STA 70400 - Quantitative Analysis for Business Decisions: Machine Learning

Fall 2021

Instructor	Prof. Rahnama Rad
Lecture Days & Time	Mo 12-2pm
Room	Online
Email	kamiar.rahnamarad@baruch.cuny.edu
Office Hours	Mo 2PM - 245PM, and by appointment

Course Description

This course applies multiple regression techniques, including linear and logistic model fitting, inference, and diagnostics. Dimensionality reduction with applications to text mining will also be discussed.

Methods with special applicability for datasets with large number of features will be emphasized. Examples includes but are not limited to, forward-backward selection, lasso and ridge regularization. Issues of model complexity, the bias-variance tradeoff, and model validation will be studied in the context of large data sets. Methods that rely less on distributional assumptions are also introduced, including cross-validation, and nonparametric methods. Students will also get introduced to text classification and neural networks. Ethical, historical and industrial aspects of data science, and domain specific applications of machine learning in fields such as accounting, economics, finance, information systems, management, marketing, media, and sociology are also included.

Students will learn the intuition, assumptions, and trade-offs behind the methodologies with a focus towards real-world problems. All programming work will be carried out in R but students are encouraged to also use Python. After completing this course students will be able to:

- identify, describe, and explain basic theoretical concepts such as model complexity, the bias-variance tradeoff, and model validation.
- write codes in R to apply multiple regression and classification learning techniques.
- research and prepare a project, using public data sources, to showcase their data mining skills, and explain why they are doing it.
- comment critically on the ethical, historical and industrial aspects of data science.
- classify text and measure the performance.

Grading

- Homework assignments 40%: there are three homework assignments.
- Paper presentation 25%: every student presents a unique paper from the list below (or other domain specific papers using machine learning).
- Project 25%: every student applies a machine learning methodology to a unique dataset, presents the results to the class.
- Participation: 10%: during the course of the semester, every student is expected to have asked at least one “meaningful” question about the presented papers.

Office Hours

Synchronously 2-245pm Mondays, or by appointment, via one or more of the following:

- Video/audio live conferencing (e.g., Zoom or Webex)
- Live participation in a chat platform (e.g., Slack, Discord, Microsoft Teams, discussion board, email)

Late work

Late homework will lead to a 50% deduction.

Communication

We will have weekly synchronous and asynchronous lectures. Office hours are live online but you can also reach me daily via email. You will receive weekly updates on blackboard.

Textbooks

Chapters from the following books are part of the required reading. Note that all the books below except for the DLR book are available for free in the public domain.

- James, G., Witten, D. Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning Springer (ISLR)*, 2nd Edition.
- Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction (ESL)*.
- Boyd, S. and Vandenberghe, L. *Introduction to Applied Linear Algebra (IALA)*
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning (DL)*
- Julia Silge and David Robinson, *Text Mining with R*.

Papers

1. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web search. Proceedings of the National academy of sciences, 107(41), 17486-17490.
2. Bakshy, Eytan, et al. ”Everyone’s an influencer: quantifying influence on twitter.” Proceedings of the fourth ACM international conference on Web search and data mining. 2011.
3. Goel, Sharad, Justin M. Rao, and Ravi Shroff. ”Precinct or prejudice? Understanding racial disparities in New York City’s stop-and-frisk policy.” The Annals of Applied Statistics 10.1 (2016): 365-394.
4. Shmueli, G., *To explain or to predict?*, Statistical science (2010): 289-310.
5. Breiman, L. *Statistical modeling: The two cultures (with comments and a rejoinder by the author)*, Statistical science 16.3 (2001): 199-231.
6. Li, F., *The information content of forward-looking statements in corporate filings - A naive Bayesian machine learning approach*, Journal of Accounting Research (2010)
7. Loughran, T. and McDonald, B., *Textual analysis in accounting and finance: A survey*, Journal of Accounting Research (2016)
8. Gu, Shihao, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. The Review of Financial Studies 33.5 (2020): 2223-2273.
9. Bybee, Leland, et al. The structure of economic news. No. w26648. National Bureau of Economic Research, 2020.
10. Gentzkow, Kelly, and Taddy. ”Text as data.” Journal of Economic Literature 57.3 (2019): 535-74.
11. Ke, Kelly, and Xiu. Predicting returns with text data. No. w26186. National Bureau of Economic Research, 2019.
12. Kozak, Nagel, and Santosh. ”Shrinking the cross-section.” Journal of Financial Economics 135.2 (2020): 271-292.
13. Kozak, Nagel, and Santosh. ”Interpreting factor models.” The Journal of Finance 73.3 (2018): 1183-1223.

