ランダムフォレストによる因果推論と最近の展開

フォレストワークショップ 2023 - Online -

Tomoshige Nakamura

February 27, 2023

Department of Mathematics, Keio University / tomoshige.nakamura [at] gmail.com

Table of contents

イントロダクション

ランダムフォレストの直近 10 年の話題

- ランダムフォレストは、2001 年に Breiman 2001 に提案されて以来、予測や分類のタスクにおいて、様々な応用がなされてきた。
- ランダムフォレストの良さは、特に応用方面からすると、まず「手軽」に試せることである。観測されたデータさえあれば、ボタン一つで当てはめることができる。
- さらに、応用方面ではよく知られた事実として、データに相関がある場合でも、データが高次元な場合でも、うまく予測をできることが知られている。
- もう 1 つの特徴として過学習が起こりにくいことも知られている。
- 変数重要度の概念があり、予測に対して有効な変数について直観的に理解しやすい。

ランダムフォレストの直近 10 年の話題

- Delgado et al. (2014) の結果では、特にパラメータ設定などは考えずに、Default で用いる場合の分類性能の良さを UCI データセット 121 に対して、様々な機械学習の手法に対して行った結果、Random forest は優秀な結果を示したことを報告している。
- 10 年前の結果なので、現在はよりよい手法があるかもしれないが、ただし多くのユーザーの手元にある手法として優秀なものであることには変わりはない。
- これほど応用方面では優秀な結果を残してきたランダムフォレストであるが、理論解析については発展途上であった。
- しかし、ここ 10 年のランダムフォレストに対する理論解析の結果から、「ランダムフォレストが統計的手法としての位置づけ」を獲得するようになった。
- まずは Random forest についてのここ 10 年の歴史的な数理解析面での進歩について、特に大きいものを 3 つお話ししたい。

ランダムフォレストの直近 10 年の話題

ランダムフォレストの解析としての大きな進捗

- まず、大きな転換点となったのは、Scornet (2015) と、Wager and Walther (2014) による結果変数に依存した場合の一致性の証明である。ここまでの一致性の証明の多くは、Tree の Partitioning が回帰の設定において結果変数に依存しない場合に限られていた。
- 次の大きな転換点となったのは、Wager and Athey (2018) の漸近正規性の証明である。この結果は、Random forest による関数推定に信頼区間が得られるという意味で画期的なものであった。それと同時に、この論文ではランダムフォレストが今日発表する因果推論の意味で活用できることを理論的に示した。
- 最後に大きな転換は、Generalized random forest の提案 (Athey, Tibshirani and Wager, 2019) である。この拡張によって、ランダムフォレストは実質的に局所推定方程式に対する解を得られるようになった。

これらをもって、ランダムフォレストによる推定結果は、統計的な推論に用いることができるようになった。

ランダムフォレストの直近 10 年の話題

これらの結果の進歩と同時に、Causal inference の文脈としての random forest の進歩としては、

- Athey and Imbens (2016) による Tree の honest という性質と、Recursive partitioning による heterogeneous causal effect の推定法の提案
- Athey and Wager (2018) による causal forest による heterogeneous causal forest の推定法の提案
- Athey, Tibshirani and Wager (2019) による Generalized random forest による、Neyman 直行化を用いた heterogeneous causal effect の推定法の提案

などがある。これらは、先ほどの漸近正規性などの議論と密接な関係性があり、本日の発表のメイントピックである。

本日の発表について

- 本日の発表においては、統計的因果推論に対するランダムフォレストの応用をメインの話題とはするが、それと関連する重要なランダムフォレスト推定量の漸近的な性質も併せて紹介する。
- まずは、Athey and Imbens （2016）による causal tree について説明し、因果効果の推定がどのようにして Tree によって行われるのかその仕組みについて説明する。
- 次に、Athey and Wager (2018) の結果から、random forest の一致性と、漸近正規性について説明する。その結果から、causal tree を base-learner として用いた causal forest による推定量も漸近正規性を持つことを示す。
- 最後に Generalized random forest について扱い、causal forest とは違う視点からの HTE の推定法について説明し、推定方程式を用いることで推定効率が向上することを述べる（時間の都合上、推定方程式の直行化などについては割愛する）。

Causal inference の導入

Causal inference - Introduction

Causal inference

観測されたデータから、Selection Bias や Confoundings bias の影響を取り除き、施策や処置の純粋な効果（Causal Effect）を統計的に推測するための方法論

- 例えば、広告の効果測定では「広告がどの程度売り上げに貢献したか」という定量的な評価が行われる。
- 直観的に考えれば「広告を見なかった人」と、「広告を見た人」の「売り上げの差」だと考えるが、これは正しくない。
- 実際は「広告を見た人」と「広告を見た人がもし見ていなかった場合」の「売り上げの差」が、広告効果である。

これらを定式化し、仮定を置いたうえで、統計的な推論を行うのが統計的因果推論の基本的な考え方である。

ここ数年で、どのような人に広告を配信すれば効果が最大化できるかというような、条件付き因果効果の推定に実務レベルで関心が高まっている。

Causal Inference - Framework

- 統計的因果推論では、現象を捉え、記述するための Framework がある。
- potential outcome framework（潜在結果変数モデル）／ counterfactural framework（反実仮想モデル）／ Neyman-Rubin causal model（ネイマンルービンの因果モデル）(Neyman, 1923; Rubin, 1974; Imbens and Rubin, 2015; Hernan and Robins; 2020)
- The causal diagram（因果ダイアグラム）(Pearl, 2009)
- これら 2 つは、数学的にはつながりがある (Richardson and Robins, 2013) が、それぞれ発展してきた目的や、計算アルゴリズム、さらには応用分野が分かれている。
- この発表は、potential outcome framework をベースとして議論を展開する。

Causal Inference - Potential outcome framework

- $Y_{a=1}, Y_{a=0} \in \mathbb{R}$: 潜在結果変数
- X_j : p 次元の pre-treatment 共変量 $(j = 1 : p)$
- $A = \{0, 1\}$: 処置変数
- $\pi(x) = \Pr(A = 1|X = x)$: 傾向スコア

(A1) Assumption : Consistency

$$Y = AY_{a=1} + (1 - A)Y_{a=0}$$

(A2) Assumption : Unconfoundedness

$$A \perp\!\!\!\perp Y_a|X \text{ for } a = 0, 1$$

(A3) Assumption : Posititvity

$$0 < \pi(x) < 1$$

Definition: Average Treatment Effect（ATE）

$$\theta^{ATE} = \mathrm{E}[Y_{a=1} - Y_{a=0}]$$

傾向スコアの逆数を用いた推定量を、Inverse probability Weighting 推定量という

Inverse probability Weighting Estimator(IPW 推定量)

$$\theta^{IPW} = \mathrm{E}\left[Y\left(\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right)\right]$$

回帰モデルについて、仮定 (A2) のもとで以下の関係式が成り立つ。

$$g_a(x) = \mathrm{E}[Y|A=a, X=x] = \mathrm{E}[Y_a|A=a, X=x] = \mathrm{E}[Y_a|X=x]$$

この結果から、以下の関係式が成り立つことがわかる。

$$\mathrm{E}[Y_{a=1}] = \mathrm{E}[g_{a=1}(X)]$$

この結果を用いた推定量を、regression estimator と呼ぶ

Inverse probability Weighting Estimator(IPW 推定量)

$$\theta^{reg} = \mathrm{E}\left[g_{a=0}(X) - g_{a=0}(X_i)\right]$$

Def: heterogeneous treatment effect (HTE)

$$\theta^{HTE}(x) = \mathrm{E}[Y_{a=1} - Y_{a=0} | X = x]$$

- HTE は、共変量が与えられたもとでの因果効果である。例えば「男性」「40 代」に絞った場合の広告の効果などを表し、効果のある集団を見つけるための重要な指標である。
- 例えば、個人に対する因果効果を、興味のある変数のみに絞った回帰モデルによって周辺化 (marginalization) することにより推定することができる。

$$\operatorname{argmin} \sum_{i=1}^{N} \left\{ \left(\frac{A_i}{\pi(x_i)} - \frac{1 - A_i}{1 - \pi(x_i)} \right) Y_i - \mu(x_i; \beta) \right\}^2$$

