ランダムフォレストによる因果推論

慶應義塾大学　研究員
中村　知繁

2020年12月1日

目次.

■導入

因果推論への期待と問題点

- 近年、因果推論の考え方（potential outcome framework, Rubin 1974など）の実データ解析における有用性に注目が集まっている。
- ビジネスなどでは、平均処置効果及び処置群における平均処置効果の推定などに興味があり、傾向スコアによる重み付け推定量などが実際に活用されている。
- さらに、ビジネスにおける施策などの効率性を高める目的で、共変量を条件づけたもとでの因果効果（subgroup内の因果効果）を推定することにも注目が集まり始めた（例えば、Zhao and Harinen 2019）

- しかし、重み付け推定量は傾向スコアのモデル誤特定に対してのロバスト性に乏しく、これらの推定を行わずに直接的に因果効果を推定できないか。という考え方に至るのは自然である。
- また、ノンパラメトリックな傾向スコアの推定のもとでの重み付け推定量は、$\sqrt{N}$-consistencyを持たないなど機械学習を従来の因果推論と組み合わせることの課題も指摘されてきた。

■導入

因果推論とランダムフォレスト

- これらの問題に対して、Wager and Athey (2018)はRandom forestを用いることで、従来のような傾向スコアの推定を行うことなく、データからconditional average treatment effect (CATE)が推定でき、さらに推定量が漸近正規性を持つことが示した。
- また、計算可能な分散の推定量が、推定量の漸近分散に収束することも示している。

- この論文に対しては、
  - Wager and Walther(2014)によるRegression treeに対する漸近的な性質に関する研究
  - Athey and Imbens(2016)によるCausal Treeの提案
  - Efron (2014)によるinfinitesimal jackknifeの提案
- が重要な貢献を果たしている。

■導入

因果推論への機械学習の応用

- この他、因果推論に対しては、
  - Generalized random forest（Athey, Tibshirani and Wager, 2019)
  - R-Learner (Nie and Wager, 2020)
  - Orthogonal random forest (Oprescu, Syrgkanis and Wu, 2019)
- などが近年は注目されている。
- 本日の発表は、causal forestに対する結果を紹介しつつ、近年の展開について述べる予定である。

発表における説明は

causal tree → causal forest →　GRF

目次.

Regression tree and causal tree

記号

- 本発表では、以下の記号を用いる。
  - $i = 1,2,...,N$：標本に対応するindex
  - $(Y_i(1), Y_i(0))$：標本iに対する潜在結果変数
  - $\tau_i = Y_i(1) - Y_i(0)$：標本iに対する因果効果
  - $W_i \in \{0,1\}$：標本iに対する処置変数
  - $Y_i = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 1 \\ Y_i(1) & \text{if } W_i = 0 \end{cases}$ ： 観測された結果変数Y_i
  - X_i：処置からの影響を受けないPre-treatment p次元共変量ベクトル
  - (Y_i, W_i, X_i)：標本iに対する観測データベクトルで、母集団からのi.i.d.な標本であるとする。

■Regression tree and causal tree

強く無視可能な割り付けの仮定と条件付き平均処置効果

Unconfoundedness Assumption仮定を置く

$$\left\{Y_i(1), Y_i(0)\right\} \perp\!\!\!\perp W_i \,|\, X_i$$

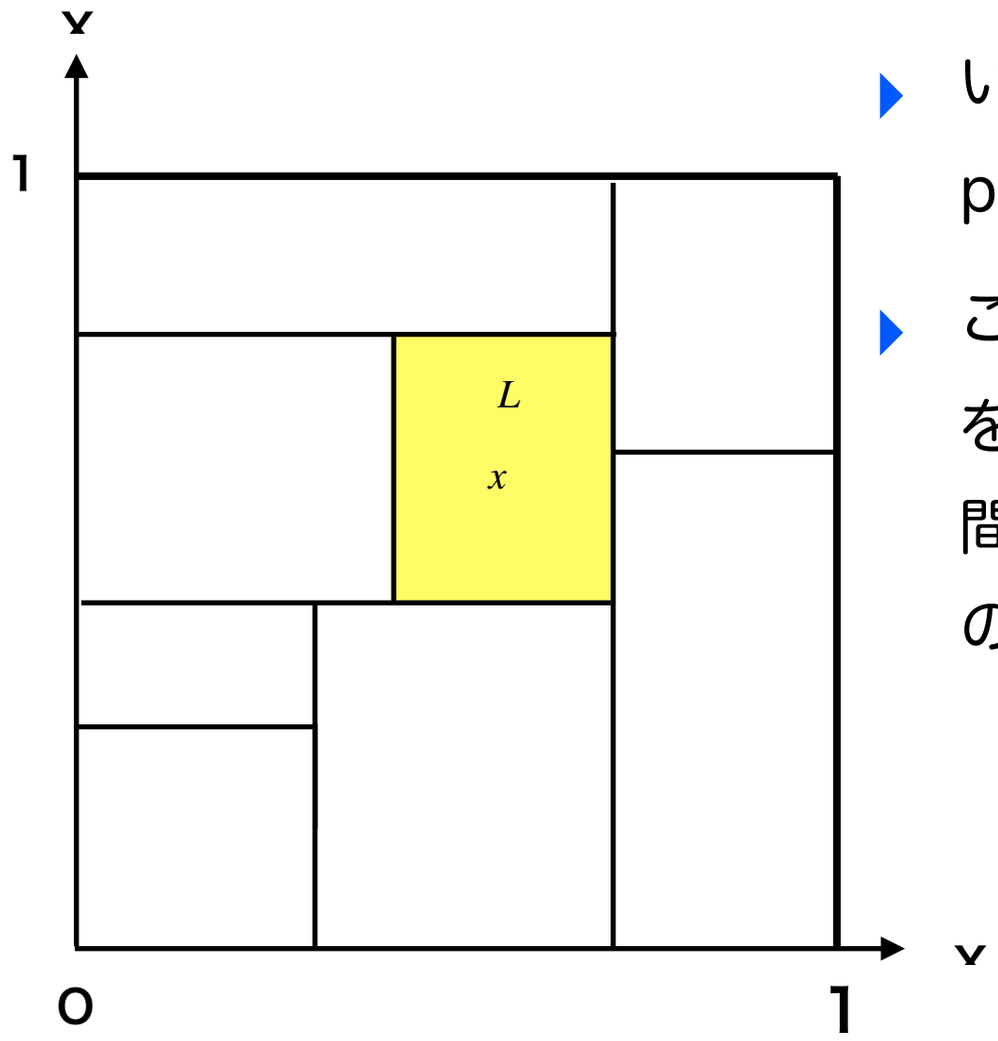
この条件のもとで、この発表での目標は、conditional average treatment effect (CATE)

