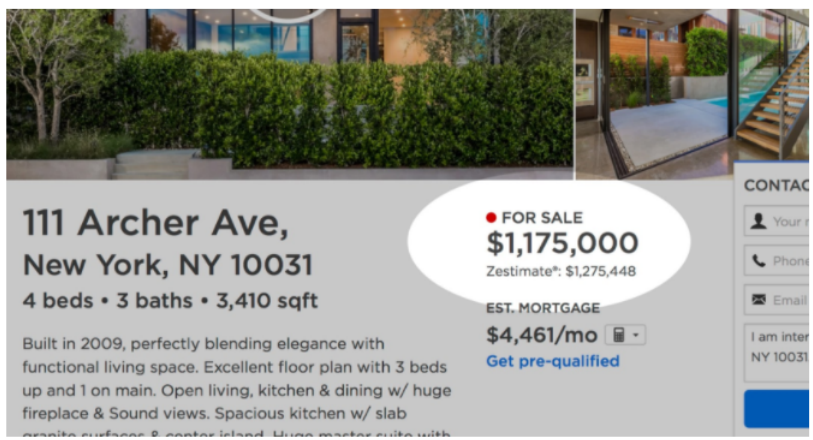


$1M Question

https://www.kaggle.com/c/zillow-prize-1

I'm excited to share the launch of Zillow Prize: Home Value Prediction (Zestimate) Competition. In this million-dollar competition, participants will develop an algorithm that makes predictions about the future sale prices of homes.

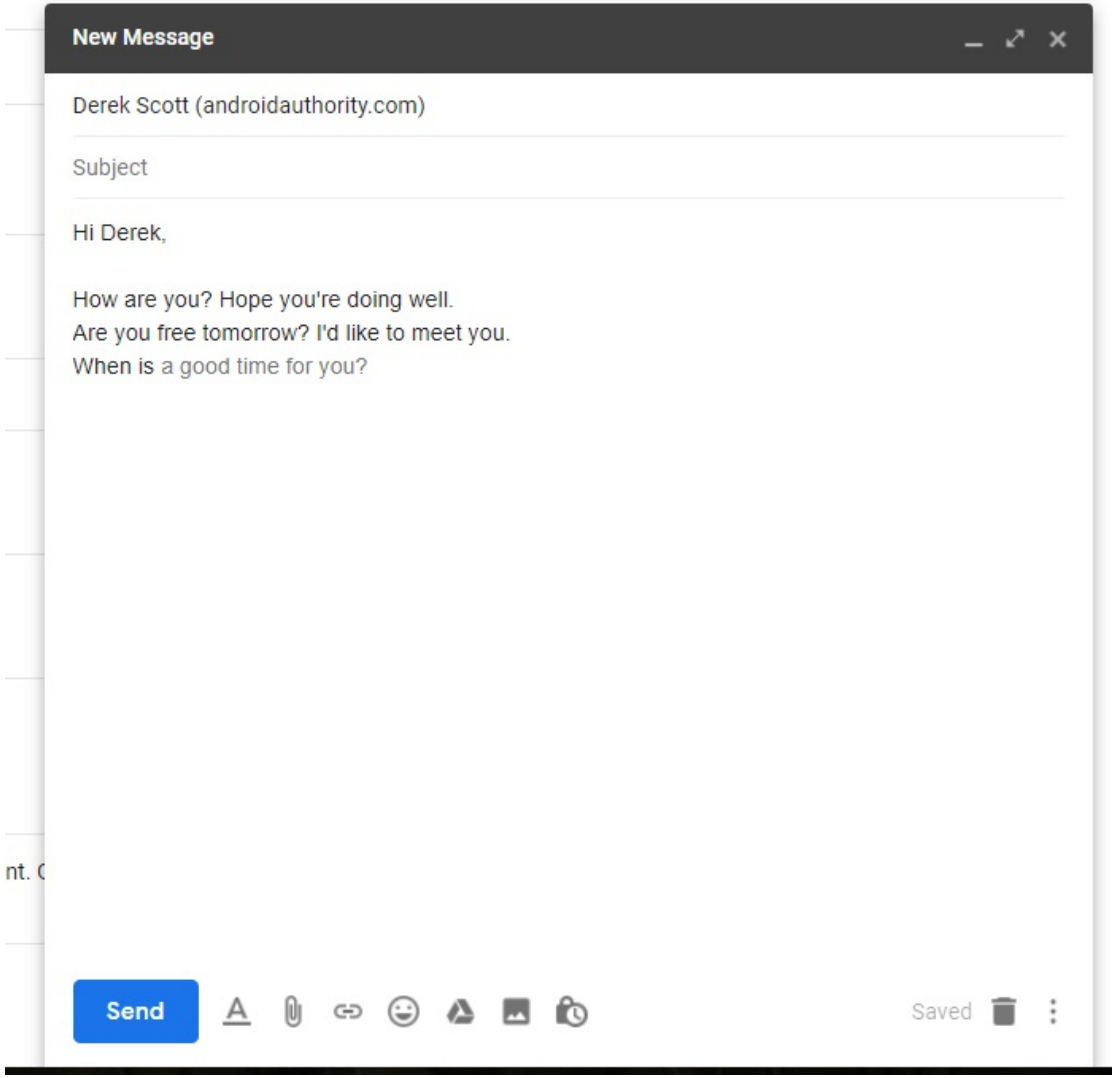
Zillow's Zestimate home valuation shook up the U.S. real estate industry when it was first released 11 years ago. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of information at no cost.