Causal Inference - 傾向スコアに依存しない因果効果の推定

- 近年、Heterogeneous treatment effect の推定に興味があるケースも増えている。
- しかし、従来の傾向スコア $\pi(x)$ の逆数による重みづけは、傾向スコアに対してパラメトリックモデルを仮定し推定する手法であり、モデルの誤特定の問題がある。
- また、ノンパラメトリックな推定を行った結果を、推定量に代入すると漸近正規性の前提となる仮定が崩れるので、直接用いたくはない。
- 傾向スコアを推定する理由がない限りは、$\pi : \mathcal{X} \mapsto (0, 1)$ の経由を避ける方が良く、これは ATE であっても、HTE であっても同じである。
- これらの問題に対して、causal tree, causal forest, generalized random forest は新たな推定方法として注目されている。

honest Trees

honest Trees - Introduction

- Causal Trees (Athey and Imbens, 2016) は、recursive partitioning を用いて、Heterogeneous causal effect を推定する手法である。
- この論文においては、causal tree の他に重要な tree における考え方 honest が提案されており、これが causal forest や GRF の漸近正規性の証明において中心的な枠割を果たす。
- また、honest 性を満たす tree は従来の CART と比較して過学習を起こしにくいという性質もある

まずは、tree の honest 性について説明する。

- まとめてしまえば、honest 性とは Tree の当てはめにおいて「Partitioning を生成するために用いるサンプル」と「Tree の Leaf 毎の推定量の計算に用いるサンプル」に別のものを用いることで、2 つが独立になった Tree のことである。
- わかりにくいので、Tree を構成する 2 つの要素、特徴空間を分割する”Partitioning” と、分割された空間の予測値”tree の推定量” について説明する。

honest trees - partitioning and tree estimator

特徴量と結果変数の組 $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, 2, ..., N$ が観測されたもとで、$\mu(x) = \mathsf{E}[Y_i | X_i = x]$ に対する Tree 推定量について考える。

Partitioning と Tree 推定量

特徴空間 $\mathcal{X}$ を背反に分割する partitioning を

$$\Pi = \{\ell_1, \ell_2, ..., \ell_{\#(\Pi)}\}$$

ただし

$$\cup_{j=1}^{\#(\Pi)} \ell_j = \mathcal{X} \ \text{ and } \ \ell_j \subset \mathcal{X}$$

で定義する。このとき、Partitioning Π が与えられたもとでの、条件付き平均関数 $\mu(x; \Pi)$ を

$$\mu(x; \Pi) \equiv \mathsf{E}[Y_i | X_i \in \ell(x; \Pi)] = \mathsf{E}[\mu(X_i) | X_i \in \ell(x; \Pi)]$$

さらに、サンプル $\mathcal{S}$ のデータを用いて構成した Tree 推定量 $\hat{\mu}(x, \mathcal{S}; \Pi)$ を、

$$\hat{\mu}(x, \mathcal{S}; \Pi) = \frac{1}{\#\{i \in \mathcal{S} : X_i \in \ell(x; \Pi)\}} \sum_{\{i \in \mathcal{S} : X_i \in \ell(x; \Pi)\}} Y_i$$

で定義する。ただし、$\ell(x; \Pi) \in \Pi$ は x を含む partitioning Π の元である。

honest trees - Partitioning と Tree 推定量

- Partitioning Π とは、$[0,1]^2$ を分割する部分空間 1 つ 1 つを指す。
- 点 x が黄色の部分空間に属する場合、この部分空間が $\ell(x;\Pi)$ で表される。
- この部分空間に対する Tree 推定量が $\hat{\mu}(x,\Pi)$ である。

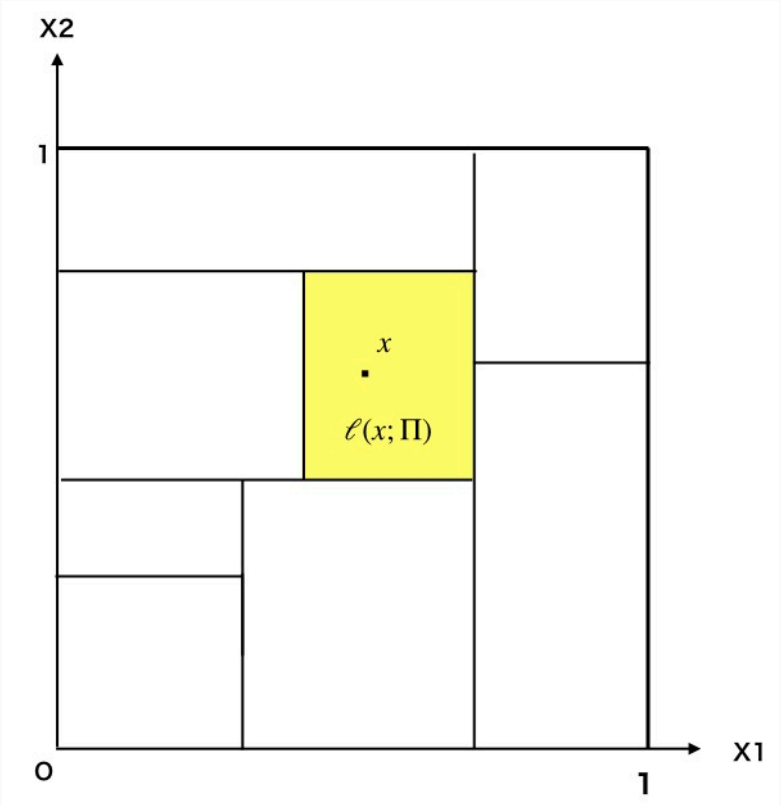

Figure 1: Partitioning と Tree 推定量

honest Trees - honest な推定とは

- honest な推定について考えるために、Tree の損失関数として一般的な平均二乗誤差を考える。
- Partition Π が与えられたもとで、Estimation sample $\mathcal{S}^{est}$ を用いて推定された条件付き平均と、テストデータ $\mathcal{S}^{te}$ の平均二乗誤差を

$$\mathrm{MSE}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = \frac{1}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \left\{ \left(Y_i - \hat{\mu}(X_i; \mathcal{S}^{est}, \Pi) \right)^2 - Y_i^2 \right\}$$

  と定義する。
- また、平均二乗誤差を $\mathcal{S}^{est}$ と $\mathcal{S}^{te}$ に対して期待値をとったものを

$$\mathrm{EMSE}(\Pi) \equiv \mathrm{E}_{\mathcal{S}^{te}, \mathcal{S}^{est}} \left[\mathrm{MSE}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) \right]$$

  と定義する。

honest Trees - honest vs adaptive

- honest な tree においては、次の関数を最大化するように Partition $\Pi(\mathcal{S}^{tr})$ を作る。

$$Q^H(\Pi) = -\mathrm{E}_{\mathcal{S}^{te},\mathcal{S}^{est}\mathcal{S}^{tr}}\left[\mathrm{MSE}(\mathcal{S}^{te},\mathcal{S}^{est},\Pi(\mathcal{S}^{tr}))\right]$$