$$\tau(x) \equiv E[Y_i(1) - Y_i(0) \,|\, X_i = x]$$

を推定することが目標である。

- これに対して、Treeを用いてCATE推定量に対する近似的な推定量を構成できるようにしたのが、Athey and Imbens(2016)である。
- 彼らは、regression treeを拡張し、$\tau(x)$に対する推定量を求めるcausal treeを提案した。

■Regression tree and causal tree
概念とcausal treeに対する補足



- いま、特徴空間$\mathbb{X} = [0,1]^2$に対して、その領域をpartitioningしたものが左図である。
- このとき、$x \in [0,1]^2$に対する因果効果の推定量を、Causal treeでは、xを含む特徴空間の部分空間（Leaf L）に属する処置群/対照群の標本平均の差によって計算する。

$$\frac{1}{\#\{i : W_i = 1, X_i \in L\}} \sum_{\{i:W_i=1, X_i \in L\}} Y_i - \frac{1}{\#\{i : W_i = 0, X_i \in L\}} \sum_{\{i:W_i=0, X_i \in L\}} Y_i$$

- ただし、ここから説明するcausal treeでなければ、この後構成するcausal forestが漸近正規性を持たないというわけではない。

■Regression tree and causal tree
適切なpartitioningとはなにか

- ただし、PartitioningについてAthey and Imbens(2016)では、一般的なのtreeが用いているMean-squared error(MSE)の最小化によるtreeの構成が適切かどうかについては考える必要があると述べている。
- 実際、一般的なtreeでは、partitioningと生成されたLeafにおける推定量の計算に、同じ訓練データを用いて行う。
  - これは、推定量の構成を、推定量の構成に使うサンプルを用いて行っているので、過学習を起こす可能性がある。
  - また理論的にはpartitionと、得られる推定量の独立性が失われるため、漸近正規性などを示すのが難しくなってくる。

- そこで、Athey and Imbens(2016)では、tree対して「honest」という概念を導入する。これは、partitioningに用いた情報を、leafにおける推定には用いないというものである。
  - 以降は、honestについて説明し、そのもとで導かれるcriterion functionが因果効果の推定に対してどのように影響するのかについて述べる。

■Regression tree and causal tree

Regression Treeについて

- ここではまず、共変量と結果変数の組$\{(X_i, Y_i), i = 1,2,...,N\}$が観測されたもとで、regression treeによる$\mu(x) = E[Y_i | X_i = x]$の予測について考える。
- いま、特徴空間$\mathbb{X}$を分割するpartition/treeをΠ、partitionによって生成される部分空間の数を$\#(\Pi)$とする。

$$\Pi = \{\ell_1, \ell_2, \ldots, \ell_{\#(\Pi)}\}, \qquad \cup_{j=1}^{\#(\Pi)} \ell_j = \mathbb{X}$$

- また、$\ell(x;\Pi)$を$x \in \ell$となる、部分空間$\ell \in \Pi$とする。
- ここで、分割を生成するアルゴリズムをπは、partitionの空間を$\mathbb{P}$、母集団からの標本データの空間を$\mathbb{S}$とするとき、$\pi : \mathbb{S} \to \mathbb{P}$である。
- ここで、PartitionΠが与えられたもとでの、条件付き期待値の関数$\mu(x;\Pi)$は

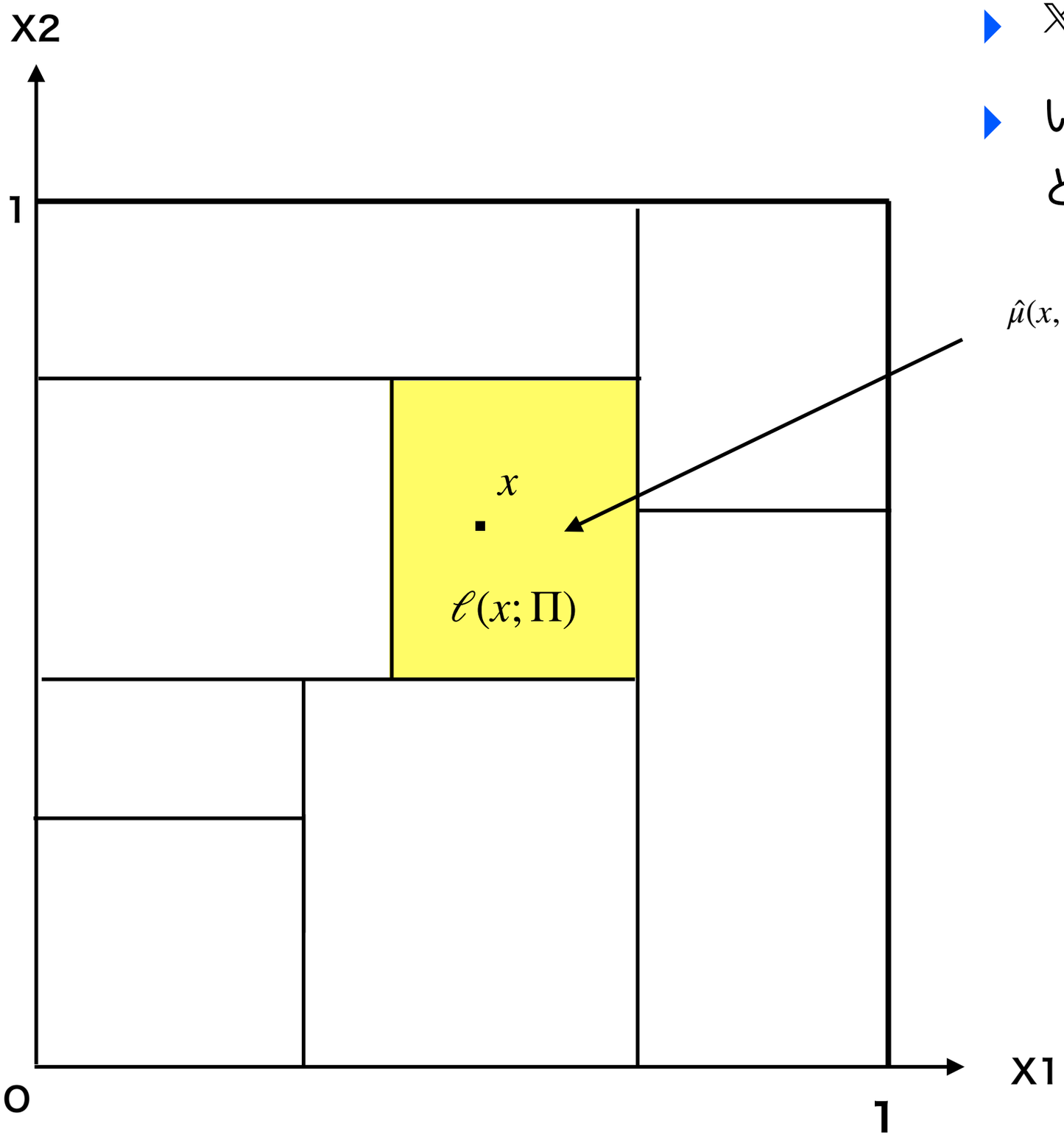
$$\mu(x;\Pi) \equiv E[Y_i | X_i \in \ell(x;\Pi)]$$

