HPS/Pl 102: Seminar in Ethics, Statistics, and Law
California Institute of Technology
Spring 2018

Instructor: Boris Babic
Email: bbabic@caltech.edu
Office: **B118 Dabney**

Time: T/Th 10-11:30am
Location: Deck 142
Office Hours: Th 1-2:30pm

NOTE: This syllabus has been modified from the actual course to reflect some things I learned about how best to teach such interdisciplinary material. For example, there is now almost a full week devoted to setting up students' computers and the introductory sections have been expanded throughout. Still, the material presumes quite a lot of formal background which was not an issue at Caltech, but I would further modify this class for humanities/ philosophy majors and teach it as an applied introduction to statistical thinking rather than a research seminar.

Course Description

This course will be an advanced seminar focusing on recent research in fairness and other ethical notions arising in statistics and machine learning. The goal of the seminar will be to have you produce a final project at the level of something that could be submitted to leading conferences in the field (e.g., ICML, NIPS, UAI, etc.). In particular, there are now a number of statistics and computer science conferences devoted exclusively to ethics in AI. For example, FATML (which is co-located with ICML) and ACM-FAT.

The first few weeks will be introductory. While I will not presuppose any background in philosophy, you will need some background in statistics, machine learning, and relevant computational tools. The material below mostly uses R, but if you are more comfortable with Python or some other software feel free to use that instead.

After the introductory material, we will investigate ethical issues arising in classification, learning, and data collection, focusing in particular on case studies in recidivism prediction. We will then look at racial profiling and discrimination testing. Next, we will transition to bias in advertising, captioning, and word embeddings. Finally, we will finish up by examining several prediction algorithms for US Supreme Court cases.

The aim of the course is to help you learn to carefully and fruitfully apply statistics and machine learning to large-scale public policy problems.

Materials

I will post all course readings to Moodle. No textbook is required. Should you like to consult a textbook, I recommend James, Witten, Hastie, and Tibshirani, *An Introduction to Statistical Learning*; Mcelreath, *Statistical Rethinking*; and Kruschke, *Doing Bayesian Data Analysis*. You may also find a number of R reference books helpful, such as Grolemund and Wickham, R for Data Science; Teetor, The R Cookbook; and Chang, The R Graphics Cookbook.

Assignments

Research project (minimum 4000 words) (60%), outline/draft (20%), participation (20%).

Deadlines: Outline, Tuesday May 8 in class. Final Draft, Thursday May 31 (for graduating seniors), Thursday June 7 (everyone else).

Schedule

1 Background

04/03 **Introduction: Ethics in AI + Setting up Your Computer**
Grolemund and Wickham (2017): R for Data Science (excerpts)
R Documentation
JupyterLab Documentation

04/05 **Setting up Your Computer**
R, https://cran.r-project.org/
Anaconda, https://conda.io/docs/user-guide/index.html
Rkernel for Jupyter Notebooks, https://irkernel.github.io/installation/

04/10 **Data Analysis and Visualization**
Wickham (2014), Tidy Data (excerpts)
Wickham (2016), ggplot2 (excerpts)
Tidyverse Documentation
Chang (2012): The R Graphics Cookbook (as needed)

04/12 **Linear Models**
Faraway (2004) Linear Models with R (excerpts)
Faraway (2006) Extending the Linear Model with R, chs 1-2
Klein (2003): Survival Analysis (Ch 8.1, Cox Proportional Hazards Model)

04/19 **Classification: Introduction**
James (2013): Chapter 4 (excerpts)
scikit-learn, scikit-learn.org/ (if using Python)

04/24 **Classification: Ensemble Methods**
James (2013): Chapter 8 (excerpts)

04/26 **Classification Practice**
James (2013): Exercise 4.7.10 (a-d)
James (2013): Exercise 4.7.11 (a,b,c,f)
James (2013): Exercise 8.4.14 (random forests only)

05/01 **Bayesian Data Analysis**
Mcelreath (2016): Statistical Rethinking (excerpts)
Kruschke (2015): Doing Bayesian Data Analysis (excerpts)

05/03 **Practice Bayesian Data Analysis**
Kruschke (2015): Ch. 8-9 Exercises
RStan quick start guide, http://mc-stan.org/users/interfaces/rstan
rstanarm, https://cran.r-project.org/web/packages/rstanarm/index.html
PyStan (if using python), http://pystan.readthedocs.io/en/latest/

