MAC6916

Machine Learning, Causality and the Time Tree

Luis Moneda

About me

Work

- Data Scientist at Nubank

Education

- MSc Computer Science student (IME-USP)
- Bachelor in Computer Engineering (Poli-USP)
- Bachelor in Economics (FEA-USP)

Twitter - @lgmoneda
LinkedIn
E-mail
Blog: lgmoneda.github.io/
Datasets (Kaggle)

Outline

1. Data Science Tasks
2. Prediction Vs Causal
3. The causal inference challenge
4. A couple of causal Inference approaches
5. The clash of the worlds
6. Why is it interesting to have a graphical probabilistic models background?
7. Causal Inference on observational data
8. Causal Forest
9. ML limitations
10. Invariance
11. Invariant Causal Prediction (ICP)
12. Invariant Risk Minimization (IRM)
13. Time, invariance and the Time Forest

Data Science Tasks

Data Science Tasks

Description	Prediction	Causal Inference
- Computing proportions - Aggregation metrics - Clustering - Visualizations	Mapping inputs (X) to output y.	Using data to calculate certain feature of the world if the world had been different: **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 - Examples

Description	Prediction	Causal Inference
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 taking the drug A reduce, on average, the risk of stroke in women with certain characteristics?

Data Science Tasks - Confusion Matrix

Approach you're using \ What you're trying to do	Description	Prediction	Causal Inference
Description	You're able to provide a snapshot of your data.	Nice benchmark, poor performance.	Misleading results, bad decisions.
Prediction	Why predict if you have the actual?	Low error predictions.	Biased estimations.
Causal Inference	Cost ineffective, but cool!	Cost ineffective, not the best performance.	Unbiased estimation of actions' effects.

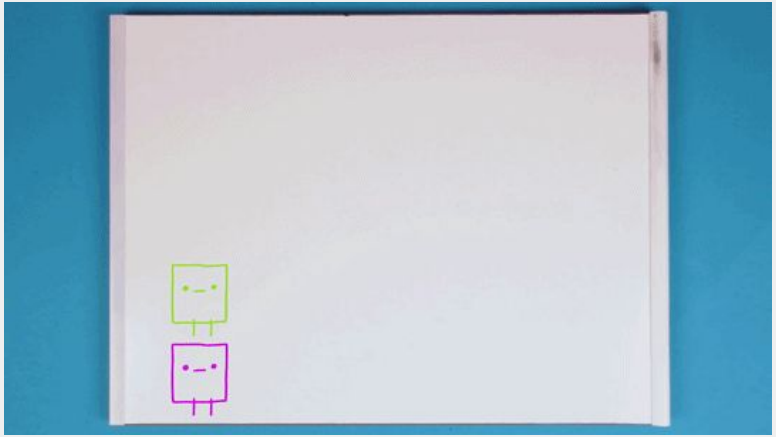
Prediction Vs Causal

Prediction

Most of successful applications today in DS are merely predictive!

Why?

1) A large dataset with inputs and outputs;
2) An algorithm that establishes a mapping between inputs and outputs;
3) A metric to assess the performance of the mapping.



All the information required is in the data!

Causal

Confusion

- Spurious correlation
- Anecdote
- Science reporting

It's hard!

- Definition is tricky
- Causal inference requires untestable assumptions
- I can only observe one potential outcome for each case

Am I facing a prediction or causal problem?

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial \pi}{\partial X_0}(Y) + \frac{\partial \pi}{\partial Y}\frac{\partial Y}{\partial X_0}$$

π: Pay-off function
X_0: Decision
Y: Outcome
Illustrative example: Umbrella x Rain dance

Reference: Prediction Policy Problems. By Jon Kleinberg. Jens Ludwig. Sendhil Mullainathan. Ziad Obermeyer.

The clash of the worlds

(A part of) The Data Science world

Predictive

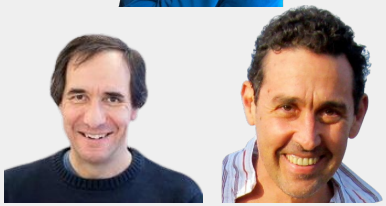
Causal Inference

Deep Learning

Yann LeCun, Geoffrey Hinton and Yoshua Bengio

Modern classification and regression

Robert Tibshirani and Trevor Hastie



Boosting and XGBoost

Robert Schapire, Yoav Freund and Tianqi Chen

Fairness

Shakar Mohamed, Timnit Gebru

ML limitations

Alex D'amour and Joaquin Quinonero-candela

Epidemiology

Miguel Hernan

Graphical Causal Models

Judea Pearl and Elias Barenboim

Invariance

Jonas Peters, Bernhard Scholkopf and Martin Arjovsky

New and old school Econometrics

Susan Athey and Joshua Angrist

Social Sciences and the potential outcomes

Guido Imbens and Donald Rubin

(A part of) The Data Science world

Predictive

Deep Learning

Yann LeCun, Geoffrey Hinton and Yoshua Bengio

Modern classification and regression

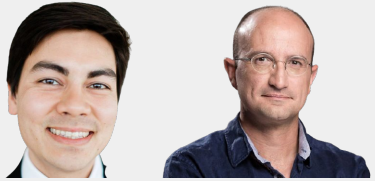
Robert Tibshirani and Trevor Hastie

Boosting and XGBoost

Robert Schapire, Yoav Freund and Tianqi Chen

Fairness

Shakar Mohamed, Timnit Gebru



ML limitations

Alex D'amour and Joaquin Quinonero-candela

Causal Inference

Epidemiology

Miguel Hernan

Graphical Causal Models

Judea Pearl and Elias Barenboim

Invariance

Jonas Peters, Bernhard Scholkopf and Martin Arjovsky

New and old school Econometrics

Susan Athey and Joshua Angrist

Social Sciences and the potential outcomes

Guido Imbens and Donald Rubin

Causality as a necessary part of predictive world

Predictive

Deep Learning

Yann LeCun, Geoffrey Hinton and Yoshua Bengio

Modern classification and regression

Robert Tibshirani and Trevor Hastie

Boosting and XGBoost

Robert Schapire, Yoav Freund and Tianqi Chen

Fairness

Shakar Mohamed, Timnit Gebru

ML limitations

Alex D'amour and Joaquin Quinonero-candela

Causal Inference

Epidemiology

Miguel Hernan

Graphical Causal Models

Judea Pearl and Elias Barenboim

Invariance

Jonas Peters, Bernhard Scholkopf and Martin Arjovsky

New and old school Econometrics

Susan Athey and Joshua Angrist

Social Sciences and the potential outcomes

Guido Imbens and Donald Rubin

The causal data science world

Epidemiology

Miguel Hernan

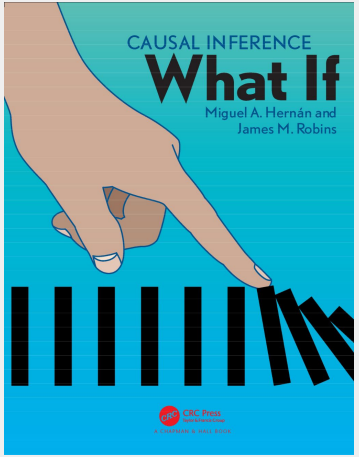
The epidemiologists

- Represent assumptions with Causal DAGs
- Use potential outcomes to estimate parameters and effects
- Randomized Controlled Trial challenges

Causal Diagrams: Draw Your Assumptions Before Your Conclusions

Learn simple graphical rules that allow you to use intuitive pictures to improve study design and data analysis for causal inference.