- これに対して、一般的な CART では次の関数を最大化するように $\Pi(\mathcal{S}^{tr})$ を作る。

$$Q^A(\Pi) = -\mathrm{E}_{\mathcal{S}^{te},\mathcal{S}^{tr}}\left[\mathrm{MSE}(\mathcal{S}^{te},\mathcal{S}^{tr},\Pi(\mathcal{S}^{tr}))\right]$$

- ここで、、テストデータが $\mathcal{S}^{te}$ である。
- 一般的な CART では訓練データは $\mathcal{S}^{tr}$ であるのに対して、honest な tree では訓練データを 2 分割し $\mathcal{S}^{est}$ と $\mathcal{S}^{tr}$ とした上で当てはめを行う。
- これによって honest 性を持つ Tree では、partition Π と、$\ell(x;\Pi)$ における推定量は独立となる。この性質を honest と呼ぶ。

honest Trees - honest vs adaptive

- honest な tree においては、Π は推定量 $\hat{\mu}$ とは独立であるため、EMSE を最適化していると考えることができる。
- そこで、Π を条件づけた下で、EMSE に対する $\mathcal{S}^{tr}$ を用いた不偏推定量を構成し、従来の CART の損失関数と比較することで、honest 性の利点を明らかにする。
- EMSE を展開すると

$$
\begin{aligned}
-\mathrm{EMSE}(\Pi) &= -\mathrm{E}_{(Y_i,X_i),\mathcal{S}^{est}}[(Y_i - \mu(X_i;\Pi)^2 - Y_i] \\
&\quad -\mathrm{E}_{X_i,\mathcal{S}^{est}}\left[(\hat{\mu}(X_i;\mathcal{S}^{est};\Pi) - \mu(X_i;\Pi))^2\right] \\
&= \mathrm{E}_{X_i}[\mu^2(X_i;\Pi)] - \mathrm{E}_{\mathcal{S}^{est},X_i}[\mathrm{Var}(\hat{\mu}(X_i;\mathcal{S}^{est};\Pi))]
\end{aligned}
$$

  となる。
- これに対して、$\mathcal{S}^{tr}$ から不偏推定量を構成すると、

$$
\widehat{\mathrm{EMSE}}(\mathcal{S}^{tr},\Pi) = \frac{1}{N^{tr}}\sum_{i\in\mathcal{S}^{tr}}\hat{\mu}^2(X_i;\mathcal{S}^{tr},\Pi) \underbrace{-\frac{2}{N^{tr}}\cdot\sum_{\ell\in\Pi}S^2_{\mathcal{S}^{tr}}(\ell)}_{penalty}
$$

  となる。ただし、$S^2_{\mathcal{S}^{tr}}(\ell)$ は $\ell \in \Pi$ における leaf 内分散である。

honest Trees - honest vs adaptive

- 一方で、従来の CART においてはこのような罰則は存在せず、

$$-\mathrm{MSE}(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\mu}^2(X_i; \mathcal{S}^{tr}, \Pi)$$

  を最大化する。
- この基準量は分割を行えば行うほど改善するために、過学習が起こる。そのため cross-validation などを用いた枝狩りが従来の CART では必要となる。
- 一方で、honest な性質を持つ Tree においては、EMSE が罰則を含むため過学習が起こりにくくなるというメリットがある。
- 第 1 項は Tree 推定量の当てはまりを改善するが、第 2 項はそれによる Tree による階段関数の分散（複雑性）が大きくなることに対する罰則が存在する。
- 実際には、$\ell \in \Pi$ に含まれるサンプルの数が小さくない限りは、第 1 項の改善の影響が大きく、従来の CART と同じ挙動を示す。
- 一方でサンプル数が小さくなると、分散の影響が大きくなり、分割が止まるという風に動作する。

Causal Trees

Causal Trees - causal tree に対する損失関数

- 因果推論は、$(Y_i, X_i, W_i) \in \mathbb{R} \times \mathbb{R}^p \times \{0, 1\}$ が観測されたもとで、$\theta^{HTE}(x) = \mathrm{E}[Y_{a=1} - Y_{a=0} | X = x]$ を推定するという問題である。
- Partitioning Π が与えられたもとで、共変量 x と処置 a のもとでの母集団平均を以下で定義する。

$$\mu(a, x; \Pi) \equiv \mathrm{E}[Y_a | X \in \ell(x; \Pi)]$$

- Partitioning Π のもとでの、$\ell \in \Pi$ における因果効果を $\tau(x; \Pi)$ とする。

$$\tau(x; \Pi) \equiv \mathrm{E}[Y_{a=1} - Y_{a=0} | X \in \ell(x; \Pi)]$$

- 次に、partitioning Π が与えられたもとでの、処置 a、共変量 x に対する訓練データ $\mathcal{S}$ を用いた推定量を $\hat{\mu}(a, x; \mathcal{S}, \Pi)$ とする。

$$\hat{\mu}(a, x; \mathcal{S}, \Pi) = \frac{1}{\#(\{i \in \mathcal{S}_a : X_i \in \ell(x; \Pi)\})} \sum_{\{i \in \mathcal{S}_a : X_i \in \ell(x; \Pi)\}} Y_i$$

- ただし、$\mathcal{S}_a$ はデータ $\mathcal{S}$ のうち処置 a のもののみの集合である。このとき、観測データから計算される因果効果の Tree 推定量は

$$\hat{\tau}(x; \mathcal{S}, \Pi) = \hat{\mu}(1, x; \mathcal{S}, \Pi) - \hat{\mu}(0, x; \mathcal{S}, \Pi)$$

Causal Trees - causal tree

- HTE の推定における Tree の損失関数は以下のようになる。

$$\mathrm{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = \frac{1}{\#(\mathcal{S}^{te})} \sum_{i \in \mathcal{S}^{te}} \left\{ \left(\tau_i - \hat{\tau}(X_i; \mathcal{S}^{est}, \Pi) \right)^2 - \tau_i^2 \right\}$$

- 先ほどの regression の場合と同様に、EMSE を以下で定義する。

$$\mathrm{EMSE}_\tau(\Pi) \equiv \mathrm{E}_{\mathcal{S}^{te}, \mathcal{S}^{est}} \left[\mathrm{MSE}_\tau \left(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi \right) \right]$$

- 式の展開においては、τ_i が観測されないため、この部分を不偏推定量で置き換えることにする。

Causal Trees - causal tree (adaptive)

- Partitioning Π が与えられたもとでの処置効果について、次の関係性が成り立つ

$$\mathrm{E}_{\mathcal{S}^{te}}\left[\tau_i | i \in \mathcal{S}^{te} : i \in \ell(x; \Pi)\right] = \mathrm{E}_{\mathcal{S}^{te}}\left[\hat{\tau}(x; \mathcal{S}^{te}, \Pi)\right]$$

- 一般的な CART を用いた因果効果推定の MSE に対する unbiased estimator は、以下のようになる。

$$\mathrm{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi) = -\frac{2}{N^{tr}} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}(X_i; \mathcal{S}^{te}, \Pi) \cdot \hat{\tau}(X_i; \mathcal{S}^{tr}, \Pi) + \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi)$$

- となるから、同様にして $\mathcal{S}^{tr}$ を用いた訓練データで、テストデータを置き換えた損失関数

$$-\mathrm{MSE}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi) \tag{1}$$

を最適化することになる。

Causal Trees - causal tree (honest)

- honest tree の場合、最適化するのは EMSE であり、展開すると

$$-\mathrm{EMSE}_\tau(\Pi) = \mathrm{E}_{X_i}\left[\tau^2(X_i;\Pi)\right] - \mathrm{E}_{\mathcal{S}^{est},X_i}\left[\mathrm{Var}(\hat{\tau}^2(X_i;\mathcal{S}^{est},\Pi)\right]$$

