Machine Learning and Economics

Methods and Applications of Machine Learning

Elisa Tosetti

2023-2024

Outline

- Machine learning vs econometrics
- The bias-variance trade-off
- Framework for using machine learning in applied econometrics

Resources:

- Athey, S., (2019), The Impact of Machine Learning on Economics, National Bureau of Economic Research
- Mullainathan S. and Spiess J. Machine Learning: An Applied Econometric Approach
- Hastie et al/Gareth et al

New data types in economics

Until twenty years ago, data on economic activity was relatively scarce

In just a short period of time, this has changed dramatically

Data are now available faster, have greater coverage and scope, and include new types of observations and measurements that previously were not available

Modern data sets also have much less structure, and are more complex than the traditional cross-sectional, time-series or that we teach in econometrics classes

New data types in economics

Three main types of new data available:

- *Tall* data sets include not so many variables, k, but many observations, N, with $N >> k$. This is for example the case with tick by tick data on selected financial transactions

- *Fat* data sets have instead many variables, but not so many observations, $k >> N$
  - Large cross-sectional databases
- Huge datasets, with very large k and N

New data types in economics

How this data revolution has affected economic research?

- Better measurements of economic effects and outcomes
- More granular and comprehensive data can help to pose new sorts of questions and enable novel research
- New data may end up changing the way economists approach empirical questions and the tools they use to answer them: shift away from the single covariate causal effects framework. . . .

A tale of two cultures....

Breiman (2001) ponders the state of statistics and sees two cultures..

- One culture assumes to know the model that supposedly generated the data. Emphasis is on model interpretability and validation, if done at all, is done through goodness-of-fit
- The other culture uses algorithmic models and treats the data-generating process as unknown. Choose the model with the highest predictive validation accuracy

Breiman (2001) argues that commitment to the first culture:

"has lead to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems"

Machine learning

Machine learning (ML) is a field that develops algorithms designed to be applied to data sets, with the main areas of focus being regression, classification, and clustering or grouping tasks

These tasks are divided into two main branches, **supervised** and **unsupervised** ML:

- **Unsupervised ML** consists of finding clusters of observations that are similar in terms of their covariates, and thus can be interpreted as "dimensionality reduction"; it is commonly used for marketing analysis, and for dealing with video, images, and text data
- **Supervised ML** typically consists of using a set of features or covariates (X) to predict an outcome (Y)

Supervised ML

A variety of ML methods for supervised learning:

- **Global/parametric methods**: Regularized regression (more on this next week)
- **Local/non parametric methods**: Decision/regression trees, random forests, kernel regression
- **Combined predictors**: Gradient Boosting, ensemble methods (use several separate algorithms and then take average)

What leads us to categorize these methods as ML methods rather than traditional econometric or statistical methods?

One common feature of many ML methods is that they use **data-driven model selection**:

The analyst provides the list of covariates or features, but the functional form is at least in part determined as a function of the data

The prediction problem: an example

Data on 10,000 randomly selected owner-occupied units from the 2011 metropolitan sample of the American Housing Survey. **150 covariates**

Performance of Different Algorithms in Predicting House Values

	Prediction performance (R^2)		*Relative improvement over ordinary least squares by quintile of house value*				
Method	*Training sample*	*Hold-out sample*	1st	2nd	3rd	4th	5th
Ordinary least squares	47.3%	41.7% [39.7%, 43.7%]	–	–	–	–	–
Regression tree tuned by depth	39.6%	34.5% [32.6%, 36.5%]	−11.5%	10.8%	6.4%	−14.6%	−31.8%
LASSO	46.0%	43.3% [41.5%, 45.2%]	1.3%	11.9%	13.1%	10.1%	−1.9%
Random forest	85.1%	45.5% [43.6%, 47.5%]	3.5%	23.6%	27.0%	17.8%	−0.5%
Ensemble	80.4%	45.9% [44.0%, 47.9%]	4.5%	16.0%	17.9%	14.2%	7.6%

Note: The dependent variable is the log-dollar house value of owner-occupied units in the 2011 American Housing Survey from 150 covariates including unit characteristics and quality measures. All algorithms are fitted on the same, randomly drawn training sample of 10,000 units and evaluated on the 41,808 remaining held-out units. The numbers in brackets in the hold-out sample column are 95 percent bootstrap confidence intervals for hold-out prediction performance, and represent measurement variation for a fixed prediction function.

