Syllabus for Machine Learning and Computational Statistics

Course name: Machine Learning and Computational Statistics
Course number: DS-GA 1003
Course credits: 3
Year of the Curriculum: one

Course Description: The course covers a wide variety of topics in machine learning and statistical modeling. While mathematical methods and theoretical aspects will be covered, the primary goal is to provide students with the tools and principles needed to solve the data science problems found in practice. This course will serve as a foundation of knowledge on which more advanced courses and further independent study can build.

Course Instructors: Julia Kempe and David Rosenberg

Academic Term in which course is given: Spring

Contact Hours: 14-week semester. Each week comprises 100 minutes of lectures and 50 minutes of lab session (in a classroom format). Course staff will be available for office hours at least 3 hours per week. Course staff will also be available online through our Piazza page (https://piazza.com/nyu/spring2019/dsga1003/home).

Course aims and objectives:

- Teach intermediate topics in machine learning
- Provide a basis for advanced study of machine learning and statistical modeling

Prerequisites:

- Introduction to Data Science (DS-GA 1001), or equivalent
- Statistical and Mathematical Methods (DS-GA 1002), or equivalent
- **Solid mathematical background**, equivalent to a 1-semester undergraduate course in each of the following: linear algebra, multivariate calculus, probability theory, and statistics (DS-GA 1002 covers the necessary material)
- **Python programming required** for most homework assignments
- *Recommended:* Computer science background up to a course in data structures and algorithms
- *Recommended:* At least one advanced, proof-based mathematics course
- Some prerequisites may be waived with permission of the instructor

Tentative List of Topics By Week (this may still change somewhat):

- Week 1: statistical learning theory framework, excess risk decomposition
- Week 2: excess risk decomposition, stochastic gradient descent,
- Week 3: L1/L2 regularization, Lasso, and Elastic Net, subgradient methods
- Week 4: loss functions, Duality, SVM, Representer theorem
- Week 5: intro to kernel methods,
- Week 6: Maximum Likelihood, Conditional Probability Methods
- Week 7: Midterm

Spring Break

- Week 8: Baysian Methods
- Week 9: Bayesian Conditional Methods, Multiclass
- Week 10: Classification and Regression Trees, Intro to Bootstrap
- Week 11: Bagging and Random Forests, Gradient Boosting
- Week 12: k-Means, Gaussian Mixture Models
- Week 13: general EM algorithm
- Week 14: intro to Neural Networks, Backpropagation

Time permitting, we may be able to cover some of the following additional topics: natural exponential families, generalized linear models, ranking problems, collaborative filtering, sparse Bayesian models (RVM), model selection, bandit problems (Thompson sampling and UCB methods), or learning-to-rank. All of these are accessible topics for a class at this level.

Method of assessment:

- **Homework**: There will be roughly 7-8 homework assignments with both written and programming components. Some homework problems are designated “**optional**”. These problems will be graded, but have **no effect** on the overall homework score (but see below). Homework will be accepted for 48 hours after the time it is due, but will have a 20% penalty.
- **Extra Credit**: Many homework assignments will have problems designated as “optional”. At the end of the semester, strong performance on these problems may lift the final course grade by up to half a letter grade (e.g. B+ to A- or A- to A), especially for borderline grades. Strong performance on optional problems would be noted in recommendation letters. You should view the optional problems primarily as a way to engage with more material, if you have the time. Along with the performance on optional problems, we will also consider significant contributions to Piazza and in-class discussions for boosting a grade.

Grading: The final numerical score will be the weighted average of homework score (40%), the Midterm (30%), and the Final Exam (30%).

Bibliography and other resources:

- Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*, Second Edition, Springer-Verlag, 2009.
- Shalev-Shwartz and Ben-David, *Understanding Machine Learning: From Theory To Algorithms*, 2014.
- David Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.
- James, Witten, Hastie, Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.
- Christopher Bishop, *Pattern Recognition and Machine Learning,* Springer, 2007.

Instructor/course evaluation: Students will complete an anonymous survey electronically at the end of the term. The tabulated results will be reviewed by the instructor, the director of the program, and the chair of the home department of the instructor. Issues will be identified and managed to successful remediation.

Academic Integrity Policy: The course conforms to NYU's policy on academic integrity for students: (http://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/academic-integrity-for-students-at-nyu.html

This policy prohibits plagiarism and cheating.

- Plagiarism: presenting others' work without adequate acknowledgement of its source, as though it were one's own. Plagiarism is a form of fraud. We all stand on the shoulders of others, and we must give credit to the creators of the works that we incorporate into products that we call our own. Some examples of plagiarism:
    - a sequence of words incorporated without quotation marks
    - an unacknowledged passage paraphrased from another's work
    - the use of ideas, sound recordings, computer data or images created by others as though it were one's own
- Cheating: deceiving a faculty member or other individual who assess student performance into believing that one's mastery of a subject or discipline is greater than it is by a range of dishonest methods, including but not limited to:
    - bringing or accessing unauthorized materials during an examination (e.g., notes, books, or other information accessed via cell phones, computers, other technology or any other means)
    - providing assistance to acts of academic misconduct/dishonesty (e.g., sharing copies of exams via cell phones, computers, other technology or any other means, allowing others to copy answers on an exam)
    - submitting the same or substantially similar work in multiple courses, either in the same semester or in a different semester, without the express approval of all instructors
    - submitting work (papers, homework assignments, computer programs, experimental results, artwork, etc.) that was created by another, substantially or in whole, as one's own
    - submitting answers on an exam that were obtained from the work of another person or providing answers or assistance to others during an exam when not explicitly permitted by the instructor
    - submitting evaluations of group members' work for an assigned group project which misrepresent the work that was performed by another group member
    - altering or forging academic documents, including but not limited to admissions materials, academic records, grade reports, add/drop forms, course registration forms, etc.

Authors of Syllabus: David Rosenberg, Yann LeCun, David Sontag, Julia Kempe