- これに対する不偏推定量を $\mathcal{S}^{tr}$ を用いて構成すると、

$$-\widehat{\mathrm{EMSE}}_\tau(\mathcal{S}^{tr};\Pi) = \frac{1}{N_{tr}}\sum_{i\in\mathcal{S}^{tr}}\hat{\tau}^2(X_i;\mathcal{S}^{tr},\Pi) \underbrace{-\frac{2}{N_{tr}}\sum_{\ell\in\Pi}\left(\frac{S^2_{\mathcal{S}^{tr}_{treat}}(\ell)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(\ell)}{1-p}\right)}_{penalty}$$

- ただし、$\ell \in \Pi$ における処置 a における $\hat{\mu}(a,x;\Pi)$ の分散を

$$S^2_{\mathcal{S}^{tr}_{treat}}(\ell) = \mathrm{Var}(\hat{\mu}(a,x;\Pi)|A=a, X\in\ell)$$

とした。また、$\ell \in \Pi$ における処置群の割合を p とした

$$p = \frac{\#\{i : A_i = 1, X_i \in \ell\}}{\#\{i : X_i \in \ell\}}$$

Causal Trees - causal tree (honest)

- honest 性を持つ causal tree の当てはめは、(i) Leaf 間での処置効果の差の最大化、および (ii) 分割によって上昇する処置群と対照群の結果変数の分散の最小化、の 2 つのトレードオフの最適化となる。
- 因果効果は 2 つの潜在結果変数の差として定義される。そのため因果効果の異質性が分割によって上昇することと、leaf 間の処置群と対照群の分散が小さくなることは、regression の場合のように比例関係があるわけではない。
- 罰則の意味が、regression の場合と causal inference では異なってくる点に注意が必要である。

Asymptotics of random forests and causal forests

Random forest

- Athey and Wager (2018) は、honest 性を満たす tree から構成された random forest の一致性と、漸近正規性を示した。
- また、この結果が HTE を推定する random forest である causal forest へも拡張されることを示した。
- まずは、$(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R} (i = 1, 2, ..., n)$ が観測されたもとでの、条件付き平均関数

$$\mu(x) := \mathsf{E}[Y_i | X_i = x]$$

を推定する問題について考える。

random forest - original RF

- 最もよく用いられている random forest は、Breiman (2001) によって提案された random forest である。
- Breiman's random forest では、サイズ n の訓練データから、サイズ s ($< N$) のブートストラップサンプルを B 回とる。
- ブートストラップサンプル $b = 1, 2, ...B$ に対して、CART を当てはめる。
- 新たな入力 x に対して、B 本の Tree それぞれの結果を平均し、random forest の推定量として返す

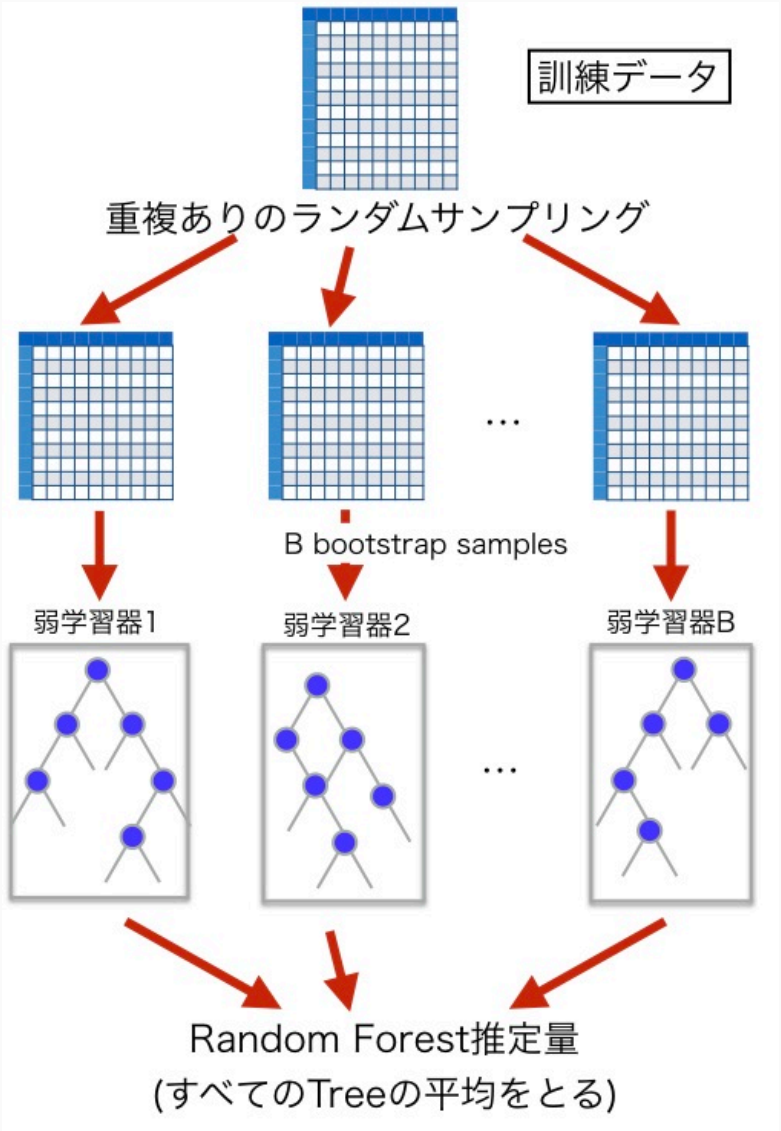

Figure 2: Breiman's RF

random forest - Breiman's RF との相違点

Wager and Athey (2018) で用いる random forest のアルゴリズムは、Breiman が提案したものと以下の 3 つの点で異なる。

- ランダムフォレストを構成する際に用いるのは、ブートストラップサンプルではなく、重複なしのサンプルである。
- adaptive tree を用いるのではなく、honest 性を持つ Tree である。すなわち、Partitioning Π と、Tree 推定量 $\hat{\mu}$ が独立になるような Tree を用いる。
- Tree におけるノード分割に、α-regular 性と、random-split 性の 2 つを持つようにさせる。

ここでは、このような条件を満たすランダムフォレストの漸近正規性について述べる。

Wager and Athey (2018) は、honest 性を満たす Tree として”**Double sample tree**” を提案している。

Double sample tree

Input : データ $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, 正則パラメータ α, 最小葉サイズ k.

1. サイズ s のサブサンプルを添え字重合 $\{1, 2, ..., n\}$ から重複なしでとり、サイズ $|\mathcal{J}| = \lfloor s/2 \rfloor$ 及び、$|\mathcal{I}| = \lceil s/2 \rceil$ となるように、背反な集合 $\mathcal{I}, \mathcal{J}$ に分割する。
2. EMSE を最小にするようにノードの分割を繰り返し行う。このとき、ノード分割（partitioning Π の生成）には、$\mathcal{J}$ に含まれるサンプルと、$\mathcal{I}$ に含まれるサンプルのうち X_i の情報のみを用いて分割を行う（honest 性）。$\mathcal{I}$ に含まれるサンプルのうち Y_i の情報は、分割には用いない。
3. 点 x に対する Tree 推定量を、$\mathcal{I}$ のデータを用いて行う。

$$\frac{1}{\#\{i \in \mathcal{I}, x \in \ell(x; \Pi)\}} \sum_{\{i \in \mathcal{I}, x \in \ell(x;\Pi)\}} Y_i$$

random forest - definition

Def: random forest estimator

- $Z_i = (X_i, Y_i), i = 1, 2, ..., n$ からの、サイズ s の重複なしサブサンプルを $\mathcal{D}_s = \{Z_{i_1}, ..., Z_{i_s}\}$ とする。
- 点 x に対する、サブサンプル $\mathcal{D}_s$ に基づく、ランダムネス ξ を含む double sample tree 推定量を

$$T(x; \xi, Z_{i_1}, ..., Z_{i_s})$$

とする。

- このとき、サブサンプルサイズ s で、base-learner T のランダムフォレスト推定量を

$$RF(x; Z_1, Z_2, ..., Z_n) = \binom{n}{s}^{-1} \sum_{1 \le i_1 < i_2 < ... < i_s \le n} \mathsf{E}_{\xi \sim \Xi} \left[T(x; \xi, Z_{i_1}, ..., Z_{i_s}) \right]$$

