PLSC 40601

Week 1: Course orientation, potential outcomes framework.

Molly Offer-Westort

Department of Political Science,
University of Chicago

Spring 2024

Housekeeping

Housekeeping

- Sign up for papers/discussion

Housekeeping

- Sign up for papers/discussion
- Fork and create a PR for the repo

Housekeeping

- Sign up for papers/discussion
- Fork and create a PR for the repo
- https://bookdown.org/halflearned/ml-ci-tutorial/

Identification.

Identification

- $Y_i(0), \ldots, Y_i(K)$ are *potential* outcomes

Identification

- $Y_i(0), \ldots, Y_i(K)$ are *potential* outcomes; typically, we only get to see an individual outcome under one version of treatment.

Identification

- $Y_i(0), \ldots, Y_i(K)$ are *potential* outcomes; typically, we only get to see an individual outcome under one version of treatment.

- **Fundamental problem of causal inference**: we can't see counterfactual potential outcomes for a given unit *at the same time*.

Identification

- $Y_i(0), \ldots, Y_i(K)$ are *potential* outcomes; typically, we only get to see an individual outcome under one version of treatment.
- **Fundamental problem of causal inference**: we can't see counterfactual potential outcomes for a given unit *at the same time*.
- How do we move from what we observe to what we would like, ideally, to measure?

Identification

- Identification: We call a parameter **identifiable** if in the case that we had infinite data, we could approximate the true parameter value to arbitrary precision.

Identification

- Identification: We call a parameter **identifiable** if in the case that we had infinite data, we could approximate the true parameter value to arbitrary precision.
    - e.g., if you have **infinite data** randomly sampled from a distribution, by taking the empirical mean, you get the mean of that distribution with arbitrary precision.

Identification

- Identification: We call a parameter **identifiable** if in the case that we had infinite data, we could approximate the true parameter value to arbitrary precision.
  - e.g., if you have **infinite data** randomly sampled from a distribution, by taking the empirical mean, you get the mean of that distribution with arbitrary precision.
  - If we're dealing with a **finite population**, we can think about identifying a target quantity about that population, as if we were to repeat the procedure through we observed data about that population, averaging *across repetitions*, we could approximate the true quantity to arbitrary precision.

Identification

- Identification: We call a parameter **identifiable** if in the case that we had infinite data, we could approximate the true parameter value to arbitrary precision.
  - e.g., if you have **infinite data** randomly sampled from a distribution, by taking the empirical mean, you get the mean of that distribution with arbitrary precision.
  - If we're dealing with a **finite population**, we can think about identifying a target quantity about that population, as if we were to repeat the procedure through we observed data about that population, averaging *across repetitions*, we could approximate the true quantity to arbitrary precision.
  - We can consider parameters to be **point identified** or **interval identified**.

Identification

- **Consistency/ Stable Unit Treatment Value Assumption (SUTVA)**: what we observe is interpretable in terms of potential outcomes.

Identification

- **Consistency/ Stable Unit Treatment Value Assumption (SUTVA)**: what we observe is interpretable in terms of potential outcomes.
  - no unobserved multiple versions of the treatment
  - no "interference between units"

Identification

- **Consistency/ Stable Unit Treatment Value Assumption (SUTVA)**: what we observe is interpretable in terms of potential outcomes.
  - no unobserved multiple versions of the treatment
  - no “interference between units”

- If unit i is assigned $W_i = 1$, $Y_i = Y_i(1)$.

Identification

- **Consistency/ Stable Unit Treatment Value Assumption (SUTVA)**: what we observe is interpretable in terms of potential outcomes.
  - no unobserved multiple versions of the treatment
  - no "interference between units"

- If unit *i* is assigned $W_i = 1$, $Y_i = Y_i(1)$.

- In econometrics, this is often expressed in terms of the "switching" equation.

$$Y_i = \mathbb{1}\{W_i = 1\} Y_i(1) + (1 - \mathbb{1}\{W_i = 1\})\, Y_i(0) +$$

Identification

- **Consistency/ Stable Unit Treatment Value Assumption (SUTVA)**: what we observe is interpretable in terms of potential outcomes.
  - no unobserved multiple versions of the treatment
  - no "interference between units"

- If unit i is assigned $W_i = 1$, $Y_i = Y_i(1)$.