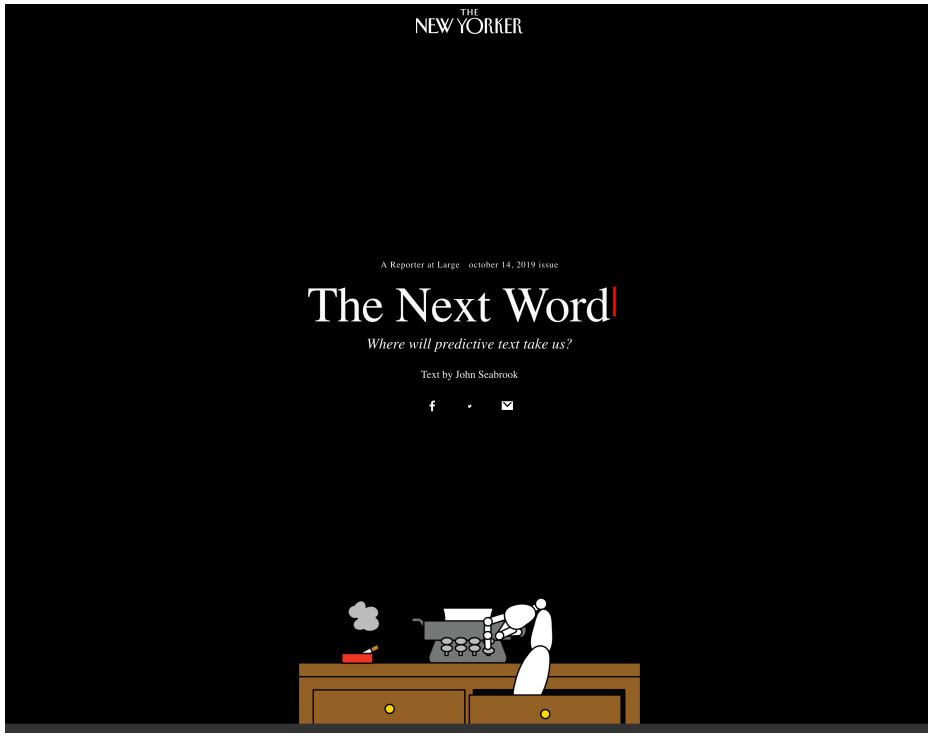
This million dollar contest is structured into two rounds. In the qualifying round, opening today, you'll be building a model to improve the Zestimate residual error. The top 100 ranking teams in this round will advance to the final round. In the final round, competitors will be challenged with building a home valuation algorithm from the ground up, using external data sources to help engineer new features that give your model an edge over the competition. The first place prize in the final round is $1,000,000 USD.

