Introduction to Machine Learning DATA221) Spring 2023 (DRAFT Rev. Mar 16)

Introduction to Machine Learning

Instructor:

William Trimble (wltrimbl@uchicago.edu)

Teaching Assistants:

Richard Huang <rrhuang@uchicago.edu>
Melissa Adrian <@uchicago.edu> (STAT PhD)

Location: Ryerson 251

Time: MWF 9:30- 10:50

Office Hours:

Week 1: 12:30-2:30 Monday, 2:00-3:30 Friday

12:30-2:00 Wednesday and Fridays starting week 2

Class materials (slides, notebooks, datasets if they are small):

https://github.com/wltrimbl/uchicago-data221

In-person instruction Spring 2023.

Catalog description:

This course introduces topics in current applications of machine learning for Data Science students. Topics include machine learning models, supervised and unsupervised learning, loss functions, risk, empirical risk and overfitting, regression and classification, clustering, gradient boosting, decision trees and random forests, and (time permitting) a brief introduction to Neural Networks and deep learning.

Prerequisite(s): MATH 13300 or 15300 or 16300 or DATA 21200, DATA 11900 or CMSC 12200 or CMSC 15200 or CMSC 16200.

Understanding of probability, calculus, and linear algebra will be extremely helpful.

(BUT I will let anyone who asks into the class.)

Objectives and prerequisites:

The objectives of this course are:

• To introduce students to the main mathematical and computational tools for teaching machines to make decisions.

• The class will be about half theory / discussion and about half programming in the python/numpy/matplotlib/pandas/scikit-learn environment.

• The core of the class is feeding example datasets, mostly real ones, into sklearn and tensorflow-implemented solvers. Expect to get your fingers dirty with data that aren't as described, don't make sense, or run you over your disk quota.

Textbooks:

I expect to be referring to these textbooks. You don't really need to purchase them, but they may be helpful.

- Burkov, Andriy "The hundred-page machine learning book" http://themlbook.com/ (This is available for purchase or for download, but is very succinct)
- Graeth, James, An Introduction to Statistical Learning: with Applications in R. Springer (This is a classic text covering the DATA119 material at a slightly more advanced level.)
- Hastie, Trevor. Elements of Statistical Learning. Springer. (2001)
- Goodfellow, Ian and Courville, Aaron. Deep Learning. (2016) MIT Press. https://www.deeplearningbook.org/
- MacKay, David. Information Theory, Inference, and Learning Algorithms Cambridge University Press (2003) http://www.inference.org.uk/mackay/itila/book.html (This is also on the web courtesy of the publisher. It's a great book, but tough, and only a third of it is relevant to us.)

Grading:

The final grade has the following components with the approximate weights:

• Homework (70%). Due end of the day (11:59pm) on Fridays. The lowest graded homework will be dropped with no questions asked. :)

• Class Project (group) (30%). Consider this as something you are preparing for an audience, let's say, outside of this class.

A Pass/Fail grade may be given upon written request to the instructor before the reading period. The grade of P will be awarded only for work of C- quality or better.

If you are receiving a degree this quarter, please inform the instructor (to make arrangements to get your grades in early)

The grade of W needs to be requested from and discussed with your Academic Adviser. If you are considering withdrawing from the class after putting some weeks of work into it, please contact the instructor.

Homework:

There will be weekly homework assignments usually due on Friday morning.

You may discuss homework problems with other students but you should code and write solutions independently. You should acknowledge any help in writing.

The official course policy is: no late homework will be accepted for grading. You receive a grade of zero for late assignments. The lowest homework score will be dropped.

Canvas, Discussion, Python and Jupyter Notebooks:

We will use Canvas to post the slides and notebooks from the class, as well as the homework. There is an Ed Discussion forum on Canvas.

Auditors:

I will share the course materials with anyone with a cnet ID who asks.

Preliminary Syllabus SPRING 2023

Week 1

What is Machine Learning?

The fitting paradigm; loss functions / risk / overfitting

Linear regression review

Week 2

Mathematical formulation; Supervised and Unsupervised learning

Curse of dimensionality, optimization

Naive Bayes Classifier

Week 3

Linear classifiers

Logistic regression

Week 4

KNN classifier

Decision trees

Week 5

Support vector machines

Random forests, ensemble models

Week 6

Clustering; kmeans, expectation-maximization

Gradient descent, simulated annealing, stochastic optimization

Feature engineering

Week 7

Neural networks

Evaluation & reporting, Algorithm choice

Week 8

Overfitting & regularization; autoencoders

Week 9

Intro to deep learning; language models, transformers; AI and art