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 depend on the following:

Homework	40%
Case Study	20%
Final Paper	40%

Homework

You will get assignments (nearly) each class of various lengths. You should consider your colleagues a resource and I encourage you to discuss the problem sets with your them. That said, each person must turn in his or her own, original answers to the homework problems.

Case Study

You will get a dataset and some questions to answer. This will be a less directed project than the homework wherein you will be asked to generate some insights in whatever ways you find most useful.

Final Paper

The class culminates in a final paper that will be written in the form of published papers you have read. You will need to do some analysis and justify the analytical process, interpret the analyses and answer the question you posed. There is no formal requirement for length, but let me suggest a couple of things. First, your literature review shouldn't be more than three or four pages. Separately from the literature review, you should present your theory - the way in which you think the conceptual pieces fit together. Part of this discussion should be a formal presentation of hypotheses. You should describe the data you're using - where you got it, what you did to it after you got it and how you think the variables you are using match the concepts they are meant to measure. You should talk about the procedure you use for testing the hypotheses along with the strengths and possible weaknesses of the procedure for this purpose. You need to discuss the results and then conclude by putting the results back in context and highlighting the most important results.

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 HOLM.
- *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
   - HOLM Preface

2. Regression and Classification (9/17)
   Reading:

   - ISLR Chapter 2
   - HOLM 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

   Homework: Problem Set 1 (due 9/23/2020, 5PM)

3. Bootstrapping and Cross-validation (9/24)
   Reading:

   - ISLR Chapter 5
   - HOLM sections 2.1-2.4

   Homework: Problem Set 2 (due 9/30/2020, 5PM)

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

   - ISRL Chapter 6
   - HOLM Chapters 3 and 6

   Homework: Problem Set 3 (due 10/7/2020, 5PM)

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

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

   Homework: Problem Set 4 (due 10/21/2020, 5PM)

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

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

   Homework: Problem Set 5 (due 10/28/2020, 5PM)

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

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

   Homework: Problem Set 6 (due 11/11/2020, 5PM)

8. Ensemble Predictions (11/12)
   Reading:

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

   Homework: Problem Set 7 (due 11/18/2020, 5PM)

9. Unsupervised Learning I: PCA/SVD (11/19)
   Reading:

   - ISLR Chapter 10, sections 10.1-10.3
   - HOLM Chapters 17-18

   Homework: Problem Set 8 (due 11/25/2020, 5PM)

10. Unsupervised Learning II: Clustering (11/26/2020)
    Reading:

    - ISLR Chapter 10, section 10.3
    - HOLM Chapters 20-22

    Problem Set 8 (due 12/2/2020, 5PM)

11. Wrap-up, Presentations (12/3)