From Linear Least-Squares to regression trees

Applying OLS to this problem requires making some choices

- For the OLS all 150 covariates have been included. But why not include interactions between variables? For example: the effect of the number of bedrooms may depend on the area of the unit; the added value of a fireplace may be different depending on the number of living rooms
- Simply including all pairwise interactions would be unfeasible for OLS estimation. We would therefore need to hand-curate which interactions to include in the regression
- Machine learning instead **searches for these interactions automatically**

Some features of machine learning

- Flexible, rich, data-driven models than typical linear regression model
- Can work with very high-dimensional data
- Limit expressiveness to avoid overfit (regularization)
- Learn how much expressiveness to allow (tuning)

The prediction problem

Machine Learning (ML) literature strongly focuses on **the prediction problem**

- The prediction problem: to accurately guess the value of some output variable Y from a set of input variables X
- The relationship between input and output is modelled in very general terms by some function

$$Y = f(X) + u$$

where u represents all that is not captured by information obtained from X via the mapping f. We say that error u is irreducible

Differences betwen ML and econometrics

- In applied econometrics, we often wish to understand an object like $f(X)$ in order to perform exercises like evaluating the impact of changing one covariate while holding others constant. This is not an explicit aim of ML modelling
- Instead, the goal of ML is to achieve goodness of fit in an independent test set by minimizing deviations between actual outcomes and predicted outcomes
- Economists typically abandon the goal of accurate prediction of outcomes in pursuit of an unbiased estimate of a causal parameter of interest

Differences betwen ML and econometrics

Another difference derives from the key concerns in different approaches:

- In predictive models, the key concern is the trade-off between expressiveness and overfitting, and this trade-off can be evaluated by looking at goodness of fit in an independent test set (more on this later)
- In contrast, there are several distinct concerns for causal models:
  1. Whether the parameter estimates from a particular sample are spurious
  2. How to handle the issue of the uncertainty over parameter estimates
  3. Whether the assumptions required to "identify" a causal effect are satisfied

Differences betwen ML and econometrics

Prediction: What is the predicted income of someone with a certain years of schooling s and other characteristics X?

$$Y = f(s, X) + \epsilon$$

- We are interested in predicting Y given s and X
- We are interested in finding the best possible $\hat{f}$ that maps s and X into Y
- Don't care about the functional form as long as it delivers best predictions
- Want prediction errors $\hat{\epsilon}$ to be small (i.e., precisely predicted)
- Use stat. learning tools based on CV
- Ground truth about Y is known: can always train on one sample and test how it performs out of sample

Causal Inference: What is the effect of years of schooling on income?

$$Y = \beta s + g(X) + \epsilon$$

- We are interested in estimating β
- Do care about one particular parameter
- Want expected standard errors of $\hat{\beta}$ to be small (e.g., precisely estimated), and $E(\hat{\beta}) = \beta$
- Use inference tools
- Ground truth about β is unknown: Need asymptotic theory, requires quasi-random variation in s

Differences betwen ML and econometrics

Object	Data science	Econometrics
Task	Prediction	Inference
Measure of success	Out-of-sample MSE	Estimator's standard errors
Asymptotic theory	Not needed	Essential
Cross-validation	Essential	Not needed
Data	High-dimensional	Low-dimensional
Variables selection	Data-driven	Theory-driven
Assumptions	Stable environment	Random assigment*

*:.. more on this later

Explanation vs prediction

Causal and predictive modelling differs along 4 main dimensions:

1. Causation (X causes Y) vs. association (X is associated to Y)
2. Theory (model) vs. data (data-driven approach to establish how X is related to Y)
3. Retrospective (test an existing set of hypothesis) vs. prospective (predict new observations)
4. Bias (focus on minimizing bias to get the correct impact of X on Y) vs. bias-variance trade off (balance bias and variance to get the best predictions)

An example

Imagine to have a data set that contains data about prices and occupancy rates of hotels

- Imagine first that a hotel chain wishes to form an estimate of the occupancy rates of competitors, based on publicly available prices. This is a **prediction problem**: the goal is to get a good estimate of occupancy rates, where posted prices and other factors (such as events in the local area, weather, and so on) are used to predict occupancy
- Imagine that a hotel chain wishes to estimate how occupancy would change if the hotel raised prices across the board. This is a question of **causal inference**

Prediction problem set-up

Given n data points $(x_1, y_1), ..., (x_n, y_n)$, the goal of ML is to obtain $\hat{f}$ such that our predictions $\hat{y}_i = \hat{f}(x_i)$ are "close" to the true outcome values y_i given some criterion. To formalize this, we follow these three steps:

Modelling: Decide on some suitable class of functions that our estimated model may belong to. In machine learning applications the class of functions can be very large and complex (e.g., decision trees, forests, high-dimensional linear models, etc). Also, we must decide on a loss function that serves as our criterion to evaluate the quality of our predictions

Fitting: Find the estimate $\hat{f}$ that optimizes the loss function chosen in the previous step

Evaluation: Evaluate our fitted model $\hat{f}$. That is, if we were given a new, yet unseen, input and output pair (x_0, y_0), we'd like to know if $y_0 \approx \hat{f}(x_0)$ by some metric

Measuring the quality of fit

- In order to evaluate the performance of a statistical method on a given data set, we need some way to measure how well its predictions actually match the observed data
- That is, we need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.
  - In regression setting, the most commonly used measure is the mean squared error (MSE), $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$
  - In classification setting, use of BIC, AIC or AUC
- In general, we want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a previously unseen test observation not used to train the algorithm (test MSE)

The bias-variance trade off

It is possible to show that the **expected MSE** for a given value x_0 can always be decomposed into three fundamental quantities:

$$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(u)$$

where

- $Var[f(x_0)]$ is a measures of the amount by which $\hat{f}$ would change if we estimated it using a different training data set
- $Bias(\hat{f}(x_0))$ is the error that comes from simplifying a complex Data Generating Process with a simple model
- $Var(u)$ is the irreducible error

See Gareth et al page 34

The bias-variance trade off

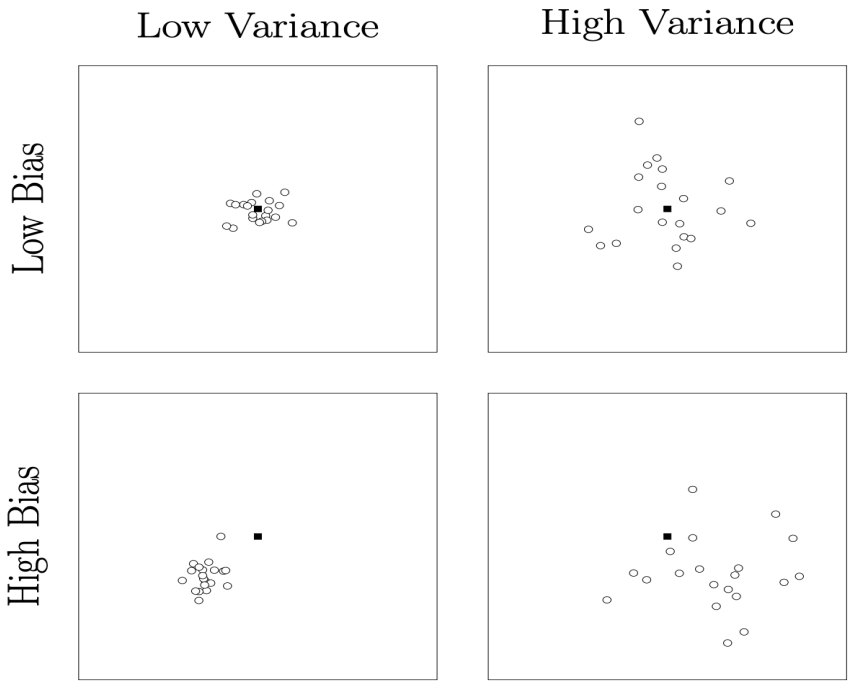
- Model flexibility affects variance and bias
- More flexible models will have higher variance, but a lower bias
- For example, linear regression assumes that there is a linear relationship between Y and $X_1, X_2, ..., X_k$. If the true f is non-linear, no matter how many training observations we are given, performing linear regression will undoubtedly result in some bias in the estimate of f

