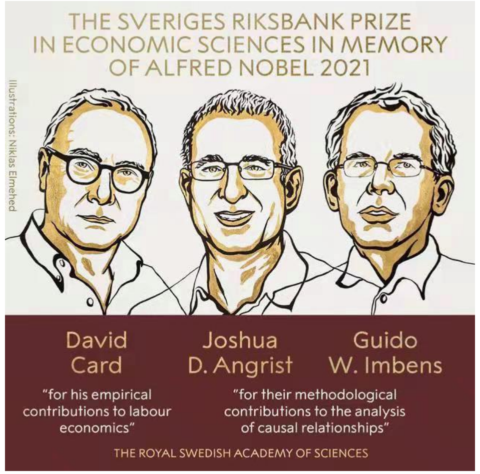
Causal AI

CausalAI 核心算法:
因果森林及其相关研究

by 龚鹤扬@1587causalai

Why Causal Forest?

- What if 我是 Imbens



Why Causal Forest?

因果树(Causal Tree）夫妻合作文章, PANS 美国科学院院报

Recursive partitioning for heterogeneous causal effects

Susan Athey[a,1] and Guido Imbens[a]

[a]Stanford Graduate School of Business, Stanford University, Stanford, CA 94305

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 20, 2016 (received for review June 25, 2015)

[Submitted on 5 Apr 2015 (v1), last revised 30 Dec 2015 (this version, v3)]

Why Causal Forest?

- 增益建模中的重要算法
- 论文 “[WWW 2022 | 结合高效整数规划求解，快手提出多元因果森林模型，智能营销效果显著]”

LBCF: A Large-Scale Budget-Constrained Causal Forest Algorithm

Meng Ai[1], Biao Li[1], Heyang Gong[1,2], Qingwei Yu[1], Shengjie Xue[1], Yuan Zhang[1], Yunzhou Zhang[1], Peng Jiang[1]
[1]Kuaishou Inc., China
[2]University of Science and Technology of China, China
{aimeng, libiao, gongheyang03, yuqingwei, xueshengjie, zhangyuan13, zhangyunzhou, jiangpeng}@kuaishou.com

Causal Forest 相关文献

[Recursive Partitioning for Heterogeneous Causal Effects - arXiv]

[Submitted on 5 Apr 2015 (v1), last revised 30 Dec 2015 (this version, v3)]

[Estimation and Inference of Heterogeneous Treatment Effects ...]

[Submitted on 14 Oct 2015 (v1), last revised 10 Jul 2017 (this version, v4)]

[Generalized Random Forests - arXiv]

[Submitted on 5 Oct 2016 (v1), last revised 5 Apr 2018 (this version, v4)]

[Machine Learning Methods Economists Should Know About]

[Submitted on 24 Mar 2019]