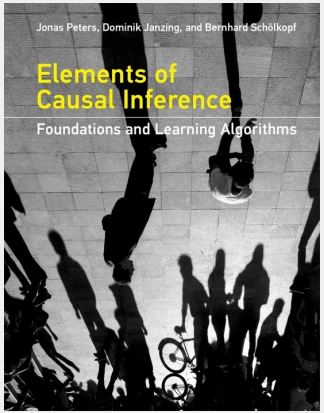
HARVARD UNIVERSITY



The causal data science world

Invariance

Jonas Peters, Bernhard Scholkopf and Martin Arjovsky



Inserting causal requirements in learning algorithms, improving prediction by making models causal

- Invariant risk minimization
- Invariant Causal Prediction
- Causality for Machine Learning

The causal data science world

New and old school Econometrics

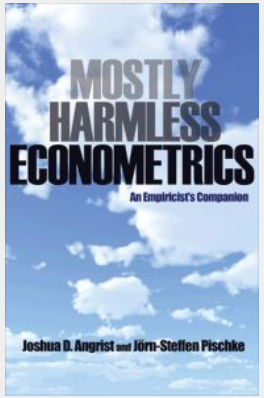
Susan Athey and Joshua Angrist

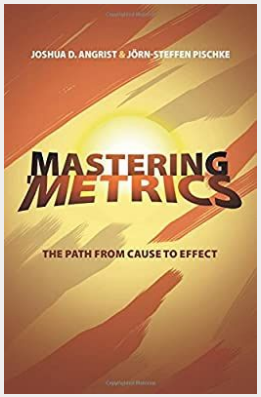
Old School

Trying to convince new generations the good and old econometrics can solve their problems

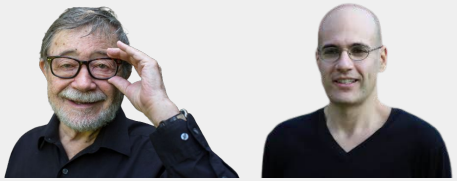
New School

- Economists working in tech companies
- Bridge between ML and econometrics, trying to use the first to answer question that matter for the second
- Heterogeneous causal effect, multi-armed bandits





The causal data science world

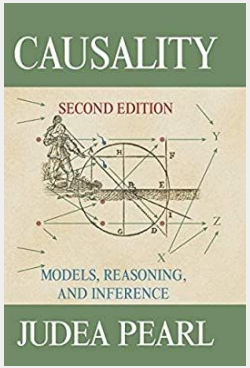


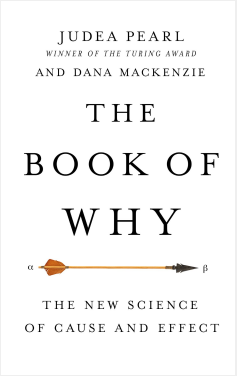
Graphical Causal Models

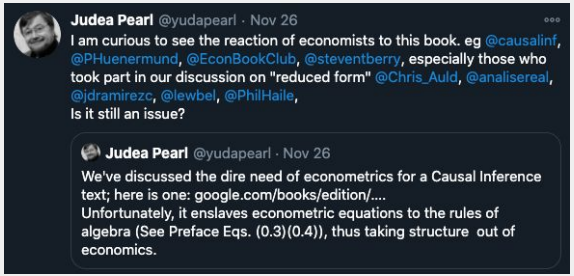
Judea Pearl and Elias Barenboim

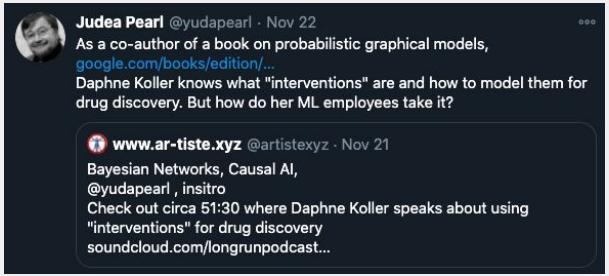
Causal DAGs and Do calculus solves everything!

- Explain all ML limitations with DAGs
- 24/7 telling ML people they should be using DAGs
- "All the impressive achievements of deep learning amount to just curve fitting"









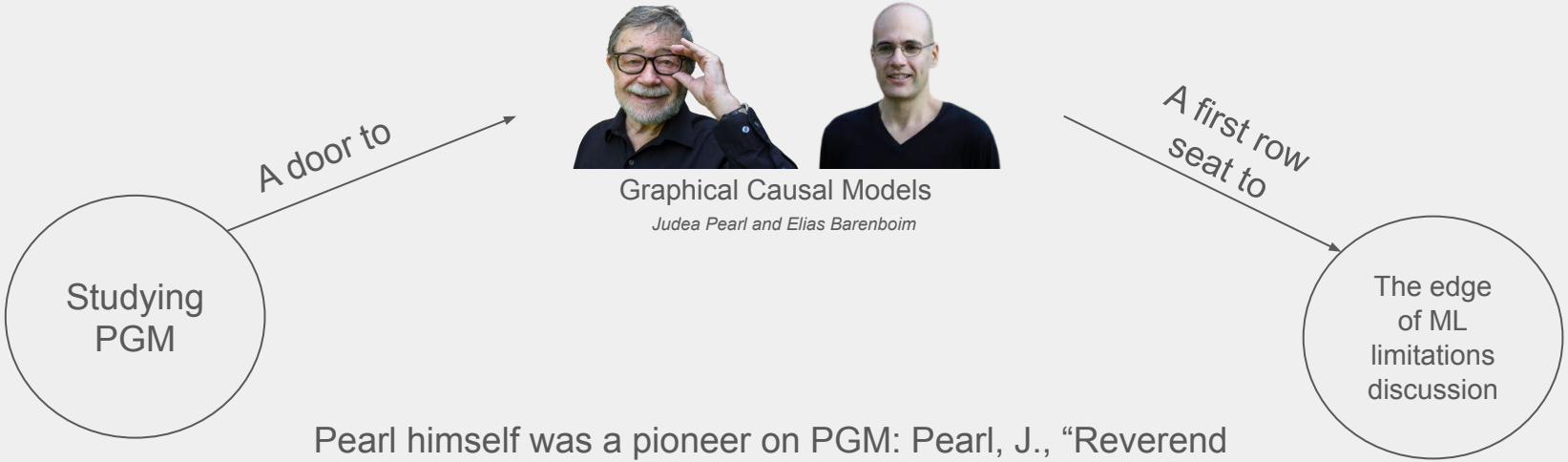
Why is it interesting to have a graphical probabilistic models background?

