Causal Inference and Machine Learning

In the search of causality in Machine Learning days

Luis Moneda

Data Scientist at Nubank

São Paulo, 2018

Outline

1. Introduction
   - Causal Inference in Economics
   - Prediction Vs Estimation
   - Data Science tasks
   - The challenge

2. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
   - What is it about?
   - Definitions
   - Procedures
   - Let's try it!

Outline

1 Introduction
- Causal Inference in Economics
- Prediction Vs Estimation
- Data Science tasks
- The challenge

2 Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- What is it about?
- Definitions
- Procedures
- Let's try it!

Economics

Wikipedia: Economics focuses on the behaviour and interactions of economic agents and how economies work.

Economists like to theoretically come up with a curve that relate two important variables and justify its shape with some storytelling. Then use data to find evidence that they really behave like they think it should.

They're usually engaged in first explaining why things happen that way and then being able to modify it to make things behave in a desired way.

$$Y = \beta_0 + \beta_1 * X_1 ...$$

Economics

- How does advertising increase the sales?
- How does fertilizer affect crop yields?
- How does education affect income?

Problems!

- Confounded variables
- Selection bias

Reference: Varian, H. "Causal inferece in economics and marketing", article link.

Causal Inference Identity

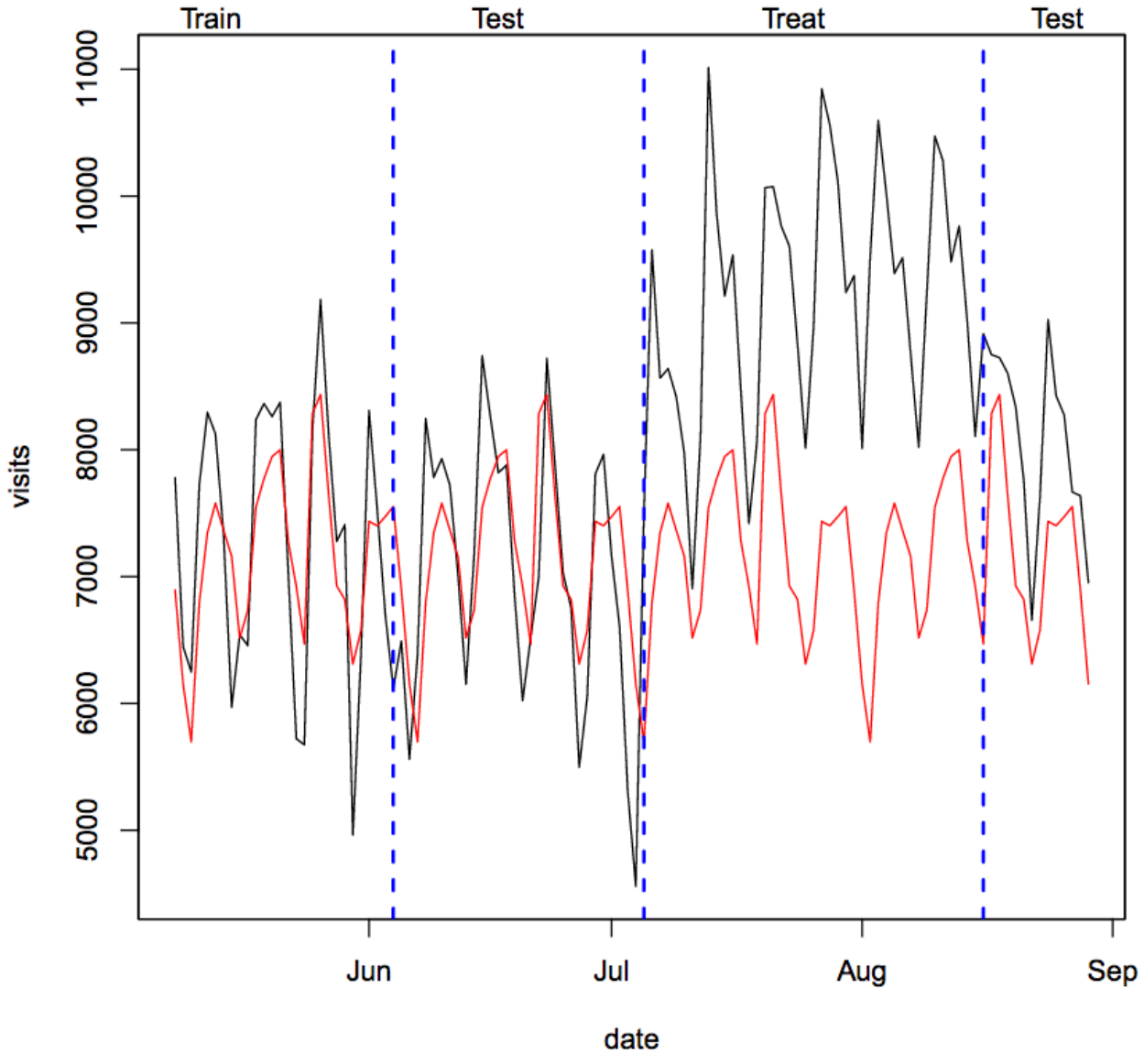
Outcome for treated - Outcome for untreated =
Outcome for treated - Outcome for treated if no treated