- In econometrics, this is often expressed in terms of the "switching" equation.

$$Y_i = \mathbb{1}\{W_i = 1\} Y_i(1) + (1 - \mathbb{1}\{W_i = 1\}) \, Y_i(0) +$$

- But once we impose that $Y_i(W_i)$ is well defined, this is largely already implied.

Identification: randomization

- When we aren't in control of assigning treatment, we say the data is **observational.**

Identification: randomization

- When we aren't in control of assigning treatment, we say the data is **observational.**

- SUTVA is not enough to get us to identification with observational data.

Identification: randomization

- When we aren't in control of assigning treatment, we say the data is **observational.**

- SUTVA is not enough to get us to identification with observational data.

- **Random assignment** gives us:

Identification: randomization

- When we aren't in control of assigning treatment, we say the data is **observational.**

- SUTVA is not enough to get us to identification with observational data.

- **Random assignment** gives us:
    - $(Y_i(1), Y_i(0)) \perp\!\!\!\perp W_i$ (independence of potential outcomes and treatment)
    - $0 < \Pr[W_i = 1] < 0$ (positivity)

Identification: randomization

- When we aren't in control of assigning treatment, we say the data is **observational.**

- SUTVA is not enough to get us to identification with observational data.

- **Random assignment** gives us:
  - $(Y_i(1), Y_i(0)) \perp\!\!\!\perp W_i$ (independence of potential outcomes and treatment)
  - $0 < \Pr[W_i = 1] < 0$ (positivity)

- This does get us identification:

$$\mathrm{E}\left[Y_i | W_i = 1\right] = \mathrm{E}\left[Y_i(1)\right]$$

Identification: randomization

- In this class, we will not spend a lot of time on identification, but it may be worth considering how various methods fare when these (or similar) assumptions are violated.

Identification: randomization

- In this class, we will not spend a lot of time on identification, but it may be worth considering how various methods fare when these (or similar) assumptions are violated.

- In particular, if you don't have identification, fancy estimating procedures **will not save you**.

Causal inference as a missing data problem

Units	Covariates X_i	Treatment W_i	$Y_i(1)$	$Y_i(0)$	Observed Y_i
1	1	1	1	?	1
2	0	0	?	0	0
3	1	1	0	?	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	0	0	?	0	0

Machine learning.

What is machine learning?

> *…we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!).*

Murphy (2012)

What is machine learning?

Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term “Machine Learning ” in 1959 while at IBM. He defined machine learning as “the field of study that gives computers the ability to learn without being explicitly programmed” [1] . Other sources also attribute the definition of machine learning to Arthur Samuel [2] .

Learn more:

1. geeksforgeeks.org
2. nzfaruqui.com
3. geeksforgeeks.org

+3 more

What is machine learning?

1 Answer

Sorted by: Highest score (default)

9

The exact quote exists in neither the 1959 paper nor the 1967 paper (second version).

These are the closest quotes from the 1959 paper:

> A computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program.

And

> Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort.

Also, Wiki page of Arthur Samuel states that:

> He coined the term "machine learning" in 1959

and references the 1959 paper.

Either the quote is created as a gist of Arthur Samuel's 1959 paper, or it is said but not written by him. In my opinion, the former is more probable, since it is not even remotely mentioned in the 1967 paper.

Share Improve this answer Follow

edited Mar 7, 2019 at 17:06

answered Mar 7, 2019 at 16:56

Esmailian

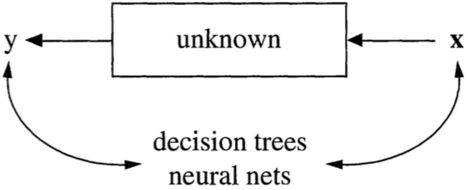
8,967 • 2 • 30 • 46

Add a comment

What is machine learning?

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Breiman (2001)

What is machine learning?