因果树简要介绍

- 因果算法的理解
- 因果树应用 EndToEnd toy example

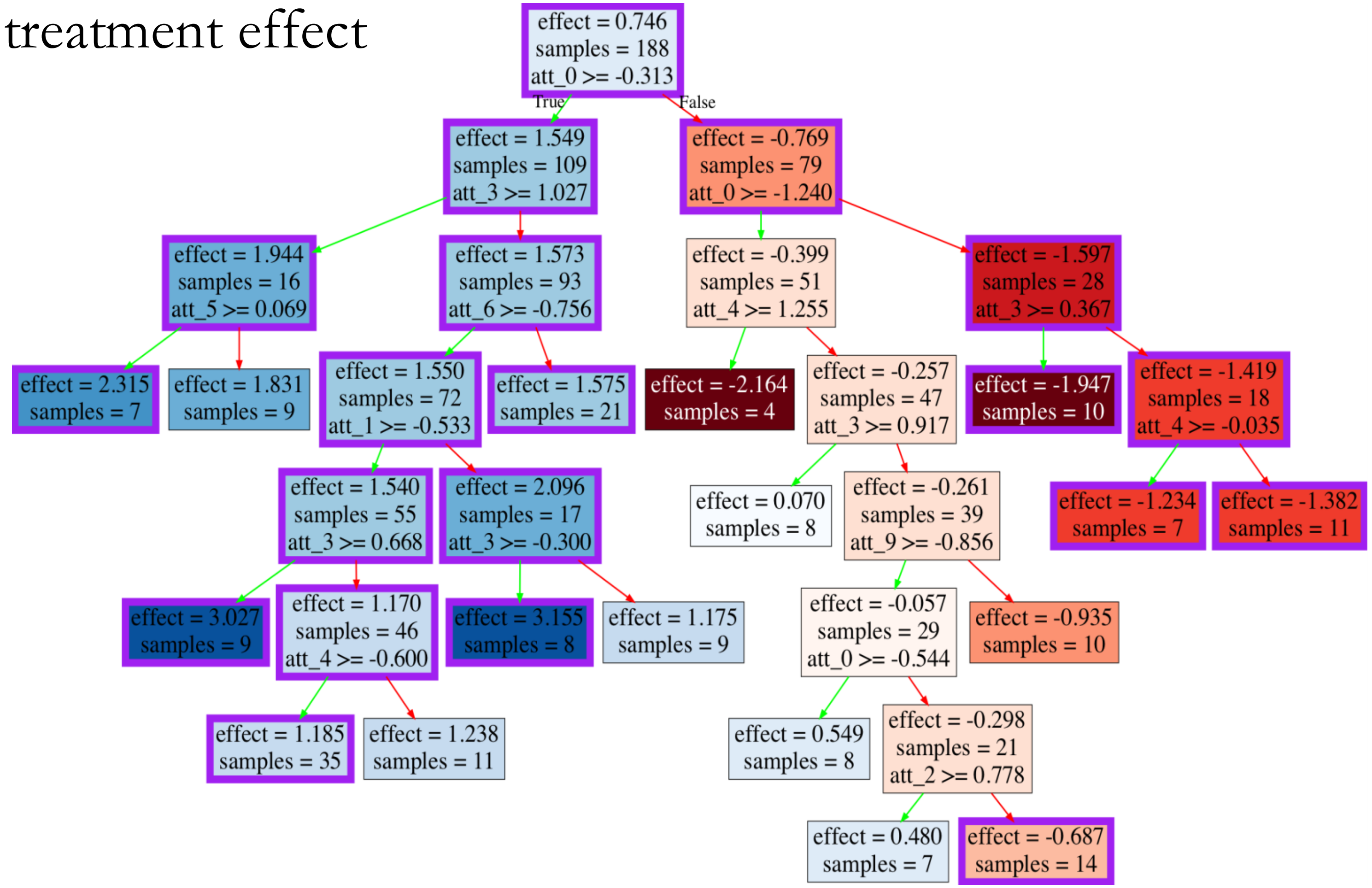
因果树模型是什么?

Estimator for Conditional average treatment effect

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$$

分裂准则是:

$$\Delta = \frac{1}{\#children} \sum_{k=1}^{\#children} \hat{\tau}_k^2$$



怎么用：一个简单的代码实现

[Quasi-oracle estimation of heterogeneous treatment effects]

$$y = w \cdot \tau(X_i) + b(X_i) + \epsilon_i \tag{1}$$

$$X_i \sim N(0, I_{d\times d}), e(X_i) = 0.5 \tag{2}$$

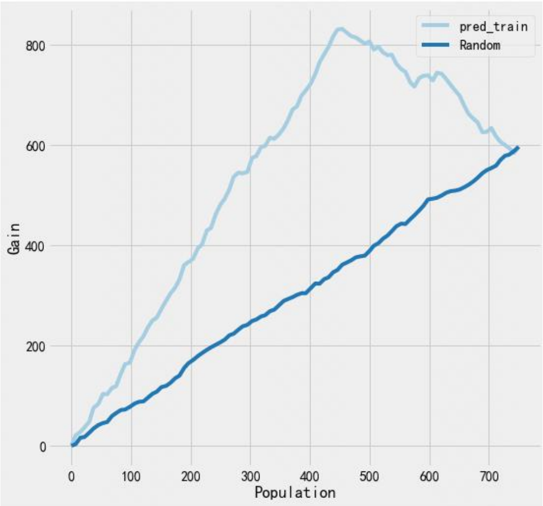
$$b(X_i) = \max\{X_{i1} + X_{i2}, X_{i3}, 0\} + \max\{X_{i4} + X_{i5}, 0\}$$

