

Machine Learning Methods for Economists

Introduction

Stephen Hansen
University of Oxford

What is Machine Learning?

The most generic definition of machine learning is the study of algorithms that allow machines to improve their performance in some given task as new data arrives.

Such algorithms can be familiar to economists, e.g. the OLS regression model, but others are quite distinct from econometric models.

There is also a difference in emphasis with econometrics:¹

- ▶ Less emphasis on statistical inference.
- ▶ Less emphasis on 'true model' that generates data.
- ▶ More emphasis on prediction.
- ▶ More emphasis on computation and optimization.

The disciplinary boundaries between machine learning, artificial intelligence, and data science are not always clear.

¹Breiman (2001, Statistical Science) "Statistical Modeling: The Two Cultures".

Basic Setup

Suppose we have N observations within a dataset of the form (y_i, \mathbf{x}_i) for $i = 1, \dots, N$:

- ▶ y_i : dependent variable, or response variable.
- ▶ \mathbf{x}_i : P -dimensional vector of independent variables, covariates, or features. Potentially $P \gg N$, e.g. genetic or text data.

Supervised Learning: Learn a mapping from \mathbf{x}_i to y_i , typically to successfully predict unseen response variables given features. When y_i is categorical, this is sometimes called a *classification* problem. When y_i is continuous, sometimes called a *regression* problem.

Unsupervised Learning: Learn some structure within the \mathbf{x}_i observations, e.g. place them into related clusters.

Evaluating Algorithms: Cross Validation

A typical way to evaluate the performance of a supervised algorithm:

1. Divide the observations into non-overlapping test and training sets.
2. Use the training-set observations to learn a model for y_i given \mathbf{x}_i .
3. Use this model to generate predicted values for the response in the test set.
4. Compared the predicted vs. actual values to assess model accuracy.

Focus on out-of-sample rather than within-sample accuracy to guard against over-fitting.

We usually want to repeat the steps above on different splits of the data into training and test sets.

Evaluating Algorithms: Cross Validation

A typical way to evaluate the performance of a supervised algorithm:

1. Divide the observations into non-overlapping test and training sets.
2. Use the training-set observations to learn a model for y_i given \mathbf{x}_i .
3. Use this model to generate predicted values for the response in the test set.
4. Compared the predicted vs. actual values to assess model accuracy.

Focus on out-of-sample rather than within-sample accuracy to guard against over-fitting.

We usually want to repeat the steps above on different splits of the data into training and test sets.

Somewhat less clear how to evaluate the output of unsupervised models without using external information.

Concerns with Cross Validation

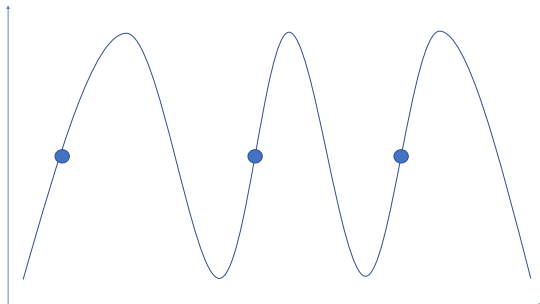
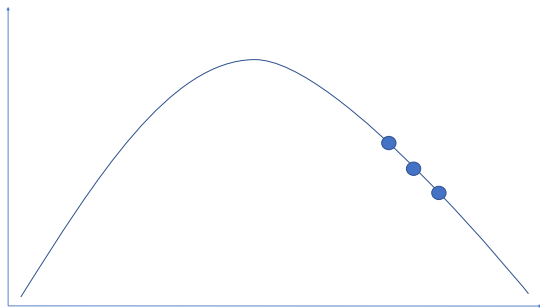
Economists usually care about causal inference and counterfactual reasoning, and supervised learning algorithms that perform well in out-of-sample prediction may not be useful for this task.

A nice example² is hotel prices and occupancy rates. In the data, these are strongly positively correlated, but what would be the expected impact of a hotel raising its prices on a given day?

Even if our goal really is just prediction, we need to ensure that we have enough data to capture the relevant shape of the mapping from \mathbf{x}_i to y_i .

² Athey (2017, Science) “Beyond prediction: Using big data for policy problems.”.

Dangers of Unrepresentative Data



Applications in Economics: Prediction

Some empirical work in economics involves an element of pure prediction.

One example in macro forecasting, e.g. Stock and Watson (1999)³ explain US monthly inflation from 1959/01 through 1997/09 with $P = 168$ economic indicators. When P is high, machine learning is useful.

Another is the classic two-step instrumental variables models. In the first step, we form a prediction for the endogenous variable given the instruments. With many instruments, or complex functional forms, machine learning is useful.⁴

³ JME, "Forecasting inflation".

⁴ Belloni et. al. (2012, ECMA), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain"; Hartford et. al. (2017, ICML) "Deep IV: A Flexible Approach for Counterfactual Prediction."


Applications in Economics: Causal Inference

A classic problem in empirical economics is treatment effect estimation.

Often the set of controls to include is not clear, and many separate specifications are run. Machine learning can allow us to select the appropriate model, while conducting valid inference.⁵

With large datasets, we might also be interested in assessing conditional treatment effects at a very granular level. Machine learning can help estimate statistically meaningful differences across subpopulations.⁶

⁵ Belloni et. al. (2014, JEP) "High-Dimensional Methods and Inference on Structural and Treatment Effects".

⁶ Athey and Imbens (2016, PNAS) "Recursive partitioning for heterogeneous causal effects". 

Applications in Economics: Unstructured Data

Unstructured data like text, satellite data, and online search histories are rich in information, but have nonstandard formats.

Hansen et. al. (2018)⁷ use a corpus of verbatim FOMC transcripts from the era of Alan Greenspan to study transparency:

- ▶ 149 meetings from August 1987 through January 2006.
- ▶ A document is a single statement by a speaker in a meeting (46,502).
- ▶ Baseline data has 6,249,776 total words and 26,030 unique words.

Challenge: how to extract information from data with no explicit labels?

⁷QJE, "Transparency and Deliberation on the FOMC: a Computational Linguistics Approach".

Applications in Economics: Label Imputation with Unstructured Data

In some cases, we need to label observations to test an economic theory, and unstructured data can help.

Prominent example is Gentzkow and Shapiro (2010)⁸ which explains variation in newspaper ideology.

Problem: we do not observe newspaper ideology.

Solution: estimate mapping from political speeches to political party, then apply the mapping to newspapers to impute ideology.

Another example is Jean et. al. (2016)⁹ which uses satellite data to measure poverty.

⁸ ECMA, "What Drives Media Slant? Evidence from US Daily Newspapers".

⁹ Science, "Combining satellite imagery and machine learning to predict poverty".

Course Outline

1. Penalized regression: LASSO and friends; applications to causal inference.
2. Text as data: bag-of-words model, probability models for discrete data, Bayesian estimation.
3. Unsupervised learning: non-parametric and parametric models, EM algorithm.
4. Graphical models, latent Dirichlet allocation, Bayesian approximation.
5. Generative models for text regression.

Programming

The course will focus on statistical ideas, some of which are implemented with example code in Python and R.

The only way you will really learn to use machine learning in your empirical work is to program algorithms yourself.

There are advantages to using fully fledged—and open source—programming languages rather than statistical packages, and these certainly are popular in machine learning.

Such languages also have many, many resources for self-study and online support, e.g. <https://learnpythonthehardway.org/>.