Machine Learning for Political Science

Danilo Freire

April 2021

E-mail: danilofreire@gmail.com
Web: danilofreire.github.io
Office Hours: TBA
Class Hours: TBA
Office: TBA
Classroom: TBA

Course Description

This course provides an overview of the recent developments in the machine learning literature and their applications in political science. First, students will learn how to summarise and visualise data, create reproducible documents, and use version control software. Then we will consider OLS for regression and generalised linear models for classification problems. Next, the course will discuss multiple imputation algorithms, feature engineering, and feature selection. This will be followed by classes on supervised learning covering support vector machines, decision trees, random forests, and gradient boosting. Unsupervised methods such as PCA, clustering, and manifold learning will be considered next. The following section of the course will cover deep learning and neural networks. Lastly, students will be introduced to causal discovery algorithms.

Course Information

The course prerequisites are basic probability and statistics, high-school linear algebra and introductory calculus. Some familiarity with computer programming also helps. The course will use R, an open source statistical language. R is the *de facto* standard language for quantitative analysis and is widely used by academics and firms like Google, Facebook and Amazon to gain insights from data. R has about 14,000 packages that extend its core functionalities and it is free to download, use or modify. Compared to Stata or SPSS, R has a steeper learning curve, but its power and flexibility greatly overweight the costs.

It is very important that you read the assigned readings and do the problem sets before class. It is also necessary that you bring your laptop to every session.

All information about the course will be available at http://danilofreire.github.io. The syllabus will be updated periodically according to the progress of the class. Please remember to visit the website regularly.

Office Hours

I am very flexible when it comes to office hours, but it is easier to contact me via email. Feel free to send me a message any time at danilofreire@gmail.com. I will reply in a few hours. You can also meet me in the afternoon at my office. If possible, please send me an email before coming to my office just to make sure two students will not book the same time slot.

Community Standards

I am committed to full inclusion of all students. Please inform me early in the term if you have a disability or other conditions that might require accommodations or modification of any of these course procedures. You may speak with me after class or during office hours. Students in need of short-term academic advice or support can contact one of the deans in the Dean of the College office.

English Language Learners

The university welcomes students from around the country and the world, and the unique perspectives international and multilingual students bring enrich the campus community. To empower multilingual learners, an array of support is available including language and culture workshops and individual appointments. No student will be penalised for their command of the English language.

Academic Integrity

Students will write five homework assignments and an essay for this course. All writing should be your own work, and I take plagiarism very seriously. I am happy to provide any help you may require with your lessons as long as you are committed to the course. It is also important to cite other people's work whenever necessary, and if in doubt, mention your sources.

Special Needs

If you have any special needs, please contact me. I'm happy to make necessary arrangements so you can follow this course.

Requirements and Grading

Participation: 10%. Students should be active participants in the course. Feel free to ask any question you may have, help others if you know how, and make suggestions or comments you believe are interesting. I hope we create a friendly, open environment for learning and students are the most important part of it.

Five Homework Assignments: 50%. Students will have five homework assignments during the course, each of them covering a section of the syllabus. They will be based on writing up the results of performing the commands learned during the lectures.You will have to write them using `Rmarkdown` and send me a pdf file via email.

Final Project: 40%. In the final project, students will have the opportunity to apply the methods introduced in the course to address a problem relevant to political science or public policy. The goal of this exercise is to demonstrate that you have the ability to conduct research in computational social science. This research paper can be an individual or group project (up to 3 people). Students have to submit a research idea by the third week of the course, which will then be reviewed by other 2 colleagues. Then students will submit a five-page summary of their research after three weeks. After receiving further feedback from the instructor, students will write a first draft of their paper and present it to class. Lastly, students will hand-in a final project on the last day of the course. The papers should be about 4,000 words in length including tables and references but not including code.

Materials

These books will help you further your understanding of the material:

- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling, New York: Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning, New York: Springer.
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning, New York: Springer.
- Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective, Cambridge: MIT Press.

Schedule

Week 1: Introduction and Course Overview

There are no required readings assigned for this class. Students will learn how to install R and use GitHub for version control.

Recommended Readings

- Grimmer, J. (2015). We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together, PS: Political Science & Politics, 48(1), 80-83.
- Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives, 31(2), 87-106.
- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good Enough Practices in Scientific Computing, PLoS Computational Biology, 13(6), e1005510.
- Jones, Z. M. (2013). Git/GitHub, Transparency, and Legitimacy in Quantitative Research, The Political Methodologist, 21(1), 6-7.
- Rainey, C. (2019). How I Use Git and GitHub for Political Science Research.

Week 2: Importing and Visualising Data

Required Readings

- Grolemund, G. (2019). The Tidyverse Cookbook. Chapters 1-4.

- Wickham, H., & Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, New York: O'Reilly Media, Inc. Chapters 1-3.
- Soltoff, B. (2019). The Grammar of Graphics.

Recommended Readings

- Lantz, B. (2015). Machine Learning with R. Birmingham: Packt Publishing Ltd. Chapters 1 and 2.
- Healy, K. (2018). Data Visualization: A Practical Introduction, Princeton University Press.
- Tufte, E. R. (2001). The Visual Display of Quantitative Information, Cheshire, CT: Graphics press.

Week 3: Introduction to Supervised Learning: Linear and Non-Linear Models for Regression

Required Readings

- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling, New York: Springer. Read 3.1-3.2, 3.5-3.8, and chapter 6.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning, New York: Springer. Chapter 3 and 7.