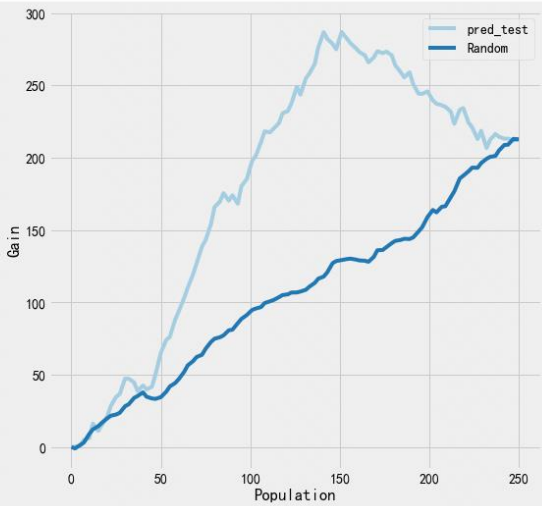
$$\tau(X_i) = X_{i1} + \log(1 + e^{X_{i1}})$$

```
import pandas as pd
import CTL
from CTL.causal_tree_learn import CausalTree
from sklearn.model_selection import train_test_split
import numpy as np
from causalml.metrics.visualize import *

n, d = 1000, 10
x = np.random.randn(n, d)
w = np.random.binomial(1, 0.5, n)
eps = np.random.randn(n, 1)
b = [max([x[i][0]+x[i][1],x[i][2],0]) + max([x[i][3]+x[i][4],0]) for i in range(x.shape[0])]
tao_x = np.array([x[i][0] + np.log(1+ np.exp(x[i][0])) for i in range(x.shape[0])])
y = w*tao_x + b
x_train, x_test, y_train, y_test, treat_train, treat_test = train_test_split(x, y, w)

# CTL及其评估
ct = CausalTree(honest=True)
ct.fit(x_train, y_train, treat_train)
ct.prune()

def eval_uplift_ranking(uplift_model, train_data, test_data, features=None):↔
def numpy_to_dataframe(x_train, y_train, treat_train, cols=None):↔

train_data = numpy_to_dataframe(x_train, y_train, treat_train)
test_data = numpy_to_dataframe(x_test, y_test, treat_test)
eval_uplift_ranking(ct, train_data, test_data)
```

```
Auuc of train is 0.9001237265458116, and test is 0.8428228466526692
```

模型的学习能力非常有限





因果树模型的主要贡献是什么?

- Rzepakowski and Jaroszewicz (2012) 的增益树模型的欧式距离等价于因果树模型

$$\Delta_{gain} = D_{after\_split}(P^T, P^C) - D_{before\_split}(P^T, P^C)$$

Kullback, Euclidean and Chi-Squared, defined as:

$$KL(P:Q) = \sum_{k=Left,Right} p_k log\frac{p_k}{q_k}$$

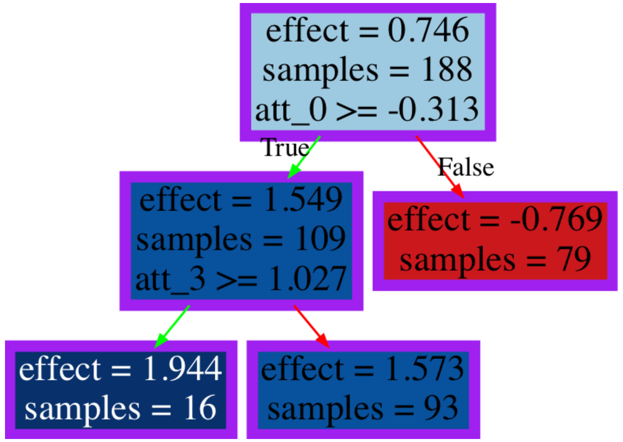
$$E(P:Q) = \sum_{k=Left,Right} (p_k - q_k)^2$$