The bias-variance trade off

In order to minimize the expected MSE, we need to select a statistical method that simultaneously achieves **low variance** and **low bias**

As model complexity increases, bias decreases, while variance increases. By understanding the trade-off between bias and variance, we can manipulate model complexity to find a model that well predicts well on unseen observations

The bias-variance trade off



- The squared points indicate the true value and round points represent estimates
- A high bias/low variance estimator may yield predictions that are on average closer to the truth than predictions from a low bias/high variance estimator

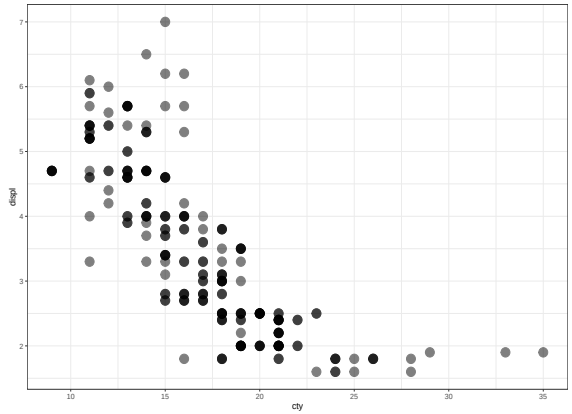
Example

Data on a sample of 234 cars and their features (taken from the *ggplot2* package in *R*)...

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact

Suppose we are interested in studying the relationship between car efficiency *cty* (our Y) and engine size *displ* (our X)

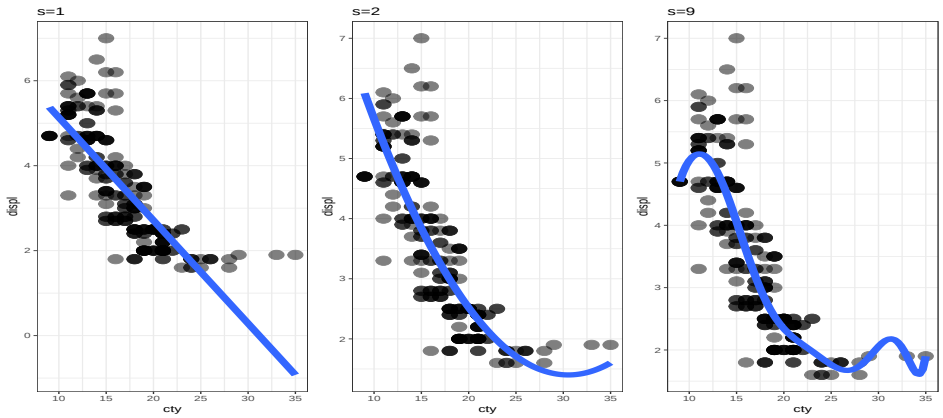
Example



Polynomial regression: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_s X^s + u$

