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

Teaching Assistant:  
Kai Zhou  
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 programmed 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](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](https://www.safaribooksonline.com))
- Kuhn, Johnson: Applied predictive modeling (APM) (available for free at [Springer Link](https://www.springer.com))
- Provost / Fawcett: Data Science for Business (DSfB)
- Tibshibani, Hastie, Friedman: Elements of Statistical Learning (ESL)
- Goodfellow, Bengio, Courville: Deep Learning (DL) (Available 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%
Homework Assignments = 60%
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?

<table>
<thead>
<tr>
<th>Reading Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMLP Ch 1, APM Ch 1-2</td>
</tr>
</tbody>
</table>

Class 2 (September, 13th). Software Infrastructure: *Python and Jupyter Notebooks. Pandas. Matplotlib and visualization. (Git and Github extra credit posted)*
Class 3 (September, 20th). Introduction to supervised learning, basic model selection. Linear models for Regression. Note: HW 1 is due.

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

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

Class 6 (October, 11th). Decision Trees and Random Forests. Gradient Boosting and Calibration. [ONLINE CLASS THIS WEEK.]

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


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

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

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

Class 12 (November, 29th). Neural Networks; Convolutional neural networks for image classification
Class 13 (December, 6th). Even more on Neural Networks. Note: HW 4 is due.

Reading Assignments
- Stanford CNN course notes, Module 2, Feature Visualization

Take Home Final Exam Due by December 13th