The traditional approach in econometrics ... is to specify a target, an estimand, that is a functional of a joint distribution of the data.

⋮

In contrast, in the ML literature, the focus is typically on developing algorithms ... The goal for the algorithms is typically to make predictions about some variables given others or to classify units on the basis of limited information, for example, to classify handwritten digits on the basis of pixel values.

Athey and Imbens (2019)

What is machine learning?

In the context of this class:

What is machine learning?

In the context of this class:

- A culture/perspective,

What is machine learning?

In the context of this class:

- A culture/perspective, of a research community that has developed a set of tools to tackle some common objectives,

What is machine learning?

In the context of this class:

- A culture/perspective, of a research community that has developed a set of tools to tackle some common objectives, which will inform how we think about framing what our research problems are, and how we go about addressing these problems.

Some common tasks we might address with ML methods

Some common tasks we might address with ML methods

- Prediction for the next observation

Some common tasks we might address with ML methods

- Prediction for the next observation
  - Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, and potentially X_{N+1}, formulate a method to predict $\hat{Y}_{N+1}$, to minimize

$$\left(Y_{N+1} - \hat{Y}_{N+1}\right)^2$$

Some common tasks we might address with ML methods

- Prediction for the next observation
  - Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, and potentially X_{N+1}, formulate a method to predict $\hat{Y}_{N+1}$, to minimize

$$\left(Y_{N+1} - \hat{Y}_{N+1}\right)^2$$

- Conditional means estimation

Some common tasks we might address with ML methods

- Prediction for the next observation
  - Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, and potentially X_{N+1}, formulate a method to predict $\hat{Y}_{N+1}$, to minimize

$$\left(Y_{N+1} - \hat{Y}_{N+1}\right)^2$$

- Conditional means estimation (often in the form of augmented regression)

Some common tasks we might address with ML methods

- Prediction for the next observation
  - Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, and potentially X_{N+1}, formulate a method to predict $\hat{Y}_{N+1}$, to minimize

$$\left(Y_{N+1} - \hat{Y}_{N+1}\right)^2$$

- Conditional means estimation (often in the form of augmented regression)
  - Estimate $\hat{g}(x)$,

$$g(x) = \mathrm{E}\left[Y|\mathrm{X}_i = x\right]$$

Some common tasks we might address with ML methods

- Prediction for the next observation
  - Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, and potentially X_{N+1}, formulate a method to predict $\hat{Y}_{N+1}$, to minimize

$$\left(Y_{N+1} - \hat{Y}_{N+1}\right)^2$$

- Conditional means estimation (often in the form of augmented regression)
  - Estimate $\hat{g}(x)$,

$$g(x) = \mathrm{E}\left[Y|\mathrm{X}_i = x\right]$$

  - Is this different from the prediction problem?

Some common tasks we might address with ML methods

- Prediction for the next observation
  - Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, and potentially X_{N+1}, formulate a method to predict $\hat{Y}_{N+1}$, to minimize

$$\left(Y_{N+1} - \hat{Y}_{N+1}\right)^2$$

- Conditional means estimation (often in the form of augmented regression)
  - Estimate $\hat{g}(x)$,

$$g(x) = \mathrm{E}\left[Y|\mathrm{X}_i = x\right]$$

  - Is this different from the prediction problem?
  - Do we have a lot of covariates? We may want to regularize, select a subset of covariates.

Some common tasks we might address with ML methods

- Prediction for the next observation
  - Given some data $(Y_1, \mathrm{X}_1), \ldots, (Y_N, \mathrm{X}_N)$, and potentially X_{N+1}, formulate a method to predict $\hat{Y}_{N+1}$, to minimize

$$\left(Y_{N+1} - \hat{Y}_{N+1}\right)^2$$

- Conditional means estimation (often in the form of augmented regression)
  - Estimate $\hat{g}(x)$,

$$g(x) = \mathrm{E}\left[Y | \mathrm{X}_i = x\right]$$

  - Is this different from the prediction problem?
  - Do we have a lot of covariates? We may want to regularize, select a subset of covariates.
  - Do we think the mean is not linear in covariates, and we would like to allow it to take a flexible form?