- 階段関数による$\mu(x)$に対する近似である。さらに、標本$\mathcal{S} \in \mathbb{S}$が与えられたもとでの推定量$\hat{\mu}(x, \mathcal{S}, \Pi)$は、$\mu(x;\Pi)$に対する不偏推定量である。

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

■Regression tree and causal tree

Regression Treeについて



- $\mathbb{X} = [0,1]^2$
- いま、partition Πが与えられたもとで、点xに対する予測は、

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

■Regression tree and causal tree

Honest or adaptive ?

- 一般的に、予測のためにregression treeを用いる場合、Mean-squared error(MSE)を最小にするように、treeを学習する。
- しかし、Treeにおけるpartitionの生成と、leafにおける予測に同じサンプルを用いる場合、外れ値の影響などがleaf内の推定量に大きく出る。
- そこで、honestなtargetを構成することでこの問題を回避する。
- Partition Πが与えられたもとで、テストサンプル$\mathcal{S}^{te}$に対する、Leafごとの出力値の推定に用いるサンプル$\mathcal{S}^{est}$から得られる条件付き平均の推定量のMSEを、以下で定義する。

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

- また、これに対して、テストサンプル$\mathcal{S}^{te}$および、推定に用いるサンプル$\mathcal{S}^{est}$で期待値をとったMSEを、以下で定義する。

$$\mathrm{EMSE}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) \equiv E_{(\mathcal{S}^{te}, \mathcal{S}^{est})}[\mathrm{MSE}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi)]$$

■Regression tree and causal tree
Honest or adaptive ?

- honest treeは、以下の基準を最大化するように学習させる。

$$Q^H(\pi) \equiv -E_{(\mathcal{S}^{te},\mathcal{S}^{est},\mathcal{S}^{tr})}[\mathrm{MSE}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi(\mathcal{S}^{tr}))]$$

- この基準では、treeによるpartitionの学習に用いるサンプルと、treeの予測値を計算するために用いるサンプルが異なっていることに注意する。

- 一方で、一般的なCARTでは、partitionの構成に用いるサンプルと、treeの予測値を計算するサンプルは同じである（adaptiveな場合）。

$$Q^C(\pi) \equiv -E_{(\mathcal{S}^{te},\mathcal{S}^{tr})}[\mathrm{MSE}(\mathcal{S}^{te}, \mathcal{S}^{tr}, \Pi(\mathcal{S}^{tr}))]$$

- ここで、$Q^C(\pi)$のcriterionは、新しいsplitによって常に改善されるため、この基準量による学習は非常に小さいleafを生成し、そのために推定量の分散が大きくなるという問題がある。（この問題に対処するために、実際は深さに対するペナルティーを入れるなどして層を浅くする）

- 一方で、$Q^H(\pi)$のcriterionによる学習では、データを分割する必要があり、そのために学習に使用できるデータが少なくなるという問題がある。しかしながら、次の結果から$Q^H(\pi)$による学習はpenaltyを内包することがわかる。

■Regression tree and causal tree
EMSEに対する不偏推定量の構成

- $EMSE(\Pi)$を展開すると、以下の関係式があることがわかる。

$$-\text{EMSE}(\Pi) = E_{X_i}[\mu^2(X_i, \Pi)] - E_{\mathcal{S}^{est}, X_i}[(\hat{\mu}(X_i\,; \mathcal{S}^{est}, \Pi) - \mu(X_i, \Pi))^2]$$

- ここで、この量は$\mathcal{S}^{est}$を用いています。また、$\mathcal{S}^{tr}$と$\mathcal{S}^{est}$のサンプルサイズが同じであると仮定すると、partitionを構成するためのデータ$\mathcal{S}^{tr}$を用いて、$-$EMSEに対する不偏推定量は、以下のように得られる。

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

- ※ここで、$S^2_{\mathcal{S}^{tr}}(\ell)$はLeaf ℓにおける$\mathcal{S}^{tr}$の分散である。
  - 細かい計算はAthey and Imbens(2016)を参照してください。

