**Syllabus for Fall 2018**

**MACHINE LEARNING FOR THE SOCIAL SCIENCES (G5073)**

Time: Thurs. 4:10pm-6:00pm

Location: 327 Seeley W. Mudd Building

Michael D. Parrott

[mp3675@columbia.edu](mailto:mp3675@columbia.edu)

Office Hours: IAB 509f, Friday, 3-5

**Teaching Assistant:**

Kai Zhou

[kz2329@columbia.edu](mailto:kz2329@columbia.edu)

Office Hours: TBA

**Course Goals**

Social scientists need to fully engage with machine learning approaches that are found in computer science, engineering, AI, tech and in industry. This course will provide a comprehensive overview of machine learning as it is applied in a number of domains. Every effort will be made to draw comparisons and contrasts between this machine learning approach and more traditional regression-based approaches in the social sciences. Emphasis will also be on opportunities to synthesize these two approaches. The basis of this course comes from the W4995 - Applied Machine Learning course taught by Andreas Mueller.

The course will start with an introduction to Python, Jupyter Notebooks, and the scikit-learn package.

After that, there will be some discussion of data exploration, visualization in matplotlib, preprocessing, feature engineering, variable imputation and feature selection.

Supervised learning methods will be considered, including OLS models, linear models for classification, support vector machines, decision trees and random forests, and gradient boosting.

Calibration, model evaluation and strategies for dealing with imbalanced datasets will be considered next.

This will be followed by unsupervised techniques: PCA, clustering and cluster evaluation, and manifold learning.

Lastly, we will consider neural networks, convolutional neural networks for image classification and recurrent neural networks.

Prerequisites are basic probability and statistics, basic linear algebra and calculus. The course will use Python, and so if students have programed in at least one software language, that will make it easier to keep up with the assignments.

You will learn by doing. In class work will require the use of a laptop with the latest version of Python 3 using the Anaconda distribution. ***Please bring a laptop to each class.***

**Course Expectations**

Attendance and Class Participation. Your attendance and participation are necessary at every meeting.

Exams. We will have two take home exams that will ask you to apply what you have learned in lectures and homework assignments.

Homework Assignments. Students will have four homework assignments due throughout the semester. They will be based on writing up the results of performing the commands learned during the lectures. Specific instructions, format and deadlines will be given as the semester progresses.

Plagiarism and Academic Dishonesty: Students must do all their work within the boundaries of acceptable academic norms. See the Academic Honesty page of the CU website regarding college policy on plagiarism and other forms of academic dishonesty - <http://www.columbia.edu/cu/history/ugrad/main/handbook/academic_honesty.html>. Students found guilty of plagiarism or academic dishonesty will be subject to appropriate disciplinary action, which may include reduction of grade, a failure in the course, suspension or expulsion.

Late Assignments. Students will lose points for handing in late assignments, at the discretion of the instructor and teaching assistant.

Textbooks. The following books will help you further your understanding of the material:

* Müller, Guido: Introduction to machine learning with python (IMLP) (available **for free** for Columbia Students via [Safari Books Online](http://proquest.safaribooksonline.com.ezproxy.cul.columbia.edu/book/programming/machine-learning/9781449369880)) lots of code examples
* Kuhn, Johnson: Applied predictive modeling (APM) Models for DM & ML

(available **for free** at [Springer Link](http://link.springer.com.ezproxy.cul.columbia.edu/book/10.1007%2F978-1-4614-6849-3)

* Provost / Fawcett: Data Science for Business (DSfB) basic stats
* Tibshibani, Hastie, Friedman: Elements of Statistical Learning (ESL)
* Goodfellow, Bengio, Courville - Deep Learning (DL) (Available **for free** online [here](http://www.deeplearningbook.org/))

The course will closely follow IMLP, which also comes with Python code and uses scikit-learn (as we will). APM goes into more detail than IMLP but only contains R code. We will not use any R code in this course. DSfB focusses on a more high-level perspective and the practical impact of data science, while ESL contains a rigorous mathematical treatment of the machine learning methods.

Additional Materials. Other articles and materials will be distributed via Courseworks that cover additional topics in more depth.

