Machine Learning (DS-GA 1003)

Term: Spring 2019

Instructor: He He,
Christopher Policastro

Section Leader: Shubham Chandel
Joshua Meisel
Yiqiu Shen

Grader: Raghav Jajodia, Peeyush Jain, Ieshan Vaidya Chinmay Singhal, Aniket Bhatnagar, Sarthak Agarwal

Contact: Public/Private correspondences should be sent through Piazza

Websites: NYU Classes: https://newclasses.nyu.edu/portal
JupyterHub: https://dsga-1003.rcnyu.org
GitHub: https://github.com/cp71/DS-GA-1003-SPRING-2020-PUBLIC
Piazza: https://piazza.com/
Gradescope: https://www.gradescope.com/
CodaLab: https://codalab.org/

Lectures: Tuesdays from 5:20pm-7:00pm
at GCASL Room C95
(Global Center for Academic and Spiritual Life, 238 Thompson St.)

Labs: Wednesdays from 6:45pm-7:35pm
at GCASL Room C95
(Global Center for Academic and Spiritual Life, 238 Thompson St.)

Instructor Office Hours: Christopher Policastro Thursday’s 9:30-10:30am
(60 Fifth Ave., Room 650 and NYU Classes via Zoom Conferencing)

Section Leader Office Hours: TBD

Grader Office Hours: TBD

Course Description

The course introduces statistical and computational methods in machine learning with a focus on supervised learning (regression and classification) and unsupervised learning (dimension reduction and clustering). Students will formulate problems and implement algorithms to make predictions and inferences about data.

The principles and techniques covered in the course will help prepare students for courses on deep learning, causal inference, natural language processing, or optimal control. The practical experience with the tools needed for implementing algorithms will enable students to solve problems with data.

Course Objectives

Students will study principles and techniques for exploring patterns in data, predicting unknown information from known information, and estimating likelihoods. The techniques include both concepts and methods. Concepts teach practices about modelling. Methods teach implementation of models through programming. Learning outcomes include

- Determining Models
    - Bias and Variance
    - Error and Noise
    - Regularization and Validation
- Exploring Patterns
    - Clustering
    - Latent Variables
    - Kernels
- Making Predictions
    - Linear/Logistic Regression
    - Support Vector Machine
    - Neural Networks
- Quantifying Uncertainty
    - Boosting
    - Bagging
    - Expectation Maximization
- Gaining experience
    - Programming in Python
    - Querying in SQL
    - Typesetting in Latex

Throughout the semester, the course will be conducted in Python. However, the course will not focus on the specifics of Python. Instead, students will learn programming skills that should apply to different programming languages.

Course Requirements

The instructor will hold lecture once a week for 100 minutes. Lecture will combine instructional lessons and interactive activities. The section leader will conduct section once a week for 50 minutes. Section will include both discussion and lab. Discussions will give the class the opportunity to review the lecture guided by questions about the material. Labs will help prepare students for homework. Working in groups, students will solve problems related to the homework assignments alongside the section leader. Students will be expected to spend time studying outside of class. Grading will be determined by assignments, exams, and a project.

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.
- Kevin Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2010.
- Christopher Bishop, *Pattern Recognition and Machine Learning,* Springer, 2007.

Course Prerequisites

The prerequisites are Introduction to Data Science (DS-GA 1001) and Statistical and Mathematical Methods (DS-GA 1002). Equivalent experience is admissible with permission of instructors.

Students should have experience with programming in Python. Experience with algorithm and data structures could be helpful.

Students must have the equivalent of one semester undergraduate course in linear algebra, multivariate calculus, probability theory, and statistics. Experience with proofs could be helpful.

Course Policies

The grade will be based on assignments, exams, and a project:

- Homework
    - Students will submit assignments through Gradescope

    - 8 assignments combining code and calculations. 1 assignment (Homework 0) outlining steps for submission along with formatting conventions
- Exams
    - Students will complete pencil-and-paper exams in class.
    - *Midterm*: Held the 7th week of the semester. If a student misses the midterm, then the final will be worth 45% (see distribution below).
    - *Final*: Held the 16th week of the semester. Time and place to be determined by the registrar.
- Project
    - Students will work together in groups on a project. The project will take the form of a competition hosted on CodaLab. However students will not be evaluated in comparison with their classmates. Moreover the project will not be graded on technicality or novelty. The project is an opportunity for groups to explore applications.
    - *Project Proposal*: due the 8th week of classes. 1/2 page write-up addressing the plans for the project.
    - *Project Milestone*: due the 12th week of classes. 1 page write-up addressing methodology, experiments and relevant datasets.
    - *Project Poster*: due the 16th week of classes. Notebook detailing datasets, features, models, and algorithms along with evaluation of results. 1 page pdf poster summarizing the notebook. Please see instructions on JupyterHub for guidelines and template.

