Lecture 01
Course overview

20 January 2016

Taylor B. Arnold
Yale Statistics
STAT 365/665

Yale

STAT 365/665: Data Mining and Machine Learning

Techniques for data mining and machine learning are covered from both a statistical and a computational perspective, including support vector machines, bagging, boosting, neural networks, and other nonlinear and nonparametric regression methods. The course gives the basic ideas and intuition behind these methods, a more formal understanding of how and why they work, and opportunities to experiment with machine-learning algorithms and apply them to data.

I would not say that description is inaccurate, but it is incredibly vague. I will concentrate in particular on:

- computation aspects; both theory and practical considerations
- machine learning as applied data analysis

This will be quite different than the way the course was taught in recent years.

Classes will typically consists of about 30 minutes of lecture followed by interactive coding, simulations, and data analysis.

Suggested Prerequisites:

- Introductory statistical theory
- Exposure to applied data analysis
- Familiar with R; proficient with R or Python

A rough outline of the course:

- 3 weeks - linear smoothers and support vector machines
- 3 weeks - introduction to neural networks
- 4 weeks - applications to computer vision
- 3 weeks - applications to natural language processing

References:

- Ian Goodfellow, Aaron Courville and Yoshua Bengio. *Deep Learning*. Book in preparation for MIT Press. `http://www.deeplearningbook.org/`.
- Jerome Friedman, Trevor Hastie and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, Berlin: Springer Series in Statistics, 2011.
- Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Book in preparation. `http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/`.

Some datasets that we will look at include:

- Taxi Data: `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`
- Million Song Dataset: `http://labrosa.ee.columbia.edu/millionsong/`
- MNIST: `http://yann.lecun.com/exdb/mnist/`
- CIFAR-10/CIFAR-100: `https://www.cs.toronto.edu/~kriz/cifar.html`
- The Street View House Numbers (SVHN) Dataset: `http://ufldl.stanford.edu/housenumbers/`
- ILSVRC: `http://image-net.org/challenges/LSVRC/2015/`
- Microsoft Common Images in Context (MS COCO): `http://mscoco.org/dataset/`
- Wikipedia

Course website:

A copy of the whole course syllabus, including a more detailed description of topics I plan to cover are on the course website.

`http://www.stat.yale.edu/~tba3/stat665/`

This is where all lecture notes, homeworks, and other references will appear.

TA’s:

Yu Lu, Jason Klusowski

Office hour times, formats, locations to be determined.

STAT 365 vs. STAT 665

Same requirements and assignments; however grading schemes may be different, even week to week.

Make sure you sign up for the correct course number!

Problem Sets:

There will be approximately 9 problem sets assigned throughout the semester, due on Thursdays. These will consist of both building custom implementations of machine learning algorithms, as well as applying established libraries to machine learning problems. You should expect to become comfortable working simultaneously in a number of programming languages. All submissions will be made electronically on the ClassesV2 site.

Grading:

Course grades will be determined based on scores from the problem sets. I want to make the grading extremely transparent, so these will all be graded on an 10 point scale (with the possibility of up to one additional point for truly exceptional work or extra credit questions). The final grade will be calculated by dropping the lowest grade, rounding the average of remainder to the nearest integer and reading off of the following table:

Grading scale

Numeric Score	Final Grade	
10	A	H
9	A-	H
8	B+	HP
7	B	HP
6	B-	HP
5	C+	P
4	C	P
3	C-	P
2	D	F
1	F	F
0	F	F

About me

Joint appointment at Yale Statistics and AT&T Labs Research

- Research focus on large-scale data analysis (think, many petabytes)
- One focus is on encoding sparsity through penalized estimation
- Applications to humanities and social sciences through with analysis of image, text, and video corpora

About you

1. Name
2. Undergraduate/Graduatate; Major/Department/School; Year in Program
3. Prior stats and computer science courses taken at Yale
4. Do you work with R, Python, or both?
5. Why are you looking to get out of the course?
6. Any applications you are particularly excited about?