UNIVERSITY OF PENNSYLVANIA
SCHOOL OF SOCIAL POLICY & PRACTICE
Data Analytics for Social Policy Certificate

MSSP 608: Practical Machine Learning Methods

Course No. MSSP-608-001 (online)

Fall 2021

Instructors
Parijat Dube
TA: Bano Abbas
TA: Adam Alavi

Instructor Contact Information
pdube@upenn.edu
s.shahbanoabbas@gmail.com
adam2318@seas.upenn.edu

Course Description/Purpose
This course prepares students with no background in machine learning or data science to use tools from those fields effectively in applied contexts. By programming with libraries, students will build skills including feature representations of complex real-world datasets; training classification and regression models for prediction tasks; evaluation of machine learning model accuracy and error analysis; and reasoning about predictive models and making tradeoffs like bias vs. variance, granularity of annotations and predictions, and complexities of real-world labeled training data and systems, including the ethical application of predictive modeling to human-centered data.

Educational Objectives
After this course, students will be comfortable with the following skills:

- Breaking down real-world problems into machine learning tasks, inputs, and outputs.
- Using Python libraries to train classification, regression, and clustering models.
- Optimizing models through model selection, feature design, and hyperparameter tuning.
- Evaluating dataset inter-annotator reliability, trained model accuracy, and error analysis.
- Recognizing the capacity and limitations of models for fairness and explainability.
- Describing results of machine learning experiments with real-world implications.

Course Requirements and Expectations

Class Delivery: The class is fully online delivered over Zoom sessions. The scheduled time for synchronous class meetings is Thursday 8:30 AM till 10:00 AM. The first class is on Thursday September 2, 2021.

Psychological or Disability Needs: In addition to any questions or needs beyond what is listed in the rest of this syllabus, if you need any guidance or assistance with counseling, accommodations for disabilities, or other services, resources are available through Student Disabilities Services and Counseling and Psychological Services. If these resources are not meeting your needs as they relate to this course, please let the instructors know and we can determine what can be done to help you.

Attendance: This course is fully online and we will be following a flexible "flipped" model, where recorded lectures and materials are made available **before** they are covered in synchronous class periods. Our synchronous class meetings will take place each week at the scheduled time for the course, and they will be recorded.

Watching all the material is vital to learning the overall content of the course, but the specific time and place that you choose to do so is up to you. If you are not able to attend a particular synchronous session, that is okay; not using the materials at all, however, is a serious problem, and may be resolved through individual meetings with the instructors or your educational advisor, and could lead to course failure, depending on circumstances.

Participation: When you do attend synchronous sessions, you are expected you to be consistently engaged with what's going on, and to ask questions when you are confused. But if things are going on at home or in your apartment, feel free to step away for a few minutes; the recorded videos will still be there when you come back. Please mute your sound and turn off your video if you think there will be an extended interruption.

We may occasionally post check-in polls and/or requests for feedback that will occur throughout the semester, distributed through Canvas. These will not be for credit but will shape the direction of the course, and we hope you take them seriously and help us improve the course delivery by letting us know your opinions.

Assignments: This course will have weekly quizzes, short take-home activities, and five technical assignments, each of which gives you a chance to demonstrate new fundamental programming skills. Each assignment will consist of quantitative questions to be answered with data and code, and you will submit both your answers and your source code for how you answered the questions. Every class (starting from the second week) will have a quiz based on the material taught in the previous class. The quiz will happen towards the end of the lecture and students will take it on canvas. There will be 15 short take-home activities.

Later in the semester, you will be working on a final project; most students will be expected to work in pairs. First, you will submit a project proposal where you choose a dataset to work on and investigate in-depth with programming. At the end of the semester, you'll submit a final report, which will have both a code and a written component. This project is the last step of the course; there is no final exam.

The project proposal and project report will be graded primarily on your written report: your ideas, your presentation of data, and the way you integrate your technical results from your coding. However, the proposal and report will also be evaluated for professional presentation, and therefore your grade will also reflect how well you accomplished aspects of publishing your work for a broader audience, and can include mechanical issues like readability and formatting. Projects that were completed as part of a previous course, or that you are already working on for another course this semester, should not be proposed or submitted.

If you need assistance with writing, resources are available at the Weingarten Learning Center, Marks Family Writing Center, and SP2 Academic & Writing Support. If you need further assistance on writing or technical presentation, you should contact the instructors directly and explain what additional help you need.

