S&DS 365/565 - Applied Machine Learning and Data Mining

Syllabus (DRAFT)

Spring 2018

Instructor:
Prof. Xiaofei (Susan) Wang (xiaofei.wang@yale.edu)
Office: 24 Hillhouse Rm. 206
Office Hours: TBD (or by appointment)

Teaching Fellows:
Natalie Doss (natalie.doss@yale.edu)
Xin Xu (xin.xu@yale.edu)
Office Hours: TBD

Prerequisites:

1. Coursework in probability and statistics (e.g. STAT 241 + 242 or equivalent)
2. Some linear algebra and multivariable calculus
3. Some computing experience (preferably R, but proficiency in Matlab, Python, C++ should help you learn R relatively quickly)
4. Experience (or curiosity) with data analysis

Course Description:
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 will give 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.

Textbook (free):

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: Springer.

URL: http://www-bcf.usc.edu/~gareth/ISL/index.html

Expectations:

Each class will involve a combination of lecture and working through examples in statistical software. You are expected to bring a laptop to each class so that you can work along.

Grading:

40% Homework
25% Midterm (Week of Feb. 26)
25% Final Exam (Week of Apr. 26, last day of class)
10% Participation (via Piazza)

Homework:
Homework makes up 40% of your grade. There will be approximately one homework assignment due every week and a half (so about 6 or 7 homeworks total throughout the semester). Assignments will consist of conceptual problems about techniques/statistical ideas in addition to applications of methods to data analysis. All coding will be done in R statistical software.

Collaboration on homework assignments with fellow students through discussion of ideas is encouraged. However, you may not share written work or code; solutions must be written entirely by yourself in your own words. Any collaboration should be clearly acknowledged, by listing the names of the students with whom you have had discussions concerning the assignment.

Lateness policy: We allow for a 1 hour grace period after the deadline to account for last minute submission glitches (page lagging out, etc.). After the 1 hour grace period, assignments submitted up to 24 hours after the deadline will receive a 10% penalty; assignments submitted between 24 hours to 48 hours after the deadline will receive a 20% penalty; assignments submitted between 48 hours to 72 hours after the deadline will receive a 30% penalty; homework will not be graded if submitted more than 72 hours after the deadline.
Exceptions may be considered if a Dean’s excuse is obtained.

Exams:
There will be 2 in-class pen-and-paper exams of equal weight, each lasting 70 minutes. These exams will assess your conceptual understanding of the material. A practice exam will be provided the week before each exam.

Tentative List of Topics:

- Linear methods for regression
- Linear methods for classification
- Model selection and regularization

- Polynomial regression and splines
- Tree-based methods
- Support vector machines
- Topic models
- Neural networks
- Introduction to deep learning

Participation:

We will use the online discussion forum Piazza for questions, answers, and discussion about all aspects of the class. If you make a post/follow-up about once every other week, you'll get an easy 9 out of 10 participation points. Going out of your way to post insightful thoughts/answers every so often will earn you the full 10 out of 10 points.

http://piazza.com/yale/spring2018/sds365565