Machine Learning Methods: Course Overview

A. Colin Cameron
Univ.of California - Davis

April 2024

1. Prediction

- We wish to **predict** y given $\mathbf{x}$ using fitted function $\widehat{f}(\mathbf{x})$.
- We could use various **nonparametric methods**
  - kernel regression such as local linear, nearest neighbors, sieves
  - but these perform poorly if $\mathbf{x}$ is high dimensional
    - the curse of dimensionality.
- **Machine learning uses different algorithms** that may predict better
  - including lasso, random forests and neural networks
  - these require setting tuning parameter(s)
    - just as e.g. kernel regression requires setting bandwidths.

2. Terminology

- The term **machine learning** is used because the machine (computer) determines the model $\widehat{f}(\mathbf{x})$ using only data
  - compared to a modeler who e.g. specifies $\mathbf{x}$ and $y = \mathbf{x}'\boldsymbol{\beta} + u$.
- **"Big data"**: The data may be big or small
  - typically $\dim(\mathbf{x})$ is large but n can be small or large.
- Many different fields developed ML methods
  - leading to the same method having different names
  - and some names can seem strange to economists.

- **Supervised learning = Regression**
  - We have both outcome y and regressors (or **features**) $\mathbf{x}$
  - 1. **Regression**: y is continuous
  - 2. **Classification**: y is categorical.
- **Unsupervised learning**
  - We have no outcome y - only several $\mathbf{x}$
  - 3. **Cluster Analysis**: e.g. determine five types of individuals given many psychometric measures.
- Focus on 1. as this is most used by economists.

y as well
as x?

Yes

Supervised
learning

y continuous
Regression

y categorical
Classification

No
Unsupervised
learning
Cluster analysis

3. Model Fitting

- Machine learning algorithms use best predictive ability
  - often **minimize mean squared error**.
- Recall MSE = Variance + Bias-squared.
- So **allow for some bias** in the prediction if this reduces the variability of the prediction
  - unlike traditional econometrics
- **Shrinkage** (or **regularization**) uses many predictors but shrinks estimated coefficients towards zero
  - examples are **LASSO** and **Ridge regression**.
- Other popular machine learning algorithms are
  - regression trees and **random forests**
  - **neural networks**.

4. Model Selection

- Econometricians often determine the model using economic theory, existing models for similar data, and statistical significance.
- Machine learners use model fit so need to control for inherent **overfitting** of the data.
- One approach uses **model complexity penalties** such as AIC, BIC.
- More often machine learners use **cross-validation**
  - this is out-of-sample predictive ability
  - new to econometrics.
- Terminology
  - **training data set** (or **estimation sample**) is used to train a model
    - in some cases this is further split into
      training data which is used to estimate a model
      validation data to evaluate how well model fits out-of-sample
  - **test data set** (or **hold-out sample set**) is additional data used to determine how good is the model fit of the preferred model after training.

5. Partial Effects

- Machine learning literature traditionally focused purely on **prediction**
  - sometimes useful in microeconomics applications
  - e.g. predict one-year survival following hip transplant operation.
- Empirical microeconomics emphasizes estimating a **partial effect**.
- In principle can perturb an x to get $\Delta\widehat{f}(\mathbf{x})$
  - but very black box especially if $\widehat{f}(\mathbf{x})$ is very nonlinear
  - statistical inference following machine learning is a problem
  - it is noncausal.
- Instead economists impose more structure.
  - e.g. estimate β in the partial linear model $y = \beta x_1 + g(\mathbf{x}_2) + u$.

6. Causal Analysis for Partial Linear Model

- Estimate β in the partial linear model $y = \beta x_1 + g(\mathbf{x}_2) + u$.
- A causal interpretation of β as giving dy/dx_1 is possible if
  - $\mathrm{E}[u|x_1, \mathbf{x}_2] = 0$
  - selection-on-observables assumption.
- The assumption is more plausible the better is $g(\mathbf{x}_2)$.
- So use a machine learning method to determine best $\widehat{g}(\mathbf{x}_2)$.
- Subsequently estimate β in a way that allows for valid inference on $\widehat{\beta}$ controlling for the data mining used to get $\widehat{g}(\mathbf{x}_2)$
  - ideally $\widehat{\beta}$ is consistent and asymptotically normal.