Grade Distribution. The distribution of the parts for your grade is as follows:

Two Exams = 30% mid-term and final

Homework Assignments = 60% 4 to 6

Attendance and Participation = 10%

Changes: There may be adjustments in the scheduling of assignments, exams, and classrooms. Changes will be posted on Courseworks along with other announcements.

**Calendar of Class Sessions and Assignments**

Class 1 (September, 6th). Introduction; How can Machine Learning help social scientists?

|  |
| --- |
| **Reading Assignments** |
| IMLP Ch 1 (Intro to ML & Python) , APM Ch 1-2 |

Class 2 (September, 13th). Software Infrastructure: *Python and Jupyter Notebooks. Pandas. Matplotlib and visualization. (Git and Github extra credit posted)*

|  |
| --- |
| **Reading Assignments** |
| IMLP Ch 1 |

Class 3 (September, 20th). Introduction to supervised learning, basic model selection. Linear models for Regression. Note: **HW 1 is due.**

|  |
| --- |
| **Reading Assignments** |
| IMLP p25-44(KNN),  APM Ch 4-4.3 (Model Tuning), IMLP p251-262 (Model Evaluation),  APM Ch 4.4-4.8 (Model Selection), |
| IMLP p45-68 (Linear Models), APM Ch 6 (Linear regression models) |

Class 4 (September, 27th). Linear models for Classification. Preprocessing and feature engineering. Note:

|  |
| --- |
| **Reading Assignments** |
| IMLP p45-68 (Linear Models),  APM Ch 12.1-12.2 (Logistic Regression), 12.5 (Penalized Logistic),  IMLP p132-140 (Preprocessing: Scaling),  IMLP p211-220 (Feature engineering) ,  APM Ch 3 (Data Preprocessing) |

Class 5 (October, 4th). Imputation and Feature Selection. Support Vector Machines. Note: **HW 2 is due.**

|  |
| --- |
| **Reading Assignments)** |
| IMLP p236-241 (Feature Selection), APM Ch 19 (Feature Selection),  IMLP p92-103(SVM), APM Ch 13.4 (SVM) |

Class 6 (October, 11th). Decision Trees and Random Forests. Gradient Boosting and Calibration.

**[ONLINE CLASS THIS WEEK.]**

|  |
| --- |
| **Reading Assignments** |
| IMLP p70-88 (Decision Tree),  IMLP p89-92 (Ensemble Decision Tree),  APM Ch 14.1-14.4 (Bagging and Random Forest),  APM Ch 14.5 (Boosting) |

Class 7 (October, 18th). **Mid-Term Exam**

Class 8 (October, 25th). Ensemble Models (Adaboost/XGBoost). Model evaluation and imbalanced datasets.

|  |
| --- |
| **Reading Assignments** |
| IMLP p275-302 (Evaluation Metrics),  APM Ch 16 (Class Imbalance) |

Class 9 (November, 1st). Dimensionality reduction using PCA, Clustering, Manifold Learning.

|  |
| --- |
| **Reading Assignments** |
| IMLP p140-156 (PCA), p163-168 (Manifold Learning),  APM p35-40 (Data Reduction and Feature Extraction) |

Class 10 (November, 8th). Resampling strategies for Imbalanced Data.

|  |
| --- |
| **Reading Assignments** |
| APM Ch16, SMOTE, Easy Ensembles, (Class imbalance)  IMLP p168-208 (Clustering) |

Class 11 (November, 15th). Working with Text as Data. Note: **HW 3 is due**

|  |
| --- |
| **Reading Assignments** |
| IMLP p323-336 (text data) |

Class 12 (November, 29th). Neural Networks; Convolutional neural networks for image classification

|  |
| --- |
| **Reading Assignments** |
| IMLP p104-109, DL Ch 6, Ch 7.8; DL Ch 7.12, Ch 9, keras docs |

Class 13 (December, 6th). Even more on Neural Networks. Note: **HW 4 is due.**

|  |
| --- |
| **Reading Assignments** |
| [Stanford CNN course notes, Module 2](http://cs231n.github.io/), Feature Visualization |

**Take Home Final Exam Due by December 13th**