$$\chi^2(P:Q) = \sum_{k=Left,Right} \frac{(p_k - q_k)^2}{q_k}$$

主要贡献是 the "honest" approach

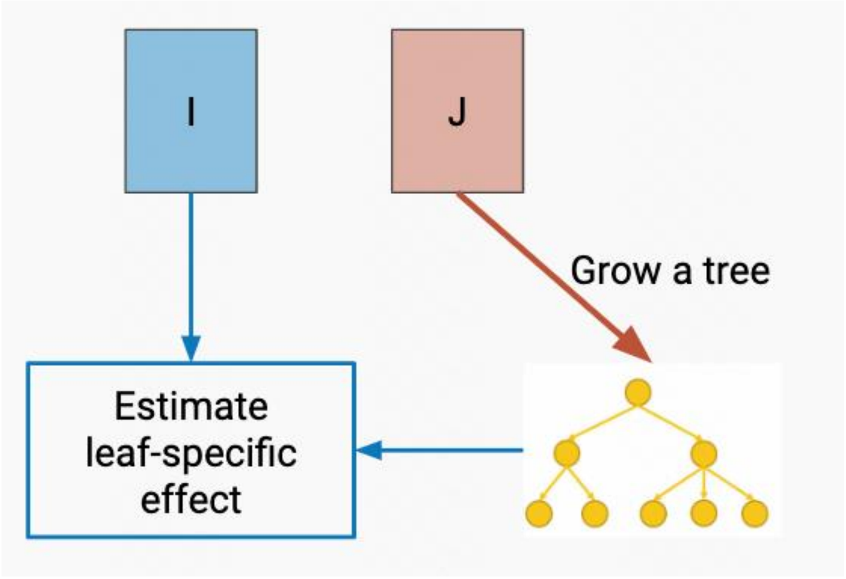
The Honest Approach 是什么?

- 因果树模型两个核心问题:
  - 1）如何分割特征空间，
  - 2）如何估计子空间上的因果效应



$$\Delta = \frac{1}{\#children} \sum_{k=1}^{\#children} \hat{\tau}_k^2$$

$$\hat{\tau}^{CT}(X_i) = \frac{\sum_{j:X_j \in \mathbb{X}_l} Y_i^{\text{obs}} \cdot W_i / \hat{e}(X_i)}{\sum_{j:X_j \in \mathbb{X}_l} W_i / \hat{e}(X_i)} - \frac{\sum_{j:X_j \in \mathbb{X}_l} Y_i^{\text{obs}} \cdot (1 - W_i)/(1 - \hat{e}(X_i))}{\sum_{j:X_j \in \mathbb{X}_l} (1 - W_i)/(1 - \hat{e}(X_i))}.$$



一半的样本用于树的生成，一半的样本用于树的叶子结果预测

The Honest Approach vs. The Adaptive Approach

train data

estimation data

学习算法
1) partition
2) estimates

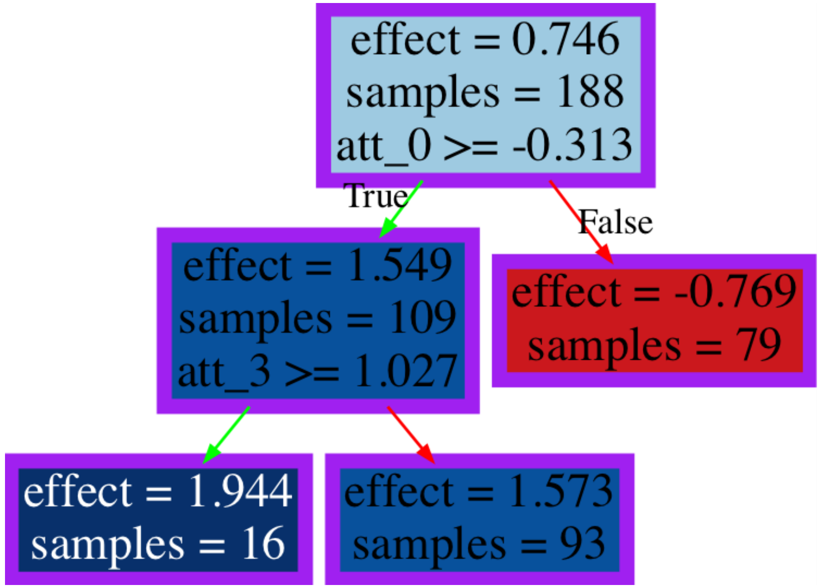
$$Q^H(\pi) \equiv -E_{S^{te}, S^{est}, S^{tr}}[MSE_\mu(S^{te}, S^{est}, \pi(S^{tr}))]$$

train data

train data

学习算法
1) partition
2) estimates

$$Q^C(\pi) \equiv -\mathbb{E}_{\mathcal{S}^{te}, \mathcal{S}^{tr}}[\mathrm{MSE}_\mu(\mathcal{S}^{te}, \mathcal{S}^{tr}, \pi(\mathcal{S}^{tr}))]$$



cannot be used directly for constructing confidence intervals because the methods are "adaptive":
"honest" 可以看作一种 cross validdation （被用于控制模型复杂度）的推广，见其论文 arxiv v2

为什么honesty instead of cross-validation?

