PSC 700: Data Science and Machine Learning

Robert Shaffer

January 31, 2021

rbshaffer0@gmail.com
Office Hours: Th 2:30-4pm
Office: 512 Eggers Hall

https://rbshaffer.github.io
Lectures: MW 3:45-5:05pm
Classroom: SOM 101

1 Overview

Data science, as a field, is concerned with drawing inferences from large, unstructured datasets. In this class, we'll explore and implement the strategies and methods that have come to compose the data science workflow from an applied social science perspective. Along the way, we'll pay special attention to the *limits* of data science techniques, including statistical, practical, and ethical challenges. Data science and machine learning techniques are powerful, but limited by the data and research designs that form their inputs. We'll spend a substantial amount of time "opening the black box", and work to understand the ideas that underlie the techniques we'll cover as well as their failure modes.

The class will be organized into three main components. We'll begin by considering our goals and defining the metrics we might use to evaluate algorithms and estimators, ranging from out-of-sample prediction and cross-validation to interpretability and fairness. Next, we'll consider a range of supervised machine learning techniques, and evaluate these techniques on the criteria we've defined. Finally, we'll move to dimensionality reduction and unsupervised learning, and explore both the opportunities and the additional challenges that unsupervised techniques can offer.

2 Grading

- **Homeworks (50%)**: There will be approximately one problem set every 2-3 weeks. In problem sets, you'll do some theoretical work, but you'll primarily *explain* your data analysis and modeling choices, and *implement* your chosen steps. I believe that learning to implement an algorithm is both the best way to understand the mechanics of that algorithm and the easiest way to gain hands-on experience that will be useful in your professional endeavors.
- **Final Paper (50%)**: In addition to the problem sets, you'll write a final paper of

no more than 15 double-spaced pages, in which you'll describe and implement an original research design based on the methods covered by this course. In addition to a manuscript, you'll be required to submit a replication file (5% of your grade) and give a short presentation on your project (10% of your grade).

3 Readings

As a text for the course, we'll primarily use Elements of Statistical Learning (ESL), which is both free and a great resource for most machine learning topics. I'll also occasionally assign readings from Murphy's Predictive Machine Learning (PML), which is a great but more mathematically-oriented resource. Finally, we'll read a series of academic papers throughout the course, which will help illustrate concepts in a more applied context.

4 Software

Programming and data management are an important part of the data science workflow, and this class is no exception. The main language we'll be using in the class is `R`, which we'll use for in-class demonstrations, assignments, and exam. You can get `R` for free here:

`https://www.r-project.org/`

To write and edit `R` code, I recommend `RStudio`, which you can also get for free here:

`https://rstudio.com/products/rstudio/download/`

It will be helpful (though not essential) to have some background familiarity with `R` before taking the class. Hadley Wickham's *R for Data Science* book is a great place to look if you need a refresher. However, we'll regularly work through `R` examples in class, and there will be several dedicated `R` workshops scheduled throughout the semester to ensure that everyone is up to speed.

5 COVID Guidelines

Health guidelines permitting, I will be holding in-person lectures throughout the semester. These lectures will also be live-streamed, and I'll interact with both the in-person and virtual audiences during the lectures. Students should feel free to attend virtually or in person.

To ensure that we meet social distancing guidelines, I'll circulate a poll before the start of the semester to gauge students' plans regarding in-person lecture attendance. Our in-person social distancing capacity is 36, so if more than 36 students are interested in attending in-person, I'll create a rotation to allow everyone to attend.

6 Late Submissions

Late submissions will be penalized **five percentage points** from the grade they would have received for each day after the deadline that they are submitted. For example, a memorandum submitted two days after the deadline that would have received an 85% will instead receive a 75%.

If unforeseen circumstances arise and you need an extension on an assignment, **reach out to me**! I'm happy to work with you if personal circumstances prevent you from submitting an assignment, but I need to talk to you to figure out an arrangement.

7 Academic Integrity

See the Syracuse Academic Integrity Policy for university-wide expectations regarding academic honestly. In sum: don't plagiarise, or otherwise violate the academic integrity expectations. Please reach out to me if you have questions!

8 Schedule

1. Introduction

   - **2/8:** Class Overview

     Gelman, Andrew and Aki Vehtari. "What are the most important statistical ideas of the past 50 years?" *arXiv preprint 2012.00174*: 2021.

   - **2/10:** Model Evaluation I: Terminology and the Bias-Variance Tradeoff

     – **ESL**, Ch. 2.6, 7.1-7.6.

   - **2/15:** Model Evaluation II: Cross-Validation and Bootstrapping

     ESL, Ch. 7.10-7.11, 8.1-8.4.