Collaboration: Programming for technical assignments must be completed individually. You must be the one writing your own code. That being said, getting help from online resources like StackOverflow and Kaggle is a normal part of data science, and students are encouraged to talk with one another about the problems if it is helpful. When you submit homework where you reused code that you found online, you must include links to your sources.

Student should not collaborate on a homework except when it is explicitly mentioned. When you work collaboratively on any homework, you should clearly state in your submission who you worked with and how the work was separated. For students who submit similar code without acknowledging collaborative work, all students may be penalized, regardless of who wrote the code first. In general, code is subject to the same academic integrity policies as other work.

Students are expected to work in pairs for the final project. Variations from the pair structure may be allowed -- either individual work, or groups of 3 -- but only with the instructor's permission. The project proposal should clearly state the division of labor and the tasks or work that each student will be working on. Each student must complete programming work as part of the final project; student pairs are not permitted to split work responsibilities between a technical (programming) student and non-technical (report writing) student.

Projects completed in larger groups will be held to a higher standard than individual projects; substantially more work will need to be done, with more results, more programming, and a more thorough report, in order to receive the same grade as peers who worked in pairs.

Academic Integrity

Students are expected to adhere to the University's Code of Academic Integrity, available at https://catalog.upenn.edu/pennbook/code-of-academic-integrity/. Care should be taken to avoid

academic integrity violations, including plagiarism, fabrication of information, and multiple submissions (see descriptions below).** Students who engage in any of these actions will be referred to the Office of Student Conduct, which investigates and decides on sanctions in cases of academic dishonesty.

1. Plagiarism: using the ideas, data, or language of another person or source without specific or proper acknowledgment. Example: copying, in part or in its entirety, another person's paper, article, or web-based material and submitting it for an assignment; using someone else's ideas without attribution; not using quotation marks where appropriate; etc.

2. Fabrication: submitting contrived or altered information in any academic exercise. Example: making up data or statistics, citing nonexistent articles, contriving sources, etc.

3. Multiple submissions: submitting, without prior permission, any work submitted to fulfill another academic requirement.

**It is students' responsibility to consult the instructor if they are unsure about whether something constitutes a violation of the Code of Academic Integrity.

Recommended Text

There are no required texts for this course; however, we will be drawing heavily from the following texts for content and explanations.

Moline, S. (2019). *Hands-On Data Analysis with Pandas*. Packt Publishing.

- Available on Amazon for $34.19 (print) or $17.19 (ebook). https://www.amazon.com/Hands-Data-Analysis-Pandas-visualization/dp/1800563450/
- Available direct from the publisher https://www.packtpub.com/product/hands-on-data-analysis-with-pandas/9781789615326 for $44.99 (print) or $31.99 (ebook)

Raschka, S. and Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow 2, 3d edition. Packt Publishing.

- Available on Amazon for $36.96 (print) or $17.19 (ebook). https://www.amazon.com/Python-Machine-Learning-scikit-learn-TensorFlow-dp-1789955750/dp/1789955750
- Available direct from the publisher https://www.packtpub.com/product/python-machine-learning-third-edition/9781789955750 for $37.99 (print) or $27.99 (ebook)

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

- Available on Amazon for rent: $22.98 (print) or $16.43 (ebook)
- Available on Amazon for sale: $48.96 (print) or $43.44 (ebook)
  https://www.amazon.com/Data-Mining-Practical-Techniques-Management-dp-0128042915/dp/0128042915/

Getting Help / Office Hours

We will have a course Piazza for asking questions that can be answered for the whole class to see. You can also address private questions directly to the instructors. Use this link to access Piazza for the course: piazza.com/upenn/fall2021/mssp6080012021c

The instructors will do their best to respond promptly by Piazza or by email, if necessary. Depending on your question or concern, the instructors may schedule individual phone or Zoom calls to resolve more complex requests.

All the instructors (the professor and the two TAs) will be holding weekly office hour. The hours will be scheduled in the first week of class based on a student survey of available times. Office hours with each TA will begin in week 2.