• Crucially, we anticipate that second-stage estimates of treatment effects will be unbiased in each leaf, because they will be performed on an independent sample.

• Through a simulation study, we show that for our preferredmethod honest estimation results in nominal coverage for 90% con-fidence intervals, whereas coverage ranges between 74% and 84%for nonhonest approaches.

因此因果森林具备良好的统计性质

因果森林算法简单介绍

主要贡献：

To our knowledge, this is the first set of results that allows any type of random forest, including classification and regression forests, to be used for provably valid statistical inference.

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0, 1) \ \textit{for a sequence}\ \sigma_n(x) \to 0.$$

如何理解该结果的重要性?

一些背景知识

- Leo Breiman 统计建模：两种文化
  - The statistics community has by and large accepted the Machine Learning (ML) revolution that Breiman ...
  - 从统计和机器学习的关系，反思数据科学，指出未来方向
- 机器学习 VS 统计
  - ML literature has focused heavily on out-of-sample performance
    - cross-validation

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

  - The ability to do inference
    - e.g. Consitency & Normality

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0, 1) \ \textit{for a sequence}\ \sigma_n(x) \to 0.$$

那么统计大样本性质到底重要吗?

Eeverything Starts with a research problem

- 搜索/广告/推荐
  - 推荐一个不喜欢的 item 并不会造成大影响
- 因果推断和医疗
  - ATE 估计显然需要知道置信区间

具体的领域问题决定了算法需要哪些性质

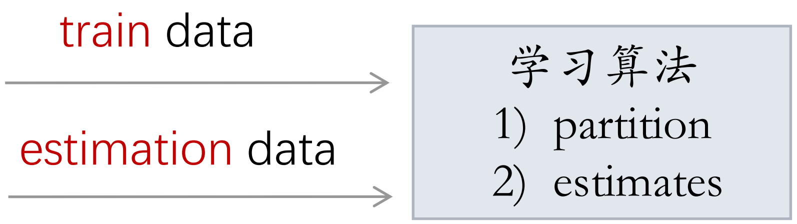
We argue that in the future, as ML tools are more widely adopted, researchers should articulate clearly the goals of their analysis and why certain properties of algorithms and estimators may or may not be important. --- Susan Athey

机器学习算法适应问题的改进

Often the ML techniques require careful tuning and adaptation to effectively address the specific problems economists are interested in. Perhaps the most important type of adaptation is to exploit the structure of the problems, e.g. the causal nature of many estimands, the endogeneity of variables, the configuration of data such as panel data, the nature of discrete choice among a set of substitutable products, or the presence of credible restrictions motivated by economic theory, such as monotonicity of demand in prices or other shape restrictions

广义随机森林算法框架介绍

- honest estimation




```
procedure GeneralizedRandomForest(set of examples 𝒮, test point x)
    weight vector α ← Zeros(|𝒮|)
    for b = 1 to total number of trees B do
        set of examples ℐ ← Subsample(𝒮, s)
        sets of examples 𝒥₁, 𝒥₂ ← SplitSample(ℐ)
        tree 𝒯 ← GradientTree(𝒥₁, 𝒳)                ▷ See Algorithm 2.
        𝒩 ← Neighbors(x, 𝒯, 𝒥₂)                      ▷ Returns those elements of 𝒥₂ that fall into
                                                        the same leaf as x in the tree 𝒯.
        for all example e ∈ 𝒩 do
            α[e] += 1/|𝒩|
    output θ̂(x), the solution to (2) with weights α/B
```