Surveys

The instructors will ask students to complete three anonymous surveys. Surveys are accessible through links on NYU Classes. Survey 1 will be posted in Week 1 to learn the background and interests of the class. Survey 2 will be posted in Week 4 to gather suggestions about the class. Survey 3 will be posted in Week 9 to follow up on the suggestions and to gauge the midterm exam.

Access

Students will use NYU Classes, Gradescope, Piazza and JupyterHub throughout the semester.

- *NYU Classes*
    - Syllabus, Calendar, Zoom Conference, Link to Piazza, Link to Gradescope, Link to JupyterHub
    - Week 1 to Week 16 agendas containing plan for lecture/section along with references and links to materials
- *Piazza*
    - Public and Private correspondence with students and instructional staff
- *Gradescope*
    - Submission of Homework and Projects. Note that Labs are not submitted through Gradescope.

    - With the exception of mid-semester and final grades, the instructional staff will not maintain grades on NYU Classes.
- *JupyterHub*
    - Students can access JupyterHub with their NYU credentials at https://dsua-112.rcnyu.org
    - JupyterHub fetches from the GitHub repository https://github.com/cp71/DS-GA-1003-SPRING-2020-PUBLIC
        - Materials for Lecture and Section
        - Homework and Datasets
    - If you plant to use JupyterHub for class, then access it before Lecture and Section to start your server. Students will be logged out after half an hour of idleness. Students will be logged out after three hours of use.
- *CodaLab*
    - Submission of Project

Collaboration

Students can collaborate on homework and labs. However, students are responsible for mentioning their collaborators' contributions in their submission. Homework or labs without acknowledgements violates course policies. With the exception of packages, students should avoid including duplicating code in their homework, labs, and project. If students duplicate code in their programs, then they must provide comments about the source with attributions.

Late Assignments

For homework and projects, each student gets 5 extension days.

- Extensions are rounded up to the nearest day. For example, 1 minute late means 1 extension day.
- After 5 extension days are used, any homework handed in late will be marked off 20% per day late, rounded up to the nearest number of days.
- No homework will be accepted more than 2 days late.
- Any regrade requests must be submitted through Gradescope within 2 days of release of grades.

Assignments will be **due before 12PM** on the day of the deadline.

Grades

The following weights will be used in the assignment of final grades:

Homework	40%
Midterm	20%
Final	25%
Projects	15%

Instructional staff will increase a letter grade through demonstration of participation

- Participation in Lecture and Section
- Participation in Instructor, Section Leader and Grader Office Hours
- Participation on Piazza

For example, a grade of B will increase to a grade of B+.

If you have questions about your grades, then please come to grader office hours rather than instructor office hours or section leader office hours

Schedule of Classes

The check the Weekly Agenda along with the Calendar on NYU Classes for the schedule.

Week	**Topic**
Week 1	Regression and Clustering
Week 2	Classification and Gradient Descent
Week 3	Regularization
Week 4	Support Vector Machines
Week 5	Kernel Methods
Week 6	Multiclass Classification
Week 7	Midterm
Week 8	Spring Break
Week 9	Conditional Probability Models
Week 10	Bayesian Methods
Week 11	Trees and Bootstrap
Week 12	Boosting
Week 13	Neural Networks
Week 14	PCA
Week 15	GMM and EM
Week 16	**Final**

University Policies

Academic Integrity

Work you submit should be your own. Please consult the CAS academic integrity policy for more information: https://cas.nyu.edu/content/nyu-as/cas/academic-integrity.html – penalties for violations of academic integrity may include failure of the course, suspension from the University, or even expulsion.

Observances and Sick Days

As a nonsectarian, inclusive institution, NYU policy permits members of any religious group to absent themselves from classes without penalty when required for compliance with their religious obligations. The policy and principles to be followed by students and faculty may be found here: The University Calendar Policy on Religious Holidays (http://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html)

If you are unwell, then please do not attend lecture, section or office hours. Please contact the instructional staff about the circumstances. If the absence impacts your completion of an activity, then the instructional staff will work with you to find an alternative time.

Disability Disclosure Statement

Academic accommodations are available for students with disabilities. The Moses Center website is www.nyu.edu/csd. Please contact the Moses Center for Students with Disabilities (212-998-4980 or mosescsd@nyu.edu) for further information. Students who are requesting academic accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance.