Machine Learning and Algorithms for Computational Social Science
ECON GA 4005

Thomas Sargent

Spring Semester 2022

3 credits

E-mail: cc7768@gmail.com
Web: **http://www.chasegcoleman.com**
E-mail: spencerlyon2@gmail.com
Web: **http://www.spencerlyon.com**
E-mail: ts43@nyu.edu
Web: **http://www.tomsargent.com/**

Office Hours: TBD
Class Hours: TBD

Course Description

This courses teaches students both foundational and frontier algorithms and models used to analyze social science data. Analytical goals are data description, data reduction, and detection of relationships among variables, and ways to interpret those relationships in terms of underlying economic and social forces. The course teaches how estimated (i.e., "fit") models can be used for prediction, forecasting, and possibly inference about cause and effect.

Course Materials

- **QuantEcon Datascience** Lectures from the QuantEcon datascience sequence at **https://datascience.quantecon.org**
- **Python Data Science Handbook** by Jake Vanderplas **https://jakevdp.github.io/PythonDataScienceHandbook/**
- **Statistical Rethinking** by Richard McElreath
- **Deep Learning** by Ian Goodfellow, Yoshua Bengio and Aaron Courville
- **Deep Learning with Python** by François Chollet

Prerequisites

We assume that students are comfortable writing programs in Python using the core data-analytics packages `numpy` and `pandas`. We also assume that students have a solid mathematical background that includes at least one course in calculus, one course in linear algebra, and exposure to probability theory and statistics.

Motivated students without a strong mathematical background can still succeed in this course, provided they are willing to work hard. This course is practical and application oriented. We aim to arm students with understandings of how the methods and algorithms work. Some of these understandings are grounded in the mathematics, but they are substantially strengthened through exploration and practice. To explore data, it always help to have some theory as a guide.

Course Objectives

This course provides students with a hands-on introduction to widely used algorithms and models used for understanding social science data. It is expected that students will be able to apply these algorithms to new datasets and problems as well explain (at a high level) how the algorithms arrive at the answers the produce.

Successful students will be able to do the following after completing this course

- Apply *exploratory data analysis* (EDA) techniques to understand a new dataset
- Communicate with subject matter experts about important relationships among variables in a dataset and key aspects of the data collection process[1]
- Determine what class of algorithm should be applied to new problems
- Implement cross-validation procedures for model evaluation and selection
- Visualize and interpret model outputs with the aim of understanding how they relate to model inputs
- Understand the concept of overfitting and how to alter model structure or apply regularization techniques to control its effects
- Utilize high-performance libraries such as `tensorflow` and `pytorch` to implement deep learning models
- Understand *Alpha-go* techniques and be able to apply them to a simpler problem domain

[1] That understanding will then inform the model-construction process.

Course Structure

Class Structure

This course will meet once a week for 3 hours.

Classes will contain mixtures of lecture time and lab time. Students should bring, and expect to use, their laptops every time the class meets.

Assessments

This course will use a mixture of homework assignments, in-class quizzes, exams, and a final project to evaluate students.

Homework:At the beginning of the course, homework will be assigned almost every week. Later in the course, there will be less frequent assignments in order to assure that you have time to work on your class project. Your two lowest homework grades will be dropped.
In-class quizzes: There will be 4 in-class quizzes during the semester. These quizzes will be open book and the lowest score will be dropped.
Exams: There will be 2 take-home exams.
Project: There will be a class project aimed at helping you apply the tools that you have learned to a "real-world problem."

Other than for quizzes and exams, we highly encourage students to work together. We have found that groups of 3-4 seem to work best. We believe that collaborative work is the best way to learn the type of material that we cover. We advise students not to rely on others to do work that you do not understand.

Grading Policy

The assignments just described will be the main inputs to the grade for the course. Assignments will be weighted evenly within groups and overall according to the following decision rule:

- Homework assignments: 25%
- In-class quizzes: 15%
- Tests: 20%
- Project: 40%

This weighting reflects our opinion that the most important skills to be acquired in this class are communicated by one's ability successfully to apply the tools that you learn to an interesting question in the social sciences.

Grading policy: A =100-90%; A- = 90-85%; B+ = 84-80%; B = 79 - 70%; C = 69 - 50%; D = 49 and lower

Schedule and weekly learning goals

The schedule is tentative and subject to change so that we can adapt material to incorporate new developments in the fast-moving fields of AI and machine learning. Several of the modules below will occupy more than one week. The learning goals target key concepts to be mastered after each module. Later modules build on earlier ones.