广义随机森林的临近样本估计

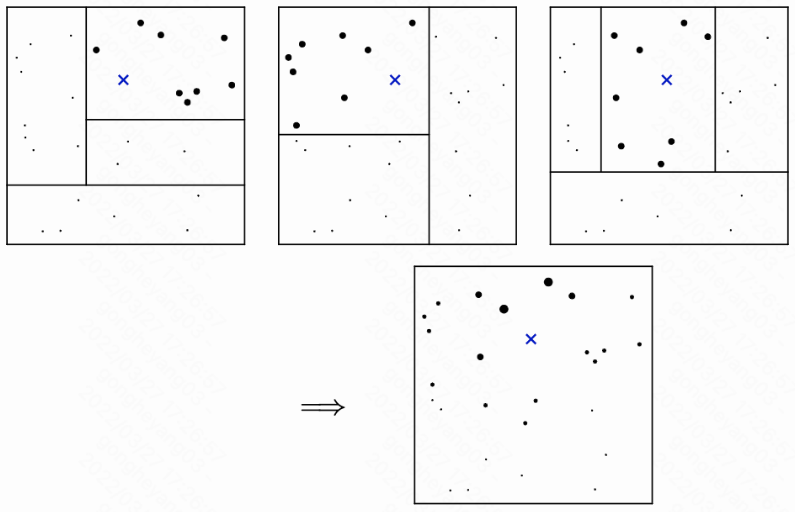


Fig 1: 基于森林的权重.

procedure GeneralizedRandomForest(set of examples $\mathcal{S}$, test point x)
 weight vector $\alpha \leftarrow$ Zeros($|\mathcal{S}|$)
 for $b = 1$ to total number of trees B **do**
  set of examples $\mathcal{I} \leftarrow$ Subsample($\mathcal{S}$, s)
  sets of examples $\mathcal{J}_1$, $\mathcal{J}_2 \leftarrow$ SplitSample($\mathcal{I}$)
  tree $\mathcal{T} \leftarrow$ GradientTree($\mathcal{J}_1$, $\mathcal{X}$) ▷ See Algorithm 2.
  $\mathcal{N} \leftarrow$Neighbors(x, $\mathcal{T}$, $\mathcal{J}_2$) ▷ Returns those elements of $\mathcal{J}_2$ that fall into the same leaf as x in the tree $\mathcal{T}$.
  for all example $e \in \mathcal{N}$ **do**
   $\alpha[e]$ += $1/|\mathcal{N}|$
 output $\hat{\theta}(x)$, the solution to (2) with weights α/B

$$\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \operatorname{argmin}_{\theta,\nu} \left\{ \left\| \sum_{i=1}^{n} \alpha_i(x)\, \psi_{\theta,\nu}(O_i) \right\|_2 \right\} \qquad (2)$$

利用得分函数进行估计，例如工具变量的结构信息及其得分函数如下:

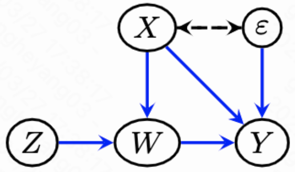
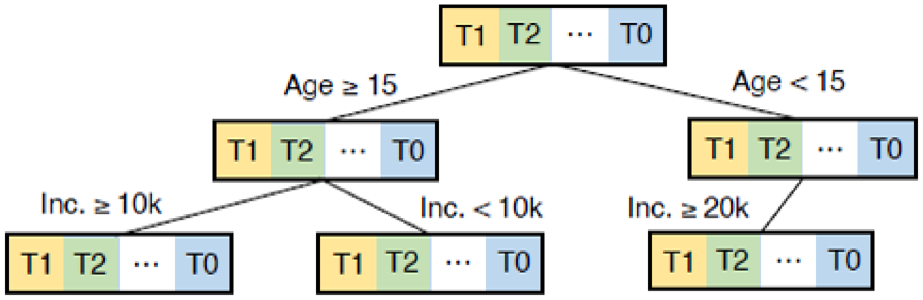


Fig 1: 工具变量法的因果图

$$\begin{cases} \tau(x) & \leftarrow E[Y_i(W_i = 1)|X_i = x] - E[Y_i(W_i = 0)|X_i = x] \\ \psi_{\tau(x),\mu(x)}(O_i) & \leftarrow (Z_i(Y_i - W_i\tau(x) - \mu(x)), Y_i - W_i\tau(x) - \mu(x)) \end{cases}$$