Causality is at the center of most ML criticism

A ctrl+f for "causal" in a couple of famous papers about ML limitations

- Underspecification Presents Challenges for Credibility in Modern Machine Learning (17 matches)
- Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift (2 matches)
- Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence (4 matches)
- Invariant Risk Minimization (6 matches)

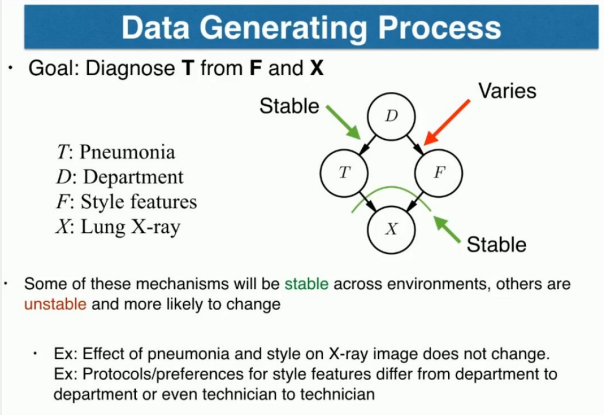
PGM is the base of Graphical Causal Models

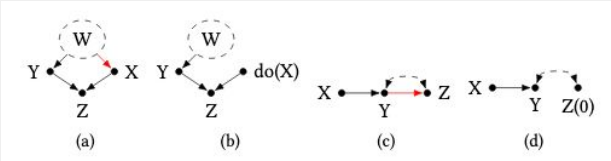


Pearl himself was a pioneer on PGM: Pearl, J., “Reverend Bayes on inference engines: A distributed hierarchical approach,” Proceedings, AAAI-82, 1982.

Many suggested solutions use PGM language

Famous researchers are turning their attention to causality also, like Yoshua Bengio and Bernhard Scholkopf.





The images link to the papers.

How to deal with causal questions?

Causal - Core Concepts, Notation

W: Treatment assignment

X_i: Features / Characteristics

Y: Observed outcome

Y^1: Outcome that would be observed if treated

Y^0: Outcome that would be observed if not treated

Causal - Core concepts

Potential outcome

The outcome we would see under each possible treatment option (Y^n).

Counterfactual

Slightly different than potential outcomes, but often used interchangeably.

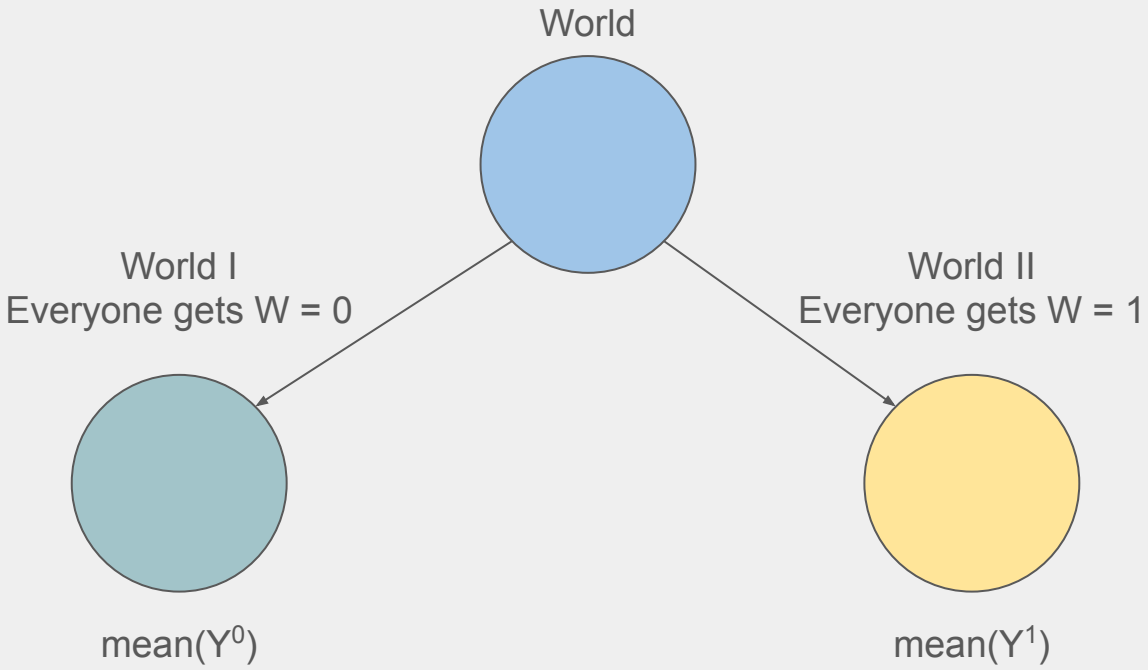
What would have happened had the action been different?

Before treatment decision is made, any outcome is a potential outcome: Y^1 or Y^0 .
After treatment there's an observed outcome Y^A and a counterfactual one Y^{1-A}.

Confounding

Anything that can impact both W and Y.

Causal - Causal Effect



Average Causal Effect = $E[Y^1 - Y^0]$

Causal - Randomized Controlled Trial (RCT)

It's almost like having two new worlds!

- Golden standard;
- Solves all our problems!
- It has its own challenges, but once solved the results are robust;
- People in academia are used to do it.

The challenge

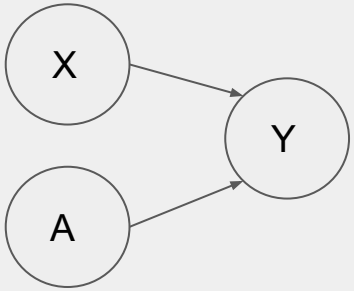
If random testing is a way to avoid all the difficulties of estimating causal effect, why do we even bother?

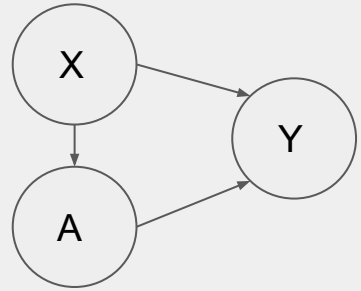
- It may not be **ethical**
- It can be **costly**

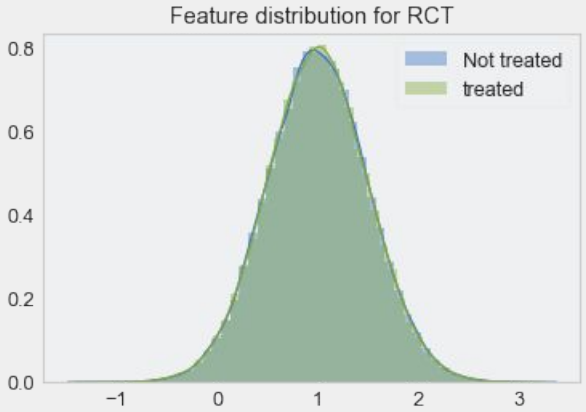
The challenge is estimating causal effect using either just **observational data** or using it with some random test data.

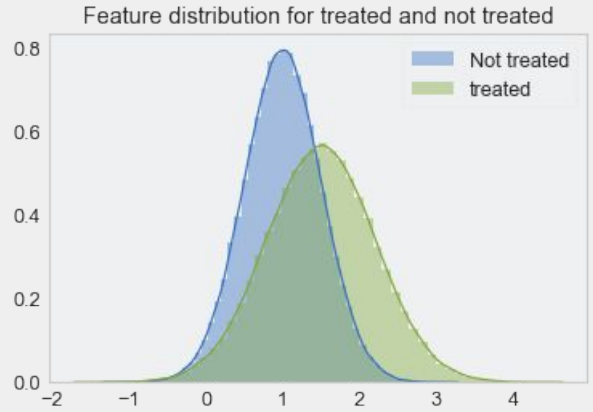
How to solve causal inference problems with observational data?

