Machine Learning Methods for Social Science Research

(Spring 2022; January 13, 2022)

This course introduces PhD students in Management to Machine Learning methods. The first half discusses underlying principles and applications of many predictive models. The second half discusses Bayesian networks to develop and estimate probabilistic models. The students will learn from discussion of the underlying theories of the methods, discussion of research articles in social sciences that apply these methods, and programming/estimation assignments modeled after current research in this area.

Instructor: Nachiketa Sahoo

Hours and location: Thursday 12:30pm to 3:15pm; HAR 412.

Grade breakdown:

1. Assignments: 40
2. Project: 30
3. Participation: 20
4. Reading summary: 10

Pre-requisite:

Students should take introductory statistics or econometrics courses (e.g., Statistics for Economists (EC 507) and Econometrics (EC 508) or their equivalent) before taking this course. Familiarity with Python programming language is recommended. Do the pre-assignment at the end of syllabus to learn Python enough to get ready for the class.[1]

Typical class structure:

This is a PhD seminar: the learning mode is in class discussion by students guided by the instructor. You will be asked to lead the discussion of 2-3 papers (assigned ahead of time) during the semester. These two components are worth 20 points towards grade.

You should read the assigned background material before class. A ≈250-word summary of the material in your own words is due at the start of each class.[2] It'll be graded as ✓, ✓+, ✓-. All ✓s get 10 points.

---

[1] You may use a different language such as R or MATLAB to do the class assignments. However, I'll provide most of the solutions in Python.

[2] Except for the first class, which you can submit before the start of the second.

Bring your laptop to class; we will work through some programming examples together. You will explore the examples on your own afterwards.

Assignments:

There will be 3-4 assignments during the course. They will have theoretical components asking to derive or show certain results, and applied components asking program implementation.

Project:

The goal is for you to select a setting that interests you and apply some of the ideas discussed in the class to explore it. Here are some possible examples:

1. Develop a predictive model: pick a dataset, briefly state what you want to predict and why, propose a probabilistic model for it, estimate, and evaluate.
2. Evaluate different tools for a prediction exercise. Apply different prediction methods discussed in class to an important task and summarize their strengths and weaknesses for the task.
3. Estimate heterogenous treatment effects: pick a high-dimensional dataset; state the effect you want to measure and why it is important to consider its heterogeneity; apply one of the techniques discussed in class and discuss the results.

Textbooks:

1. Bishop: Pattern Recognition and Machine Learning—Christopher M. Bishop
2. Murphy: Machine Learning: A Probabilistic Perspective—Kevin Murphy
   (Feel free to skip the *'ed sections while reading for class, unless explicitly included.)

Weekly Schedule:

1. Introduction to Machine Learning

Objective: Understand differences between prediction and causal inference, learn different types of machine learning methods, and review fundamental probability concepts.

   a. Supervised vs unsupervised learning
   b. Parametric vs non-parametric methods, k-nearest neighbor
   c. Inference vs prediction
   d. Review of probability theory: expectation, Bayes rule, conditional independence
   e. Discrete and continuous probability distributions
   f. Information theory (entropy, KL divergence, mutual information)

Background:

- Murphy 1, 2.2, 2.8

- (Optional) Murphy 2.3-2.6; Bishop 1.1–1.4, 1.6, 2

Conceptual papers:

- Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 25(3), 289-310.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491-95.

Other Recommended:

- Machine Learning and Prediction in Economics and Finance lecture by Sendhil Mullainathan (YouTube link).

2. Introduction to model estimation and comparison

Objective: Learn common model estimation and evaluation techniques.

   a. MLE, MAP, and predictive distribution
   b. Common distribution estimation; Naïve Bayes classifier
   c. Model comparison: AIC/BIC, Bayes Factor, cross validation, ROC curve.
   d. Decision theory
   e. Optimization techniques for parameter estimation: Quasi Newton methods and automated differentiation.

Background:

- Murphy 3, 5
- (Optional) Bishop 1.5, 3.4, 3.5

Other Recommended:

- Practical Optimization by David Duvenaud and Dougal Maclaurin (week 2 folder on QuestromTools (QT))

At home before next class:

- Read the scikit-learn getting started guide and do the examples therein.
- Take a look at HW1.

