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 Instructor: David Rosenberg, dr129@nyu.edu

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) or supervised research activities. 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/spring2016/dsga1003/home).

Course aims and objectives:

- Teach intermediate topics in machine learning
- Provide hands-on experience in designing and programming data science algorithms
- 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:

- Week 1: statistical learning theory framework, stochastic gradient descent, matrix/vector differentiation
- Week 2: excess risk decomposition, L1/L2 regularization, Lasso algorithms
- Week 3: loss functions, convex optimization, SVM
- Week 4: subgradient methods, feature engineering
- Week 5: representer theorem, kernel methods, regression trees

- Week 6: classification trees, **test 1**
- Week 7: bootstrap, bagging, random forest

Spring Break

- Week 8: AdaBoost and gradient boosting
- Week 9: structured prediction
- Week 10: maximum likelihood, conditional probability models
- Week 11: **test 2**
- Week 12: bayesian networks, class-conditional models, naïve Bayes
- Week 13: k-means clustering, gaussian mixture models, intro to EM
- Week 14: general EM algorithm
- Week 15: project poster session

Time permitting, we may be able to cover some of the following additional topics: natural exponential families, generalized linear models, bayesian methods, bayesian linear models ranking problems, collaborative filtering, sparse Bayesian models (RVM), Bayesian model selection, feed-forward neural networks, and bandit problems (Thompson sampling and UCB methods). All of these are accessible topics for a class at this level. In any case, *ambitious students are encouraged to seek my guidance in pursuing these topics on their own*.

Method of assessment:

- **Homework**: There will be roughly 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 is due at **6pm** on the date specified. Homework will still be accepted for 48 hours after this time but will have a 20% penalty.
- **Tests**: There will be two tests. Each will cover material from lectures, lab sessions, homework, and assigned readings up to the week before the exam. The first test will be one-hour long, and the second test will be two hours.
- **Final Project**: Final projects will be done in groups of two or three students. Each group will be assigned to a senior data scientist from industry who will serve as an adviser. The project will typically involve either a new data source, or doing something new with a well-known data source. More methodological or theoretical projects are also possible. In any case, the project must have some degree of "figuring out the approach", rather than just implementing or comparing known methods.
- **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. 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 first test (15%), the second test (25%), and the final project (20%).

Bibliography and other resources:

- Hastie, Tibshirani, Friedman, *Elements of Statistical Learning*, Second Edition, Springer-Verlag, 2009.
- 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.
- Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.
- Boyd and Vandenberghe, *Convex Optimization*, Cambridge University Press, 2009.

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