Machine Learning for Humanities and Social Science

Spring 2023
Tue Thu 10:30-11:45AM
N4, School of DHCSS at KAIST

Instructor: Taegyoon Kim, Ph.D. in Political Science and Social Data Analytics

- Email: taegyoon.research@gmail.com
- Office hours: Mon Wed 11:45–12:30PM & Fri 9:00–11:00AM & By appointment
- Personal webpage: https://taegyoon-kim.github.io
- Course webpage: https://github.com/taegyoon-kim/machine_learning_dhcss

Course Overview: Research in humanities and social science is now often conducted using data that is larger and more complex than the data for which conventional statistical approaches were designed. Examples of such data include information on large-scale individual-level consumer behavior, streams of social media content, and historical archives of government documents. On one hand, the data contains rich information to make inferences about unseen data, providing abundant opportunities for prediction/forecasting. On the other hand, the data is often so complex and high-dimensional that it is difficult to specify a theory-driven model using conventional statistical approaches. Machine learning is well-suited to deal with theses challenges as well as to make best of the opportunities, as it is capable of learning model structure, selecting variables, and producing accurate predictions. The three broad objectives in this course are that students will develop:

- An in-depth understanding of machine learning concepts and algorithms that have proven most useful in the study of humanities and social science.
- Command of software tools for machine learning (R & Python).
- Awareness regarding research objectives that are best-suited to investigation with machine learning & practical experience in conducting research using machine learning.

Prerequisites: Students in this course should have background in basic descriptive and inferential statistics. This includes an understanding of descriptive statistics, hypothesis testing, regression analysis, and some experience with a software.

Readings: The main reference used in the course is James, Witten, Hastie and Tibshirani (2021), which is available for free (link). This book will be mainly used to familiarize yourself with the mathematical foundation of machine learning approaches introduced in the course. In addition, one or two research articles applying machine learning are assigned each week to help students develop literacy of machine learning in the context of applied research in digital humanities and computational social science. Students are expected to read articles before class and be prepared to actively engage in discussion.

Software: Programming in the course will be conducted in both R and Python. All in-class scripts will be provided in both languages. Students can use either R or Python for *Replication Assignment*, *Methods Tutorial*, and *Research Paper* (see below). Students who are already comfortable with R are encouraged to learn Python as many cutting-edge machine learning packages are based on Python.

Major Tasks: Students are expected to complete the following tasks.

- *Replication Assignment:* There will be at least one assignment covering each of the top-level topics listed in the course schedule. Worth 30% of the final grade.
- *Method Tutorial:* Each student will be responsible for presenting a detailed tutorial of one of the methods covered in the course. Each week, one student will introduce the tutorial to the class to demonstrate their mastery of the method and help others practically use the method. Worth 20% of the final grade.
- *Application Review:* Each student will be responsible for writing a review of (1–2 pages), and leading in-class discussion for, one of the application papers. Worth 10% of the final grade.
- *Research Paper:* Students are required to complete an original research paper and present the proposal (Week 9) and paper (Week 16). Students are highly encourage to orient their effort on Research Paper toward developing and/or completing their master thesis. The research paper and presentation are worth 40% of the final grade.

Grading Scale: Grade values will not be rounded. That is, any grade value that is greater than or equal to 'Lower' and less than 'Upper' will receive the respective grade.

Grade	Lower	Upper
A	92	101
A-	90	92
B+	88	90
B	82	88
B-	80	82
C+	78	80
C	72	78
C-	70	72
D+	68	70
D	62	68
D-	60	62
F	0	60

Course Schedule:

1. 2/28 & 3/1, Introduction to Machine Learning
   - Course introduction
   - Application: Rheault, Rayment and Musulan (2019)
2. 3/7 & 3/9, Explanation vs. Prediction
   - Shmueli (2010)
   - Application: Toft and Zhukov (2012); Cranmer and Desmarais (2017)
3. 3/14 & 3/16, Logistic Regression & Naive Bayes
   - James et al. (2021) Ch. 4.1–4.5
   - Application: Chenoweth and Ulfelder (2017); Rossini, Stromer-Galley and Zhang (2021)
4. 3/21 & 3/23, Support Vector Machine
   - James et al. (2021) Ch. 9.1–5
   - Application: Pan (2019)
5. 3/28 & 3/30, Tree-based Models

- James et al. (2021) Ch. 8.1–8.2
- Application: Streeter (2019); Gohdes (2020)

6. 4/4 & 4/6, Neural Networks
   - James et al. (2021) Ch. 10.1–10.8
   - Application: Lagazio and Russett (2004)
7. 4/11 & 4/13, Model Comparison and Selection
   - James et al. (2021) Ch. 5.1–2
   - Application: Harden and Desmarais (2011)
8. 4/18 & 4/20, Linear Regression
   - James et al. (2021) Ch. 3.1–3.3
   - Application: Golder, Golder and Siegel (2012); Shorrocks and Grasso (2020)
9. 4/25 & 4/27, Proposal Presentation
10. 5/2 & 5/4, Regularization
    - James et al. (2021) Ch. 6.1–6.3
    - Application: Wilf (2016)
11. 5/9 & 5/11, Principal Component Analysis
    - James et al. (2021) Ch. 12.1–12.2
    - Application: Peters (2015); Michaud, Carlisle and Smith (2009)
12. 5/16 & 5/18, Clustering
    - James et al. (2021) Ch. 12.4
    - Application: Harris (2015)
13. 5/23 & 5/25, Text Embedding
    - Rodriguez and Spirling (2022)
    - Application: Rodman (2020)

14. 5/30 & 6/1, Topic Models
    - Blei, Ng and Jordan (2003)
    - Application: Saraceno (2020); Rothschild, Howat, Shafranek and Busby (2019)
15. 6/6 & 6/8, Memorial Day & Individual Session
16. 6/13 & 6/15, Final Presentation

Instruction Mode: The instruction mode is in-person. However, depending on the public health challenges caused by the COVID-19 pandemic, some classes might be offered remotely. Any change to the mode of instruction will be announced in advance.