Some common tasks we might address with ML methods

Some common tasks we might address with ML methods

- Supervised classification

Some common tasks we might address with ML methods

- Supervised classification

- Unsupervised classification

Some common tasks we might address with ML methods

- Supervised classification
- Unsupervised classification
- What should we do next?

What are we actually asking from the data, and where can we find connections between machine learning and causal inference?

ML in CI

- Estimating average treatment effects.

ML in CI

- Estimating average treatment effects.

$$\mu(w, x) = \mathrm{E}[Y_i | W_i = w, \mathrm{X}_i = x]$$
$$e(x) = \mathrm{E}[W_i | \mathrm{X}_i = x]$$

ML in CI

- Estimating average treatment effects.

$$\mu(w, x) = \mathrm{E}[Y_i | W_i = w, X_i = x]$$
$$e(x) = \mathrm{E}[W_i | X_i = x]$$

$$\tau = \mathrm{E}\left[\mu(1, X_i) - \mu(0, X_i)\right]$$

ML in CI

- Estimating average treatment effects.

$$\begin{aligned} \mu(w, x) &= \mathrm{E}\left[Y_i | W_i = w, X_i = x\right] \\ e(x) &= \mathrm{E}\left[W_i | X_i = x\right] \end{aligned}$$

$$\begin{aligned} \tau &= \mathrm{E}\left[\mu(1, X_i) - \mu(0, X_i)\right] \\ &= \mathrm{E}\left[\frac{Y_i W_i}{e(X_i)} - \frac{Y_i(1 - W_i)}{1 - e(X_i)}\right] \end{aligned}$$

ML in CI

- Estimating average treatment effects.

$$\begin{aligned} \mu(w,x) &= \mathrm{E}\left[Y_i | W_i = w, \mathrm{X}_i = x\right] \\ e(x) &= \mathrm{E}\left[W_i | \mathrm{X}_i = x\right] \end{aligned}$$

$$\begin{aligned} \tau &= \mathrm{E}\left[\mu(1,\mathrm{X}_i) - \mu(0,\mathrm{X}_i)\right] \\ &= \mathrm{E}\left[\frac{Y_i W_i}{e(\mathrm{X}_i)} - \frac{Y_i(1-W_i)}{1-e(\mathrm{X}_i)}\right] \\ &= \mathrm{E}\left[\frac{\left(Y_i - \mu(1,\mathrm{X}_i)\right) W_i}{e(\mathrm{X}_i)} - \frac{\left(Y_i - \mu(0,\mathrm{X}_i)\right)(1-W_i)}{1-e(\mathrm{X}_i)}\right] \\ &\quad + \mathrm{E}\left[\mu(1,\mathrm{X}_i) - \mu(0,\mathrm{X}_i)\right] \end{aligned}$$

ML in CI

- Many of the tools we use for estimating conditional means can be used for estimating conditional average treatment effects; but may need to account for optimizing for τ rather than $\mu(\cdot)$ in e.g., parameter selection.

ML in CI

- In particular, this form of the estimator will come back when we read Chernozhukov et al. (2018); Schuler and Rose (2017):

$$= \mathrm{E}\left[\frac{(Y_i - \mu(1, X_i))\, W_i}{e(X_i)} - \frac{(Y_i - \mu(0, X_i))\,(1 - W_i)}{1 - e(X_i)}\right] + \mathrm{E}\left[\mu(1, X_i) - \mu(0, X_i)\right]$$

ML in CI

- In particular, this form of the estimator will come back when we read Chernozhukov et al. (2018); Schuler and Rose (2017):

$$= \mathrm{E}\left[\frac{(Y_i - \mu(1, \mathsf{X}_i))\, W_i}{e(\mathsf{X}_i)} - \frac{(Y_i - \mu(0, \mathsf{X}_i))\,(1 - W_i)}{1 - e(\mathsf{X}_i)}\right] + \mathrm{E}\left[\mu(1, \mathsf{X}_i) - \mu(0, \mathsf{X}_i)\right]$$