+

[Outcome for treated if not treated - Outcome for untreated]
= Impact of treatment on treated + selection bias.

The good thing about randomized trials is that the selection bias would be zero. But controlled experiments and randomized trials are not always possible.

Train-test-treat-compare



Discontinuity

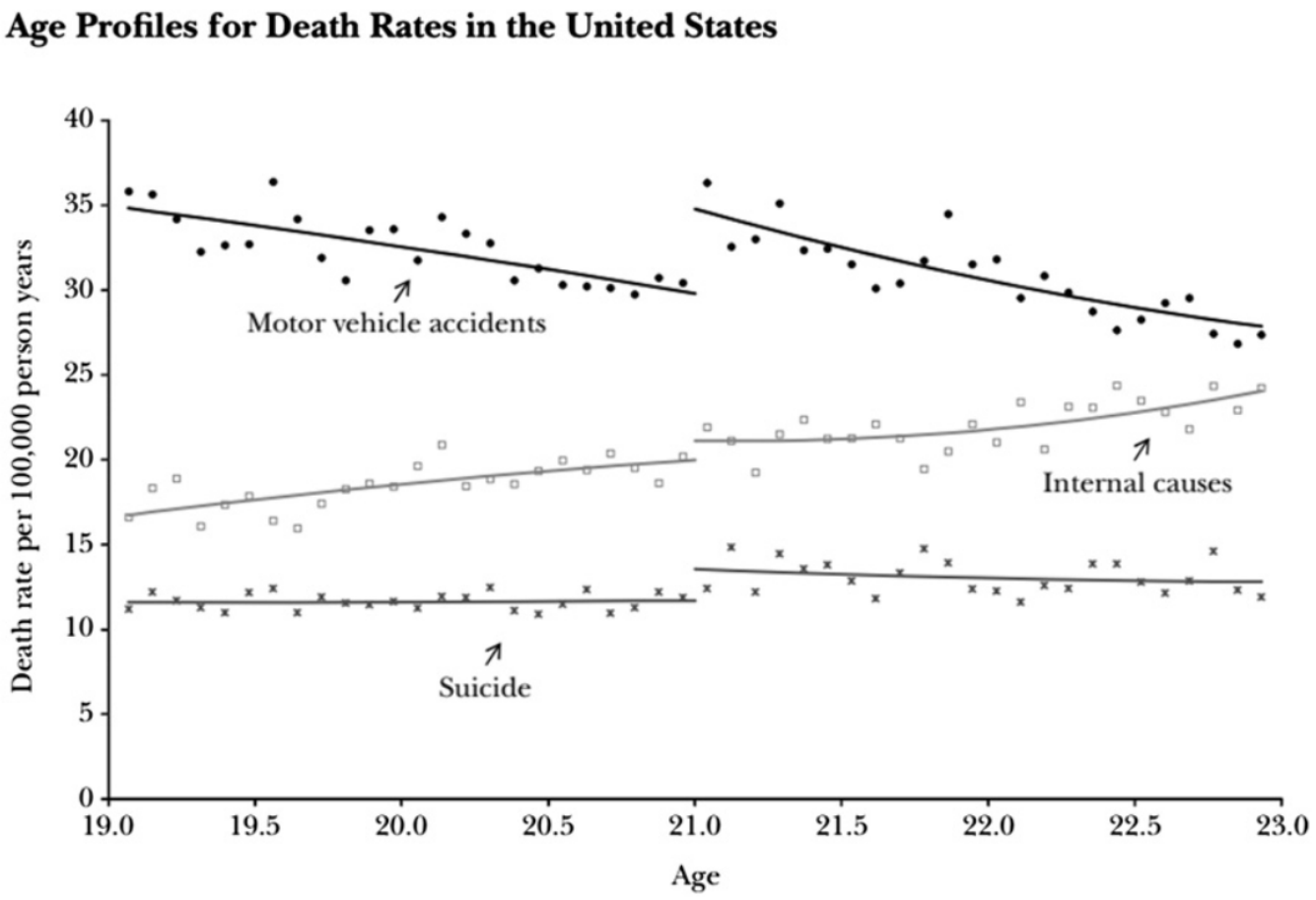


Fig. 2. Death rates by age by type of death. Reprinted with permission from ref. 13.

Difference in differences

s_{TA} = sales after ad campaign for treated groups

s_{TB} = sales before ad campaign for treated groups

s_{CA} = sales after ad campaign for control groups

s_{CB} = sales before ad campaign for control groups

We assemble these numbers into a 2×2 table and add a third column to show the estimate of the counterfactual.

The counterfactual is based on the assumption that that the (unobserved) change in purchases by the treated would be the same as the (observed) change in purchases by the control group. To get the impact of the ad campaign, we then compare the predicted counterfactual sales to the actual sales:

Period	Treatment	Control	Counterfactual
Before	s_{TB}	s_{CB}	s_{TB}
After	s_{TA}	s_{CA}	$s_{TB} + (s_{CA} - s_{CB})$

Effect of treatment on treated $= (S_{TA} - S_{TB}) - (S_{CA} - S_{CB})$

Outline

1 Introduction
- Causal Inference in Economics
- Prediction Vs Estimation
- Data Science tasks
- The challenge

2 Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- What is it about?
- Definitions
- Procedures
- Let's try it!

Prediction Vs Estimation

Example: Hotel data about price and occupancy.

