MASTER of SCIENCE PROGRAM
Mathematics in Finance

NEW YORK UNIVERSITY
Courant Institute of Mathematical Sciences

Machine Learning & Computational Statistics (Half Semester)
MATH-GA.2071-001
Thursdays, 7:10 p.m. - 9 p.m.
CIWW Room:102
Updated January 19, 2025

Instructors:
Ivailo Dimov: iid201@nyu.edu
Grader:
Chao Cheng: cc7990@nyu.edu
Online office hours: TBA

Special dates
Thursday, January 23, 2024 - first lecture
Thursday, March 6, 2024 - final lecture and exam (1 hr exam)

Prerequisites: Multivariate calculus linear algebra, and calculus-based probability. Students should also have working knowledge of basic statistics and machine learning (such as what is covered in the course "Data Science and Data-Driven Modeling"). Experience with Python.

Course description: This half-semester course (a natural sequel to the course "Data Science & Data-Driven Modeling") examines techniques in machine learning and computational statistics in a unified way as they are used in the financial industry.

We cover supervised learning (regression and classification using linear and nonlinear models), specifically examining splines and kernel smoothers, bagging and boosting approaches; and how to evaluate and compare the performance of these machine learning models. Cross-validation and bootstrapping are important techniques from the standard machine learning toolkit, but these need to be modified when used on many financial and alternative datasets. In addition, we discuss random forests and provide an introduction to neural networks.

Hands-on homework forms an integral part of the course, where we analyze real-world datasets and model them in Python using the machine learning techniques discussed in the lectures.

It is important that students taking this course have good working knowledge of multivariate calculus, linear algebra and calculus-based probability. Students should also know basic statistics and machine learning (such as what is covered in the "Data Science & Modeling" course at NYU Courant) and be familiar with the standard "Python stack."

Course requirements: Students are expected to attend lectures and actively participate. The main work for students outside of class includes homework based mathematical,

computational and data driven modeling.

Throughout the semester there will be a short in-class quiz assessing the understanding of the lectures and assigned readings.

There is an exam at the end of the half-semester.

Each student will be graded based on their performance in (1) homework assignments (40%); (2) Quiz (20%); Exam (40%).

Late homework policy: Unless a student obtains approval from the instructors to submit a deliverable late, there is a penalty according to this schedule:

- Baseline penalty for late homework: 20%
- More than a week: 40%
- More than two weeks: 60%
- More than a month, or if other students' corrected homework has been returned: 100%

Academic Integrity: Students must make a serious commitment to academic integrity. The consequences of cheating are serious for the cheater and for NYU as a whole. Students are required to immediately report cheating incidents they observe to the instructors.

A student caught cheating on an assignment may have their grade for the class reduced by one letter at the first offense and their grade reduced to an F, or dismissal from NYU, for repeated offenses. During exams and quizzes, students may not communicate in any way, nor use any materials or technology not explicitly permitted. No cellphone or other electronic devices may be used during the exam; they should be stored away. Students may not look at each other's test and/or screen during the exam. Cheating on an exam or quiz will immediately result in an F for the class.

Students are not permitted to share course materials outside of class without written permission of the instructors. Students are not allowed to record (photography, audio and/or video) any lectures or sessions without written permission of the instructors.

All disciplinary actions will be taken in accordance with the policies of the

NYU Graduate School of Arts and Science: GSAS Statement on Academic Integrity (nyu.edu)

Communication & Zoom: Announcements, homework and other course related material will be posted on Brightspace. The NYU Brightspace site has a discussion forum for the course. Please post any questions related to lectures, homework and group projects there in the appropriate thread.

This course will be fully in person. Whenever personal instruction is unavailable, zoom lectures will be provided with links and recordings will be available in the Zoom section of the course site.

Textbook readings & suggested readings:
[D2L.ai] https://d2l.ai - Deep Dive to Deep Learning online textbook

[Golub] G. H. Golub and C. F. Van Loan, *Matrix Computations*, JHU Press (2012)

[Hastie] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning,* Springer Series in Statistics (2008).
https://web.stanford.edu/~hastie/Papers/ESLII.pdf

[Hayashi] Hayashi, Fumio. *Econometrics*, Princeton University Press (2000).

[James] James et al., *An Introduction to Statistical Learning*, Springer Texts in Statistics (2013).
http://www.stat.berkeley.edu/~rabbee/s154/ISLR_First_Printing.pdf

[Goodfellow] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning (Adaptive Computation and Machine Learning series)* Artificial Intelligence and Semantics (2016).
https://www.deeplearningbook.org

Article readings:

[Breiman: RForests] L. Breiman, "Random Forests," Machine Learning, v 45, Issue 1, pp 5-32 (2001).

[LeCun: Backprop] Y. LeCun, et. al. "Efficient backprop," Lect. Notes Comput. Sci.(including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 7700 (1998): 9-48.

Software:

- Jupyter notebook, Python
- Python stack: anaconda, numpy, pandas, matplotlib, bqplot, scikit-learn. statsmodels, nlp toolkit, beautifulsoup

Software will be introduced in class. Download links and instructions will be provided as part of class handouts.

Tentative schedule:

Week 1 Supervised learning: Linear regression and classification
Course Overview and Syllabus
Lectures:

- Review of the bias-variance tradeo-off, loss functions and Linear Models in regression
- Linear classification I: Logistic regression, linear discriminant analysis

Reading: Lecture slides, Hastie 3.3 and 3.6-3.9, 4.1, 4.4

Week 2 Linear classification II. Intro to non-linear models: feature map regression
Lectures:

- Demo of Elastic Nets, Classification and evaluating of classifier models via cross-validation
- Evaluating classifiers: Classification accuracy, confusion matrix, metrics computed from a confusion matrix
- Demo of Latent Semantic Analysis and classification

Reading: Lecture slides, Parts of Hastie Ch 5-6 and demos based on Géron

Week 3 Intro to Feature Map Regression; Tree based models I
Lecture:

- Feature map regressions, their universality and regularization
- Demo of splines and kernel smoothing
- Intro to decision and regression trees, boosting and additive models

Reading: Lecture slides, Hastie Ch. 9-10 and demos based on Géron

Week 4 Tree based models II and SVMs
Quiz (students need to attend the live online lecture to take the quiz)
Lecture:

- Demo of trees
- Bagging and random forests
- Demo of boosting and random forests

Reading: Lecture slides, Hastie Ch. 12, Ch.15 and demos based on Géron

Week 5 Cross-validation, SVMs and intro to neural networks
Lecture:

- Support Vector Machines
- Review of k-fold cross-validation, other CV approaches, CV for time-series data
- Single layer perceptron, logistic regression and the XOR problem
- Multilayer perceptrons, universality and backpropagation

Reading: Lecture Slides, Hastie Ch. 7.10-7.11, parts of Ch. 11, Géron Ch. 10

Week 6 Neural networks II
Lecture:

- Deep networks design, training and regularization
- Demos of PyTorch network design and training

Reading: Lecture slides, Géron Ch. 10-11