- An important question is what is s, the degree of the polynomial: this parameter controls the complexity of the model
- One may imagine that more complex models are better, but that is not always true, because a very flexible model may try to simply interpolate over the data at hand, but fail to generalize well for new data points
- To illustrate, we try with $s = 1, 2$ and 9

Example



When s is too small, we permit only a very simple model that may suffer from misspecification bias. We call this **underfitting**

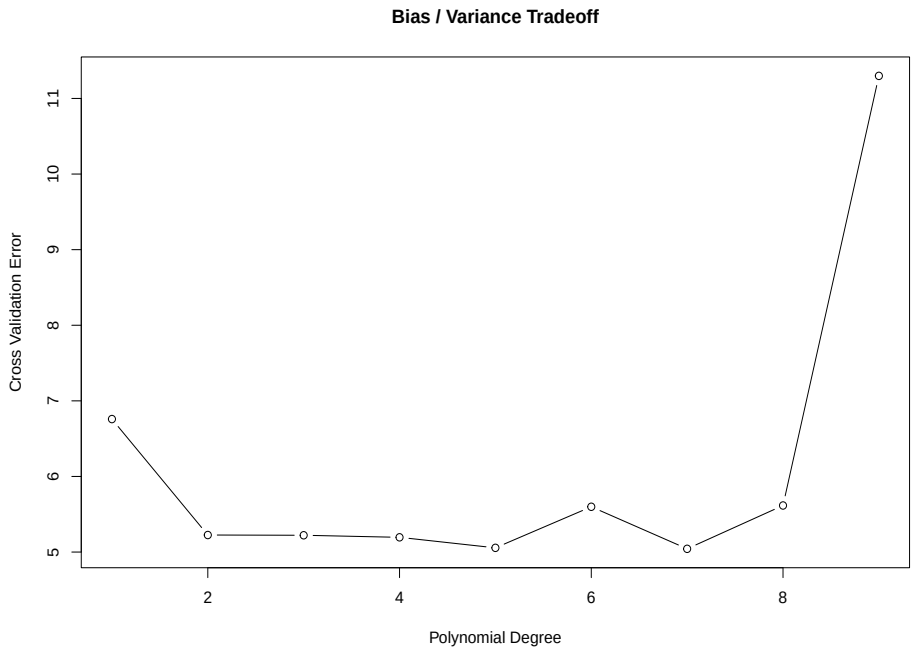
- The main feature of underfitting is **high bias** – the selected model just isn't complex enough to accurately capture the relationship between input and output variables

When s is too large, we permit a more complex model that may suffer from **overfitting**

- The main feature of overfitting is **high variance**, in the sense that, if we were given a different data set of the same size, we'd likely get a very different model

Example

Compute the average square error on $K = 5$ out-of-sample observations for $s = 1, 2, ..., 9$ and plot it



This tension is called the bias-variance trade-off: simpler models underfit and have more bias, more complex models overfit and have more variance

ML vs traditional econometrics

Many economic applications revolve around parameter estimation: produce good estimates of parameters that underlie the relationship between X and Y

- Which of the features actually matter in determining the outcome?
- Can we say something about the causal relationship between the X and the Y? $\longrightarrow$ separating correlation from causality

What economists do:

- Run OLS or IV regression
- Try a lot of functional forms
- Make a lot of assumptions without a great way to test them

On the other hand, machine learning algorithms are often pointed as **black boxes**

Can we make black boxes explainable?

Importance of interpretability

Especially in areas such as life and social sciences, interpretability is key to scientific discoveries

Interpretability is necessary in research to help identify causal relationships and increase the reliability and robustness of machine learning algorithms

Interpretable Machine Learning

Recent advent of the so-called **interpretable (or explainable) machine learning**: a field of ML that attempts to make the inner workings and reasoning behind the predictions of complex predictive modelling systems *more transparent*:

Some important tools:

- **Variable Importance**
- **Partial Dependence Plots**
- Accumulated Local Effects, Shapley Values, etc.