快手多元因果森林介绍

- 适应业务特别的改造
- 分配算法的速度优化
- 适应问题的评估方法



Unified Model

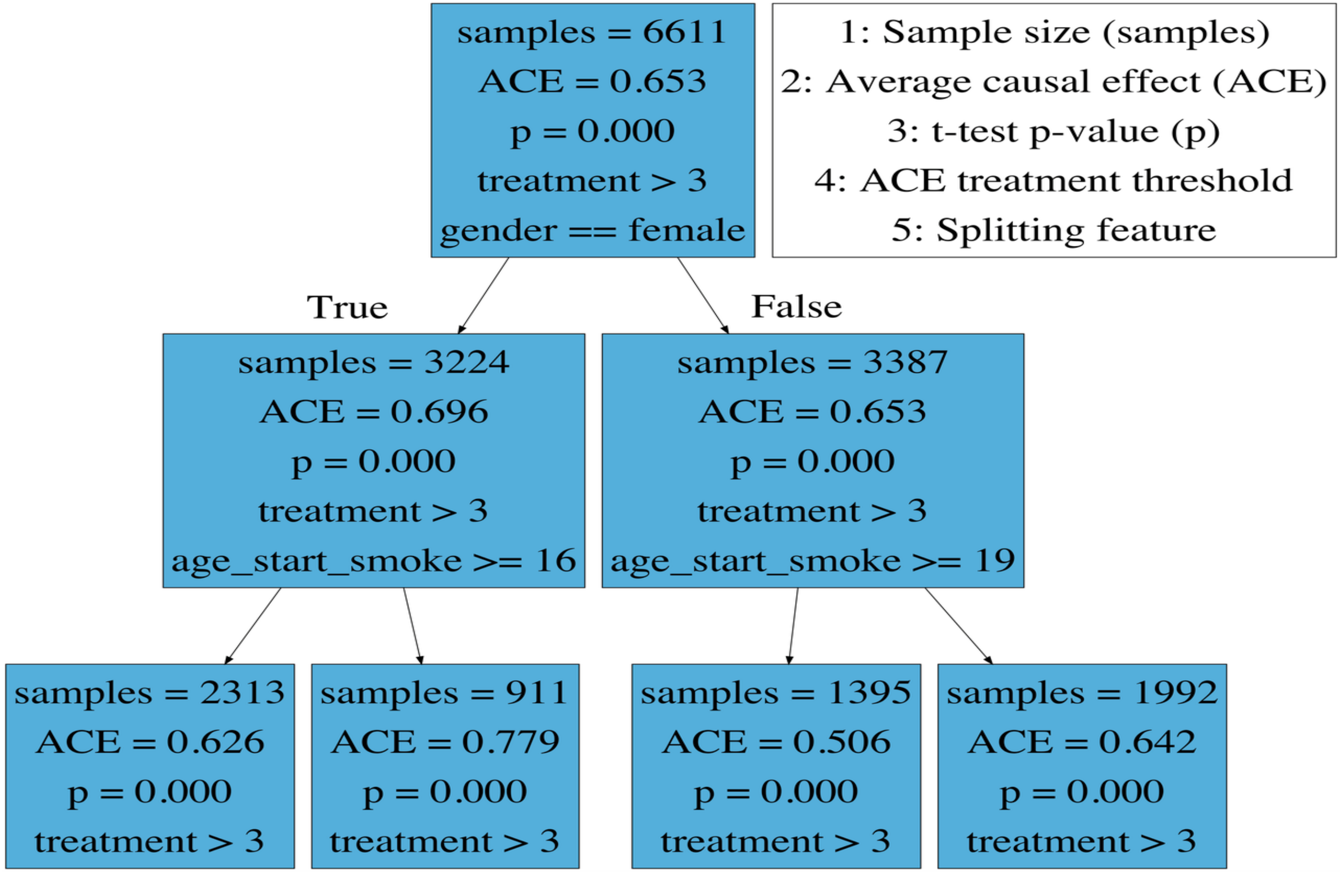
- Inter Split：叶子之间的敏感度曲线之间的异质性

$$\tilde{\Delta}_{inter}(\phi_1, \phi_2) = \sum_{l=1}^{2} \frac{1}{|\{i : \mathbf{x_i} \in \phi_l\}|} \sum_{j=1}^{K} \Big(\sum_{\{i:\mathbf{x_i} \in \phi_l\}} \rho_{ij} \Big)^2.$$

- Intra Split：敏感度曲线上不同档位的异质性

$$\tilde{\Delta}_{intra}(\phi_1, \phi_2) = \sum_{l=1}^{2} \sum_{j=1}^{K} (\hat{\theta}_{\phi_l}^{(j)} - \bar{\theta}_{\phi_l})^2,$$

Learning Triggers for Heterogeneous Treatment Effects



Tran C, Zheleva E. Learning triggers for heterogeneous treatment effects[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 5183-5190.