3. Some common linear classifiers
   Objective: Learn some of the linear models for classification; evaluate how well they work for a dataset.
   a. Gaussian discriminant analysis
   b. Logistic regression and choice models
   c. Bayesian treatment of logistic regression

Background:

- Bishop 4, (Optional) 4.1

- Murphy 4.2–4.4
- (Optional) Murphy 8.1–8.4
- Conditional Logit Analysis of Qualitative Choice Behavior, by: Daniel Mcfadden, edited by: P. Zarembka, In *Frontiers in econometrics* (1974), pp. 105-142

Application paper:

- McFadden, D. (1986). The choice theory approach to market research. *Marketing science*, 5(4), 275-297.

4. Model Tuning

   Objective: Learn how to identify a small subset of features that contains most of the information for predicting the output.

   a. Feature selection
   b. Regularization
   c. Active learning

Background:

- Murphy 13 (skip 13.4)

Application Papers:

- Active Feature-Value Acquisition—Maytal Saar-Tsechansky, Prem Melville, Foster Provost (2009) *Management Science*

5. Support Vector Machines

   Objective: learn how kernel representation helps classify datasets with highly non-linear separating boundaries in the original feature space.

   a. Kernels functions
   b. Maximum margin classifiers
   c. SVM for regression and classification

Background:

- Andrew Ng's lecture notes on Support Vector Machines (Week 5 folder on QT)
- Bishop 6.1–6.3, 7.1
- (Optional) Murphy 14.1–14.6

Application Papers:

- Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science*, 24(4), 595-615.
- Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science*, 56(7), 1146-1160.

Other recommended:

- Andrew Ng's lecture on SVM at Coursera: https://www.coursera.org/learn/machine-learning/home/week/7 (Also on YouTube, Lectures 12.*, ≈ 1.5h)

At home before next class:

- Install econml (on a shell terminal: `pip install econml`). Import package and print version (on a python prompt: `import econml as eml; print(eml.__version__)`): should print 0.12.0 or higher.

6. Tree-based Methods and ensembles
Objective: non-linear classification techniques with implicit feature selection.
    a. Decision trees
    b. Bagging, Boosting, and Random Forest
    c. Causal trees

Background:

- Bishop 14.1–14.4
- Murphy 16.1–16.4, 16.6–16.8
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353-7360.
    - Watch Susan Athey's lecture on estimating heterogenous treatment effect.

7. Causal inference using Machine Learning

Objective: understand the challenges to causal inference in the presence of high dimensional data and learn how to overcome them using machine learning techniques.

    a. Estimating treatment effect with high dimensional datasets
    b. Heterogenous treatment effects
    c. Incorporating instrumental variables

Background With Applications:

- Sections Estimation Methods under Unconfoundedness and Estimation Methods with Instruments from EconML user guide. Experiment with the sample code.
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. "High-dimensional methods and inference on structural and treatment effects." Journal of Economic Perspectives 28.2 (2014): 29-50.
- Hartford, Jason, et al. "Deep IV: A flexible approach for counterfactual prediction." International Conference on Machine Learning. PMLR, 2017.
- Athey, Susan, Julie Tibshirani, and Stefan Wager. "Generalized random forests." The Annals of Statistics 47.2 (2019): 1148-1178 (skip Sections 3 and 4).

8. Introduction to Probabilistic Graphical Models
Objective: learn how to graphically specify conditional independences in probabilistic models and exploit them to simplify estimation/inference.
    a. Semantics: conditional independence, d-separation
    b. Exact inference, variable elimination
    c. Structure learning

Background:

- Bishop 8
- Paper: Michael I. Jordan. Graphical Models. *Statistical Science* 19(1):140-155, 2004.
- (Optional) Murphy 10, 20, 26

Application Papers:

- Machine learning for direct marketing response models: Bayesian networks with evolutionary programming—G Cui, ML Wong, HK Lui (2006) *Management Science*

Recommended:

- Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688. See the version with discussions by other scholars.

9. Latent Variable Models
Objective: understand how/why latent variables are used in many common probabilistic models and estimation challenges they pose.
    a. Mixture models
    b. Factor analysis, PCA
    c. Expectation Maximization framework for estimation

Background:

- Bishop 9
- Murphy 11, 12
- (Optional) Bishop 12, 14.5
- (Optional) Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society*. Series B (methodological), 1-38.