7. Double / Debiased Machine Learning

- Consider partial linear model $y = \beta x_1 + g(\mathbf{x}_2) + u$.
- Estimation uses **double/debiased machine learning**.
- **Orthogonalization**
  - estimation of β is based on an orthogonalized moment condition
  - one where first stage estimation of $g(\mathbf{x}_2)$ does not affect the subsequent second step estimation of β
  - leads to consistency and asymptotic normality.
- **Cross fitting** (or sample splitting)
  - use part of the data to determine $\widehat{g}(\cdot)$
  - use a separate part of the data to determine $\widehat{\beta}$
  - reduces bias.
- This approach is general
  - not just for partial linear model (e.g. for ATE in binary treatment)
  - a variety of machine learners can be used (not just LASS0).

8. More on Machine Learning

- Data carpentry or data wrangling creates y and $\mathbf{x}$
  - web scraping, text mining, digitizing images, SQL, ...
- Machine learning methods entail many decisions
  - how are features converted into x's, tuning parameter values, which ML method to use, ...
- For commercial use this may not matter
  - all that matters is that predict well enough.
- For published research we want reproducibility
  - at the very least document exactly what you did
  - provide all code (and data if it is publicly available)
  - keep this in mind at the time you are doing the project.
- For public policy we prefer some understanding of the black box
  - this may be impossible
  - and it can be misapplied
    - e.g. using credit scores to decide whether to rent house.

9. Software

- Stata provides an easy entry but serious ML generally needs R or Python.
- R has many packages for ML prediction and is easy to install
  - including *Introduction to Statistical Learning 2e* is all in R.
- Python is viewed as being best for ML prediction
  - modules scikit-learn, Pytorch, Keras, TensorFlow.
- Stata has limited built-in commands for ML prediction
  - basically Lasso, Ridge and elastic net
  - importantly Stata does have some ML for causal analysis
- Some Stata add-ons provide a front-end to R and Python commands
  - then you need to have R or Python also installed.
- Stata can directly use Python commands once Python is installed.
- Stata can directly use R given the user-written `Rcall` package
  - https://github.com/haghish/rcall
- With R and Python it can be difficult to know which package is best, and the best package will change over time.

10. Course Outline

- **1.** Variable selection and cross validation
- **2.** Shrinkage methods
  - ridge, lasso, elastic net
- **3.** ML for causal inference using lasso
  - OLS with many controls, IV with many instruments
- **4**. Other methods for prediction
  - nonparametric regression, principal components, splines
  - neural networks
  - regression trees, random forests, bagging, boosting
- **5.** More ML for causal inference
  - ATE with heterogeneous effects and many controls.
- **6.** Classification and unsupervised learning
  - classification (categorical y) and unsupervised learning (no y).

11. Key References

- ISL2: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2021), An Introduction to Statistical Learning: with Applications in R, 2nd Edition, Springer.
  - Masters level book. Free PDF from https://www.statlearning.com/ and $40 softcover book via Springer mycopy.
- ISLP: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor (2023), An Introduction to Statistical Learning: with Applications in Python, Springer.
  - Masters level book. Free PDF from https://www.statlearning.com/ and $40 softcover book via Springer mycopy.
- ESL: Trevor Hastie, Robert Tibsharani and Jerome Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd Edition, Springer.
  - PDF and $40 softcover book at https://link.springer.com/book/10.1007/978-0-387-84858-7
- Geron: Aurelien Geron (2022), Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow, Third edition, O'Reilly.

Key References (continued)

- EH: Bradley Efron and Trevor Hastie (2016), Computer Age Statistical Inference: Algorithms, Evidence and Data Science, Cambridge University Press.
- A. Colin Cameron and Pravin L. Trivedi (2022), Microeconometrics using Stata: Volume 2: Nonlinear Models and Causal Inference, Second Edition, Stata Press.
  - especially Chapter 28.
- Alex Belloni, Victor Chernozhukov and Christian Hansen (2014), "High-dimensional methods and inference on structural and treatment effects," Journal of Economic Perspectives, Spring, 29-50.
  - early causal methods with three applications.

- My website has some material
  - https://cameron.econ.ucdavis.edu/e240f/machinelearning.html
  - https://cameron.econ.ucdavis.edu/sfu2022/ has Stata code for these slides.