14. Bryzgalova, Pelger, and Zhu. ”Forest through the trees: Building cross-sections of stock returns.” Available at SSRN 3493458 (2019).
15. Erel, Isil, et al. Selecting directors using machine learning. No. w24435. National Bureau of Economic Research, 2018.
16. Adamopoulos, P., Ghose, A., and Todri, V. (2018). The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. Information Systems Research, 29(3), 612-640
17. de Matos, Ferreira, and Krackhardt (2014). Peer influence in the diffusion of the iPhone 3G over a large social network. Management Information Systems Quarterly, 38(4), 1103-1134.
18. Gong, Abhisek, and Li (2018). Examining the Impact of Keyword Ambiguity on Search Advertising Performance: A Topic Model Approach, MIS Quarterly, 42(3), 805-829.
19. Jian, Yang, Ba, Lu, and Jiang (2019). Managing the Crowds: The Effect of Prize Guarantees and In-Process Feedback on Participation in Crowdsourcing Contests. Management Information Systems Quarterly, 43, 97-112.
20. Mousavi, and Gu (2019). The Impact of Twitter Adoption on Lawmakers? Voting Orientations. Information Systems Research, 30(1), 133-153.
21. Goel et al. ”The structural virality of online diffusion.” Management Science 62.1 (2016): 180-196.
22. Shmueli, and Koppius. Predictive analytics in information systems research. Robert H. Smith School Research Paper No. RHS(2010): 06-138.
23. Moon, and Russell. Predicting product purchase from inferred customer similarity: An autologistic model approach. Management Science 54.1 (2008): 71-82.
24. Bertsimas et al. ”Algorithmic prediction of health-care costs. Operations Research 56.6 (2008): 1382-1392.
25. Wei et al. ”Credit scoring with social network data.” Marketing Science 35.2 (2015): 234-258.
26. Dzyabura and Hauser. ”Active machine learning for consideration heuristics.” Marketing Science 30.5 (2011): 801-819.
27. Cui, Wong, and Lui. ”Machine learning for direct marketing response models: Bayesian networks with evolutionary programming.” Management Science 52.4 (2006): 597-612.
28. Liu, Xiao, Dokyun Lee, and Kannan Srinivasan. ”Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning.” Journal of Marketing Research 56.6 (2019): 918-943.

Paper Presentation

For each week in which we have “Paper Presentation and Discussion,” three students present three papers. Each paper presentation is 30 minutes, followed by a 10 minute discussion.

Recommended

- Chang, W. *R Graphics Cookbook*, O’Reilly, 2013
- Adler, J. *R in a Nutshell: A Desktop Quick Reference.* O’Reilly Media, 2010.
- Zumel, N. and Mount, J. *Practical Data Science with R* Manning Publication, 2014.
- Van De Geer, S.A., *Least Squares Estimation (LSE)*, Volume 2, pp. 1041-1045.

The first textbook above is strongly recommended for creating informative visuals using R.

Academic dishonesty

Academic dishonesty is unacceptable and will not be tolerated. Cheating, forgery, plagiarism and collusion in dishonest acts undermine the college’s educational mission and the students’ personal and intellectual growth. Baruch students are expected to bear individual responsibility for their work and to uphold the ideal of academic integrity. Any student who attempts to compromise or devalue the academic process will be sanctioned. Please see the Baruch College Website for Further Information:

http://www.baruch.cuny.edu/academic/academic_honesty.html

Schedule

- Week 1: Statistical Learning, Linear Algebra for Machine Learning and Text Mining.
- Week 2: Supervised Learning. Bayesian Binary Classification. Hyperplanes, half-spaces, distances from hyperplanes. Algorithmic approach versus the data modeling approach. KNN.
- Week 3: Supervised Learning. MLE of logistic regression. R example of text mining. Quantifying risk in classification problems.
- Week 4: Paper Presentation and Discussion.
- Week 5: Model Selection and Regularization. Reading: ISLR chapter 5,6 and ESL chapter 7. Homework 1.
- Week 6: Paper Presentation and Discussion.
- Week 7: Moving Beyond Linearity. Reading: ISLR chapter 7. Homework 2.
- Week 8: Paper Presentation and Discussion.
- Week 9: Tree-Based Methods. Reading: ISLR chapter 8. Homework 3.
- Week 10: Paper Presentation and Discussion.
- Week 11 and 12: Deep Feedforward Networks. Reading: DLR chapter 1-4
- Week 13: Project Presentation
- Week 14: Project Presentation

Students who participate in this class with their camera on or use a profile image are agreeing to have their video or image recorded solely for the purpose of creating a record for students enrolled in the class to refer to, including those enrolled students who are unable to attend live. If you are unwilling to consent to have your profile or video image recorded, be sure to keep your camera off and do not use a profile image. Likewise, students who un-mute during class and participate orally are agreeing to have their voices recorded. If you are not willing to consent to have your voice recorded during class, you will need to keep your mute button activated and communicate exclusively using the ”chat” feature, which allows students to type questions and comments live.