其他相关算法

- Evidence-Based Policy Learning（CLeaR 2022 Oral）
  - An algorithm to find subgroups with statistically significant treatment effects in randomized-trial data
- Generalized Causal Tree for Uplift Modeling
- A tree-based federated learning approach for personalized treatment effect estimation from heterogeneous data sources
- Causal transfer random forest: Combining logged data and randomized experiments for robust prediction
- ……

从因果树到广义随机森林

- Causal Tree
  - honest estimation: 非训练样本估计模型的局部参数 $\mu_L(\cdot)$
- Causal Forest
  - An estimator with provably valid statistical inference

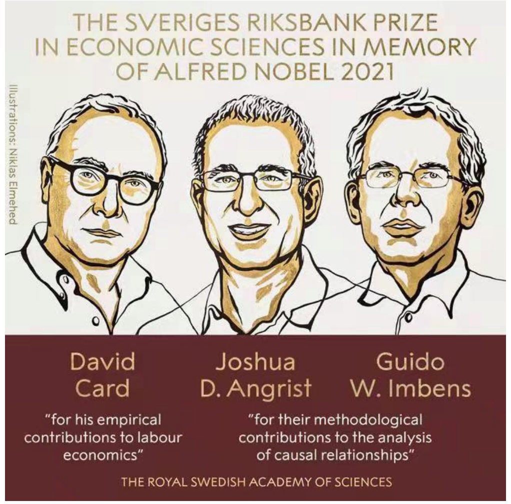
$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0,\,1)\ \ for\ a\ sequence\ \sigma_n(x) \to 0.$$

- General Random Forest
  - 结合所关注特定问题结构的机器学习树类模型算法框架

因果森林系列工作是 CausalAI 研究的一个核心算法

Remark of Susan&Imbens's Work

- 聚焦于具体的异质因果效应(HTE)等参数估计问题



Imbens: There have not been similar applications of the DAG framework, and more papers discussing toy models will not be sufficient to convince economists to use this framework.

Imbens G W. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics[J]. Journal of Economic Literature, 2020, 58(4): 1129-79.

因果之梯: 结构因果模型文献推荐

- Pearl J. The seven tools of causal inference, with reflections on machine learning[J]. Communications of the ACM, 2019, 62(3): 54-60.
- Schölkopf B. Causality for machine learning[M]//Probabilistic and Causal Inference: The Works of Judea Pearl. 2022: 765-804.
- On Pearl's Hierarchy and the Foundations of Causal Inference
- Bongers, Stephan, et al. "Foundations of structural causal models with cycles and latent variables." The Annals of Statistics 49.5 (2021): 2885-2915.

What If ?

Theorem 1. *[Causal Hierarchy Theorem (CHT), informal version] The* PCH *almost never collapses. That is, for almost any SCM, the layers of the hierarchy remain distinct.* ■

What does *almost-never* mean? Here is an analogy. Suppose (fully specified) SCMs are drawn at random from Ω. Then, the probability that we draw an SCM relative to which PCH collapses is 0. This holds regardless of the distribution on SCMs, so long as it is smooth.

The CHT thus says in a general manner that there will typically be causal questions that one cannot answer with knowledge and/or data restricted to a lower layer in the hierarchy.[29] In fact, this can be seen as the formal grounding for the intuition behind the PCH as discussed in [Pearl and Mackenzie 2018, Ch. 1]:

Corollary 1. *To answer questions at Layer i, one needs knowledge at Layer i or higher.* ■

	Layer (Symbolic)	Typical Activity	Typical Question	Example	Machine Learning
$\mathcal{L}_1$	Associational $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?	Supervised / Unsupervised Learning
$\mathcal{L}_2$	Interventional $P(y\|do(x),c)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured?	Reinforcement Learning
$\mathcal{L}_3$	Counterfactual $P(y_x\|x',y')$	Imagining	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?	

CLeaR2022 因果学习和推理会议

Topics of submission may include, but are not limited to:

- **Machine learning building on causal principles**
- **Causal discovery in complex environments**
- Efficient causal discovery in large-scale datasets
- Causal effect identification and estimation
- **Causal generative models for machine learning**
- Unsupervised and semi-supervised deep learning connected to causality
- **Machine learning with heterogeneous data sources**
- Benchmark for causal discovery and causal reasoning
- Reinforcement learning
- Fairness, accountability, transparency, explainability, trustworthiness, and recourse
- **Applications of any of the above to real-world problems**

CLeaR 2022 will be held in Eureka, CA, USA from April 11 to 13, 2022