From random to observational









Causal - Assumptions

SUTVA

The outcome Y depends only on the individual features. No interaction/interference between individuals.

Consistency

The observed outcome under the treatment W must match the potential outcome $Y = Y^W$

Conditional Ignorability

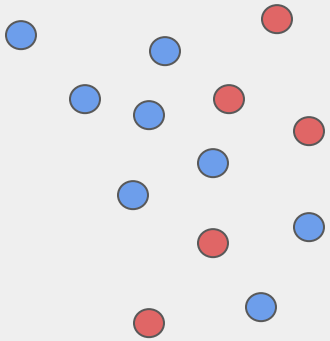
No unknown confounders: from the $Y^1 \perp Y^0 \mid W$ in RCT, to the $Y^1 \perp Y^0 \mid W, X$ in observational studies

Positivity

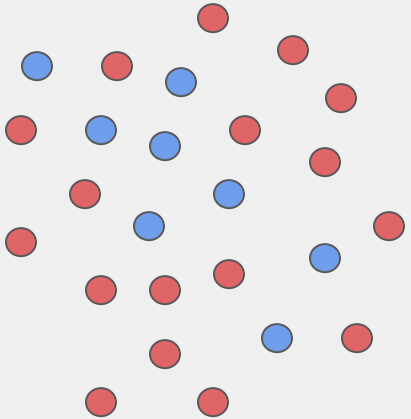
The chance of being treated is positive: $P[W = 1 \mid X = x] > 0$ for all x. The treatment can't be deterministic.

Matching

Treated

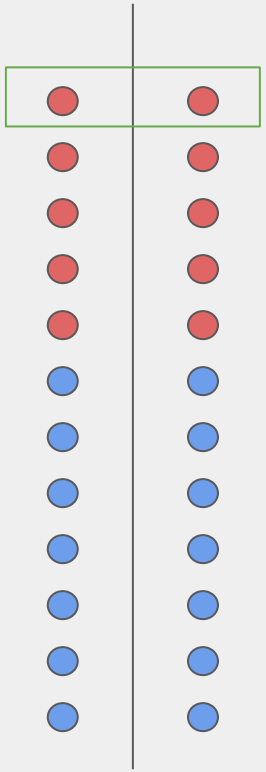


Not Treated

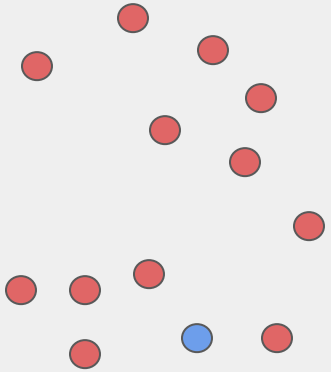


Matching

Treated



Not Treated



IPTW: Inverse Probability of Treatment Weighting

Treated

P(A = 1 | X = 1):

P(A = 1 | X = 0):

Not Treated

P(A = 0 | X = 1):

P(A = 0 | X = 0):

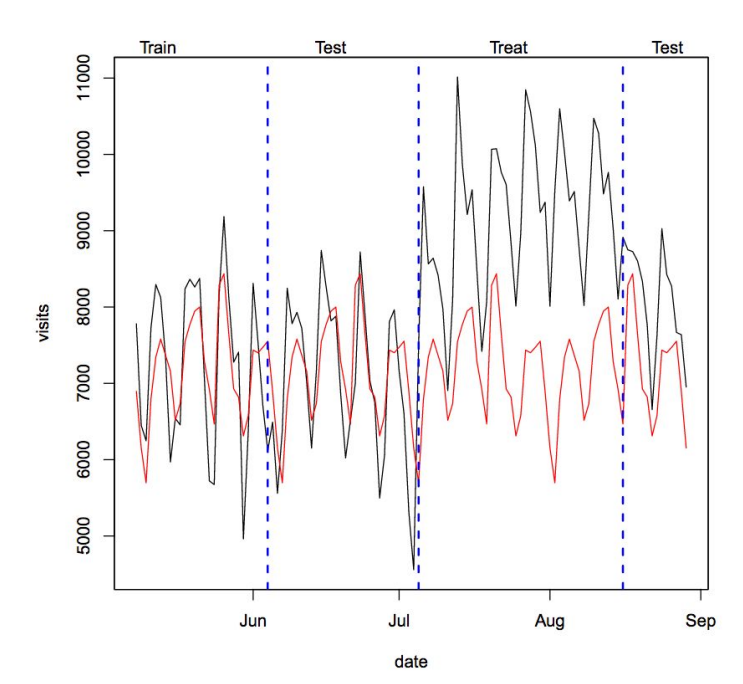
IPTW

Creates a pseudo-population where treatment assignment no longer depends on X. There's no confounding now.

$$Weight = \frac{1}{Pr[A=a|X=l]}$$

Causal Inference tricks!

Causal Inference - Train Test Treat Compare



Reference: Causal inference in economics and marketing by Hal R. Varian

Causal Inference - Regression Discontinuity

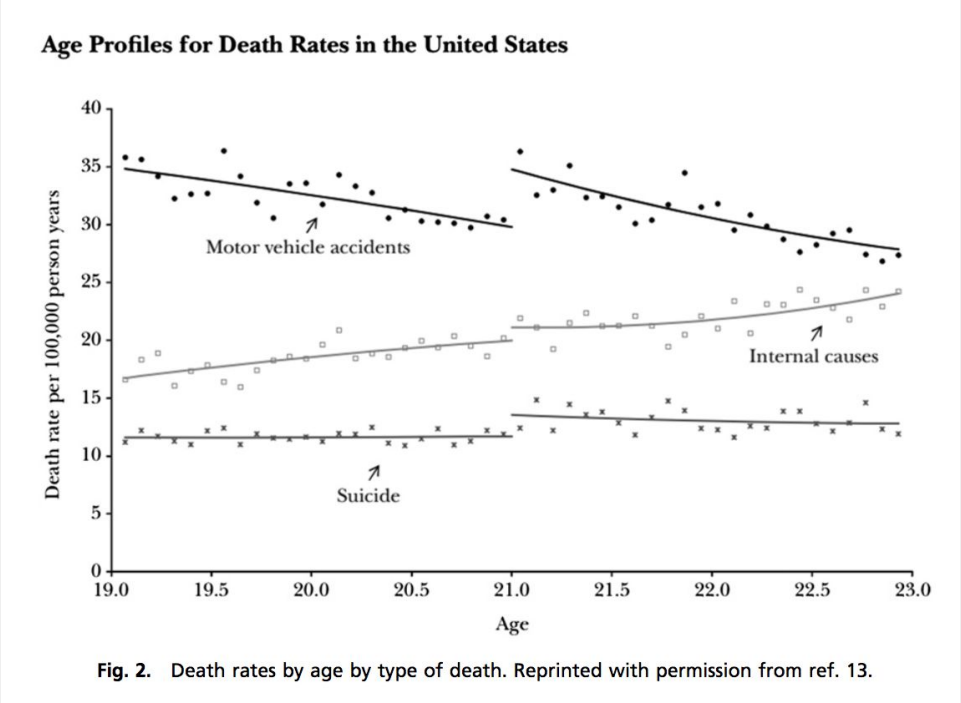


Fig. 2. Death rates by age by type of death. Reprinted with permission from ref. 13.