で定義する。

random forest - 漸近正規性のための準備

ランダムフォレストによる推定量が漸近正規性を満たすためには、
Base-learner T が次の (A1)-(A3) を満たすことが必要となる。

(A1) Honest
(double sample tree の場合) サブサンプルを背反な 2 つの集合 $\mathcal{I}, \mathcal{J}$ に分け、Partitioning Π を生成する際に $\mathcal{J}$ の X_i, Y_i の情報及び $\mathcal{I}$ の X_i の情報のみを用いて計算し、推定量の計算に $\mathcal{I}$ の Y_i の情報を用いる。

(A2) random-split
Tree におけるノード分割において、任意の $0 < \pi \leq 1$ を満たす定数を用いて、すべての変数が各ノードで分割される確率が $\pi/p (> 0)$ で下から抑えられているとき、Tree は random-split 性を持つという。

(A3) α-regular
分割によって生成される 2 つの子ノードが親ノードのデータ数の少なくとも $\alpha \in (0, 0.2]$ の比率を含むように分割を行う。また事前に設定した最小葉サイズ $k \in \mathbf{N}$ に対して、各ノードが k 以上、$2k - 1$ 未満のサンプルとなるように分割する。

このとき、次の 2 つの定理が成り立つ。

random forest - ランダムフォレストのバイアスの評価

Th 3 (Athey and Wager, 2018)

base-learner $T(x; Z_1, ..., Z_s)$ が (A1), (A2), (A3) を満たすとする。このとき、次の 3 つの仮定の下で

- $X_1, ..., X_s \sim U([0,1]^p)$ に独立に従う。
- $\mu(x)$ がリプシッツ連続である。
- $\alpha \leq 0.2$ を満たす

ランダムフォレストによる $x \in [0,1]^p$ の推定量 $\hat{\mu}(x)$ のバイアスは

$$\left\| \mathrm{E}\left[\hat{\mu}(x) - \mu(x)\right] \right\| = \mathcal{O}\left(s^{-\frac{1}{2} \frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})} \frac{\pi}{p}} \right)$$

によって評価できる。

ここで、以下の関数は α の単調増加関数である。

$$0 < \frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})} \leq \frac{\log\left((1-0.2)^{-1}\right)}{\log(0.2^{-1})} \approx 0.139$$

Th 1 (Athey and Wager, 2018)

base-learner $T(x; Z_1, \ldots, Z_s)$ が (A1), (A2), (A3) を満たすとする。さらに、Th.3 (Athey and Wager, 2019) と同様の仮定と、適当な正則条件のもとで、サブサンプルサイズ s_n

$$s_n \asymp n^{\beta} \quad \text{for some} \quad \beta_{min} := 1 - \left(1 + \frac{p}{\pi} \cdot \frac{\log(\alpha^{-1})}{\log((1-\alpha)^{-1})}\right)^{-1} < \beta < 1$$

を満たすランダムフォレスト推定量 $\hat{\mu}^{RF}(x)$ は、$\sigma_n \to 0$ を満たす列が存在して

$$\frac{\hat{\mu}_n^{RF}(x) - \mu(x)}{\sigma_n(x)} \xrightarrow{d} N(0,1)$$

を満たす。さらに、このような列 σ_n に対して、infinitesimal jacknife 推定量 (Wager et al., 2014) $\hat{V}_{IJ}(x)$ は、一致性を満たす。

$$\hat{V}_{IJ}(x)/\sigma_n(x) \to 0$$

random forest - Infinitesimal jacknife 推定量

ランダムフォレストに対する、Infinitesimal Jacknife 推定量 (Wager et al., 2014) は以下で定義される。

Def: ランダムフォレストに対する Infinitesimal jacknife 推定量

$b = 1, 2, ..., B$ の Tree の推定量を $\hat{\mu}_b^*(x)$ とし、その平均を $\bar{\mu}^*(x)$ を

$$\bar{\mu}^*(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_b^*(x)$$

とする。さらに、$N_{bi}^* \in \{0, 1\}$ を i が b 番目の tree の訓練データとして用いられたかどうかの指示関数とする。このとき、Infinitesimal Jacknife 推定量は

$$\hat{V}_{IJ}(x) = \frac{n-1}{n} \left(\frac{n}{n-s} \right)^2 \sum_{i=1}^{n} \left\{ \frac{1}{B} \sum_{b=1}^{B} (N_{bi}^* - 1)(\mu_b^*(x) - \bar{\mu}^*(x)) \right\}^2$$

で表される。

この結果から、各点 x における分散の推定量が得られるので、信頼区間を描くことができる（$B \approx 2000$ 程度で十分な精度が出ると報告されている）。

random forest - causal inference への拡張

これらの結果を、$(Y_i, X_i, A_i) \in \mathbb{R} \times \mathbb{R}^p \times \{0, 1\}$ が観測された場合の、HTE の推定へは直接的に拡張することができる。

$$\tau(x) = \mathsf{E}[Y_{a=1} - Y_{a=0} | X = x]$$

そのためには、前に定義した causal tree を double sample tree へと拡張する必要がある。さらに、honest 性や、α-regular 性についても、処置群と対照群のサンプルがあることから、拡張する必要がある。

random forest - causal inference への拡張

従来の条件付き平均の推定に用いた仮定を、因果推論の文脈では以下のように置き直す。

(A1') Honest
(double sample tree の場合) サブサンプルを背反な 2 つの集合 $\mathcal{I}, \mathcal{J}$ に分け、Partitioning Π を生成する際に $\mathcal{J}$ の X_i, A_i, Y_i の情報及び $\mathcal{I}$ の X_i, A_i の情報のみを用いて計算し、推定量の計算に $\mathcal{I}$ の Y_i の情報を用いる。

(A2) random-split
Tree におけるノード分割において、任意の $0 < \pi \leq 1$ を満たす定数を用いて、すべての変数が各ノードで分割される確率が $\pi/p (> 0)$ で下から抑えられているとき、Tree は random-split 性を持つという。

(A3') α-regular
分割によって生成される 2 つの子ノードが親ノードのデータ数の少なくとも $\alpha \in (0, 0.2]$ の比率を含むように分割を行う。また事前に設定した最小葉サイズ $k \in \mathbf{N}$ に対して、各ノードが処置群 $a = 1$ のサンプルと、対照群 $a = 0$ のサンプルが、k 以上、$2k - 1$ 未満のサンプルとなるように分割する。

random forest - causal inference への拡張

Double sample causal tree

Input : データ $\mathcal{D}_n = \{(X_i, Y_i, A_i)\}_{i=1}^n$, 正則パラメータ α, 最小葉サイズ k.