Application Papers:

- A Dynamic Analysis of Market Structure Based on Panel Data—Erdem, Tülin (1996) *Marketing Science*
- The generalized multinomial logit model: accounting for scale and coefficient heterogeneity—DG Fiebig, MP Keane, J Louviere, Nada Wasi (2010), *Marketing Science*

10. Approximate Inference (Functional)

Objective: learn how to successively approximate a desired model, which is hard to optimize numerically, via simpler functions.

a. Expectation Maximization
b. Variational approximation
c. Variational Bayes

Background:

- Bishop 10–10.4. (Optional) 10.5–10.7
- (Optional) Murphy 21, 22
- (Optional) Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2), 183-233.
- (Reference) Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families and Variational Inference. *Foundations and Trends in Machine Learning* 1(1-2):1-305, 2008.

Application Papers:

- Variational Inference for Large-scale Models of Discrete Choice—Michael Braun and Jon McAuliffe (2010), *Journal of the American Statistical Association*

At home before next class:

- Install PyMC3 and test with the first block of code in the file 11_bayes_logistic.py in QT.

11. Approximate Inference (Stochastic)

Objective: understand why this is one of the most used inference techniques, when to use and when there might be better alternatives with a little bit of work.

a. Importance sampling, MCMC, Metropolis Hasting, Gibbs
b. Convergence diagnostics
c. Tools that facilitate MCMC based estimation

Background:

- Bishop 11
- Murphy 23.1–23.4, 24

Application Papers:

- Path to purchase: A mutually exciting point process model for online advertising and conversion—L Xu, JA Duan, A Whinston (2014), *Management Science*

12. Dynamic Bayesian Networks

Objective: understand tradeoffs, independence assumptions and computational challenges, between different models of sequential data.

a. HMM, State space models, and their variants
b. Dynamic programming for inference on chains
c. Particle filtering

Background:

- Murphy 17, 18, 23.5–23.6
- Bishop 13
- (Optional) Murphy 27

Application Papers:

- A Hidden Markov Model of Collaborative Filtering—Nachiketa Sahoo, Param Vir Singh, Tridas Mukhopadhyay (2012) *Management Information Systems Quarterly*
- (Optional) Modeling Online Browsing and Path Analysis Using Clickstream Data Alan L. Montgomery, Shibo Li, Kannan Srinivasan, John C. Liechty (2004), *Marketing Science*

13. Text Mining

Objective: transform unstructured text documents to a structured form, apply some of the already discussed machine learning techniques, and learn a few specialized techniques.

a. Pre-processing, representation
b. Text classification
c. Document clustering, topic models (Latent Dirichlet Allocation)

Background:

- Ch 2.1–2.2, 6.2–6.4 of "Introduction to Information Retrieval" by Manning, Raghavan, Schütze (Free book copy by the authors).
- Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289-296)
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- (Optional) Blei, D. M., & Lafferty, J. D. (2009). Topic models. In *Text Mining* (pp. 101-124). Chapman and Hall/CRC.

Application Papers:

- Deriving the pricing power of product features by mining consumer reviews—N Archak, A Ghose, PG Ipeirotis (2011) *Management Science*

14. Text Mining II and wrap up

Objective: learn analyses that preserve the order of the words in a sentence.

a. Embeddings, word2vec

Background:

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).
- Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746-751).

Application Papers:

- Yang, M., Adomavicius, G., Burtch, G., & Ren, Y. (2018). Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, 29(1), 4-24.

Python Pre-assignments

Install the software

I recommend using `conda` package manager for installing `python` and python packages because it simplifies the process. Please install Miniconda for Python 3 for your operating system. Then, using `conda`, install the packages `numpy`, `scipy`, `matplotlib`, `ipython`, `pandas`, `statsmodel`, `seaborn`, `sympy`, and `scikit-learn`. You might find the conda cheat sheet helpful. I recommend the Integrated Development Environment (IDE) PyCharm for writing Python code (professional version is free for students).

Become familiar with Python

Please read and do the examples/exercises in the following sections of Scipy Lecture Notes (2020e):

- Chapter 1. Python scientific computing ecosystem
- Chapter 2.1–2.7. The Python language
- Chapter 4.1–4.5 NumPy: creating and manipulating numerical data
- Chapter 6.1–6.7, 6.11 (up to p 215) Scipy: high-level scientific computing
- Chapter 16.1–16.5 Statistics in Python

You should write the code yourself and experiment with variations as you work through these chapters. Let me know if you get stuck or have questions.