Week 1: Introduction to Machine Learning

Sources and tools:

- Class notes
- https://datascience.quantecon.org/applications/ml_in_economics.html

Topics to be mastered:

- Definition of Machine Learning (ML)
- Relation to econometrics
- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

Week 2: Linear Models, I

Sources and tools:

- Class notes
- https://python.quantecon.org/ols.html
- https://www.statsmodels.org/stable/index.html
- https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model

Topics to be mastered:

- Linear regression for regression tasks
- Logistic regression for classification tasks
- Regression with `statsmodels`
- Regression with `scikit-learn`

Week 3: Linear Models, II

Sources and tools:

- Class notes
- https://datascience.quantecon.org/applications/classification.html
- https://datascience.quantecon.org/applications/regression.html

Topics to be mastered:

- Loss functions
- Goodness of fit (evaluation metrics)
- Dealing with imbalanced datasets
- Homoscedasticity versus heteroscedasticity in regressions
- **Example** Fitted residual plots with height vs. weight data
- Handling categorical, sparse, temporal, or discrete data

Week 4: Overfitting and model selection, I

Sources and tools:

- Class notes
- https://scikit-learn.org/stable/model_selection.html
- Chapter 5 of Deep Learning

Topics to be mastered:

- Define overfitting
- Diagnose overfitting
- Bias-variance tradeoff
- VC dimension
- Regularization via L1 and/or L2 penalty (Lasso, Ridge, ElasticNet regression)
- Bayesian interpretation of Lasso and Ridge regression

Week 5: Overfitting and model selection, II

Sources and tools:

- Class notes
- https://scikit-learn.org/stable/model_selection.html
- Chapter 5 of Deep Learning

Topics to be mastered:

- Training, validation and test sets
- Hyperparameter tuning
- Data leakage
- Grid search
- Random search

Week 6: Overfitting and model selection, III

Sources and tools:

- Class notes
- https://scikit-learn.org/stable/model_selection.html
- Chapter 5 of Deep Learning

Topics to be mastered:

- Cross validation procedure
- Cross validation for time series
- Model ensembles, stacking, bagging and boosting
- **Example** Polynomial regression

Week 7: Data Preprocessing

Sources and tools:

- Class notes
- https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing

Topics to be mastered:

- Standardization
- Feature encoding
- Polynomial features
- Non-linear transformations

Week 8: Non-linear models

Sources and tools:

- Class notes
- https://python.quantecon.org/mle.html
- https://scikit-learn.org/stable/

Topics to be mastered:

- Decision trees
- Random forest
- k-nearest neighbors
- Support vector machines
- The "kernel trick"
- Non-linear models in `scikit-learn`
- Maximum Likelihood Estimation

Week 9: Unsupervised Learning

Sources and tools:

- Class notes
- https://scikit-learn.org/stable/unsupervised_learning.html
- https://datascience.quantecon.org/applications/working_with_text.html

Topics to be mastered:

- Principal Component Analysis
- Latent Semantic Analysis
- t-SNE

Week 10: Numerical Optimization, I

Sources and tools:

- Class notes
- https://www.tensorflow.org/api_docs/python/tf/keras/optimizers
- https://pytorch.org/docs/stable/optim.html
- Chapter 4 of Deep Learning

Topics to be mastered:

- Gradient Descent
- Stochastic Gradient Descent (and its modern variants)
- Conditioning
- Overflow and underflow

Week 11: Numerical Optimization, II

Sources and tools:

- Class notes
- https://www.tensorflow.org/api_docs/python/tf/keras/optimizers
- https://pytorch.org/docs/stable/optim.html

Topics to be mastered:

- Momentum optimizers
- Adaptive optimizers
- Evolutionary methods
- Second-order methods

Week 12: Deep learning, I

Sources and tools:

- Class notes
- https://www.tensorflow.org/
- https://pytorch.org/
- https://keras.io/
- Chapter 6 and 7 of Deep Learning

Topics to be mastered:

- Manifold hypothesis
- Linear regression and logistic regression using `tensorflow`
- Multilayer perceptron as nested linear regression
- Back-Propagation
- Regularization for Deep Learning
- Architecture search

Week 13: Deep learning, II

Sources and tools:

- Class notes
- https://www.tensorflow.org/
- https://pytorch.org/
- https://keras.io/
- Chapter 9 of Deep Learning

Topics to be mastered:

- Convolutional neural networks (CNN)
- Convolution and pooling
- Neuroscientific basis for convolutional networks
- 1-dimensional CNN for time-series data

Week 14: Deep learning, III

Sources and tools:

- Class notes
- https://www.tensorflow.org/
- https://pytorch.org/
- https://keras.io/
- Chapter 10 of Deep Learning

Topics to be mastered:

- Recurrent neural networks (RNN)
- Long Short Term Memory (LSTM) network
- RNNs for time-series

Week 15: Understanding Alpha-go

Sources and tools:

- Class notes
- https://deepmind.com/research/case-studies/alphago-the-story-so-far

Topics to be mastered:

- SARSA and Q-learning
- Game theory and alpha-beta pruning
- Monte-carlo tree search
- Dynamic programming and approximate dynamic programming
- Deep Q-Networks
- Generative adversarial networks
- **Example** Competition-style application to card or board games

Course Policies

Professional Behavior

Attend class. They say "eighty percent of success is just showing up." We have found that those who show up perform systematically better.

Arrive to class on time and stay until the end of class. Chronically arriving late or leaving class early is unprofessional and disruptive to the rest of the class.

We understand that the electronic recording of notes will be important for class and so computers will be allowed in class. Please refrain from using computers for anything but activities related to the class. Phones are prohibited as they are rarely useful for anything in the course. Eating and drinking are allowed in class but please refrain from it affecting the course. Try not to eat your lunch in class as the classes are typically active.