Machine Learning for Social Science

CRIM4012/CRIM6012/SOCI3501/SOCI6012

Course Description:

This course provides an introduction to machine learning techniques for social science researchers. The course will cover a range of techniques including supervised and, time permitting, unsupervised learning, as well as more specialized methods such as deep learning and natural language processing. The course will also discuss some ethical considerations in the use of machine learning, as well as the role of machine learning in policy and decision-making.

The aim of the course is to be focused on applications. While the class will present the formal background on the development of the machine learning methods, the class will focus on putting the tools into practice. We will use data on a variety of topics including criminal justice data (recidivism prediction in Georgia parole system), education (predicting dropout in the National Education Longitudinal Survey), and health (drug use in the National Survey on Drug Use and Health). Students completing the course will know how to apply several of the most common machine learning tools to a variety of social science problems. The course will also discuss the role of machine learning in causal inference, using machine learning methods to estimate propensity scores.

Course Goals:

Upon completing this course, students will be able to:

1. Understand the basics of machine learning and how it can be applied to social science research
2. Choose appropriate machine learning techniques for a given research problem
3. Implement machine learning algorithms in R
4. Evaluate the performance of machine learning models and interpret their results
5. Understand the role of machine learning in policy and decision-making

Course Outline:

1. Review
    a. Probability – independence, conditional probability, Bayes Theorem, common probability distributions, expected value, variance
    b. Calculus – derivatives, integrals, and their properties
    c. Linear algebra – addition, multiplication, transpose, inverse, matrix operations in R
2. Naïve Bayes classifier
    a. Estimation
    b. Evaluation
    c. Evidence balance sheets
    d. Handling missing values
3. Nearest-neighbor classifier and evaluating model quality
    a. Loss functions

b. Bias/variance tradeoff
c. Tuning parameter
d. Out-of-sample predictive performance
e. Training set bias
f. Overfitting/underfitting
g. 10-fold cross-validation

4. Linear Models
    a. Linear regression
    b. Logistic regression
        i. Brier score decomposition and calibration
    c. Splines
    d. Basis functions
    e. Ridge regression (L_2 penalty)
    f. LASSO (L_1 penalty)
5. Decision Trees
    a. CART algorithm
    b. Bagging, a tree ensemble
6. Singular value decomposition
    a. Working with more complex data objects (e.g. images, sound)
7. Boosting
    a. Generalized boosted models
    b. Application to propensity score weighting
8. Natural Language Processing
    a. Text preprocessing
    b. Sentiment analysis
    c. Topic modeling
9. Neural Networks
    a. Handwriting recognition
10. Ethical and Privacy Considerations
    a. Fairness and transparency
11. Clustering (time permitting)
    a. K-means
    b. Hierarchical clustering
    c. Density-based clustering

Assessment:

For undergraduate CRIM4012/SOCI3501 students

- Eight assignments (8 x 10% = 80)

  These assignments will involve the testing and applying of each method to real datasets in R. Students will submit their code and be prepared to present their findings in class.

- Prediction challenge (20%)

Students will be given a new dataset not used in class. The aim is to test a variety of machine learning methods and select the one offering the best out-of-sample predictive performance. I will hold out a test set and evaluate each student's machine learning model on the held out test set.

For graduate CRIM6012/SOCI6012 students

- Eight assignments (60%)
- Prediction challenge (20%)
- Final project (20%)

Students enrolled in the graduate level version of the class will be required to complete a final project. This will involve applying a machine learning method, possibly one not discussed in class, to a new method. The final report should include a technical description of the machine learning method used.

Prerequisites:

I will review the basics of probability, calculus, and linear algebra at the start of the class. This should not be the first time encountering probability or calculus, but it will be fine even if it has been several years. Students will not be grinding through mathematics, but basic understanding of differential calculus is essential for optimizing machine learning methods and integral calculus shows up regularly when characterizing the performance of machine learning methods. I do not expect students to have had any prior exposure to linear algebra. I will give some introductory linear algebra instruction and exercises and will continue to build students' linear algebra skills throughout the semester.

Familiarity with programming in R is essential. This course should not be a student's first exposure to R. Many students will have taken CRIM4002/CRIM6002/SOCI6002 before taking this course, which is more than sufficient for the R programming we will be doing in this course. If a student has not used R, but is very comfortable in conducting data science in Python, then that student should take an intensive R programming course in advance of attempting this course. I am not planning on teaching the basics of data science (loading, merging, subsetting, transforming, and cleaning datasets).