1. サイズ s のサブサンプルを添え字重合 $\{1, 2, ..., n\}$ から重複なしでとり、サイズ $|\mathcal{J}| = \lfloor s/2 \rfloor$ 及び、$|\mathcal{I}| = \lceil s/2 \rceil$ となるように、背反な集合 $\mathcal{I}, \mathcal{J}$ に分割する。
2. EMSE を最小にするようにノードの分割を繰り返し行う。このとき、ノード分割（partitioning Π の生成）には、$\mathcal{J}$ に含まれるサンプルと、$\mathcal{I}$ に含まれるサンプルのうち X_i とA_iの情報のみを用いて分割を行う（honest 性）。
3. 点 x に対する Tree 推定量を、$\mathcal{I}$ のデータを用いて行う。

$$\frac{1}{\#\{i \in \mathcal{I}_{treat}, x \in \ell(x; \Pi)\}} \sum_{\{i \in \mathcal{I}_{treat}, x \in \ell(x;\Pi)\}} Y_i$$
$$- \frac{1}{\#\{i \in \mathcal{I}_{control}, x \in \ell(x; \Pi)\}} \sum_{\{i \in \mathcal{I}_{control}, x \in \ell(x;\Pi)\}} Y_i$$

random forest - causal forest の漸近正規性

Th 11 (Athey and Wager, 2019)

base-learner $\Gamma(x; Z_1, ..., Z_s)$ が (A1'), (A2), (A3') を満たす double sample causal tree とする。さらに、Th.3 (Athey and Wager, 2019) と同様の仮定と、適当な正則条件のもとで、サブサンプルサイズ s_n

$$s_n \asymp n^{\beta} \quad \text{for some} \quad \beta_{min} := 1 - \left(1 + \frac{p}{\pi} \cdot \frac{\log(\alpha^{-1})}{\log((1-\alpha)^{-1})}\right)^{-1} < \beta < 1$$

を満たすと仮定する。このとき、base-learner $\Gamma(x)$ とする causal forest 推定量 $\hat{\tau}^{CF}(x)$ は、$\sigma_n \to 0$ を満たす列が存在して

$$\frac{\hat{\tau}_n^{CF}(x) - \tau(x)}{\sigma_n(x)} \xrightarrow{d} N(0,1)$$

を満たす。さらに、このような列 σ_n に対して、infinitesimal jacknife 推定量 (Wager et al., 2014) $\hat{V}_{IJ}(x)$ は、一致性を満たす。

$$\hat{V}_{IJ}(x)/\sigma_n(x) \to 0$$

causal forest - simulation

以下の設定で、$\tau(X)$ を推定する。

- $X \sim N(0, I_{20})$
- $W \sim \text{Bernoulli}(0.4 + 0.2 \cdot 1\{X_1 > 0\})$
- $\tau(X) = \text{pmax}(X_1, 0)$
- $Y = \tau(X) \cdot W + X_2 + \text{pmin}(X_3, 0) + N(0, 1)$

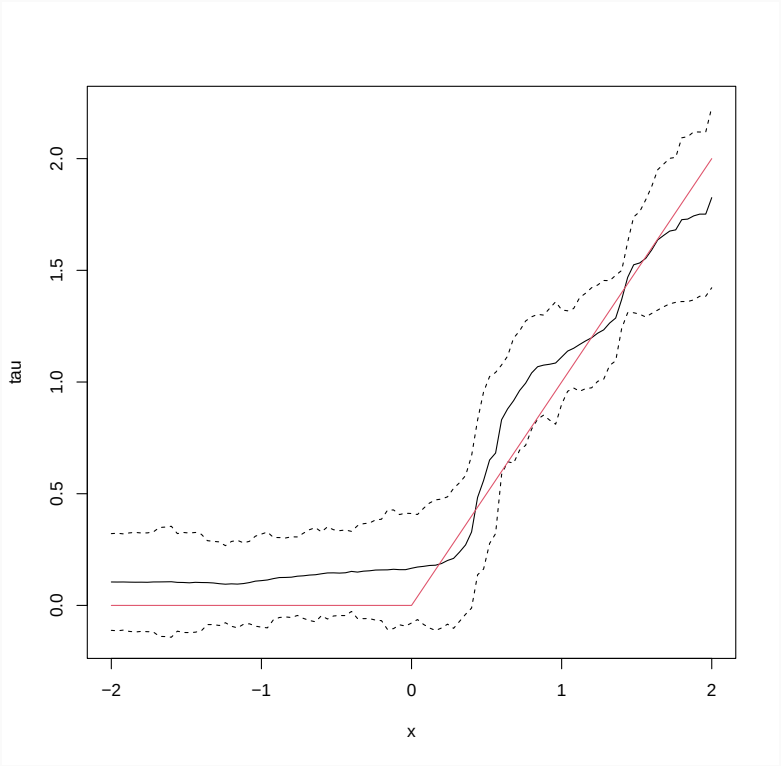

Figure 3: random forest による信頼区間

random forest - causal forest の漸近正規性

- 以上の議論から、double sample causal tree を base-learner とする causal forest は漸近正規性を持つことがわかり、その分散の推定も行えることがわかる。
- この結果、強く無視可能な割り付けなどの条件が成り立つならば、random forest を応用することで、因果効果の推定は可能である。

ただし、causal forest には欠点が存在する。

- ノードを分割する際に、処置群 $a = 1$ と対照群 $a = 0$ のサンプルを k 以上、$2k - 1$ 以下含む必要があることである。
- この条件では、処置確率が高い場所ほど、どちらかのサンプルが多くなり分割が行われにくくなり、推定精度が低下する。
- これを解消するアイデアが、Generalized random forest と、R-Learner を用いた因果効果のフレームワークである。

Generalized random forest

Generalized Random Forest

Generalized random forest (GRF; Athey, Tibshirani and Wager, 2019) は、局所推定方程式によって定義されるパラメータ $\theta(x)$ に対する forest-based な推定量を求める手法である。

$$\mathsf{E}[\psi_{\theta(x),\nu(x)}(O_i)|X_i = x] = 0 \qquad \text{for all } \ x \in \mathcal{X} \tag{2}$$

- $\psi(\cdot)$: スコア関数
- $\nu(\cdot)$: 局外パラメータ (optional)

GRF は、$\theta(x)$ の関数の推定を 2 つのステップで行う。

1. forest-based weight $\alpha_i(x)$ を計算する: テスト点 x_0 におけるパラメータ $\theta(x_0)$ に対する、i 番目の traning-example の関連性 (影響) の大きさ。
2. $\alpha_i(x)$ を重みとした、empirical version の推定方程式を解く.

$$\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \text{argmin}_{\theta,\nu} \left\{ \left\| \sum_{i=1}^{n} \alpha_i(x)\psi_{\theta,\nu}(O_i)) \right\|_2 \right\} \tag{3}$$

Generalized random forest - forest weights

- 赤い×を予測する際に、重みがかかっているサンプルを示した。
- GRF ではすべてのサンプルに対して、Tree を用いて、重み $\alpha_i(x)$ を計算し、その重みづけ推定方程式を解くことで推定量を得る。
- つまり、GRF とは forest を用いて、重みづけ kernel 関数をノンパラメトリックに推定している。