Reference: Causal inference in economics and marketing by Hal R. Varian

Causal Inference - Diff-in-diff

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}	$s_{TB} + (s_{CA} - s_{CB})$

Reference: Causal inference in economics and marketing by Hal R. Varian

Machine Learning + Causal Inference: Causal Forest

Causal Forest - What is it about?

Goal: **Heterogeneous treatment effect** using **observational data**, estimating the effect on individuals rather than the average for the whole population or subgroups.

How: trying to learn the causal effect by **grouping similar observations** in the same leaf and comparing the treated and untreated.

Why it's interesting: for decision making in causal inference problems you need confidence intervals since you can't validate in the data.

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - 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.

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - From CART to Causal

- **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:** $\hat{\tau}(x) = B_{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - Learning

1) Draw a random subsample of size s from {2, ... , n} without replacement, and then divide into two disjoint sets of size I and J, both of size s/2;
2) Grow a tree via recursive partitioning. The splits are chosen using any data from the J sample, but without using Y-observations from the I-sample;
3) Estimate leaf-wise responses using only the I-sample observations.

The splits are done maximizing the variance of the estimated effect using the J sample. **Each leaf should contain k or more I-sample observations of each treatment class.**

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - Learning

Estimation on the leaf using:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1,\, X_i \in L\}|} \sum_{\{i:W_i=1,\, X_i \in L\}} Y_i \; - \; \frac{1}{|\{i : W_i = 0,\, X_i \in L\}|} \sum_{\{i:W_i=0,\, X_i \in L\}} Y_i. \quad (5)$$

The splits maximize the estimation variance for each example in J.

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - What is happening inside it?

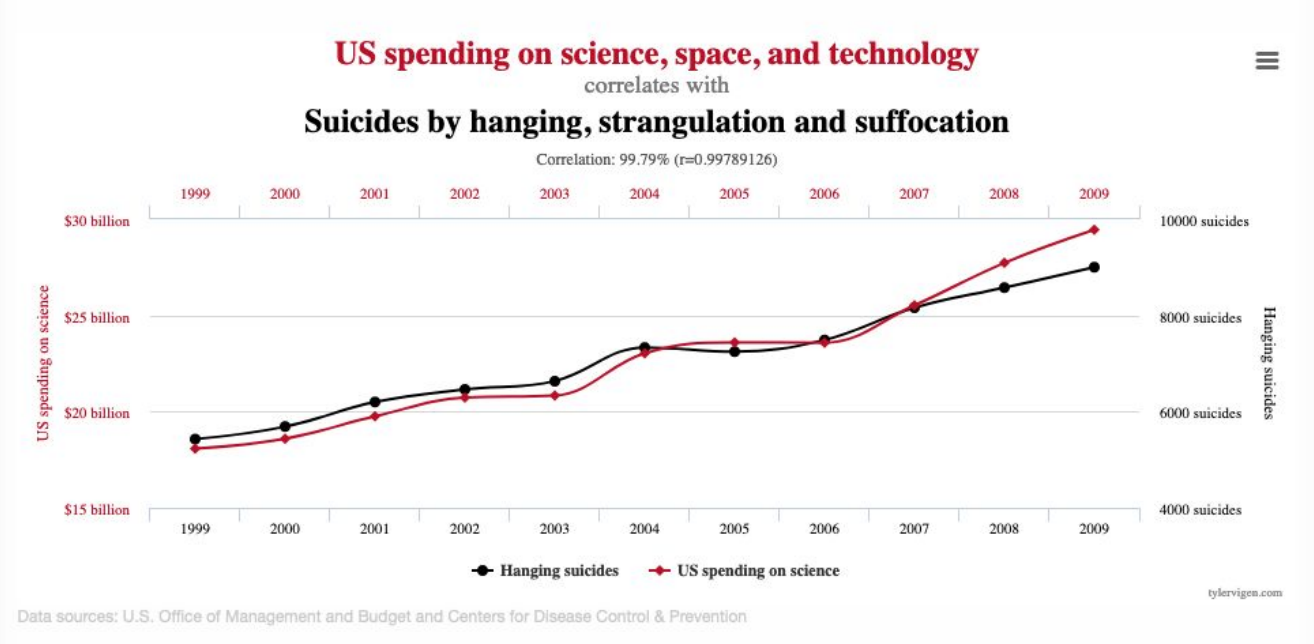
- The estimation in the leafs addresses the effect of treatment;
- The idea is that in each leaf it behaves like a random experiment in a sub group
- The restriction of having k or more examples of each treatment helps to make it closer to a random experiment (both classes equally represented) and also to prevent overfitting (at least k examples);
- We maximize the variance so it's meaningful to split into two groups for a certain feature value, it worths treating them separately;
- The more the treatment is far from a RCT, the harder it’s to work with a small k, because it may be hard to find treated and untreated examples for very specific splits (a certain space in the features space).

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/sub groups examples.

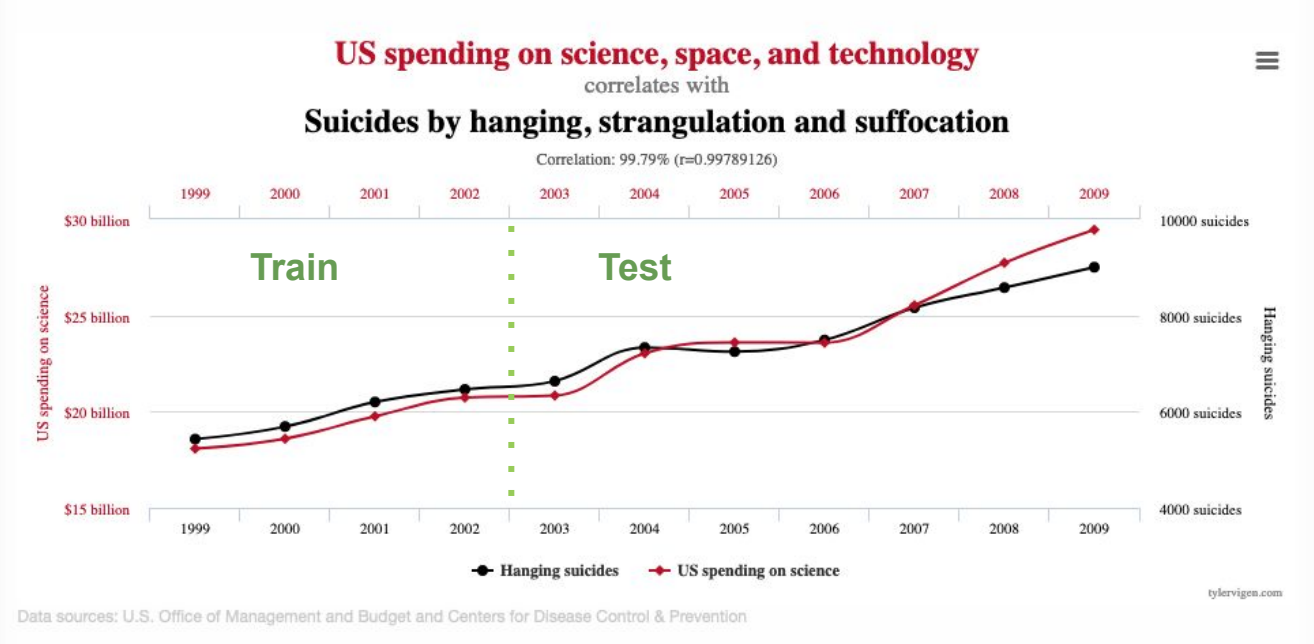
Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Predictive Machine Learning Limitations