Join the competition

Recent Deployment of ML



Language models



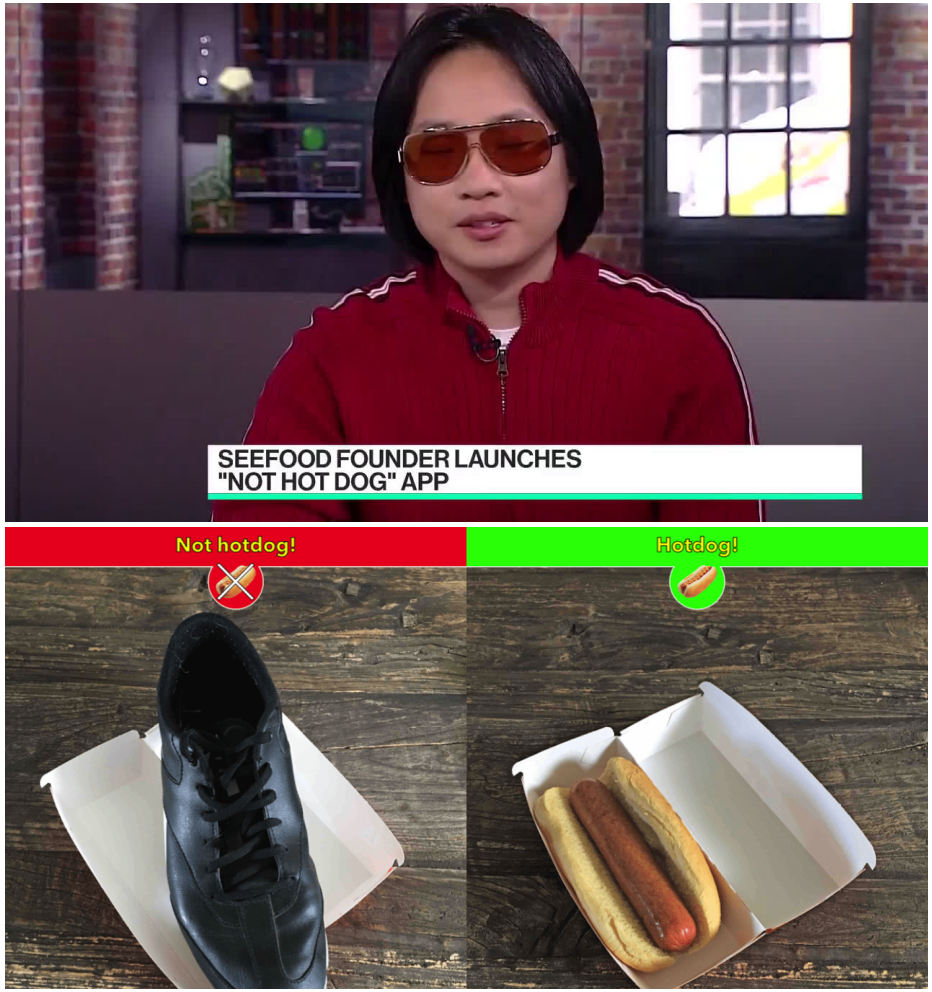
GPT-3 (July 2020) 175 billion parameters $\approx$ 700GB

GPT-3 generated blog posts (https://www.theverge.com/2020/8/16/21371049/gpt3-hacker-news-ai-blog)

---

https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker

Hot Dogs



It’s not all Gravy

TECH \ AMAZON \ ARTIFICIAL INTELLIGENCE

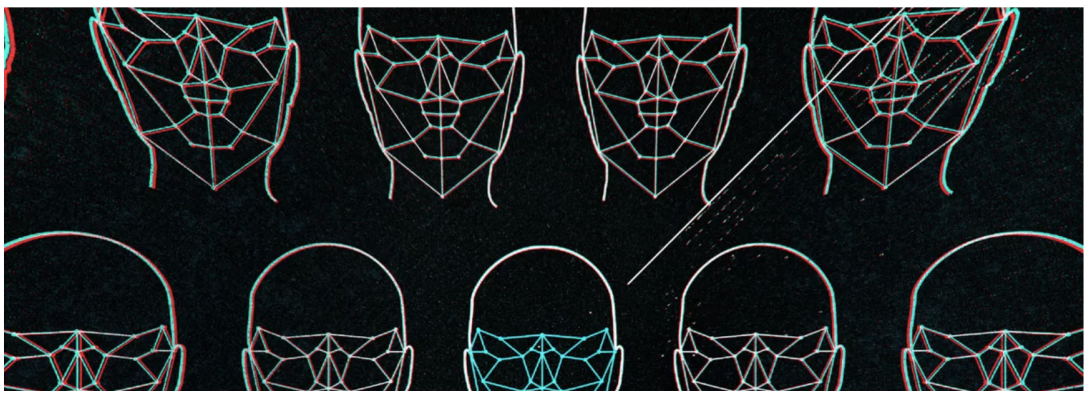
Gender and racial bias found in Amazon’s facial recognition technology (again)

17

Research shows that Amazon’s tech has a harder time identifying gender in darker-skinned and female faces

By James Vincent | Jan 25, 2019, 9:45am EST

SHARE



Syllabus

~~*Data Mining*~~ *and Machine Learning* is an introduction to some of the key ideas and techniques in statistical machine learning. Basic methodology and relevant concepts are presented in lectures, including the intuition behind the methods and a more formal understanding of how and why they work. Assignments give students hands-on experience with the methods on different types of data.

Topics include linear and nonlinear regression and classification, tree-based methods, clustering, topic models, word embeddings, and ~~recurrent~~ neural networks. Examples come from a variety of sources, including Twitter feeds, political speeches, archives of scientific articles, telescope imagery, and real estate listings. Programming is central to the course, and is based on the `python` programming language.

