Machine Learning: A Very Brief Overview

A. Colin Cameron
U.C.-Davis

September 2023

Introduction

- Machine learning is used to make predictions.
- The term **machine learning** is used because the machine learns from past data, rather than using models specified by knowledgeable experts.
- For example, initial efforts for computer-based language translation used rules of grammar. Now machine learning predictions are used instead.
- There are many different machine learning algorithms.
- In many settings they perform much better than previous prediction methods.

Regression

- Suppose we want to predict y based on potential variables $x_1, ..., x_p$.
- Traditional economics methods would use some combination of
  - specify a model such as a linear regression model and include only those variables that previous studies suggest are relevant
  - possibly then drop variables that are statistically insignificant.
- Machine learning methods instead
  - specify a vary flexible nonlinear model
  - use methods that reduce the variability of the prediction by allowing some bias
  - use out-of-sample prediction to choose the best model and guard against in-sample overfitting.

Overview

1. Terminology
2. Model selection - especially cross-validation.
3. Variance-bias trade-off and shrinkage (LASSO and Ridge)
4. Neural nets
5. Regression trees and random forests
6. Other Methods
7. Classification
8. Unsupervised learning (cluster analysis)
9. Prediction for economics
10. Causal Inference
11. References

1. Terminology

- The term **machine learning** is used because the machine (computer) figures out from data the model $\widehat{y} = \widehat{f}(\mathbf{x})$
  - compared to a modeler who e.g. specifies $\mathbf{x}$ and $y = \mathbf{x}'\boldsymbol{\beta} + u$.
- The data may be big or small
  - typically $\dim(\mathbf{x})$ is large but n can be small or large.

Terminology (continued)

- **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.
- A lot of machine learning is actually used for 2.
  - license plate recognition, Google translate, ...

Terminology (continued)

- Consider two types of data sets
  - 1. **training data set** (or **estimation sample)**
    - ★ used to fit a model.
  - 2. **test data set** (or **hold-out sample** or **validation set**)
    - ★ additional data used to determine how good is the model fit
    - ★ a test observation $(\mathbf{x}_0, y_0)$ is a previously unseen observation.

2. Model selection

- Machine learners choose $x's$ by using predictive ability
  - often **mean squared error MSE** $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$.
- Problem: models “overfit” within sample.
- Solution 1:
  - use an in-estimation-sample prediction with penalty for overfitting
    - e.g. $\bar{R}^2$, AIC, BIC, Mallows Cp
- Solution 2:
  - use out-of-estimation sample prediction (**cross-validation**)
    - new to econometrics
    - can apply to other loss functions and not just MSE.

K-fold cross-validation is standard method

- K-fold cross-validation (standard choices are $K = 5$ and $K = 10$)
  - split data into K mutually exclusive folds of roughly equal size
  - for $j = 1, ..., K$ fit using all folds but fold j and predict on fold j
- The following shows case $K = 5$

	Fit on folds	Test on fold
$j = 1$	2,3,4,5	$1 \longrightarrow \text{MSE}_{(1)}$
$j = 2$	1,3,4,5	$2 \longrightarrow \text{MSE}_{(2)}$
$j = 3$	1,2,4,5	$3 \longrightarrow \text{MSE}_{(3)}$
$j = 4$	1,2,3,5	$4 \longrightarrow \text{MSE}_{(4)}$
$j = 5$	1,2,3,4	$5 \longrightarrow \text{MSE}_{(5)}$

- The K-fold CV estimate is

$$\text{CV}_K = \frac{1}{K}\sum_{j=1}^{K} \text{MSE}_{(j)}, \text{ where } \text{MSE}_{(j)} \text{ is the MSE for fold } j.$$

- Choose the model with smallest CV_K.

3. Bias-Variance Trade-off and Shrinkage Estimation

- The goal is minimize MSE $=$ Variance $+$ Bias-squared.
- More flexible models have
  - less bias (good) and more variance (bad).
  - this trade-off is fundamental to machine learning.
- Shrinkage reduces variance and may offset increased bias.
  - e.g. $\widehat{\beta} = 0$ has reduced variance to zero.

Shrinkage Methods: Ridge Regression

- Shrinkage estimators minimize RSS (residual sum of squares)
  - but with a penalty for model size
  - this shrinks parameter estimates towards zero.
- The ridge estimator $\widehat{\boldsymbol{\beta}}_\lambda$ of $\boldsymbol{\beta}$ minimizes

$$Q_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})\}^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

  - where $\lambda \geq 0$ is a tuning parameter to be determined
- Equivalently minimize $\sum_{i=1}^{n} \{y_i - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})\}^2$ subject to $\sum_{j=1}^{p} \beta_j^2 \leq s$.
- Typically first standardize $x's$ to have mean zero and variance 1.