Prediction: I want to predict competitors occupancy using public price data. It's expected that high price will be related with high occupancy, because hotels adjust their prices dynamically accordingly to their occupancy.

Estimation: I want to check how occupancy would change if I raise my prices in 5%. It would hardly accepted as an answer that raising the prices would increase the occupancy. There's no way to estimate it using only observational data.

Reference: Athey, S. "The impact of machine learning on economics", article link.

Outline

1 Introduction
- Causal Inference in Economics
- Prediction Vs Estimation
- Data Science tasks
- The challenge

2 Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- What is it about?
- Definitions
- Procedures
- Let's try it!

Data Science tasks

- Description: computing proportions, mean, clustering and other visualizations;
- Prediction: the mapping of some inputs (X) to output(s) (y);
- Causal Inference: *"using data to calculate certain feature of the world if the world had been different (that is, causal inference is counterfactual prediction)"*.

Reference: Miguel A. Hernán, John Hsu, Brian Healy. "Data science is science's second chance to get causal inference right: A classification of data science tasks", arXiv:1804.10846v2

Data Science tasks: Prediction

Most of successful applications today in DS are merely predictive tasks because it's easily achieved when we have:

- a large dataset with inputs and outputs;
- an algorithm that establishes a mapping between inputs and outputs;
- a metric to asses the performance of the mapping.

After having the task defined, all the information required is in the data.