- 一方で、一般的なCARTにおいては、

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

■Regression tree and causal tree
honest vs adaptive

- –EMSEの最大化は、Leaf内の期待値の最大化を行いつつ、分散をできる限り小さくするというふうに解釈できる。
- ただし、予測モデルの文脈においては、
  - leafサイズ小さくなりすぎない限り、Leaf内の期待値の最大化は、Leaf内の分散の最小化との間に比例の関係がある。
  - honestかどうかは、それほど重要ではない。
- 一方、Treatment effectの推定においては、
  - Leaf内の処置効果を最大化と、Leaf内の処置群および対照群の結果変数の分散の最小化は、比例の関係があるとは限らない。
  - よって、honest treeの場合を用いると、一般的な基準で構成されるtreeとは異なる結果が得られることが期待される。

■Regression tree and causal tree

causal tree

- ここまでの議論を処置効果の推定に拡張し、観測データの組(Y_i, W_i, X_i)から、処置効果を推定するためのtreeである、causal treeについて述べる。
- ここで、$\mathcal{S}_{treat}, \mathcal{S}_{control}$をそれぞれ処置群及び対照群のサブサンプルを表す記号として用いる。また、それぞれの濃度を$N_{treat}, N_{control}$とする。
- Tree Πが与えられたもとで、任意のxと、処置wのもとでの、結果変数の母集団平均を、

$$\mu(w, x; \Pi) \equiv E[Y_i(w) \,|\, X_i \in \ell(x, \Pi)]$$

- と定義し、処置効果を

$$\tau(w, x; \Pi) \equiv E[Y_i(1) - Y_i(0) \,|\, X_i \in \ell(x, \Pi)]$$

- で定義する。また、Leaf $\ell(x, \Pi)$におけるμ, τに対する推定量は

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

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

■Regression tree and causal tree
causal tree (honest or adaptive?)

- さらに、partition Πが与えられたもとでの、処置効果に対するMSEと、EMSEを以下で定義する。

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

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

- これらは、未知のパラメータτが含まれるため、このままでは計算することができないが、これに対しては次の不偏推定量が構成できる。

$$\widehat{\mathrm{EMSE}}_\tau(\mathcal{S}^{te}, \Pi) \equiv \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}, \Pi)$$

→ **Leaf間での処置効果の差の最大化**

$$- \frac{2}{N^{tr}} \cdot \sum_{\ell \in \Pi} \left(\frac{S^2_{\mathcal{S}^2_{treat}}(\ell)}{p} + \frac{S^2_{\mathcal{S}^2_{control}}(\ell)}{1-p} \right)$$

→ **Leaf内の結果変数の最小化（通常のTreeだと含まれない）**

ただし、 $p = \dfrac{N_{treat}}{N}$

■Regression tree and causal tree
causal tree (honest or adaptive?)

- Honest Treeによる推定は、以上の議論から
  - Leaf内の結果変数の分散に対して罰則をつけたもとでの、処置効果の最大化であると解釈できる。
  - このことから、Honest Treeにおいては、
    - 通常のMSE最小化によって得られるtreatment effecの推定量よりも、推定される処置効果の分散をsplittingによって小さくなる。
    - 実際には、処置効果に影響しない共変量が結果変数の変動に影響を与えている場合には、その変数でsplittingを選択する場合が生まれる。

- この結果は、treeをbase-learnerとするcausal forestにおいても効いてくる。

predictionの問題とは異なり、処置効果を推定をする場合には
Treeを学習する上でsplitting criterionは非常に大切

目次.