   - **2/17:** Model Evaluation III: Ethics, Interpretability, and Computability

     Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models For High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5) (2019): 206-215.

     Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

   - **2/22:** `R` Workshop 1: Data Management and Programming in `R`

   - **2/24:** `R` Workshop 2: Graphical Display of Data

2. Supervised learning and model interpretation

   - **3/01:** Regularization IA: Classical Regression and Overfitting

     ESL, Ch. 3.1-3.2.

     PML, Ch. 11.1-11.2.

   - **3/03:** Regularization IB: Variable Selection and Shrinkage

     ESL, Ch. 3.3, 3.4, 3.6.

     PML, Ch. 11.3-11.5.

     Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996): 267-288.

   - **3/08:** Regularization II: Kernel Ridge Regression/KRLS

     Hainmueller, Jens, and Chad Hazlett. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach".*Political Analysis* (2014): 143-168.

Mohanty, Pete, and Robert Shaffer. "Messy Data, Robust Inference? Navigating Obstacles to Inference with bigKRLS". *Political Analysis* (2019): 127-144.

- **3/10:** Regularization III: A Bayesian Aside

  Skim **PML** Ch. 2, 7.1-7.2 for background.

  PML, Ch. 11.6. See also Tibshirani (1996), §6.

- **3/15:** Decision Trees I: Growing a Tree

  ESL, Ch. 9.2, skim 9.3-9.7.

- **3/17:** Decision Trees II: Ensembles and Random Forests

  ESL, Ch. 15.

- **3/22:** `R` Workshop 3: Opening the Black Box

- **3/24:** Neural Networks IA: The Basics

  ESL Ch. 11.3.

  Skim **PML**, Ch. 13.

- **3/29:** Neural Networks IB: Estimation and Model Fitting

  ESL Ch. 11.4-11.5.

- **3/31:** Neural Networks II: Data and Model Structures

  – Skim **PML** Ch. 14-15.

3. Unsupervised learning and applications

- **4/05:** Dimensionality Reduction I: Clustering

  ESL, Ch. 14.1, 14.3.

- **4/07:** Dimensionality Reduction II: PCA, Factor Analysis, and PCA Regression

  ESL Ch. 3.5, 14.5.1, 14.5.2, 14.7, 14.8

- **4/12:** Introduction to Text-as-Data

  Grimmer, Justin and Brandon Stewart. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents". *Political Analysis* 21, no. 3 (2013): 267-297.

  Denny, Matthew J., and Arthur Spirling. "Text Preprocessing for Unsupervised Learning: Why it Matters, When it Misleads, and What to Do About It". *Political Analysis* 26, no. 2 (2018): 168-189.

- **4/14:** Text-as-Data IA: Topic Models

Blei, David. "Probabilistic Topic Models". Communications of the ACM 55.4 (2012): 77-84.

Roberts, Margaret E. et al. "Topic Models for Open-Ended Survey Responses with Applications to Experiments". *American Journal of Political Science* 58.4 (2014): 1064-1082.

- **4/19:** Text-as Data IB: Word Embeddings

  Mikolov, Tomas, et al. "Distributed Representations of Words and Phrases and their Compositionality." *arXiv preprint arXiv:1310.4546* (2013).

  Wang, Bin, et al. "Evaluating word embedding models: methods and experimental results." *APSIPA Transactions on Signal and Information Processing* 8 (2019).

- **4/21:** Wellness Day **(no class)**

- **4/26:** Text-as-Data IC: Text Classification

  Gurciullo, Stefano, and Slava J. Mikhaylov. "Detecting Policy Preferences and Dynamics in the UN General Debate with Neural Word Embeddings." *International Conference on the Frontiers and Advances in Data Science (FADS)*: IEEE, 2017.

  Rice, Douglas, et al. "Machine Coding of Policy Texts with the Institutional Grammar". *Public Administration.* Forthcoming, 2021.

  Anastasopoulos, Jason, and Anthony M. Bertelli. "Understanding Delegation Through Machine Learning: A Method and Application to the European Union". Forthcoming, *American Political Science Review* (2021).

- **4/28:** `R` Workshop 5: Text Classification

- **5/03:** Text-as-Data IIA: Transfer Learning and Cautionary Notes

  Rodriguez, Pedro L. and Arthur Spirling. "Word Embeddings: What Works, What Doesn't, and How to Tell for Applied Research". Forthcoming, *Journal of Politics* (2021).

  Garg, Nikhil et al. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes". *Proceedings of the National Academy of Sciences* 115.16 (2018): 3635-3644.

  Bender, Emily M. et al. "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?" *Proceedings of FAccT* (2021).

4. **5/05, 5/10, 5/12:** Presentations and Final Paper Office Hours