Data Science task: Causal Inference

Example: What income would have been observed for a person if he had choosen a different course? It's hard because:

- it usually needs expert knowledge, usually in the form of unverifiable causal assumptions
- even with a well-defined causal task and acquiring relevant data, subject-matter knowledge is necessary to guide the data analysis and to prove a justification for endowing the resulting numerical estimates with a causal interpretation;
- no algorithm can guarantee the validity of causal inferences from observational data
- the validity of causal inferences depends on expert causal knowledge, which is fallible;

Causal Inference Example

Maternal smoking during pregnancy on the risk of infant mortality

- **Confounding**: pregnant women who do and do not smoke differ in many characteristics that affect the risk of infant mortality (alcohol consumption, diet, access to adequate prenatal care);
- We should adjust the analysis for them. But not all confounders should be adjusted!
- Birth-weight is strongly associated with both maternal smoking and infant mortality, but adjustment for birth weight induces bias because birth-weight is a risk factor that is itself affected by maternal smoking.
- Adjustment for birth weights induces a bias referred to as the "birth-weight paradox: low birth weight babies from mothers who smoked during pregnancy have a lower mortality than those from mothers who did not smoke during pregnancy.

Tasks table

	Data Science Task		
	Description	**Prediction**	**Causal inference**
Example of scientific question	What proportion of women aged 60-80 years had a stroke last year?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	Feature(s) to be described, e.g., diagnosis of stroke	• Output, e.g., diagnosis of stroke over the next year • Inputs, e.g., age, blood pressure, history of stroke, diabetes… at baseline	• Outcome, e.g., diagnosis of stroke over the next year • Treatment, e.g., initiation of statins at baseline • Eligibility criteria • Confounders • Effect modifiers (optional)
Examples of analytics	Estimation of sample statistics Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Stratification/Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

Outline

1 Introduction
- Causal Inference in Economics
- Prediction Vs Estimation
- Data Science tasks
- The challenge

2 Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- What is it about?
- Definitions
- Procedures
- Let's try it!

The challenge

If random testing is a way to avoid all the difficulties of estimating causal effect, why do we even bother?

1. It may not be ethical: ask random women to keep or start smoking during their pregnancy to check its effect on the mortality of their children;
2. It can be costly: doing an action randomly, like approving credit.

Since randomization is not ethical or costly, the challenge becomes to estimate causal effect using either just observational data or using it with some random test data.

Outline

1 Introduction
- Causal Inference in Economics
- Prediction Vs Estimation
- Data Science tasks
- The challenge

2 Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- What is it about?
- Definitions
- Procedures
- Let's try it!

What is it about?

- Goal: Heterogeneous treatment effect, estimating the effect on individuals rather than the average for the whole population or subgroups.
- How: using an approach called **causal forest** that is able to perform causal inference with good confidence intervals while dealing with a lot of covariates.
- Why it's interesting: for decision making in causal inference problems you need confidence intervals since you can't validate in the data.

Outline

1 Introduction
- Causal Inference in Economics
- Prediction Vs Estimation
- Data Science tasks
- The challenge

2 Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- What is it about?
- Definitions
- Procedures
- Let's try it!

Definitions

- Observed data: (X_i, Y_i, W_i)
- Unconfoundedness: $\{Y_i^1, Y_i^0\} \perp W_i \mid X_i$
- Treatment effect: $\tau(x) = \mathbb{E}[Y_i^1 - Y_i^0 \mid X_i = x]$
- Treatment propensity: $e(x) = \mathbb{P}[W_i = 1 \mid X_i = x]$

Honesty

A tree is honest if, for each training sample i, it only uses the response Y_i to estimate the within-leaf treatment effect τ or to decide where to place the splits, but not both.

Structure and effect should not share examples.