Team

- Sahand Negahban (Prof)
- Curtis McDonald
- Luke Benson
- Daniel Kim
- Ariadne Letrou

Office hours: Posted to canvas. Will have 1-1 office hours and group office hours. I will use calendly to schedule office hours. The link is on canvas.

Ed Discussions

Materials posted to Canvas. Discussion and copies of some materials on Ed Discussions

Please see canvas for access.

Use for any questions about lectures, homework, etc. rather than email! DO NOT POST YOUR SOLUTIONS AND ASK PEOPLE IF YOU ARE RIGHT.

Prerequisites

Statistics and Data Science 242; calculus and linear algebra; some computing experience (e.g., R, Matlab, Python, C++)

Prerequisites

Statistics and Data Science 242; calculus and linear algebra; some computing experience (e.g., R, Matlab, Python, C++)

Unofficially, some degree of:

- Exposure to basic statistical ideas and methods
- Exposure to basic linear algebra and multivariable calculus
- Exposure to basic algorithmic ideas and methods
- Programming experience (loops, writing functions, logic)

Course objectives

Gain (a deeper) understanding of and experience with basic statistical machine learning methodology

This is not a course on theoretical aspects (though some people feel it is)— we will not be proving theorems. (If anybody is curious about proving theorems post on Piazza and I'll give some resources)

Emphasis is on building mathematical intuition for how ML methods work by applying them to data.

But we will also discuss some of the formal properties that help explain why they work — and when they don't.

Evaluation

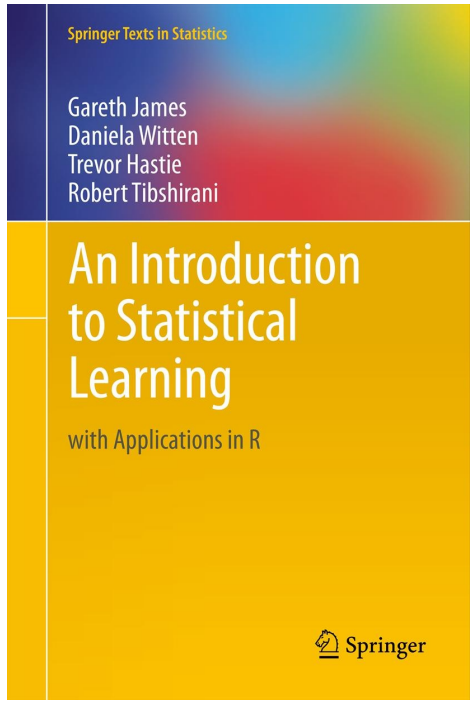
- Assignments (55%)
- Mid-semester exam (15%)
- Second exam (15%)
- Five short quizzes (10%)
- Participation (mainly Piazza) (5%)

Assignments can be turned in up to two days late. First day is a 5% penalty and next is a 10% penalty.

You can drop the lowest homework grade and quiz grade.

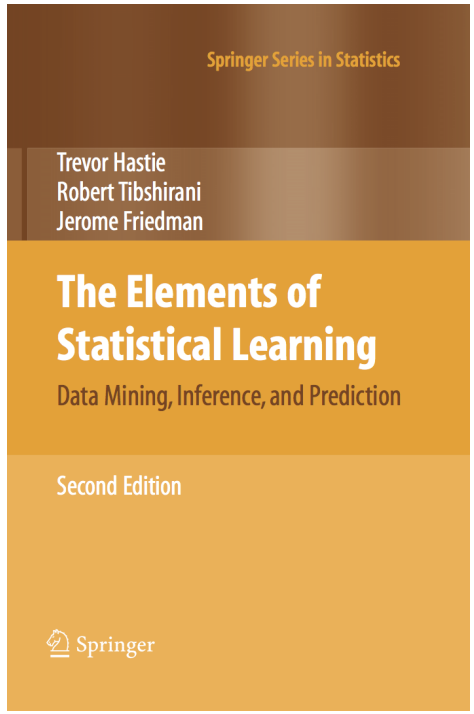
Book

- "An Introduction to Statistical Learning," by G. James, D. Witten, T. Hastie, and R. Tibshirani, Springer (2013), `http://www-bcf.usc.edu/~gareth/ISL`



More advanced version

- “Elements of Statistical Learning: Data Mining, Inference and Prediction,” by J. Friedman, T. Hastie, and R. Tibshirani



Assignments

- Roughly every 1.5 weeks
- Due at 11:59pm on the day
- Can turn in 2 days late (see above)
- Submitted using GradeScope
- Mix of problem solving and data analysis
- `python` used for computation

Collaboration

Collaboration on homework assignments with fellow students is encouraged. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had any discussions concerning the problem. You may *not* share written work or code—after discussing a problem with others, the solution must be written by yourself.