2 Ethics in Classification

05/08 **Recidivism**
Pro Publica: Machine Bias
Pro Publica: How We Analyzed the COMPAS Recidivism Algorithm
github.com/propublica/compas-analysis (original analysis)
github.com/anniejw6/compas-analysis (Annie J. Wang reassessment)

05/10 **The (Im)Possibility of a Fair Classifier**
Kleinberg (2016): Inherent Trade-Offs in the Fair Determination of Risk Scores

Corbett-Davies et al (2017): Algorithmic Decision Making and the Cost of Fairness

05/15 **Impact vs. Treatment Disparity**
Lipton et al (2017): Does Mitigating ML's Impact Disp. require Treatment Disp.?
United States v. Carolene Products Company, 304 U.S. 144 (1938)
Ricci v. DeStefano, 557 U.S. 557 (2009) (excerpts)

05/17 **Fairness Beyond Recidivism**
Hardt (2016): Equality of Opportunity in Supervised Learning
Dwork (2011): Fairness Through Awareness
Anderson (2007): Fair Opportunity in Education (excerpts)

05/22 **More on Fairness**
Locke: Second Treatise of Government (excerpts)
Rawls (2001): Justice as Fairness (excerpts)

05/24 **Still More on Fairness**
Anderson (2003: Value in Ethics and Economics (excerpts)
Anderson (1999): What is the Point of Equality (excerpts)

3 Fairness in Learning

05/24 **Bandits and Adversaries**
Joseph et al (2016): Fairness in Learning, Classic and Contextual Bandits
Wadsworth (2018): Achieving Fairness through Adversarial Learning

4 Discrimination and Racial Profiling

05/29 **When is Discrimination Wrongful?**
Becker (1957): The Economics of Discrimination
Arrow (1973): The Theory of Discrimination (excerpts)
Anderson (2003): Value in Ethics and Economics (excerpts)
Washington v. Davis, 426 U.S. 229 (1976) (excerpts)
Griggs v Duke Power Co., 401 US 424 (1971)

05/31 **Testing for Discrimination**
Knowles et al (2001): Racial Bias in Motor Vehicle Searches: Theory and Evidence
Simoiu et al (2017): The Problem of Infra-Marginality
Pierson et al (2017): Fast Threshold Tests for Detecting Discrimination
Pierson et al (2017): A Large-scale Analysis of Racial Disparities in Police Stops

5 Bias

06/05 **Captioning and Word Embedding**
Bolukbasi et al (2016): Debiasing Word Embeddings
Hendricks et al (2018): Women Also Snowboard, Overcoming Bias in Caption Models
Zhao et al (2017): Reducing Gender Bias Amplification with Corpus-Level Constraints

6 Prediction

06/07 **Predicting US Supreme Court Decisions**
Martin et al (2004): Competing Approaches to Predicting S.Ct. Decision Making
Katz et al (2017): A General Approach for Predicting the Behavior of the S.Ct.

Attendance and reading

Engaged participation is an important component of this class and I expect everyone to contribute meaningfully to class discussion. This does not mean I will reward those who speak most. And it does not mean you cannot do well on the participation component if you're less comfortable speaking up. Learning to articulate your thoughts in a professional, courteous and persuasive manner is an invaluable skill and a goal of this course is to improve your ability to do this.

While the readings are not long, they can be very difficult. As a result, you should plan to spend a fairly significant amount of time reading and re-reading the material.

Submitting assignments and Late policy

All papers must be submitted in hard-copy in class on the day they are due.

If you anticipate needing more time on an assignment, you should contact me in advance. Otherwise, late assignments will be penalized by one-third of a letter grade for each day they are late.

Students with disabilities

If you think you may need accommodation for a disability, please let me know as early as possible.

Plagiarism

Written work submitted for a grade in this course must be your own. You are responsible for making sure that none of your work is plagiarized. You should cite the sources you rely on, and err on the side of caution where necessary. Feel free to consult me if you are not sure of the appropriate format for quotations or references.

More information on plagiarism is available on the Hixon Writing Center's website:
www.writing.caltech.edu/students/plagiarism.