Shrinkage Methods: LASSO

- Instead of squared penalty use absolute penalty.
- The Least Absolute Shrinkage and Selection (LASSO) estimator $\widehat{\boldsymbol{\beta}}_{\lambda}$ of $\boldsymbol{\beta}$ minimizes

$$Q_{\lambda}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})\}^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

  - where $\lambda \geq 0$ is a tuning parameter to be determined.
- Equivalently minimize $\sum_{i=1}^{n} \{y_i - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})\}^2$ subject to $\sum_{j=1}^{p} |\beta_j| \leq s$.
- The acronym LASSO is because
  - it sets some $\beta' s$ to zero and shrinks others towards zero.

LASSO versus Ridge (key figure from ISL)

- Two regressors: ellipses are the residual sum of squares for different values of β_1 and β_2 and the green squares are constraints
- Estimate is ellipse that just satisfies the constraint
  - LASSO is likely to be at a corner where some coefficients are zero.

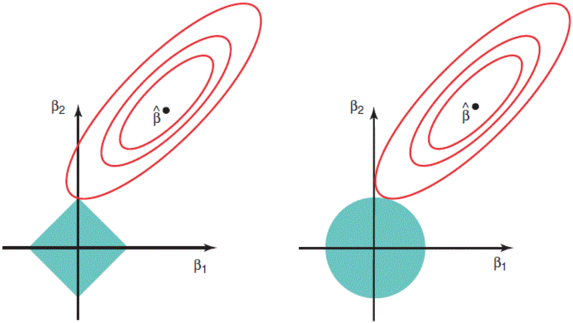
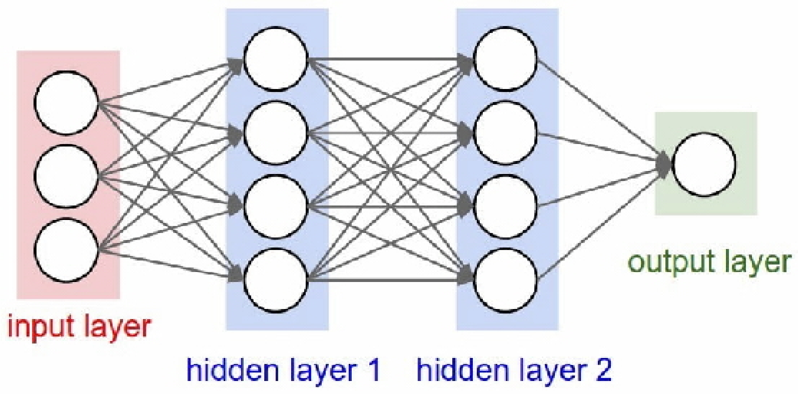


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

4. Neural Networks (deep learning)

- A neural network involves a series of nested regressions.
- A single hidden layer neural network explaining y by $\mathbf{x}$ has
  - y depends on $\mathbf{z}'s$ (a hidden layer)
  - $\mathbf{z}'s$ depend on $\mathbf{x}'s$.
- A neural network with two hidden layers explaining y by $\mathbf{x}$ has
  - y depends on $\mathbf{w}'s$ (a hidden layer)
  - $\mathbf{w}'s$ depend on $\mathbf{z}'s$ (a hidden layer)
  - $\mathbf{z}'s$ depend on $\mathbf{x}'s$.
- Neural nets are good for prediction
  - especially in speech recognition (Google Translate), image recognition, ...
  - but require much tuning and very difficult (impossible) to interpret
  - and basis for deep nets and deep learning.

Neural Network Example

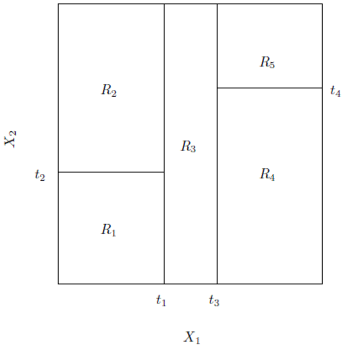


5. Regression Trees and Random Forests

- Regression trees sequentially split regressors $\mathbf{x}$ into regions that best predict y.
- Sequentially split $\mathbf{x}'s$ into rectangular regions in way that reduces RSS
  - then $\widehat{y}_i$ is the average of $y's$ in the region that $\mathbf{x}_i$ falls in
  - with J blocks RSS$= \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \bar{y}_{R_j})^2$.
- Simplest case is a single x
  - split at x^* that minimizes $\sum_{i:x_i \leq x^*} (y_i - \bar{y}_{R_1})^2 + \sum_{i:x_i > x^*} (y_i - \bar{y}_{R_1})^2$
    - ★ where $\bar{y}_{R_1}$ is average of y_i for $i : x_i \leq x^*$
    - ★ and $\bar{y}_{R_2}$ is average of y_i for $i : x_i > x^*$.
  - second split is then best split within R_1 and R_2
  - then predicted $y's$ are a step function of x.