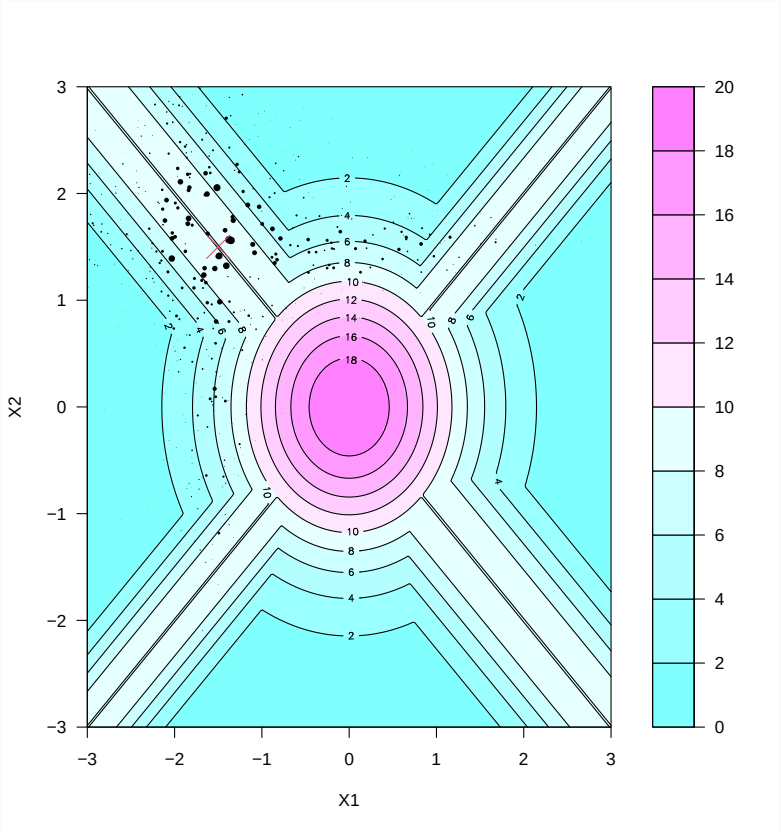

Figure 4: forest weights

Forest-based weight $\alpha_i(x)$ の計算手順は、以下の通り。

- $b = 1, 2, ..., B$ によって添え字付けされる B 個の Tree を考え、$L_b(x)$ を x を含む b 番目の Tree の Leaf が含む training example の集合とする。このとき、

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \alpha_{bi}(x), \tag{4}$$

  ただし、

$$\alpha_{bi}(x) = \frac{1\{X_i \in L_b(x)\}}{|L_b(x)|} \tag{5}$$

- 重み $\alpha_i(x)$ は、i 番目の traning example が、x を含む Leaf に何回含まれたかという頻度を測る指標である。

このことから、$\theta(x)$ の推定量の良さは Tree に依存することがわかる。

Generalized Random Forest - 重み α の推定

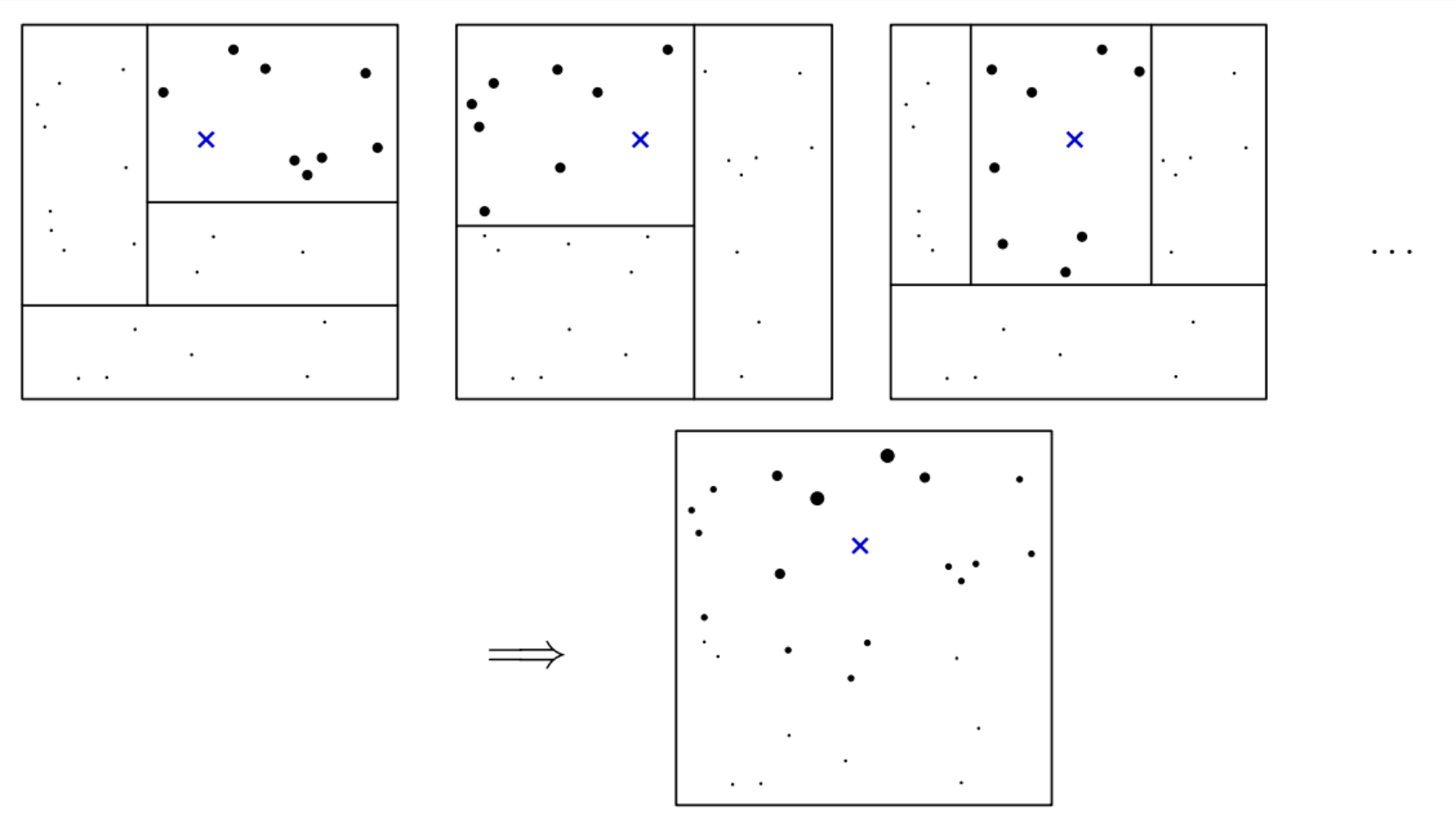

Figure 5: forest-based weights

Generalized Random Forest - 例

- 一般的な Random forest の場合: $O_i = Y_i$ としてスコア関数

$$\psi_{\mu(x)}(Y_i) = Y_i - \mu(x) \tag{6}$$

  実際、ランダムフォレストの推定量とは、以下の方程式の解である。

$$\sum_{i=1}^{n} \alpha_i(x)(Y_i - \mu(x)) = 0 \tag{7}$$

- q 分位点回帰モデルの場合、$O_i = Y_i$ として

$$\psi_{\theta(x)}(Y_i) = q \cdot 1\{Y_i > \theta(x)\} - (1-q)1\{Y_i \leq \theta(x)\}$$

- 操作変数による回帰モデルの場合: $O_i = \{Y_i, W_i, Z_i\} \in \mathbb{R} \times \{0,1\} \times \{0,1\}$ で、Z_i（操作変数）として、$Z_i \perp\!\!\!\perp e_i | X_i$ かつ $\mathrm{Cov}(Z_i, W_i | X_i) \neq 0$ の仮定の下で

$$\psi_{\tau(x),\mu(x)}(O_i) = \{Y_i - W_i\tau(x) - \mu(x)\} \begin{pmatrix} 1 \\ Z_i \end{pmatrix}$$

Generalized Random Forest - Gradient Trees

Athey, Tibshirani and Wager (2019) は、$\alpha_i(x)$ を推定する方法として、*Gradient Tree* を提案した。Gradient Tree は、$\theta(x)$ の異質性に着目してノードを分割し、Tree を構成する recursive partitioning algorithm である。