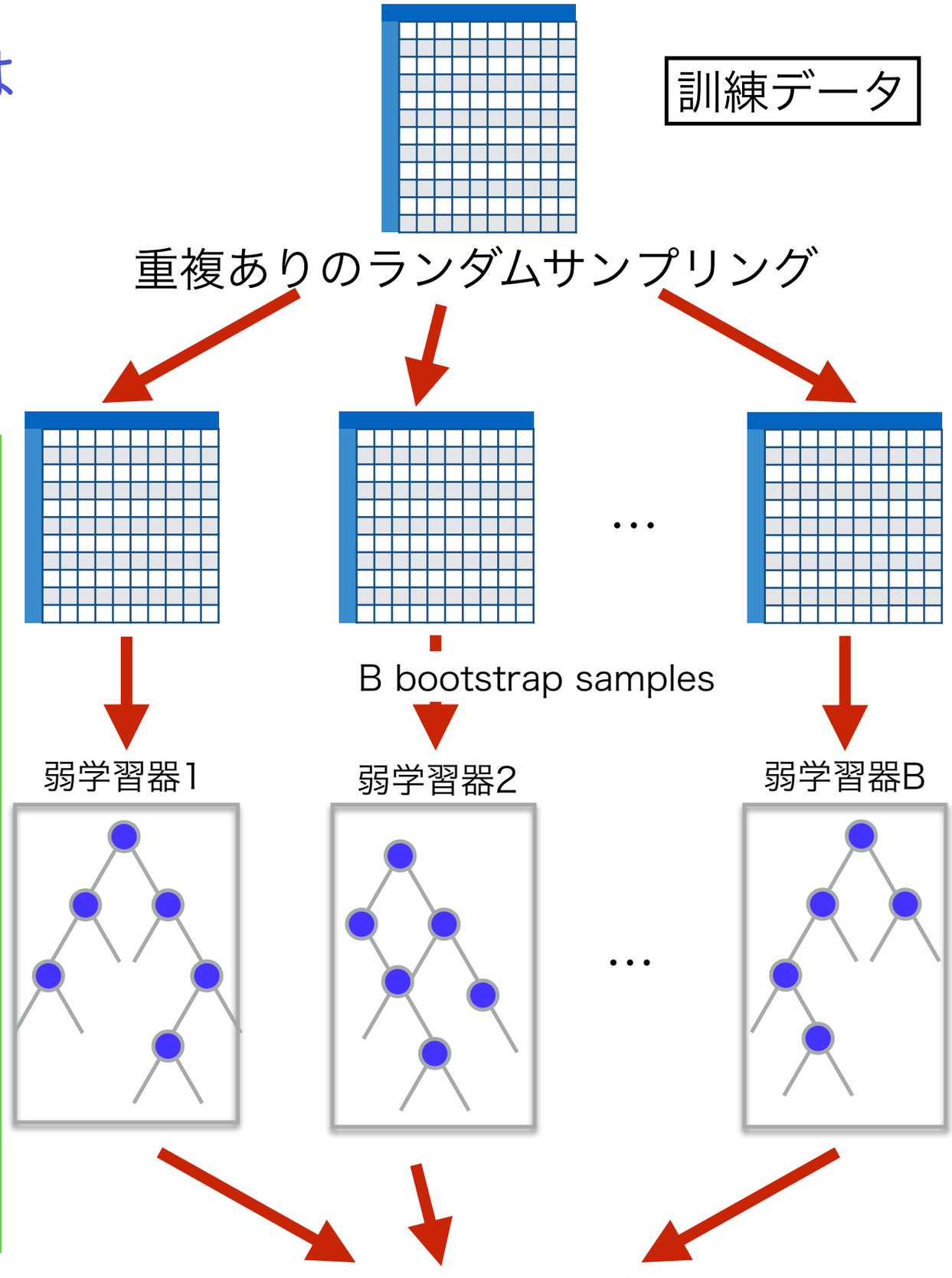
causal forest

（一般的な）random-forestとは

- 特徴X_iから結果Y_iを予測する問題を考える。

Random forest(アルゴリズム)

- サイズNの訓練データから、サイズ$s(<N)$のbootstrapサンプル（重複ありのサブサンプル）をB個取る。
- それぞれのbootstrapサンプルに対して、適当なcriterion funtionを持つCARTを学習する。
- 特徴xに対して、B本のTreeそれぞれに対して予測値を計算し、それらの平均をrandom forestの推定量として返す。



■causal forest

Random forestに対するいくつかの変更

- ここで、Wager and Athey(2018)で解析されているrandom forestは、一般的なRFといくつかの点で異なっている。
  - 一般的なRFでは、重複ありのbootstrapサンプリングを行うが、ここでは重複なしのRFである。
  - 学習に使用するTreeは、一般的なCARTではなく、causal treeの議論でも使用した、Honest性を満たすTreeである。
    - Partitioningと得られる推定量が独立である。

W&A(2018)のRandom forest(アルゴリズム)

- サイズNの訓練データから、サイズ$s(<N)$の重複なしサンプルB回取る。
- それぞれのサンプルに対して、適当な条件を満たすHonest Treeを学習する。
- 特徴xに対して、B本のHonest Treeそれぞれに対して予測値を計算し、それらの平均をrandom forestの推定量として返す。

■Causal forest

木から森へ

- Wager and Athey(2018)は、causal treeを含むようなある性質を満たすtreeに対して、次の3つを示している。
  - treeに対するバイアスの評価
  - treeをbase-learnerとしたもとでの、random-forestの漸近正規性
  - treeをbase-learnerとしたもとでの、random-forestの分散の推定量の提案

- Wager and Athey(2018)はHonest treeのアルゴリズムを2つ示している。
  - Double sample tree（予測及び処置効果の推定の両方の場合）
  - Propensity tree（処置効果の推定の場合のみ）

- これらについて述べたあとで、この他の条件について、詳しく述べる。

■Causal forest

Double sample trees

INPUT

- サイズNの訓練データ: 予測の場合(Y_i, X_i)/処置効果の推定の場合(Y_i, W_i, X_i)
- 最小Leafサイズ: $k \in \mathbb{N}$

Treeの構成

- 訓練データ$\{1,2,...,N\}$から、サイズsのsub-sampleを重複無しでとり、サイズが$|I| = \lfloor s/2 \rfloor$及び$|J| = \lceil s/2 \rceil$となる、2つの排反な集合$I, J$をとる。
- Iのデータの共変量X（及びW）の情報と、Jのデータすべてを用いて、treeによってpartitionを生成する。このときIの結果変数の情報は用いない（honest性）。また、partitionによって生成される部分空間（Leaf）に、Iのサンプルが予測の場合k以上、処置効果の推定の場合は各群のサンプルがk以上含まれるようにする。
- 生成されたLeaf毎の推定量を、Iの標本を用いて計算する。

■Causal forest

Propensity trees

INPUT

- サイズNの訓練データ:(Y_i, W_i, X_i)
- 最小Leafサイズ: $k \in \mathbb{N}$

Treeの構成

- 訓練データ$\{1,2,...,N\}$から、サイズsのsub-sampleの集合I（$|I| = s$）を重複無し構成する。
- Iのデータの共変量X及び処置変数Wの情報を用いて、treeによってpartitionを生成する。このときIの結果変数の情報は用いない（honest性）。また、partitionによって生成される部分空間（Leaf）に、各群のサンプルがk以上含まれるようにする。
- 生成されたLeaf毎の推定量を、Iの標本を用いて計算する。

■Causal forest

RFのbase-learnerに対するその他の条件①

random split

- Treeがrandom split性を持つとは、特徴空間の次元pのすべての次元が、各splitでsplittingの対象となる可能性があることである。すなわち、Treeにおけるsplitting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j $(= 1,2,...,p)$番目の変数がsplittingされる確率がπ/pで下から抑えられる。

- これは、漸近的にLeafのすべての次元の大きさが、$N \to \infty$の状況のもとで、0に収束することを意味する。
- Consistencyを示すために必要な条件である。