Grading

15% Flipped Activities

15% Weekly Quizzes

50% Programming Assignments (5 assignments each 10%)

20% Final Project

Class Schedule

Week 1 (September 2): Getting Started with Machine Learning

- Overview of course topics and goals, online schedule and structure.
- Introduction to Machine Learning
- Supervised, Unsupervised, and Reinforcement Learning
- Core machine learning tasks: classification and regression
- Machine learning terminology and pipelines
- Python for machine learning, packages for scientific computing and machine learning
- Setting up your programming environment using Anaconda
- **PARTICIPATION ACTIVITY: Create conda environment and install sci-kit learn and its dependencies; Start a Jupyter notebook and change the kernel to the installed environment.**

Week 3 (September 9): Decision Trees and K-Nearest Neighbors

- Information theory core concepts
- Rule-based models

- Decision trees
- Random Forest
- K-nearest neighbors
- **PARTICIPATION ACTIVITY: Will be announced in the class**

ASSIGNMENT 1
OUT: September 10, DUE: September 22

Week 4 (September 16): Naive Bayes, Logistic Regression, and Support Vector Machine

- Core probability concepts
- Naïve Bayes
- Logistic Regression
- Testing for linear separability
- Support Vector Machine and kernel trick
- **PARTICIPATION ACTIVITY: Will be announced in the class**

Week 5 (September 23): Training Considerations

- Data preprocessing
- Training and test data
- Overfitting and underfitting
- Bias-variance tradeoff
- Model quality metrics and model comparison, complexity vs generalization
- **PARTICIPATION ACTIVITY: Will be announced in the class**

ASSIGNMENT 2
OUT: September 24, DUE: October 6

Week 5 (September 30): Model Evaluation and Optimization

- Cross-validation
- Debugging algorithms with training and validation curves
- Feature selection and regularization
- Hyperparameter optimization, grid search, randomized, successive halving
- Performance evaluation metrics: confusion matrix, precision, recall, receiver operating characteristic (ROC)
- **PARTICIPATION ACTIVITY: Will be announced in the class**

Week 6 (October 7): Ensembles

- Learning with ensembles
- Combining predictions from different classifiers
- Bagging and boosting
- Leveraging weak learners via adaptive boosting, AdaBoost
- **PARTICIPATION ACTIVITY: Will be announced in the class**

ASSIGNMENT 3
OUT: October 8, DUE: October 27

October 14: Fall break, no class

Week 7 (October 21): Unsupervised Learning and Exploratory Analysis

- Goals and reasons for exploratory analysis
- k-Means algorithm
- Hierarchical Clustering
- Principal Component Analysis
- Evaluation metrics
- **PARTICIPATION ACTIVITY: Will be announced in the class**

Week 8 (October 28): Machine Learning Lifecycle

- Machine learning lifecycle stages
- Model drift, concept drift, data drift
- Model retraining and continual learning
- Automated machine learning
- **PARTICIPATION ACTIVITY: Will be announced in the class**

ASSIGNMENT 4
OUT: October 29, DUE: November 10

Week 9 (November 4): Deep Learning Basics

- Perceptron and its limitations
- Multi-layer feed-forward neural networks
- Activation functions and loss functions
- Training a deep neural network: stochastic gradient descent and weight updates
- Regularization techniques in deep learning
- **PARTICIPATION ACTIVITY: Will be announced in the class**

Week 10 (November 11): Project Guidance

- Guidelines for final project, project scope and grading rubric
- Expectations for machine learning reproducibility
- Example projects walkthrough
- **PARTICIPATION ACTIVITY: Create a Github Page for your final project and make a first commit**

Project Proposal
OUT: November 11, DUE: November 25

Week 11 (November 18): **Bias, Explainability and Fairness**

- Definitions of Bias and Fairness
- Conducting Audits
- Discussion of current trends, topics, and issues
- Fairness evaluation metrics and tools
- Ethical decision-making case studies
- **PARTICIPATION ACTIVITY: Will be announced in the class**

ASSIGNMENT 5
OUT: November 19, DUE: December 1

Week 12 (November 23): Text as Data

- Representations of text
- Classifiers for text data
- Natural language processing methods
- Fairness concerns in text data
- **PARTICIPATION ACTIVITY: Will be announced in the class**

Week 13 (December 2): Data Collection

- Corpus building and annotation manuals
- Measuring and improving inter-rater reliability
- Fairness concerns in data collection and labeling
- Learning with limited data
- **PARTICIPATION ACTIVITY: Will be announced in the class**

Week 14 (December 9): Images as Data

- Representation of image data
- Convolutional Neural Networks (CNN) and its building blocks
- Implementing a CNN and regularizing it
- Fairness concerns in image classification
- **PARTICIPATION ACTIVITY: Train a CNN using MNIST dataset with and without dropout and report its performance**

Final Project DUE: Dec 15