DATA SCIENCE II:
Machine Learning
MTH 9899
Baruch College

Spring 2016

Instructor:	Adrian Sisser	**Email:**	Adrian.Sisser@baruch.cuny.edu
Instructor:	Dmytro Karabash	**Email:**	Dmytro.Karabash@baruch.cuny.edu
Time:	W 18:05 – 21:00	**Room:**	9-140, (24 St & Lexington Ave)
TA:	TBA		

Office Hours: By appointment before class.

Main References: This is a restricted list of various interesting and useful books that will be touched during the course. You need to consult them occasionally.

- Trevor J. Hatie and Robert John Tibshirani and Jerome H. Friedman, *The Elements Of Statistical Learning : Data Mining, Inference, and Prediction* - Available here .

Objectives: This course should teach you machine learning with applications to finance.

Prerequisites:

- An undergraduate-level understanding of Probability, Statistics, Graph Theory, Algorithms, and Linear Algebra.
- Some knowledge of Python and Linux is highly preferable but can be picked up during the course (we are not going to teach it).

Grading Policy: Homework and quizzes (40%), Final Project (50%), instructor discretion (10%). At the end of the course, the lowest grade from any quiz or homework will be automatically dropped.

Class Policy:

- Regular attendance is essential and expected.
- We will do our best to provide a 10 minute break in the middle of class.

Academic Honesty: Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation. We take it seriously.

Overview: Machine learning is a very broad area of study that can't be completely covered in a single half-semester course. We will give very brief coverage to a lot of different algorithms, while providing more extensive focus on two that have found widespread practice inside of finance, Regression Trees and Neural Networks/Deep Learning. Lectures will be supplemented throughout with online lecture videos from various sources that will be provided throughout the class. We will then use lecture to reinforce these lectures focus more deeply on the details of certain algorithms and applications to finance. The supplemental lectures are mandatory, and the material learned in them will be used in periodic quizzes. In addition, they will be covered in homeworks and might be used in projects.

Homework: Most homework assignments will consist of using various ML algos on datasets that will be provided. The goal is to implement algorithms as well as use existing tools to understand what techniques work in different problems. Most of the analysis results and discussion of them. There will approximately 4 homework assignments over the class. Homework assignments will be due at 6PM on the day of class. For each student, we will allow 1 late (up to 24 hours) homework assignment per semester, and NO credit will be received for other late homeworks. Homeworks can be worked on collaboratively, but not copied. The names of collaborators must be included.

Quizzes: Quizzes will be given at the start of class and will include topics covered in assigned readings and supplemental materials up through that day. All quizzes will be announced at least 48 hours in advance, if you have any issues making it to class on time for the quiz, please let us know at least 24 hours in advance and we will try to make arrangements.

Projects: The class will work on a large project in assigned groups of 2 or 3. The subject for the projects will be real financial data sets, and the goal will be to apply the ML algos to build models essential to successful quantitative trading. As part of the projects, students will model the datasets using existing analysis tools in Python. At the end of the course, groups will present their findings to the class.

Optimistic Course Outline: Below is a list of topics we hope to cover. Please note that we will likely not be able to stay on this exact schedule.

- Lecture 1 - General topics
  - Supervised vs Unsupervised Learning
  - Classification vs Regression
  - Cross-Validation
  - Fit-quality Measurements
  - Bias-Variance Trade-off
  - Nearest Neighbors
  - Logistic Regression
  - Optimization Techniques - Stochastic Gradient Descent and Batch Learning
- Lecture 2 - Linear Models and More
  - Least Squares, Partial Least Squares, Lasso, Ridge, and Elastic Net Regression
  - Least Angle Regression
  - Basic Expansion
  - Principal Component Analysis and Dimensionality Reduction
  - Trees - A Primer for Classification
  - EM Algorithm - Fitting Gaussian Mixture Models and Outlier Detection
  - Support Vector Machines/Regression
- Lecture 3 to 4: Trees + Boosting/Ensemble Methods
  - ID3/C4.5/C5.0 Algorithms
  - Model-Based Trees
  - Ensemble Learning Methods
  - Boosting
  - Bootstrapping and Subsampling

- Lecture 5 to 7: Deep Learning
  - Keras - A Python Framework for Deep Learning
  - TensorFlow
  - Restricted Boltzman Machines
  - Neural Networks
  - Multilayer Perceptrons
  - Convolutional Neural Networks
  - AlexNet
  - Reccurent Neural Network (LSTM)
  - Bidirectional Reccurent Neural Network (LSTM)