STA 70400 - Quantitative Analysis for Business Decisions: Machine Learning

August 26, 2023

Instructor	Prof. Rahnama Rad
Lecture Days & Time	Tu 9-11am
Room	13-254
Email	kamiar.rahnamarad@baruch.cuny.edu
Office Hours	30 minutes after class, or by appointment

- First day of class Tuesday, Aug 29.
- Tuesday, October 10 – CONVERSION DAY – Classes follow a Monday schedule
- Last day of class Tuesday, Dec 5.

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.

A significant part of the course will cover large language models, word embeddings, sentiment analysis and all the relevant quantitative and programing backgrounds needed to apply them to large corpuses of texts.

In these lectures we focus on a few recent papers at the intersection of machine learning, text mining and business.

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 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 Python to apply multiple regression and classification learning techniques, and large language models and perform sentiment analysis
- 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 the methodologies employed for different shapes of datasets
- classify text, measure the performance.

- critically comment on the historical context in which these large language and statistical models were innovated.

The content of the course will be heavily about mathematical, computational and communication aspects of machine learning. All the lectures and homework assignments and sample codes will be in the form of a python code, a google colab file. Therefore it is important that you have google colab functioning on your system and that you familiarize yourself with basic LaTex.

Schedule

1. text mining, Zipf law, entropy for preprocessing text data, n-grams, bag of words, multinomial distribution
2. unsupervised learning: clustering: Gutenberg text data, tf-idf cosine distance between texts, hierarchical clustering, k-means, PCA
3. unsupervised learning: dimensionality reduction: pairwise distance, cosine distance, hierarchical clustering, k-means, PCA, manifold learning MNIST image data, **hwk 1**
4. **presentation**
5. supervised learning theory: data, models, loss, model complexity, optimization, **hwk 2**
6. supervised learning: regression, matrix algebra, absolute error loss, neural net optimization, real data
7. supervised learning: classification, neural net optimization, entropy loss, multi class, hyperplane distance, logistic, IMDB reviews sentiment analysis, ROC curve, imbalanced data, **hwk 3**
8. **presentation**
9. overfitting, regularization, cross validation, model selection, IMDB reviews
10. classification: SVM, random forest, bagging, boosting, MNIST, **hwk 4**
11. deep learning: feed forward neural nets, word embeddings, example IMDB reviews
12. **presentation**
13. deep learning: recurrent neural nets, example Gutenberg text data, **hwk 5**

Grading

- Homework 60%
- Presentation 20%
- Project 20%

Late work

20% deduction per day.

Pre requisites

Linear algebra of multiple linear regression and principal component analysis. Basic probability. This book might be helpful: Boyd, S. and Vandenberghe, L. *Introduction to Applied Linear Algebra (IALA)*

Course Materials

- Jurafsky and Martin (2023) Speech and Language Processing (3rd ed. draft)
- Bybee, Leland, et al. Business news and business cycles. No. w29344. National Bureau of Economic Research, 2021.
- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. "Text as data." Journal of Economic Literature 57.3 (2019): 535-574.
- Thompson et al (2022) The computational limits of deep learning.
- Kozlowski et al (2019) The geometry of culture: Analyzing the meanings of class through word embeddings American Sociological Review

- Vafa, Palikot, Du, Kanodia, Athey, Blei (2022). CAREER: Transfer Learning for Economic Prediction of Labor Sequence Data.
- Adamopoulos, Ghose, and Todri, (2018). The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. Information Systems Research, 29(3), 612-640
- James, G., Witten, D. Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning Springer (ISLR)*.
- Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction (ESL)*.
- Hassan, Tarek Alexander, et al. Sources and transmission of country risk. No. w29526. National Bureau of Economic Research, 2021.

Others Papers

1. Hassan, Tarek Alexander, et al. Sources and transmission of country risk. No. w29526. National Bureau of Economic Research, 2021.
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. Shmueli, G., *To explain or to predict?*, Statistical science (2010): 289-310.
4. Breiman, L. *Statistical modeling: The two cultures (with comments and a rejoinder by the author)*, Statistical science 16.3 (2001): 199-231.
5. Li, F., *The information content of forward-looking statements in corporate filings - A naive Bayesian machine learning approach*, Journal of Accounting Research (2010)
6. Loughran, T. and McDonald, B., *Textual analysis in accounting and finance: A survey*, Journal of Accounting Research (2016)
7. Gu, Shihao, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. The Review of Financial Studies 33.5 (2020): 2223-2273.
8. Ke, Kelly, and Xiu. Predicting returns with text data. No. w26186. National Bureau of Economic Research, 2019.
9. Kozak, Nagel, and Santosh. "Shrinking the cross-section." Journal of Financial Economics 135.2 (2020): 271-292.
10. Kozak, Nagel, and Santosh. "Interpreting factor models." The Journal of Finance 73.3 (2018): 1183-1223.
11. Bryzgalova, Pelger, and Zhu. "Forest through the trees: Building cross-sections of stock returns." Available at SSRN 3493458 (2019).
12. Erel, Isil, et al. Selecting directors using machine learning. No. w24435. National Bureau of Economic Research, 2018.
13. 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.
14. 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.
15. 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.
16. Mousavi, and Gu (2019). The Impact of Twitter Adoption on Lawmakers? Voting Orientations. Information Systems Research, 30(1), 133-153.
17. Goel et al. "The structural virality of online diffusion." Management Science 62.1 (2016): 180-196.
18. Shmueli, and Koppius. Predictive analytics in information systems research. Robert H. Smith School Research Paper No. RHS(2010): 06-138.
19. Moon, and Russell. Predicting product purchase from inferred customer similarity: An autologistic model approach. Management Science 54.1 (2008): 71-82.
20. Bertsimas et al. "Algorithmic prediction of health-care costs. Operations Research 56.6 (2008): 1382-1392.
21. Wei et al. "Credit scoring with social network data." Marketing Science 35.2 (2015): 234-258.

22. Dzyabura and Hauser. ”Active machine learning for consideration heuristics.” Marketing Science 30.5 (2011): 801-819.
23. Cui, Wong, and Lui. ”Machine learning for direct marketing response models: Bayesian networks with evolutionary programming.” Management Science 52.4 (2006): 597-612.
24. 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.

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, to learn the rules and definitions that underlie the practice of academic integrity, and to uphold its ideals. Ignorance of the rules is not an acceptable excuse for disobeying them. Any student who attempts to compromise or devalue the academic process will be sanctioned.

Please see the Baruch College Website for Further Information:

https://provost.baruch.cuny.edu/academic-affairs/teaching-and-learning/academic_honesty

Student Disability Services

It is college policy to provide Accommodations and Academic Adjustments to students with disabilities. Any student who has a disability who may need accommodations in this class should register as early as possible with Student Disability Services which is located in VC 2-272. All discussions will remain confidential. 646-312-4590.