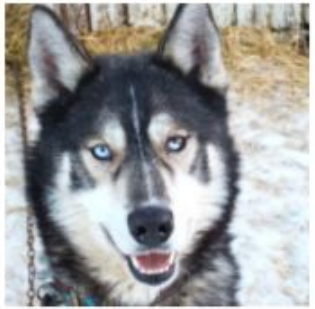
Spurious correlation



Spurious correlation



Spurious correlation



(a) Husky classified as wolf

(b) Explanation

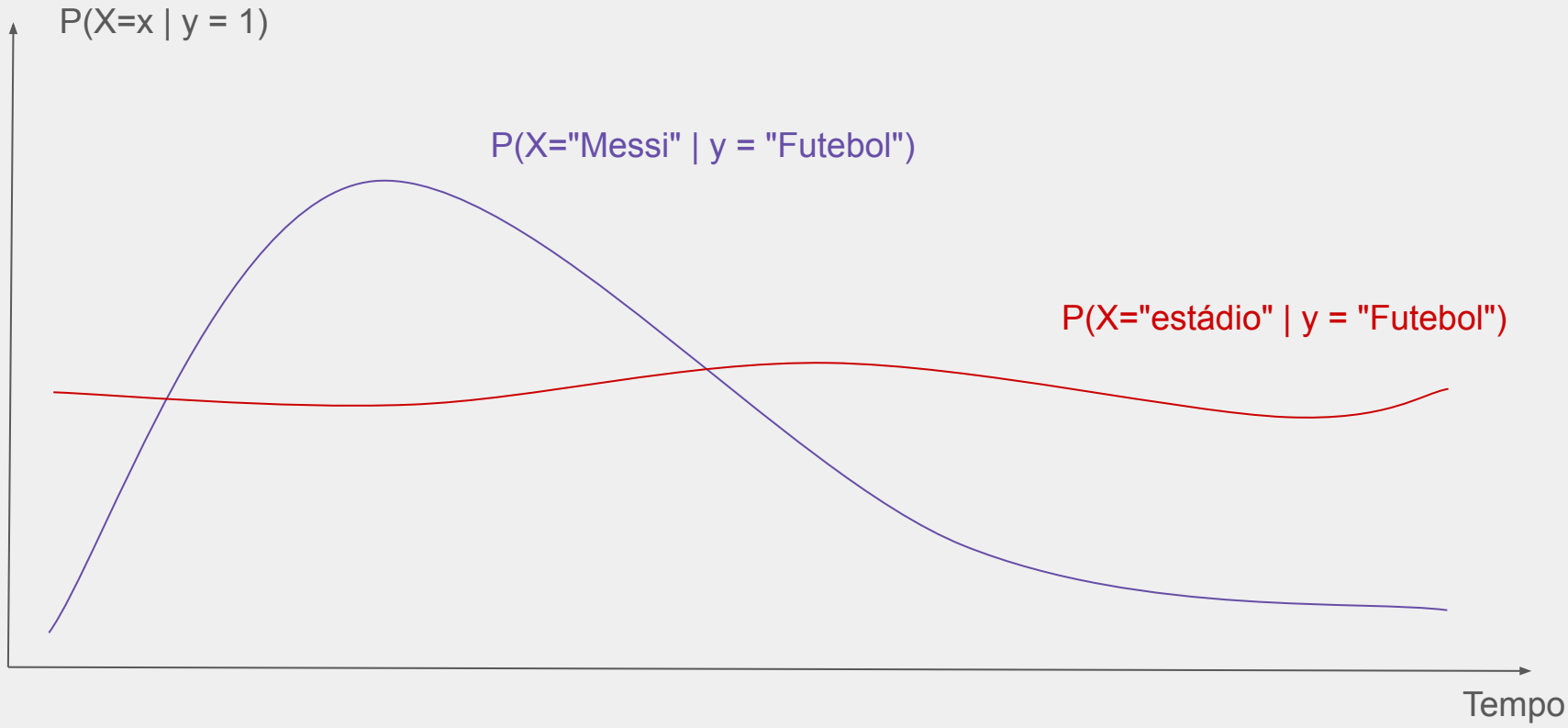
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Ribeiro et al. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD16.

Concept drift



Invariance

Causal invariance

A is a city's altitude, *T* is the average year temperature and we have a sample for a couple of countries.

$$p(a,t) = p(a \mid t)p(t) \qquad \text{T} \rightarrow \text{A}$$

$$= p(t \mid a)p(a) \qquad \text{A} \rightarrow \text{T}$$

How would we know the right direction?

Causal invariance

Now we add an identifier to each country:

$$p^{Austria}(a,t) = p^{Austria}(t \mid a)p^{Austria}(a)$$
$$p^{Suíça}(a,t) = p^{Suíça}(t \mid a)p^{Suíça}(a)$$

Hypothesis: physics is invariant on different contexts

$$p^{Austria}(t \mid a) = p^{Suíça}(t \mid a) = p(t \mid a)$$

Causal invariance - method

Assuming an Additive Noise model:

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

Where Y and the noise N are independent.

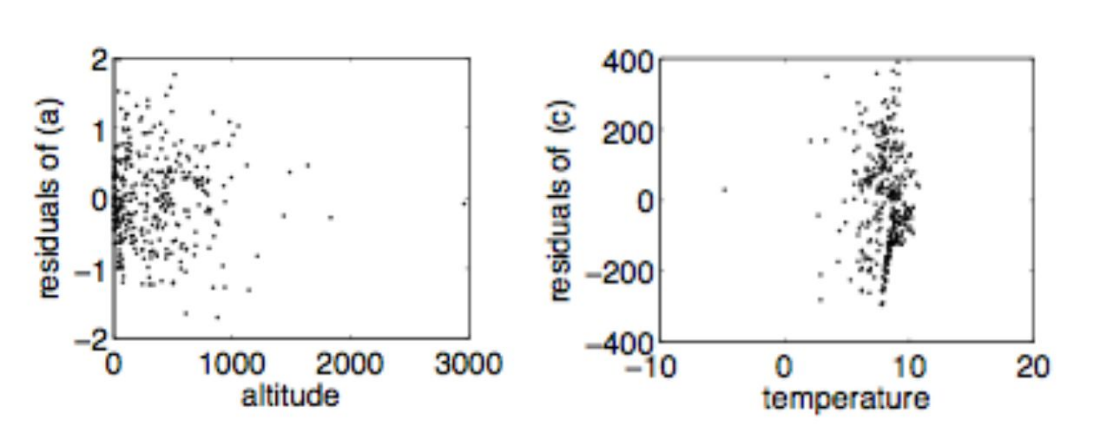
Then, there's no such model that $X = g(Y) + N_x$

And X is independent of N.