From CART to causal

Leaf estimation:

- **CART:** $\hat{\mu}(x) = \frac{1}{|\{i:X_i \in L(x)\}|} \sum_{\{i:X_i \in L(x)\}} Y_i$
- **Causal:** $\hat{\tau}(x) = \frac{1}{|\{i:W_i=1, X_i \in L(x)\}|} \sum_{\{i:W_i=1, X_i \in L(x)\}} Y_i - \frac{1}{|\{i:W_i=0, X_i \in L(x)\}|} \sum_{\{i:W_i=0, X_i \in L(x)\}} Y_i$
- **Ensemble of B trees:** $\hat{\tau}(x) = B_{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$

The predictions made by a causal forest are asymptotically Gaussian and unbiased: $\frac{(\hat{\tau}(x) - \tau(x))}{\sqrt{Var[\hat{\tau}(x)]}} \Rightarrow \mathcal{N}(0, 1)$

Outline

1 Introduction
- Causal Inference in Economics
- Prediction Vs Estimation
- Data Science tasks
- The challenge

2 Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- What is it about?
- Definitions
- Procedures
- Let's try it!

Double-Sample Trees

Procedure 1. DOUBLE-SAMPLE TREES

Double-sample trees split the available training data into two parts: one half for estimating the desired response inside each leaf, and another half for placing splits.

Input: n training examples of the form (X_i, Y_i) for regression trees or (X_i, Y_i, W_i) for causal trees, where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random subsample of size s from $\{1, ..., n\}$ without replacement, and then divide it into two disjoint sets of size $|\mathcal{I}| = \lfloor s/2 \rfloor$ and $|\mathcal{J}| = \lceil s/2 \rceil$.

2. Grow a tree via recursive partitioning. The splits are chosen using any data from the $\mathcal{J}$ sample and X- or W-observations from the $\mathcal{I}$ sample, but without using Y-observations from the $\mathcal{I}$-sample.

3. Estimate leaf-wise responses using only the $\mathcal{I}$-sample observations.

Double-sample *regression* trees make predictions $\hat{\mu}(x)$ using (4) on the leaf containing x, only using the $\mathcal{I}$-sample observations. The splitting criteria is the standard for CART regression trees (minimizing mean-squared error of predictions). Splits are restricted so that each leaf of the tree must contain k or more $\mathcal{I}$-sample observations.

Double-sample *causal* trees are defined similarly, except that for prediction we estimate $\hat{\tau}(x)$ using (5) on the $\mathcal{I}$ sample. Following Athey and Imbens [2016], the splits of the tree are chosen by maximizing the variance of $\hat{\tau}(X_i)$ for $i \in \mathcal{J}$; see Remark 1 for details. In addition, each leaf of the tree must contain k or more $\mathcal{I}$-sample observations of *each* treatment class.

Propensity Trees

Procedure 2. PROPENSITY TREES

Propensity trees use only the treatment assignment indicator W_i to place splits, and save the responses Y_i for estimating τ.

Input: n training examples (X_i, Y_i, W_i), where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random subsample $\mathcal{I} \in \{1, ..., n\}$ of size $|\mathcal{I}| = s$ (no replacement).
2. Train a classification tree using sample $\mathcal{I}$ where the outcome is the treatment assignment, i.e., on the (X_i, W_i) pairs with $i \in \mathcal{I}$. Each leaf of the tree must have k or more observations of *each* treatment class.
3. Estimate $\tau(x)$ using (5) on the leaf containing x.

In step 2, the splits are chosen by optimizing, e.g., the Gini criterion used by CART for classification [Breiman et al., 1984].

Putting together what's happening here...

- The estimation in the leafs addresses the effect of treatment;
- The idea is that in each leaf it behaves like a random experiment;
- The restriction of having k or more examples of **each treatment** helps to make it closer to a random experiment;
- The more the treatment is far from randomness the harder it's to work with a small k, the restriction of having at least k examples of each treatment;
- Point-wise estimations errors are asymptotically Gaussian then you can have confidence intervals.

At the end of the day: I'm just comparing treated and not treated examples using a tree to split it smartly and build a fair group to do this comparison for individual examples.

Outline

1 Introduction
- Causal Inference in Economics
- Prediction Vs Estimation
- Data Science tasks
- The challenge

2 Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
- What is it about?
- Definitions
- Procedures
- Let's try it!

Summary

- We need to start classifying tasks accordingly with their predictive or estimation nature.
- A lot of ML/Causal Inference new tricks popping up, we should keep an eye!