Statistical Machine Learning (W4400)

Fall 2014
https://courseworks.columbia.edu

John P. Cunningham
jpc2181@columbia.edu

Benjamin Reddy
bmr2136@columbia.edu

Course Syllabus

Description

The use of statistical models in computer algorithms allows computers to make decisions and predictions, and to perform tasks that traditionally require human cognitive abilities. Machine learning is the interdisciplinary field at the intersection of statistics and computer science which develops such statistical models and interweaves them with computer algorithms. It underpins many modern technologies, such as speech recognition, Internet search, bioinformatics and computer vision—Amazon's recommender system, Google's driverless car and the most recent imaging systems for cancer diagnosis are all based on Machine Learning technology.

The course will provide an introduction to Machine Learning and its core models and algorithms. The aim of the course is to provide students of statistics with detailed knowledge of how Machine Learning methods work and how statistical models can be brought to bear in computer systems—not only to analyze large data sets, but to let computers perform tasks that traditional methods of computer science are unable to address. Examples range from speech recognition and text analysis through bioinformatics and medical diagnosis. This course provides a first introduction to the statistical methods and mathematical concepts which make such technologies possible.

Administrative

Lecture

- Tuesday and Thursday, 4:10PM–5:25PM
  Location: 207 Mathematics

Instructor

- John Cunningham
  Office: Department of Statistics, Room 1026, 10th Floor School of Social Work, 1255 Amsterdam
  Email: jpc2181@columbia.edu

Teaching Assistant

- Benjamin Reddy
  Office Hours: Wednesdays, 3-5pm
  Office: Department of Statistics, Room 901, 9th Floor School of Social Work, 1255 Amsterdam
  Email: bmr2136@columbia.edu

Virtual Office Hours

- Owing to student schedules and the subsequent challenges of finding mutually suitable office hours, we will also use a virtual platform. Piazza is a highly regarded forum for students to discuss class questions, homework problems, and more. Discussing problems is encouraged, but full solutions should not be posted (see section on academic integrity). The tool can be found at: `https://piazza.com/class/hz034zxilyf11c`.

Many Columbia classes find this to be a much quicker and more effective tool than traditional office hours, and we encourage students to use it both to ask questions and to improve their own understanding by posting answers and comments.

Prerequisites

- A previous course in statistics, elementary probability, multivariate calculus, linear algebra and ability to do moderate coding in R.

Grading and Academic Integrity

I take the honor code very seriously; students caught cheating or otherwise in violation will face disciplinary action. Please note the Barnard honor code text:

> "We... resolve to uphold the honor of the College by refraining from every form of dishonesty in our academic life. We consider it dishonest to ask for, give, or receive help in examinations or quizzes, to use any papers or books not authorized by the instructor in examinations, or to present oral work or written work which is not entirely our own, unless otherwise approved by the instructor.... We pledge to do all that is in our power to create a spirit of honesty and honor for its own sake."
> `http://barnard.edu/node/2875`
> `https://www.college.columbia.edu/academics/academicintegrity`

Your grade will be determined by three different components:

- **Homework (30%).** Homework will contain both written and `R` data analysis elements. This is due online by the beginning of class on the due date.
- **Midterm Exam (30%).** This will be given in class during midterm week. You will be permitted use one handwritten page, front and back, of notes.
- **Final Exam (40%).** This will be given in class during the finals period. You will be permitted use one handwritten page, front and back, of notes.

Failure to complete any of these components may result in a D or F.

Late Work and Regrading Policy: No late work or requests for regrades are accepted.

Homework: Students are encouraged to work together, but homework write-ups must be done individually and must be entirely the author's own work. Homework is due at the **beginning** of the class for which it is due. **Late homework will not be accepted under any circumstances.** To receive full credit, students must thoroughly explain how they arrived at their solutions and include the following information on their homeworks: name, UNI, homework number (e.g., HW03), class (STAT W4400), and section number. All homework must be turned in online through Courseworks in two parts: 1) The written part of submitted homework must be in PDF format, have a .pdf extension (lowercase!), and be less than 4MB; and 2) the code portion of submitted homework must be in R and have a .R extension (uppercase!). Homeworks not adhering to these requirements will receive no credit.

Reading Material

No explicit readings will be assigned. Rather, students should use the following two books as supporting references. The latter is (mostly) a shortened version of the former, and students may prefer the exposition in either text.

- Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition.* Springer, 2009. Available online at `http://statweb.stanford.edu/∼tibs/ElemStatLearn/`

- James, G., Witten, D. Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning* Springer, 2014. Available online at `http://www-bcf.usc.edu/∼gareth/ISL/`

Other useful books:

- Adler, J. *R in a Nutshell: A Desktop Quick Reference.* O'Reilly Media, 2010.

- Bishop, C. *Pattern Recognition and Machine Learning.* Springer-Verlag, 2006.

- Witten, I. H., Frank, E. and Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufman, 2011.

Approximate Lecture Outline

Lecture	Content
1	Introduction and basic concepts: maximum likelihood
2	Classification: loss functions, risk, naive bayes
3	Classification: linear classifier, perceptron
4	Classification: maximum margin, support vector machines
5	Classification: support vector machines continued
6	Optimization 1
7	Optimization 2
8	Classification: multiclass, kernels
9	Classification: SVM with kernels
10	Cross Validation
11	Classification: trees
12	Classification: boosting
13	Classification: face detection
14	Classification: bagging, random forests
15	Midterm Exam
16	Regression: linear regression, linear algebra
17	Linear Algebra: eigenvalues, normal distributions
18	Shrinkage: ridge regression
19	Shrinkage: LASSO
20	Bias-Variance tradeoff
21	Unsupervised Learning: PCA
22	Unsupervised Learning: clustering
23	Expectation-Maximization
24	Exponential Family
25	Information Theory
26	Model Order Selection
27	Markov Models
28	Bayesian Models
29	Final Exam
	(Other possible subjects: topic models, sampling, ...)