Recommended Readings

- Bishop, C. M. (2006). Pattern Recognition and Machine Learning, New York: Springer. Chapter 3.
- Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective, Cambridge: MIT Press. Chapters 7 and 13.

Week 4: Linear Models for Classification, Preprocessing, and Feature Engineering

Required Readings

- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling, New York: Springer. Chapters 11 and 12.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning, New York: Springer. Chapter 4.

Recommended Readings

- Bishop, C. M. (2006). Pattern Recognition and Machine Learning, New York: Springer. Chapter 4.

Week 5: Imputation and Feature Selection

Required Readings

- Blackwell, M., Honaker, J., & King, G. (2017). A Unified Approach to Measurement Error and Missing Data: Overview and Applications, Sociological Methods & Research, 46(3), 303-341.
- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling, New York: Springer. Chapters 3.4, 18, and 19.

Recommended Readings

- Pepinsky, T. B. (2018). A Note on Listwise Deletion Versus Multiple Imputation, Political Analysis, 26(4), 480-488.
- Stekhoven, D. J., & Bühlmann, P. (2011). MissForest—Non-Parametric Missing Value Imputation for Mixed-type Data, Bioinformatics, 28(1), 112-118.
- Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection: A Data Perspective, ACM Computing Surveys (CSUR), 50(6), 94.

Week 6: Support Vector Machines

Required Readings

- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling, New York: Springer. Chapter 7.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning, New York: Springer. Chapter 9.

Recommended Readings

- Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2), 121-167.
- Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support Vector Machines in R. Journal of Statistical Software, 15(i09).

Week 7: Decision Trees, Random Forests, and Gradient Boosting

Required Readings

- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling, New York: Springer. Chapters 8 and 14.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning, New York: Springer. Chapter 8.
- Kaufman, A. R., Kraft, P., & Sen, M. (2019). Improving Supreme Court Forecasting Using Boosted Decision Trees, Political Analysis, 27(3), 381-387.

Recommended Readings

- Jones, Z., & Linder, F. (2015). Exploratory Data Analysis Using Random Forests, In: Prepared for the 73rd annual MPSA conference.

- Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests, Journal of the American Statistical Association, 113(523), 1228-1242.

Week 8: Discussion of Final Projects

No readings assigned for this session.

Week 9: Model Evaluation and Imbalanced Datasets

Required Readings

- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling, New York: Springer. Chapter 4.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning, New York: Springer. Chapter 5.
- Neunhoeffer, M., & Sternberg, S. (2019). How Cross-Validation Can Go Wrong and What To Do about It, Political Analysis, 27(1), 101-106.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique, Journal of Artificial Intelligence Research, 16, 321-357.

Recommended Readings

- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from Class-Imbalanced Data: Review of Methods and Applications, Expert Systems with Applications, 73, 220-239.
- Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data, Political Analysis, 24(1), 87-103.
- Wang, Y. (2019). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data: A Comment, Political Analysis, 27(1), 107-110.

Week 10: Dimensionality Reduction using PCA, Clustering, and Manifold Learning

Required Readings

- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling, New York: Springer. Pages 35-40.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning, New York: Springer. Chapter 10.

Recommended Readings

- Peng, R. (2016). Exploratory Data Analysis with R. Chapters 11-13.
- Gatto, L. (2019). An Introduction to Machine Learning with R. Chapter 4.

Week 11: Working with Text as Data

Required Readings

- Silge, J., & Robinson, D. (2017). Text Mining with R: A Tidy Approach, New York: O'Reilly Media, Inc. Chapters 1 to 7.
- Blei, D. (2012). Probabilistic Topic Models, In Proceedings of the 17th ACM SIGKDD International Conference Tutorials. ACM.

Recommended Readings

- Barberá, P., Boydstun, A., Linn, S., Nagler, J. & McNahon, R. (2019). Automated Text Classification of News Articles: A Practical Guide, Working paper.
- Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data, American Political Science Review, 110(2), 278-295.
- Robinson, D. (2016). Text Analysis of Trump's Tweets Confirms He Writes Only the (Angrier) Android Half.
- Robinson, D. (2016). Trump's Android and iPhone Tweets, One Year Later.

Week 12: Neural Networks, Convolutional Neural Networks for Image Classification

Required Readings

- Chollet, F., & Allaire, J. J. (2018). Deep Learning with R, New York: O'Reilly Media, Inc. Chapters 1-3, and 5.
- RStudio Team (2019). R Interface to Keras.

Week 13: More on Neural Networks

Required Readings

- Johnson, J. & Karpathy, A. (2019). Stanford CNN Course Notes, Module 2.
- Jin, H., Song, Q., & Hu, X. (2019). Auto-Keras: An Efficient Neural Architecture Search System. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, p.1946-1956. ACM.
- R TensorFlow Team (2019). Auto-Keras.

Week 14: Introduction to Causal Discovery Algorithms

Required Readings

- Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models, Frontiers in Genetics, 10.
- Freire, D. (2019). Uncovering Causal Relations in Observational Data Using Machine Learning and Graphical Models. Working paper.
- Kalisch, M., Mächler, M., Colombo, D., Hauser, A., Maathuis, M. H., & Bühlmann, P. (2014). More Causal Inference with Graphical Models in R Package pcalg.

Week 15: Final Project Presentations