Hoyer et al.: Nonlinear causal discovery with additive noise models.NIPS21, 2009
Peters et al: Causal Discovery with Continuous Additive Noise Models, JMLR 2014
Peters et al.: Detecting the Direction of Causal Time Series.ICML2009

Causal invariance - method

1. Fit a function f as a non-linear model of X on Y (assumption of noise additive model)
2. Compute the residual N = Y - f(X)
3. Check whether N and X are statistically independent



Indeed, there's a strong dependence in the anti causal direction when we look to real data.

Invariant Causal Prediction - ICP

Invariant Causal Prediction - ICP

We're going to search for a subset of features that present a stable relationship with the target.

Hipótese 1 (predição invariante): *Existe um vetor de coeficientes* $\gamma^* = (\gamma_1^*, ..., \gamma_p^*)^t$ *com suporte* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, ..., p\}$ *que satisfaz*

$$\text{para todo } e \in \boldsymbol{\varepsilon} : X^e \text{ tem uma distribuição arbitrária e}$$
$$Y^e = \mu + X^e\gamma^* + \epsilon^e, \epsilon^e \sim F_\epsilon \text{ e } \epsilon^e \perp X^e_{S^*}, \tag{4}$$

onde $\mu \in \mathbb{R}$ *é o intercepto,* ϵ^e *é um ruído aleatório com média zero, variância finita e a mesma distribuição* F_ϵ *para todo* $e \in \boldsymbol{\varepsilon}$.

Invariant Causal Prediction

For every possible subset of features:

1) Train with all data	2) Test the residues	3) Final model with stable variables
Train a model using the data from all the context and the subset of features. Calculate the residues.	Test for every context if the residues on them have the same mean and variance than each other context. Do a joint test to decide rejecting or not that subset.	In the final subset, we use the intersection of all the not rejected subsets from the previous step.

Invariant Causal Prediction

$$
\begin{aligned}
x_1 &\sim \mathbb{N}(0, \sigma(e)) \\
y &\sim x_1 + \mathbb{N}(0, 1) \\
x_2 &\sim y + \mathbb{N}(0, 1)
\end{aligned}
$$

Result:

$$S = \{x_1\}$$

$$
\begin{aligned}
x_1 &\sim \mathbb{N}(0, \sigma(e)) \\
x_3 &\sim \mathbb{N}(0, \sigma(e)) \\
y &\sim x_1 + 2x_3 + \mathbb{N}(0, \sigma(e)) \\
x_2 &\sim y + \mathbb{N}(0, 1)
\end{aligned}
$$

Result:

$$S = \{x_1, x_3\}$$

Invariant Risk Minimization - IRM

Invariant Risk Minimization - IRM

The objective function is modified to reflect the preference for a model which is optimal under different contexts.

ERM

$$L_{ERM}(\beta) = R(\beta)$$

IRM

$$L_{IRM}(\Phi, \omega) = \sum_{e \in \varepsilon_{tr}} R^e(\omega \circ \Phi) + \lambda \mathbb{D}(\omega, \Phi, e)$$

Linear regression case

$$L_{IRM}(\Phi, \omega) = \sum_{e \in \varepsilon_{tr}} R^e(\omega \circ \Phi) + \lambda \mathbb{D}(\omega, \Phi, e)$$

$$R^e(\omega \circ \Phi) = R^e(\hat{\beta}) = \frac{1}{n}(X^e\hat{\beta} - y^e)(X^e\hat{\beta} - y^e)^T$$

$$\mathbb{D}_{lin}(\omega, \Phi, e) = \left\| \mathbb{E}_{X^e}[(X^e\hat{\beta})^T(X^e\hat{\beta})]\omega - \mathbb{E}_{X^e,Y^e}[(X^e\hat{\beta})^T Y^e] \right\|^2$$

Linear regression case

Gradient descent: $\hat{\beta}_{t+1} = \hat{\beta}_t - \gamma \nabla_{\beta|\beta=\hat{\beta}_t} L(\beta)$

$$L_{IRM}(\Phi, \omega) = \sum_{e \in \varepsilon_{tr}} R^e(\omega \circ \Phi) + \lambda \mathbb{D}(\omega, \Phi, e)$$

$$-\frac{1}{N} X^T (X\hat{\beta} - y)$$

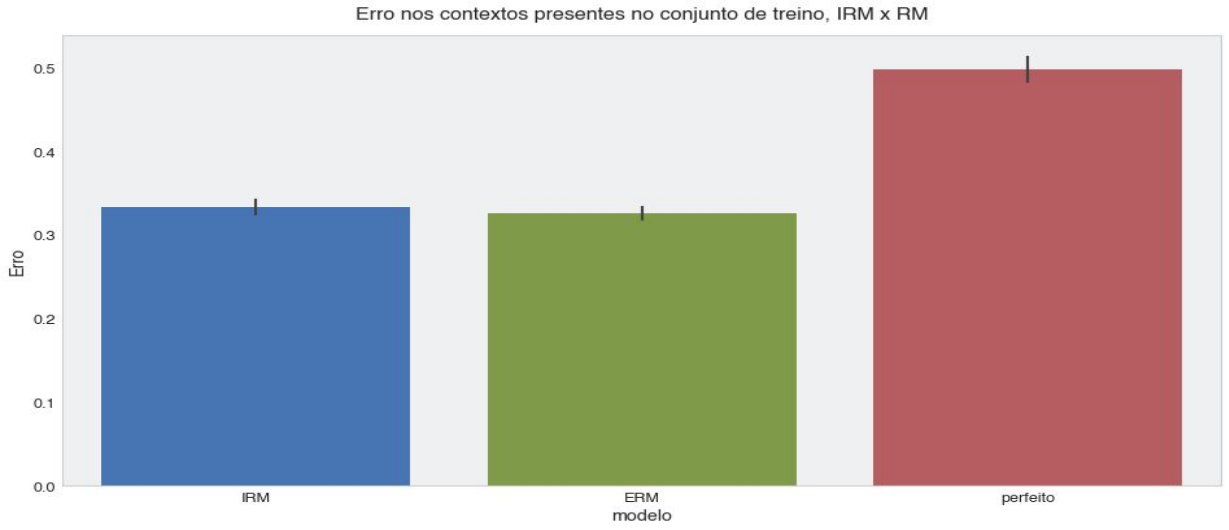
$$4\frac{1}{N}(((X^e\hat{\beta})^T X^e\hat{\beta}) - ((X^e\hat{\beta})^T y))$$
$$(X^e(X^e\hat{\beta})^T - ((X^e\hat{\beta})^T y^e))$$

Implementing paper example

Applying the previous equations for the linear regression case and using the same set of equations used in the paper as the motivational example.