■Causal forest

RFのbase-learnerに対するその他の条件②

α-regular tree

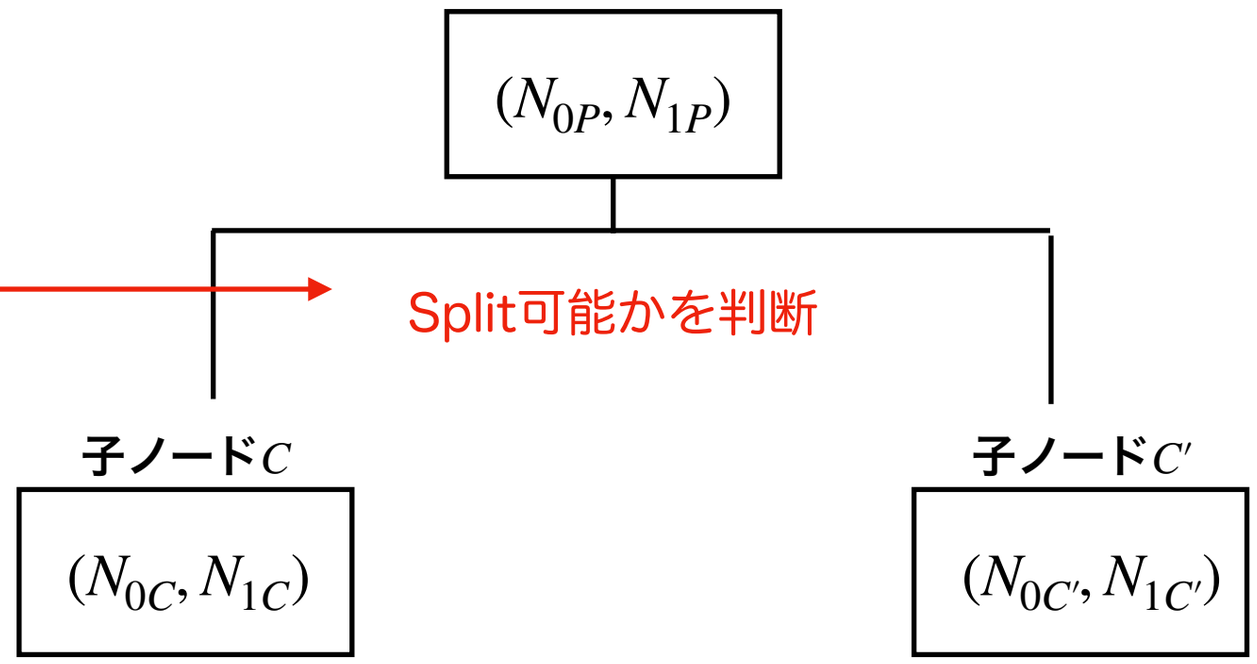
- Treeがα-regular性を持つとは、任意の深さにおいて、splitによって生成される2つのノードが、splitの対象となっているサンプルの100α%以上のサンプルを含むことであり、さらにterminal nodeにk以上および$2k-1$以下のサンプルが含まれることである。

- 処置効果の推定の場合は、さらに処置群/対照群のサンプルがそれぞれk以上、$2k-1$以下terminal nodeに含まれることである。

- Double sample treesの場合は、上の条件がIのサンプルに対して成り立つことと定義する。

$$N_{0C}, N_{0C'} \geq \alpha N_{0P}$$
$$N_{1C}, N_{1C'} \geq \alpha N_{1P}\,.$$
$$k \leq N_{1C}, N_{1C'}, N_{0C}, N_{0C'} \leq 2k-1$$

Regular性の条件



■Causal forest

Asymptotic normality for random forest

Theorem 1 of Wager and Athey 2018

- 適当な正則条件(Wager and Athey, 2018)のもとで、honest, α-regular ($\alpha \leq 0.2$)及び、random-split性を満たすTreeをbase-learnerとする。このとき$\mu(x) = E[Y_i | X_i = x]$がLipschitz連続性を満たし、サブサンプルサイズs_nのスケールが、

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

- であるようなランダムフォレストの推定量$\hat{\mu}^{RF}(x)$は漸近正規性を持つ。

$$\frac{\hat{\mu}_n^{RF}(x) - \mu(x)}{\sigma_n(x)} \to N(0,1) \qquad \text{for a sequence } \sigma_n \to 0$$

- さらに、分散σ_nに対して、infinitesimal jackknife推定量$\hat{V}_{IJ}$(Wager et al. 2014)は、次の性質を満たす。

$$\hat{V}_{IJ}(x)/\sigma_n(x) \to 1$$

■Causal forest

推定量の漸近的なバイアスに対するBound

Lemma 2 of Wager and Athey 2018

- Tがα-regular及び、random-splitであるとする。xを含むleafを$L(x)$とし、$\mathrm{diam}(L(x))$を$L(x)$が含むことのできる最大の線分の長さとする。また、$L_j(x)$を、j番目の変数の軸に沿った最大の線分とする。
- このとき、Tree TのLeafに対して、任意の$\eta \in (0,1)$と、十分大きなsのもとで、

$$\mathbb{P}\left[\mathrm{diam}_j(L(x)) \geq \left(\frac{s}{2k-1}\right)^{-\frac{0.99(1-\eta)\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}}\right] \leq \left(\frac{s}{2k-1}\right)^{-\frac{\eta^2}{2}\cdot\frac{1}{\log(\alpha^{-1})}\frac{\pi}{p}}$$

- が成り立つ。

この議論は、Treeによって生成される任意のLeaf $L(x)$の大きさがα−regularかつrandom-splitであれば0に行くことを示している。
（ηはChernoff boundの計算の過程で出てくる）

■Causal forest

推定量の漸近的なバイアスに対するBound

Theorem 3 of Wager and Athey 2018

- Lemma 2の条件に加えて、$\mu(x)$がLipschitz性を持ち、かつbase-learnerであるTree Tがhonestであるとき、$\alpha \leq 0.2$のrandom forestのバイアスは以下のように評価できる。

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

$$0 < \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \leq 1 \quad \text{は } \alpha \text{ に対する単調増加関数}$$

$$\frac{\pi}{p} \quad \text{は次元}p\text{に対する単調減少関数}$$

よって、random forestによるバイアスの収束オーダーを大きくするためには、αをできる限り大きく取る必要がある。

また、次元が高くなるとバイアスの収束オーダーは悪くなる。

■Causal forest

推定量の漸近的なバイアスに対するBound

- ここまでは(X_i, Y_i)に対するrandom forestであった。
- Causal forestが漸近正規性を持つためには、α-regular treeの定義を拡張し、処置群または対照群に属するサンプルが、各terminal nodeにおいてk以上$2k-1$以下含まれているという条件に変更する必要がある。
- このほかは、(X_i, Y_i)対して置かれる正則条件を、$(Y_i(1), X_i)$及び$(Y_i(0), X_i)$に対して成り立つと変えるだけで良い（Theorem 11 of Wager and Athey(2018)）。
- この結果から、Causal Treeをbase-learnerとするrandom forest (causal forest)は、漸近正規性を持つことが示される。

→ よって、これらの結果を用いれば各個体に対する処置効果が、傾向スコアなどの仮定を置くことなく推定することができる。

目次.