Variable importance

The permutation feature importance algorithm by Fisher, Rudin, and Dominici (2019):

1. Estimate the original model error $L(Y, \hat{f}(X))$ (e.g. mean squared error)
2. For each X_j, $j = 1, 2, ..., k$:
   a. Generate feature matrix X_{perm} by permuting the jth feature in the data X. This breaks the association between the jth covariate and outcome Y
   b. Estimate error $e_{perm} = L(Y, \hat{f}(X_{perm}))$ based on the predictions of the permuted data. Calculate permutation feature importance as quotient $PFI_j = e_{perm}/e$ or difference $PFI_j = e_{perm} - e$
   c. Sort features by descending PFI

Feature importance

Original data set

	x_1	x_2	...	x_p	y	$\hat{y}$
1	2	0.2		female	1	1
2	8	0.6		male	0	0
3	7	0.5		male	1	0
4	3	1.1		female	0	0
5	14	0.8		female	1	1
...						
n	11	0.4		male	1	1

Data set with permuted covariate x_1

	x_1	x_2	...	x_p	y	$\hat{y}$
1	14	0.2		female	1	0
2	11	0.6		male	0	1
3	2	0.5		male	1	1
4	7	1.1		female	0	1
5	3	0.8		female	1	0
...						
n	8	0.4		male	1	1

$$PFI_1 = L(y, f(X_{\text{perm},1})) - L(y, f(X))$$

Original data set

	x_1	x_2	...	x_p	y	$\hat{y}$
1	2	0.2		female	1	1
2	8	0.6		male	0	0
3	7	0.5		male	1	0
4	3	1.1		female	0	0
5	14	0.8		female	1	1
...						
n	11	0.4		male	1	1

Data set with permuted covariate x_p

	x_1	x_2	...	x_p	y	$\hat{y}$
1	2	0.2		male	1	1
2	8	0.6		female	0	0
3	7	0.5		male	1	0
4	3	1.1		male	0	1
5	14	0.8		female	1	1
...						
n	11	0.4		female	1	0

$$PFI_p = L(y, f(X_{\text{perm,p}})) - L(y, f(X))$$

Partial dependence plot

A partial dependence (PD) plot shows the marginal effect that one (or two) variable have on the predicted outcome of a ML model

A PD plot can show whether the relationship between the outcome and a variable is linear, monotonous or more complex

Let X_s be the set of features of interest and X_c the complement set which contains all other features, with $f(X) = f(X_s, X_c)$

The partial dependence function marginalizes over the feature distribution in set c

$$f_{X_s}(X_s) = \frac{1}{n} \sum_{i=1}^{n} f(X_s, X_c^{(i)})$$

This is a measure of the effect of X_s on $f(X)$ after accounting for the (average) effects of the other variables X_c on $f(X)$

PDP and causal interpretation

- Without further assumptions we cannot interpret the effects of changes in features on the model prediction as effects that would be present in the real world
- More is needed to interpret PDP as causal effects

Can ML help solve causal inference problem?

Note that many economic problems can be decomposed into **predictive** and **causal parts**:

- Can use off-the-shelf ML **only** for the predictive part: by doing this we achieve data-driven model selection while retaining the ability to do inference
- ML can be useful as an **intermediate step** in empirical work in economics $\rightarrow$ Better prediction give you better causal inference

Motivating examples

Finding prediction tasks ($\hat{y}$-tasks) that are relevant in economics:

- When using new kinds of data for traditional questions: for example, in measuring economic activity using satellite images or in extracting sentiment from financial news. Making sense of complex data such as images and text often involves a prediction pre-processing step
- The first stage of a linear instrumental variables regression is effectively a prediction step (more on this later)
- When when estimating propensity scores (more on this later)
- When when estimating heterogeneous treatment effects (more on this later)

A new literature on ML and causal inference

In the next lectures we will learn the basics of the literature on ML and causal inference:

- A framework for causal inference
- Attempt of economists to use data-driven model selection: the high-dimensional setting
- ML for causal models