Attendance: Regular attendance is critical for building on the skills and knowledge developed throughout the class. Students who participate more actively have a more complete understanding of the material presented and are more likely to succeed in the class. Given the COVID-19 pandemic, however, students will not be penalized for absences although they will be held responsible for making up lecture materials and in-class assignments they miss.

Extended Absence: During your enrollment, unforeseen challenges may arise. If you ever need to miss an extended amount of class in such a circumstance, please notify your instructor so you can determine the best course of action to make up missed work.

Late Submission Policy: A penalty of 20% will accrue for each (rounded up) day that an assignment is late.

Syllabus Change Policy: This syllabus is a guide and every attempt will be made to provide an accurate overview of the course. However, circumstances and events may make it necessary for the instructor to modify the syllabus during the semester and may depend, in part, on the progress, needs, and experiences of the students.

References

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3(Jan):993–1022.

Chenoweth, Erica and Jay Ulfelder. 2017. "Can structural conditions explain the onset of nonviolent uprisings?" *Journal of Conflict Resolution* 61(2):298–324.

Cranmer, Skyler J and Bruce A Desmarais. 2017. "What can we learn from predictive modeling?" *Political Analysis* 25(2):145–166.

Gohdes, Anita R. 2020. "Repression technology: Internet accessibility and state violence." *American Journal of Political Science* 64(3):488–503.

Golder, Matt, Sona N Golder and David A Siegel. 2012. "Modeling the institutional foundation of parliamentary government formation." *The Journal of Politics* 74(2):427–445.

Harden, Jeffrey J and Bruce A Desmarais. 2011. "Linear models with outliers: Choosing between conditional-mean and conditional-median methods." *State Politics & Policy Quarterly* 11(4):371–389.

Harris, Rebecca C. 2015. "State responses to biotechnology: Legislative action and policymaking in the US, 1990–2010." *Politics and the Life Sciences* 34(1):1–27.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2021. *An introduction to statistical learning*. Springer.

Lagazio, Monica and Bruce Russett. 2004. "A neural network analysis of militarized disputes, 1885-1992: Temporal stability and causal complexity." *The Scourge of War: New Extensions on an Old Problem. Ann Arbor, MI: University of Michigan Press (28–60)* .

Michaud, Kristy EH, Juliet E Carlisle and Eric RAN Smith. 2009. "The relationship between cultural values and political ideology, and the role of political knowledge." *Political Psychology* 30(1):27–42.

Pan, Jennifer. 2019. "How Chinese officials use the Internet to construct their public image." *Political Science Research and Methods* 7(2):197–213.

Peters, Margaret E. 2015. "Open trade, closed borders immigration in the era of globalization." *World Politics* 67(1):114–154.

Rheault, Ludovic, Erica Rayment and Andreea Musulan. 2019. "Politicians in the line of fire: Incivility and the treatment of women on social media." *Research & Politics* 6(1):2053168018816228.

Rodman, Emma. 2020. "A timely intervention: Tracking the changing meanings of political concepts with word vectors." *Political Analysis* 28(1):87–111.

Rodriguez, Pedro L and Arthur Spirling. 2022. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research." *The Journal of Politics* 84(1):101–115.

Rossini, Patrícia, Jennifer Stromer-Galley and Feifei Zhang. 2021. "Exploring the relationship between campaign discourse on facebook and the public's comments: A case study of incivility during the 2016 US presidential election." *Political Studies* 69(1):89–107.

Rothschild, Jacob E, Adam J Howat, Richard M Shafranek and Ethan C Busby. 2019. "Pigeonholing partisans: Stereotypes of party supporters and partisan polarization." *Political Behavior* 41(2):423–443.

Saraceno, Joseph. 2020. "Disparities in a flagship political science journal? Analyzing publication patterns in the journal of politics, 1939–2019." *The Journal of Politics* 82(4):e45–e55.

Shmueli, Galit. 2010. "To explain or to predict?" *Statistical science* 25(3):289–310.

Shorrocks, Rosalind and Maria T Grasso. 2020. "The attitudinal gender gap across generations: support for redistribution and government spending in contexts of high and low welfare provision." *European Political Science Review* 12(3):289–306.

Streeter, Shea. 2019. "Lethal force in black and white: Assessing racial disparities in the circumstances of police killings." *The Journal of Politics* 81(3):1124–1132.

Toft, Monica Duffy and Yuri M Zhukov. 2012. "Denial and punishment in the North Caucasus: Evaluating the effectiveness of coercive counter-insurgency." *Journal of Peace Research* 49(6):785–800.

Wilf, Meredith. 2016. "Credibility and Distributional effects of International Banking regulations: evidence from us Bank stock returns." *International Organization* 70(4):763–796.