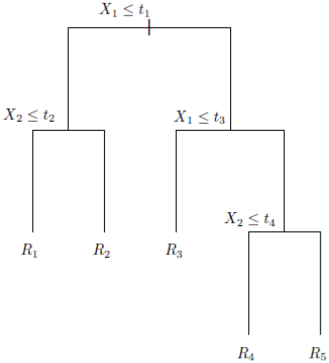
Tree example from ISL page 308

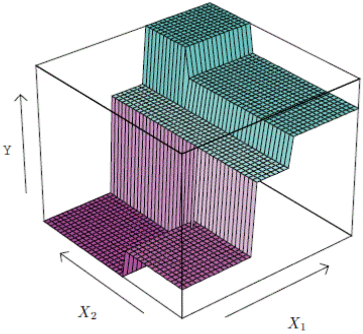
- (1) split X1 in two at t_1;
  (2) split the lowest X1 values into R1 and R2 on basis of X2 $\gtrless t_2$;
  (3) split the highest X1 values at t_3 into R3 and R4/R5;
  (4) split the highest X1 values on basis of X2 $\gtrless t_4$ into R4 and R5.



Tree example from ISL (continued)

- The left figure gives the tree.
- The right figure shows the predicted values of y.





Improvements to regression trees

- Regression trees are easy to understand if there are few regressors.
- But they do not predict as well as methods given so far
  - due to high variance (e.g. split data in two then can get quite different trees).
- Better methods are
  - bagging
    - bootstrap aggregating averages regression trees over many samples
  - random forests
    - averages regression trees over many sub-samples
  - boosting
    - trees build on preceding trees (fit residuals not y).

Random Forests

- If we bootstrap resample with replacement (bagging) the B estimates are correlated
  - e.g. if a regressor is important it will appear near the top of the tree in each bootstrap sample.
  - the trees look similar from one resample to the next.
- Random forests get bootstrap resamples (like bagging)
  - but within each bootstrap sample use only a random sample of $m < p$ predictors in deciding each split.
  - usually $m \simeq \sqrt{p}$
  - this reduces correlation across bootstrap resamples.
  - Susan Athey and coauthors are big on random forests.

6. Other Methods

- Principal components
  - - reduce from p regressors to $M < p$ linear combinations of regressors
- Basis function models
  - scalar case: $y_i = \beta_0 + \beta_1 b_1(x_i) + \cdots + \beta_K(x_i) + \varepsilon_i$
    - ★ where $b_1(\cdot)$, ..., $b_K(\cdot)$ are basis functions that are fixed and known.
  - global polynomial regression
  - splines: step functions, regression splines, smoothing splines
  - wavelets
  - polynomial is global while the others break range of x into pieces.

Nonparametric and Semiparametric regression

- Nonparametric regression is the most flexible approach
  - for $f(\mathbf{x}_0) = \mathrm{E}[y|\mathbf{x} = \mathbf{x}_0]$ borrow from observations near to $\mathbf{x}_0$
  - k-**nearest neighbors** and **kernel-weighted local regression**
  - not practical even for moderate $p = \dim(\mathbf{x})$
  - due to the curse of dimensionality
    - ★ e.g. if 10 bins in one dimension need 10^2 bins in two dimensions, ...
- Semiparametric models provide some structure to reduce the nonparametric component from many dimensions to fewer dimensions (often one).
  - partially linear models $y = f(\mathbf{x}, \mathbf{z}) + u = \mathbf{x}'\boldsymbol{\beta} + g(\mathbf{z}) + u$
  - single-index models $y = g(\mathbf{x}'\boldsymbol{\beta})$.
  - generalized additive models $y = g_1(x_1) + g_2(x_2) + \cdots$

7. Classification

- $y's$ are now categorical e.g. binary.
- Interest lies in predicting y using $\widehat{y}$ (classification)
  - whereas economist typically want $\widehat{\Pr}[y=j|\mathbf{x}]$
  - use number misclassified as loss function (not MSE).
- Some methods choose category with highest $\widehat{\Pr}[y=j|\mathbf{x}]$
  - logit, k-nearest neighbors, discriminant analysis
- Support vector machines skip $\widehat{\Pr}[y=j|\mathbf{x}]$ and directly get $\widehat{y}$
  - can do better.
- Many ML applications are to classification.