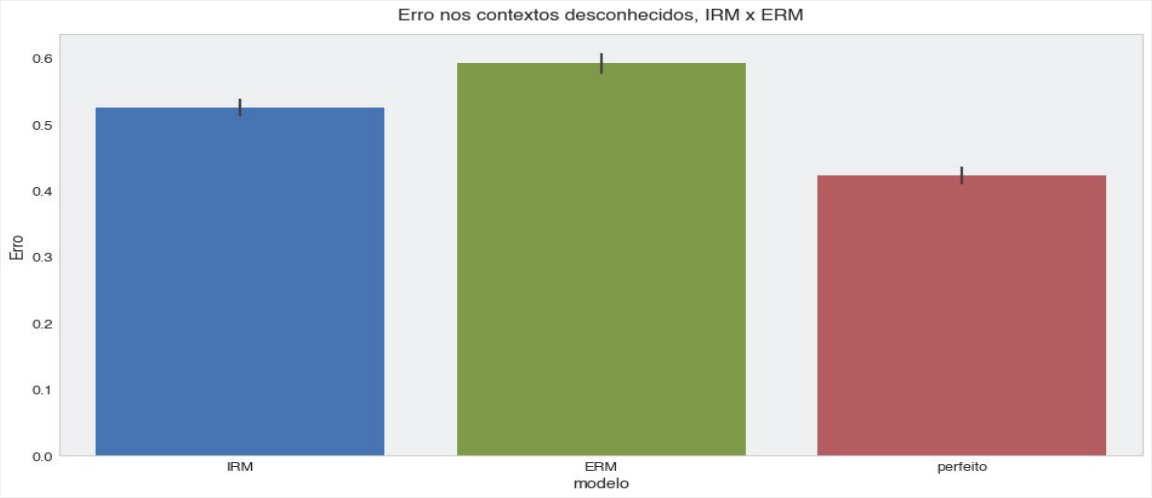
$$x_1 \sim \mathbb{N}(0, \sigma(e))$$
$$y \sim x_1 + \mathbb{N}(0, \sigma(e))$$
$$x_2 \sim y + \mathbb{N}(0, 1)$$

Results



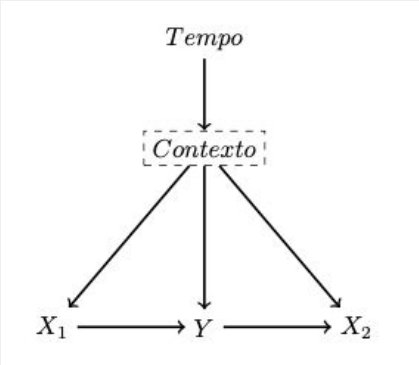
It's a test set, but the contexts are present in the training sample.

Results


Erro nos contextos desconhecidos, IRM x ERM
Erro
0.0
0.1
0.2
0.3
0.4
0.5
0.6
IRM
ERM
perfeito
modelo

Time, invariance and the Time Tree

Time and data



Tempo	Contexto
1	1
2	1
3	1
4	2
5	2
6	3
7	4
8	4
9	4
10	5

	period	x_1	y	x_2
0	1	2.470984	-15.570138	-17.971118
1	1	40.614529	4.154478	5.154888
2	1	-45.239741	-67.811901	-67.205011
3	1	-25.560394	-68.037824	-68.738988
4	1	50.446213	54.842746	53.820852

Time tree

It's very close to a decision tree using the ID3 algorithm, except that:

1) We need to decide about a time granularity to group our data (hour, days, week, month, year?)
2) A hyper parameter controls the minimum sample required in every time split on the leafs
3) When deciding on the best split, the gain is calculated by averaging the gain on all periods

By forcing examples from all periods to be present in every leaf and optimizing the average gain, we expect to force the algorithm the requirement of learning invariant in time rules.

Real data

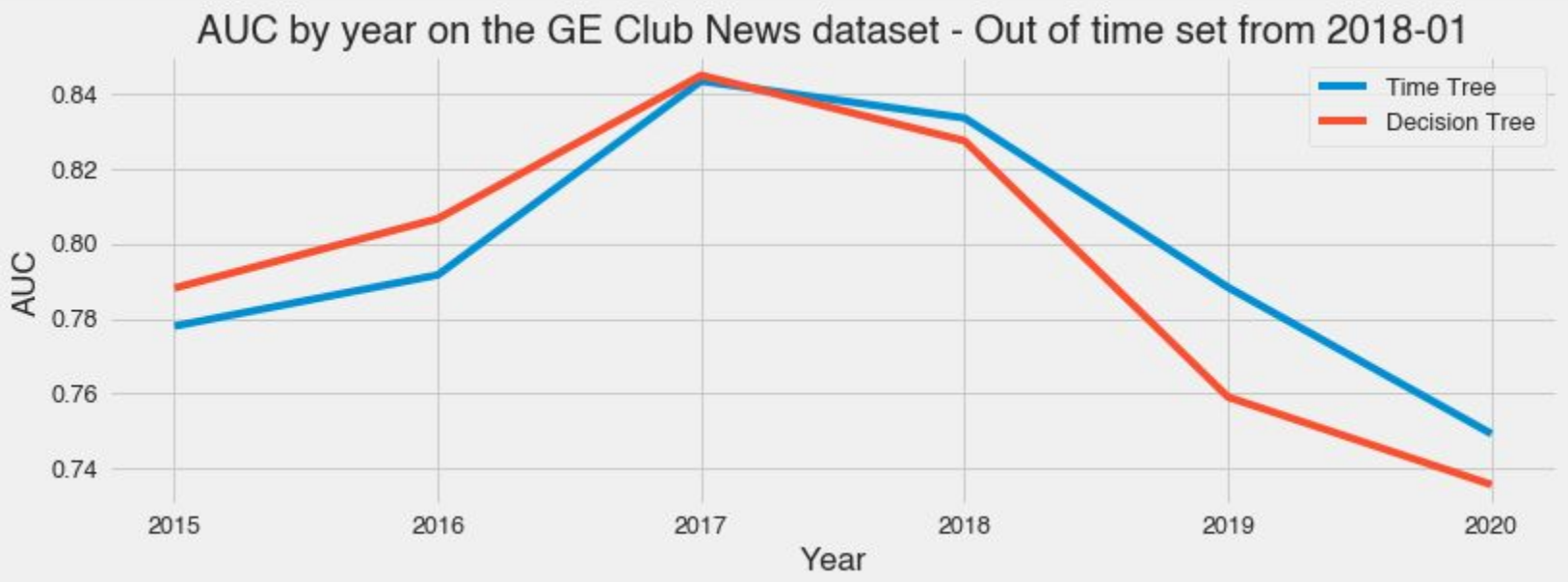
Data

- Data was extracted from GloboEsporte.com
- News from soccer clubs
- From June 2015 to November 2020
- All clubs from first league
- The label is the club name from which the page we extracted it
- Transform in a binary classification by considering one team as the positive case (Flamengo in our case)

Experiment

- Split data in time, train from 2015 to 2017 and evaluate on 2018 to 2020
- Compare time tree to decision tree

Results


AUC by year on the GE Club News dataset - Out of time set from 2018-01
Time Tree
Decision Tree
AUC
0.84
0.82
0.80
0.78
0.76
0.74
2015
2016
2017
2018
2019
2020
Year

Questions?

Contact

Twitter: @lgmoneda
E-mail: lgmoneda@gmail.com