■Generalized Random Forest(GRF)
パラメータ推定をより柔軟にするための枠組み

- ここまでの話は、モデルの仮定を一切置くことなく処置効果の推定をするという話を進めてきた。
- しかし、因果推論においては、ある程度、具体的なモデルを仮定したもとで、CATEを推定したいというモチベーションも存在する（IVモデルなどはその代表例である）。
- このような場合に、Generalized Random Forest (Athey, Tibshirani and Wager, 2019)の枠組みは有用である。

GRFにおける問題設定

- データ$\{(X_i, O_i), i = 1,2,...,N\}$が与えられたとき、スコア関数$\psi(\cdot)$及び、パラメータ$\theta(x)$および、nuisance$\nu(x)$から構成される次の局所推定方程式の解を求める。

$$E[\psi_{\theta(x),\nu(x)}(O_i) \mid X_i = x] = 0 \quad \text{for all } x \in \mathcal{X}$$

ただし、$\mathcal{X}$はX_iのサポートである。

■Generalized Random Forest(GRF)

具体的な問題

- Random forest
  - 通常のrandom forestでは、$O_i = Y_i$として、次のスコア関数を考えたものと同じになる。
  - $\psi_{\mu(x)}(Y_i) = Y_i - \mu(x)$
- 分位点回帰モデル
  - $O_i = Y_i$のとき、以下の式を満たすq分位点の関数$\theta_q(x)$の推定
  - $\psi_\theta(Y_i) = q1(\{Y_i > \theta\}) - (1-q)1(\{Y_i \leq \theta\})$
- 操作変数による回帰モデル
  - $O_i = (Y_i, W_i, Z_i)$として、W_iを2値の処置変数、Z_iを2値の操作変数とする。
  - $Z_i \perp\!\!\!\perp e_i | X_i$及び、$Cov(Z_i, W_i | X_i) \neq 0$を仮定したもとで、モデル$Y_i = \mu(X_i) + \tau(X_i)W_i + e_i$における$\tau(X_i)$を推定する。

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

■Generalized Random Forest(GRF)

kernel回帰 と GRF

- 局所推定方程式を解く方法としてよく知られているのは、kernel関数による重み付け推定量である。
- これは、xにおけるパラメータ$\theta(x)$を推定する際に、その周辺のサンプルに対して重み付けをして、推定をするというものである。

$$\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \arg\min_{\theta,\nu} \left\{ \left\| \sum_{i=1}^{N} \alpha_i(x)\, \psi_{\theta,\nu}(O_i) \right\|_2 \right\}$$

- $\alpha_i(x)$は従来の局所線形回帰モデルであれば、kernel関数である。
- しかし、Generalized random forestは、この重みをrandom forestを用いてnonparametricに推定するというアイデアである。

- 次に、Generalized random forestのアルゴリズムについて説明する。

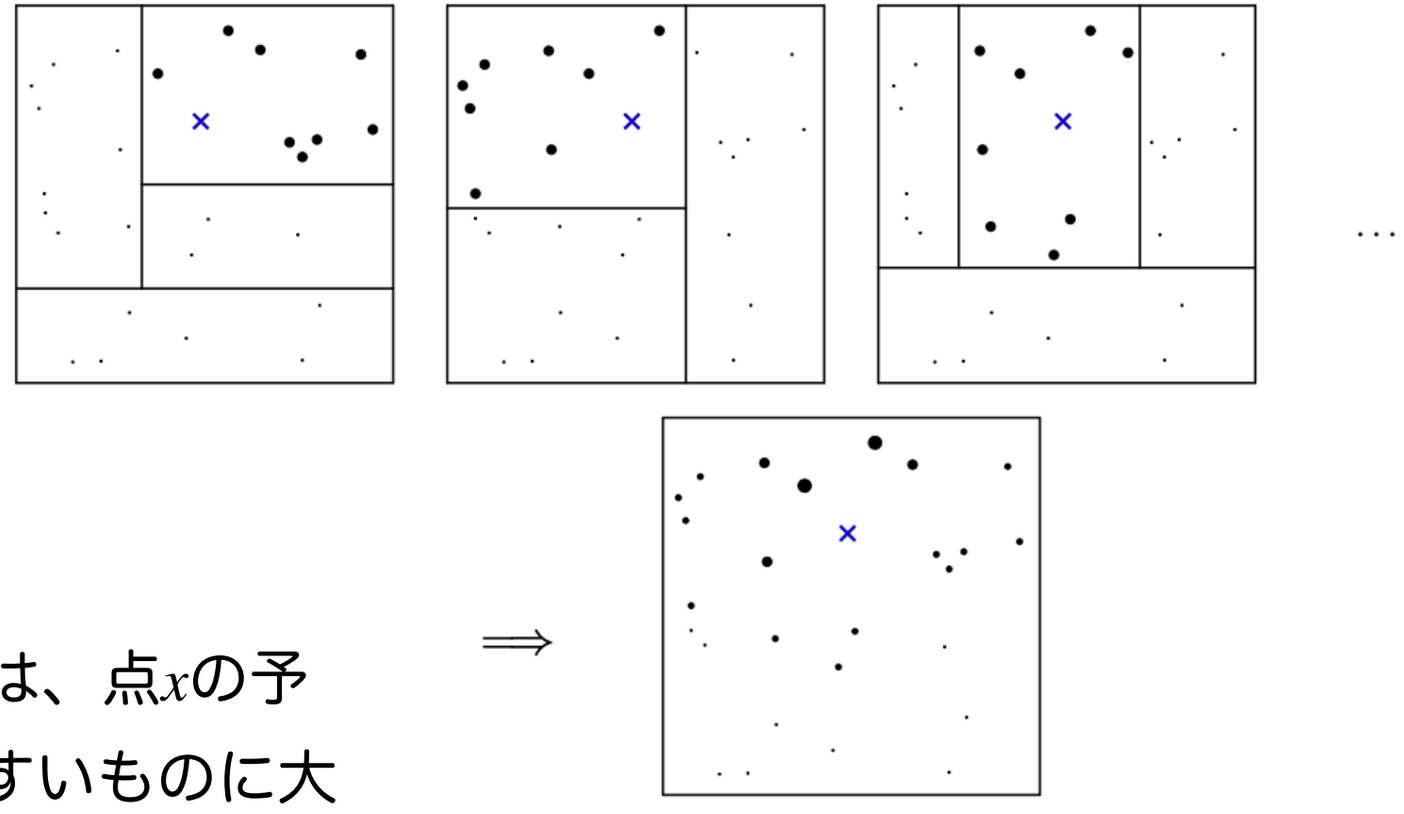
■Generalized Random Forest(GRF)