8. Unsupervised Learning: cluster analysis

- Challenging area: no y, only $\mathbf{x}$.
- Example is determining several types of individual based on responses to many psychological questions.
- Principal components analysis
  - already presented earlier.
- Clustering Methods
  - k-means clustering.
  - hierarchical clustering.

ISL Figure 10.5

- Data is $(x_1.x_2)$ with $K = 2, 3$ and 4 clusters identified.

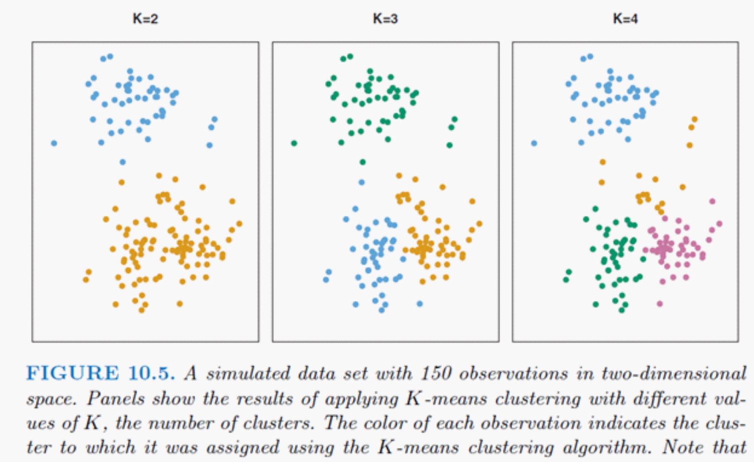


FIGURE 10.5. *A simulated data set with 150 observations in two-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that*

9. ML for Prediction for Economics

- Microeconometrics focuses on estimation of β or of partial effects.
- But in some cases we are directly interested in predicting y
  - probability of one-year survival following hip transplant operation
    - if low then do not have the operation.
  - probability of re-offending
    - if low then grant parole to prisoner.
- Sendhil Mullainathan and J. Spiess: "Machine Learning: An Applied Econometric Approach", Journal of Economic Perspectives, Spring 2017, 87-106.
  - good article to read
  - consider prediction of housing prices
  - detail how to do this using machine learning methods
  - and then summarize many recent economics ML applications.

10. ML for causal effects in partial linear model

- Microeconometrics focuses on estimation of β or of partial effects.
- Consider **partial linear model** $y = \beta x_1 + g(\mathbf{x}_2) + u$.
  - a good choice of controls $g(\mathbf{x}_2)$ makes the assumption that $Cov(x_1, u) = 0$ more plausible
  - so can give β a causal interpretation.
- So use machine learner to come up with good $g(\mathbf{x}_2)$
  - but use ML in a way that allows valid inference even though we are data mining.
- Belloni, Chernozhukov and Hansen (2014), "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, Spring, 29-50
  - provides three examples including an IV example.

ML for causal treatment effects

- Consider a binary treatment, so $x_1 = d \in \{0, 1\}$
- The preceding partially linear model $y = \beta d + \mathbf{x}_2'\delta + u$
  - restricts the same response β for each individual
  - requires that $E[u|d, \mathbf{x}_2] = 0$ for unconfoundedness.
- The heterogeneous effects approach is more flexible
  - different responses for different individuals
  - and unconfoundedness assumptions may be more reasonable.
- ML methods have been developed for estimation of the average treatment effect.

11. References

- My website has various detailed slides on machine learning
  - http://cameron.econ.ucdavis.edu/e240f/machinelearning.html
- Chapter 28 of A. Colin Cameron and Pravin K. Trivedi, *Microeconometrics using Stata: Volume 2*, Stata Press
  - covers machine learning methods for prediction and for causal inference
  - Stata provides a good introduction to machine learning
    - though more advanced ML prediction methods use Python or R.
  - see https://cameron.econ.ucdavis.edu/mus2/ for book information.
- Standard texts available free at https://www.statlearning.com/ are
  - Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2021), *An Introduction to Statistical Learning: with Applications in R*, 2nd edition, Springer.
  - Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor (2021), *An Introduction to Statistical Learning: with Applications in R*, 2nd edition, Springer.