1. Labeling step: 親ノード P のデータを用いて、$\hat{\theta}_P$ 及び $\hat{\nu}_P$ を推定する。

$$(\hat{\theta}_P, \hat{\nu}_P) \in \text{argmin}_{\theta,\nu} \left\{ \left\| \sum_{i:X_i \in P} \psi_{\theta,\nu}(O_i)) \right\|_2 \right\} \tag{8}$$

また、Γ_P をスコア関数の微分 $\nabla \mathrm{E}[\psi_{\hat{\theta}_P,\hat{\nu}_P} | X_i \in P]$ に対する一致推定量とする。例えば、

$$\Gamma_P = \frac{1}{|i : X_i \in P|} \sum_{\{i:X_i \in P\}} \nabla \psi_{\hat{\theta}_P,\hat{\nu}_P}(O_i). \tag{9}$$

これらを用いて、pseudo-outcome を構築する。

$$\rho_i = -\xi^T \Gamma_P^{-1} \psi_{\hat{\theta}_P,\hat{\nu}_P}(O_i) \in \mathbb{R} \tag{10}$$

ここで、ξ は ψ のうち、$\theta(\cdot)$ に対応する部分を抽出するベクトルである。

2. Regression step: pseudo-outcome ρ_i に対して、CART と同様の分割する。すなわち、次の基準を最大化するように親ノード P を子ノード C_1, C_2 を、変数 X を基準としてaxis-alignedに分割する。

$$\Delta(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{|\{i : X_i \in C_j\}|} \left(\sum_{\{i:X_i \in C_j\}} \rho_i \right)^2 \tag{11}$$

ここで、Athey, Tibshirani and Wager (2019) は、評価関数 Δ を最大化することが、次の error を最小化することと漸近的に同等であることを示している。

$$\sum_{j=1,2} \Pr\left[X \in C_j \middle| X \in P\right] \mathrm{E}\left[\left(\hat{\theta}_{C_j} - \theta(X)\right)^2 \middle| X \in C_j\right] \tag{12}$$

ここで $\hat{\theta}_{C_j}$ は、子ノード C_j における推定方程式の解である。

Generalized Random Forest の因果推論への応用

- (X_i, Y_i, A_i), $1, 2, ..., n$ を $i.i.d.$ なサンプル
- X : p-次元の処置前変数 $(j = 1, 2, ..., p)$
- $Y \in \mathbb{R}$: 結果変数
- $A = \{0, 1\}$: 処置変数

強く無視可能な割り付けの仮定のもとで、

$$\{Y_{a=1}, Y_{a=0}\} \perp\!\!\!\perp A \mid X$$

CATE を推定するという問題を考える。

$$\tau(x) = \mathsf{E}[Y_{a=1} - Y_{a=0} | X = x] \tag{13}$$

Generalized random forest を用いるためには、$\tau(x)$ に対する局所推定方程式を構成する必要がある。

Double Residual モデルによる因果効果の推定方程式の構築

そこで、Nie and Wager (2021) にしたがって、因果効果 $\tau(x)$ を推定する問題を、Double residual model (Robinson, 1988) を用いて、結果変数と処置変数に対する条件付き平均の関数を用いて以下のように変換する。

Def : Robinson's Double Residual 変換

$$Y_i - m(X_i) = \{A_i - \pi(X_i)\}\tau(X_i) + \varepsilon_i \tag{14}$$

ここで、

- $\pi(x) = \mathrm{Pr}(A = 1|X = x)$: 傾向スコア
- $m(x) = \mathrm{E}[Y|X = x]$: 結果変数に対する条件付き平均関数
- ε_i : 誤差変数

ただし、強く無視可能な割り付けの条件より $\mathrm{E}[\varepsilon_i|A_i, X_i] = 0$.

Double Residual モデルによる因果効果の推定方程式の構築

この結果に基づいて, $\tau(\cdot)$ を次の 2 つのステップで推定する (R-learner; Nie and Wager, 2021)。

1. $\hat{m}(x)$ 及び $\hat{\pi}(x)$ を適当な方法で推定する (e.g. random forest / XGboost / CNN)
2. $\hat{m}(x)$ 及び $\hat{\pi}(x)$ を代入して、CATE を推定する:

$$\hat{\tau}(\cdot) = \mathrm{argmin}_{\tau} \left\{ \hat{L}_n(\tau(\cdot)) + \Lambda_n(\tau(\cdot)) \right\} \tag{15}$$

ここで

$$\hat{L}_n(\tau(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} \left((Y_i - \hat{m}^{(-i)}(X_i)) - (A_i - \hat{\pi}^{(-i)}(X_i))\tau(X_i) \right)^2, \tag{16}$$

であり、$\hat{m}^{(-i)}$ と $\hat{\pi}^{(-i)}$ は i 番目のサンプルを用いずに推定した関数である (cross-fitting)。

ここで、R-learner は quasi-oracle error bound を満たすことが知られている (Nie and Wager, 2021)。

HTE 推定のための Gradient tree

R-Learner の結果から、中心化した結果変数と、処置変数を以下のように定義する。

- centered outcomes : $\tilde{Y}_i = Y_i - \hat{m}^{(-i)}(X_i)$
- centered treatments : $\tilde{A}_i = A_i - \hat{\pi}^{(-i)}(X_i)$

また、スコア関数は

$$\psi_{a(x),\tau(x)} = (\tilde{Y}_i - a(x) - \tilde{A}_i\tau(X_i))(1, \tilde{A}_i)^T. \tag{17}$$

となる。ただし、$a(x)$ は切片項であり、$\tau(x)$ は条件付き因果効果である。

任意の親ノード P と、ノード P 内の各 i サンプルに対して、

$$\rho_i = \Gamma_P^{-1}(\tilde{A}_i - \bar{A}_P)\left(\tilde{Y}_i - \bar{Y}_P - (\tilde{A}_i - \bar{A}_P)\hat{\tau}_P\right) \tag{18}$$

を定義する、ここで、$\bar{A}_P$ 及び $\bar{Y}_P$ は $\tilde{Y}_i$ 及び $\tilde{A}_i$ の親ノードでの平均であり、$\hat{\tau}_P$ 親ノード P での推定方程式の解である。

$$\Gamma_P = \frac{1}{|\{i : X_i \in P\}|} \sum_{\{i:X_i \in P\}} \left(\tilde{A}_i - \bar{A}_P\right)\left(\tilde{A}_i - \bar{A}_P\right)^T \tag{19}$$

Generalized random forest のまとめ

- 時間の都合上、紹介は割愛したが Generalized random forest による $\theta(x)$ に対する局所推定方程式の解 $\hat{\theta}(x)$ は、漸近正規性を持つ。
- この結果、random forest は推定方程式によって定義されるパラメータに対して、適当な正則条件（多くの統計の問題が満たす）のもとでは、統計的な推論が可能となる。
- GRF の応用については、まだ発展途上であるが、徐々に応用例も増え始めており、今後も発展の可能性がある。

References

- Athey, S and Imbens, G. (2016). ”Recursive partitioning for heterogeneous causal effects.” ,Proceedings of the National Academy of Sciences, 113(27):7353-7360.
- Athey et al., (2019). ”Generalized Random Forests”. Annals of Statistics, 47(2).
- Nie, X and Wager, S. (2021). ”Quasi-Oracle Estimation of Heterogeneous Treatment Effects”, Biometrika, 108(2).
- Wager, S and Athey, S. (2018) ”Estimation and Inference of Heterogeneous Treatment Effects using Random Forests.” Journal of the American Statistical Association, 113(523), 2018.
- Shiraishi, H, Nakamura, T, and Shibuki, R. (2022) ”Time series quantile regression using random forests”, https://arxiv.org/abs/2211.02273

ありがとうございました！