$\alpha_i(x)$の推定

- GRFでは、まず通常のrandom forestと同様にhonest treeを作る。これらをT_b $(b = 1,2,...,B)$とする。
- 次に、点xが与えられたときに、生成した各Tree T_bに対して、xを含むLeafを$L_b(x)$とすると、$L_b(x)$に含まれる標本iに対して、次のように重みを計算する。

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

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



ここで得られる重みは、点xの予測に対して影響しやすいものに大きな重みがかかっている

■Generalized Random Forest
GRFで用いるTree アルゴリズム

- 重みの推定に用いられるTreeを、それぞれが$\theta(x)$を推定するTreeとして構成するのが妥当である。
- そのため、Treeのcriterion functionを以下のように設定する。
- treeのParent node Pに対して、推定方程式の解を$\hat{\theta}_P, \hat{\nu}_P$とする。

$$\left(\hat{\theta}_P, \hat{\nu}_P\right) \in \arg\min_{\theta,\nu} \left\{ \left\| \sum_{X_i \in P} \psi_{\theta,\nu}(O_i) \right\|_2 \right\}$$

- 次に、Treeの分割によってPから生成される子ノードをC_1, C_2とする。
- このとき、split criterionを考える。

$$\mathrm{err}(C_1, C_2) = \sum_{j=1,2} \mathbb{P}[X \in C_j \,|\, X \in P] E[(\hat{\theta}_{C_j} - \theta(X)^2) \,|\, X \in C_j]$$

- このcriterionを最小化する分割は、$\hat{\theta}_{C_1}$と$\hat{\theta}_{C_2}$の差を最大化する分割である（Proposition 1 of Athey et al. 2019)。

■Generalized Random Forest
GRFで用いるTree アルゴリズム

- しかし、この方法は、splitの毎に毎回推定方程式を解く必要があり、計算量の意味で現実的ではない。
- そこで、最適なsplitを近似的に求める方法を提案している。この方法は、具体的には、Parent nodeの推定結果を用いて、criterion functionを近似する方法である。
- 具体的には、以下の基準を最大化するようなspliting を行う。

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

- ただし、 $\rho_i = -\xi^T A_P^{-1} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \in \mathbb{R}$

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

Athey et al. (2019)では、このsplitting criteriaと、$\mathrm{err}(C_1, C_2)$によるsplitting criteria が適当な条件のもので漸近的に同値であることを示している。

■Generalized Random Forest

GRF推定量の漸近正規性

- GRFによる推定量は、Athey et al. (2019)によるTheorem 5より、漸近正規性を持つことが証明されている。
- GRFは、Wager and Athey(2018)の結果を拡張子、因果推論の枠組みから発展して、推定方程式の解をrandom forestを用いて解くための枠組みを提案した。

- 最後に、紹介するOrthogonal random forestは、Neyman orthogonalityを用いて推定方程式に対して工夫を加えることで、推定される関数の分散することを目指すものである。

目次.

■まとめや近年の発展など

- 本発表ではRegression Treeから議論をはじめて、Causal treeについて述べ、その後forestへと議論を拡張し、最後にGeneralized random forestについて説明した。
- Random forestの近年の理論面発展は著しいものがあり、今後も高次元への対応や、データ解析応用事例などが増えてくると思われる。

- また、近年では、ノンパラメトリックな推定量の改善を、Neyman orthogonality(Chernozhukov et al. 2018) を用いて行うのが1つの主流になりつつある。
- Neyman orthogonalityを満たす推定方程式とは、nuisance paramterに関するGateaux微分が0になるような推定方程式である。
  - →nuisanceの変動によって、推定方程式が崩れにくいという性質がある。
- この性質を利用した方法には、例えば
  - R-Learner (Nie and Wager, 2020)
  - Orthogonal random forest(Oprescu, Syrgkanis and Wu, 2019)
- がある。

参考文献

参考文献

- Wager, S., and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113, 1228-1242.
- Wager, S., and Walther, G. (2016). Adaptive concentration of regression trees, with application to random forests. arXiv
- Athey, S., Tibshirani, J, and Wager, S. (2019). Generalized random forests. Ann. Statist. 47 (2019), no. 2, 1148-1178.
- Athey, S., snd Imbens, G. (2015) Recursive partitioning for heterogeneous causal effects. arXiv.
- X, Nie., and S Wager. (2020). Quasi-Oracle Estimation of Heterogeneous Treatment Effects, Biometrika. forthcoming.
- Oprescu, M., Syrgkanis, V., and Wu, Z, S. (2019) Orthogonal Random Forest for Causal Inference. arXiv.
- Chernozhukov et al . (2018). Double/debiased machine learning for treatment and structural parameters. Econometric Journal, 21.

ありがとうございました