9592: Machine Learning

11:30AM-1:30PM, Thursday (9/9/2020-12/9/2020), Online

Instructor: Dave Armstrong
Office: SSC 4142
Hours: by appointment
E-mail: dave.armstrong@uwo.ca
web: `http://www.quantoid.net/teachuwo/uwo9592`

This course is designed to get you thinking about machine learning tools and how they can be used effectively in the social sciences. This requires not only understanding the tools themselves, but thinking clearly about the suitability of the tools given the ends they are trying to reach. This includes thinking about the place of machine learning tools in inference and and prediction.

The course will be taught from a more applied, rather than mathematical, point of view. While there will certainly be some math in the course, I will try to ensure that the intuitions of the tools we discuss are conveyed in non-mathematical terms as well. As you are all graduate students, I expect that you will attend class regularly, do the readings and ask questions when something is confusing. You are ultimately responsible for knowing the material. I will do my best to teach it in a way that is likely to make sense, but if you do not understand something, you need to take responsibility for figuring it out by asking questions, either in or outside of class. If you miss class, you are responsible for learning the material you missed in a manner that proves least distracting for the other participants in the course. Late papers and assignments are not accepted (rare exceptions may be allowed on a case-by-case basis when arrangements are made before the due date).

Computing

The work in this class will all be done in R. This course is not an introduction to R and will assume that you have some familiarity with the software before the course starts. When you have work using the computer that needs to be turned in, it should be done in such a way that facilitates easy reading and evaluation. This generally means a "knittable" R Markdown document will be the product you turn in.

Grading

You final grade in the course will be an average of the grades on three short papers. This average will count for 80% of your grade. The short papers are designed to get you examining a single dataset. You get to choose the dataset that suits your own needs. The hope is that by the end of the course, you'll have something that looks like a publishable paper. There is one milestone that I need you to meet, though. Sometime in the weeks before class on 9/24, you need to have a meeting with me where we talk about your

general idea and the dataset you want to use. This is essential to being able to do the work in the course. This meeting (having it and having a dataset to show me) counts for 20% of your grade.

- Short paper 1: Assigned 10/1, Due 10/22/2020, 5PM
- Short paper 2: Assigned 10/22, Due 11/12/2020, 5PM
- Short paper 3: Assigned 11/12, Due 12/10/2020, 5PM

Textbook

- *Hands on Machine Learning with R* by Bradley Boehmke and Brandon Greenwell, which is freely available in HTML form at: `https://bradleyboehmke.github.io/HOML/`. I will refer to this book as HOML.
- *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, which is freely available as a pdf at: `http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf`. I will refer to this book as ISLR.

Outline

1. Introduction (9/10)
   Reading:
   - ISLR pp 1-9
   - HOML Preface
2. Regression and Classification (9/17)
   Reading:
   - ISLR Chapter 2
   - HOML Chapters 1, Chapter 2 sections 2.5-2.6
   - Optional: ISLR Chapter 3 (pp. 59-102) for a review of the linear model; Chapter 4 (pp. 127-138) for a review of logistic regression
3. Bootstrapping and Cross-validation (9/24)
   Reading:
   - ISLR Chapter 5
   - HOML sections 2.1-2.4

4. Feature Selection and Regularization (10/1)
   Reading:

   - ISRL Chapter 6
   - HOML Chapters 3 and 6

   Short paper 1 Assigned.

5. Non-linearity in Regression Models (10/8-15)
   Reading:

   - ISLR Chapter 7
   - HOML Chapter 7
   - Harelzak, Ruppert and Wand (HRW), Chapter 2 (pp. 15-31)
   - Stasinopoulos et al (GAMLSS), Chpaters 2 and 9 (pp. 255-293)

6. Kernel Regression and KNN, Polywog (10/22)
   Reading:

   - HOML Chapters 8-9
   - Hainmeller and Hazlett *Political Analysis* piece.
   - Kenkel and Signorino (Working Paper)

   Short paper 1 Due, Short paper 2 Assigned.

7. Tree-based Regression, Bagging and Boosting (10/29)
   Reading:

   - ISLR Chapter 8
   - HOML Chapters 10-12
   - Montgomery and Olivella *American Journal of Political Science* piece.

8. Ensemble Predictions (11/12)
   Reading:

   - HOML Chapter 15
   - Tattar *Hands On Ensemble Learning with R* Chapters 7-9.

   Short paper 2 Due, Short Paper 3 Assigned.

9. Unsupervised Learning I: PCA/SVD (11/19)
   Reading:
   - ISLR Chapter 10, sections 10.1-10.3
   - HOML Chapters 17-18
10. Unsupervised Learning II: Clustering (11/26/2020)
    Reading:
    - ISLR Chapter 10, section 10.3
    - HOML Chapters 20-22
11. Wrap-up, Presentations (12/3)
    Short Paper 3 due - 12/10/2020, 5PM.