ML in CI

- In particular, this form of the estimator will come back when we read Chernozhukov et al. (2018); Schuler and Rose (2017):

$$
\begin{aligned}
= \mathrm{E}\left[\frac{(Y_i - \mu(1, X_i))\, W_i}{e(X_i)} - \frac{(Y_i - \mu(0, X_i))\,(1 - W_i)}{1 - e(X_i)}\right] \\
+ \mathrm{E}\left[\mu(1, X_i) - \mu(0, X_i)\right]
\end{aligned}
$$

- We may want to estimate $\mu(w, x)$ and $e(x)$ to get better estimates of $\hat{\tau}$, but we don't necessarily care about how "good" our estimates of these parameters are.

ML in CI

- In particular, this form of the estimator will come back when we read Chernozhukov et al. (2018); Schuler and Rose (2017):

$$= \mathrm{E}\left[\frac{(Y_i - \mu(1, \mathrm{X}_i))\, W_i}{e(\mathrm{X}_i)} - \frac{(Y_i - \mu(0, \mathrm{X}_i))\,(1 - W_i)}{1 - e(\mathrm{X}_i)}\right] + \mathrm{E}\left[\mu(1, \mathrm{X}_i) - \mu(0, \mathrm{X}_i)\right]$$

- We may want to estimate $\mu(w, x)$ and $e(x)$ to get better estimates of $\hat{\tau}$, but we don't necessarily care about how "good" our estimates of these parameters are.
  - $\rightarrow$ "nuisance" parameters

- We can use cross-fitting and orthogonalization to get estimates of $\mu(w, x)$ and $e(x)$ that will result in an estimator that has really nice properties: **double robustness**.

Experimental design

- Reinforcement learning more broadly: what action to take to optimize an objective.

Experimental design

- Reinforcement learning more broadly: what action to take to optimize an objective. (e.g., AlphaGo)

Experimental design

- Reinforcement learning more broadly: what action to take to optimize an objective. (e.g., AlphaGo)
    - Multi-armed bandits: exploration/exploitation tradeoff.

Experimental design

- Reinforcement learning more broadly: what action to take to optimize an objective. (e.g., AlphaGo)
    - Multi-armed bandits: exploration/exploitation tradeoff. Experiments addressing a broader range of objectives.

Experimental design

- Reinforcement learning more broadly: what action to take to optimize an objective. (e.g., AlphaGo)
    - Multi-armed bandits: exploration/exploitation tradeoff. Experiments addressing a broader range of objectives.
    - (if time: discuss applications)

Some additional definitions

- Cross-validation (many variants), cross-fitting

Some additional definitions

- Cross-validation (many variants), cross-fitting

- Bootstrapping, bagging, sample splitting

Some additional definitions

- Cross-validation (many variants), cross-fitting
- Bootstrapping, bagging, sample splitting (we'll get to these next week)

Some additional definitions

- Cross-validation (many variants), cross-fitting
- Bootstrapping, bagging, sample splitting (we'll get to these next week)
- Boosting

Some additional definitions

- Cross-validation (many variants), cross-fitting
- Bootstrapping, bagging, sample splitting (we'll get to these next week)
- Boosting
- We'll talk about some methods/tools: penalized regression, trees/forests, policy learning, reinforcement learning

Some additional definitions

- Cross-validation (many variants), cross-fitting
- Bootstrapping, bagging, sample splitting (we'll get to these next week)
- Boosting
- We'll talk about some methods/tools: penalized regression, trees/forests, policy learning, reinforcement learning, but there are a lot of approaches we won't get into (neural nets, support vector machines, many tools for classification problems, language tools)

Some additional definitions

- Cross-validation (many variants), cross-fitting
- Bootstrapping, bagging, sample splitting (we'll get to these next week)
- Boosting
- We'll talk about some methods/tools: penalized regression, trees/forests, policy learning, reinforcement learning, but there are a lot of approaches we won't get into (neural nets, support vector machines, many tools for classification problems, language tools) and we will talk about some high-level approaches that can combine/be applied with multiple machine learning estimating techniques (doubly robust estimation, conformal inference).

References I

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1):C21.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Schuler, M. S. and Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73.