6/25　勉強会資料

原田悠介

Contents

- 因果推論分野の細菌の発展(引用)

  統計的因果推論とデータ解析

  https://speakerdeck.com/tomoshige_n/causal-inference-and-data-analysis

- 論文紹介

  Wager & Athey(2018)

  “ Estimation and Inference of Heterogeneous Treatment Effect using Random Forests”

  Journal of American Statistical Association vol.523　1228-1242

近年の因果推論分野の発展(引用)

- 近年、機械学習を用いてheterogeneous treatment effect の推定を行うという動きが活発化している
- これらの成果は因果効果の推定量の漸近的性質を明らかにしているものが多い⇒仮説検定、統計的推論が可能

- 大きく分けると2つのアプローチがある

➢Random forestを応用した方法(今回はこちらを紹介します)

➢結果変数/傾向スコアをML regressionで推定してから推定した値を用いて回帰の問題をMLによって解く方法

Estimation and Inference of Heterogeneous Treatment Effect using Random Forests

原田悠介

論文目次

- Introduction
- Causal Forests
- Simulation Experiments

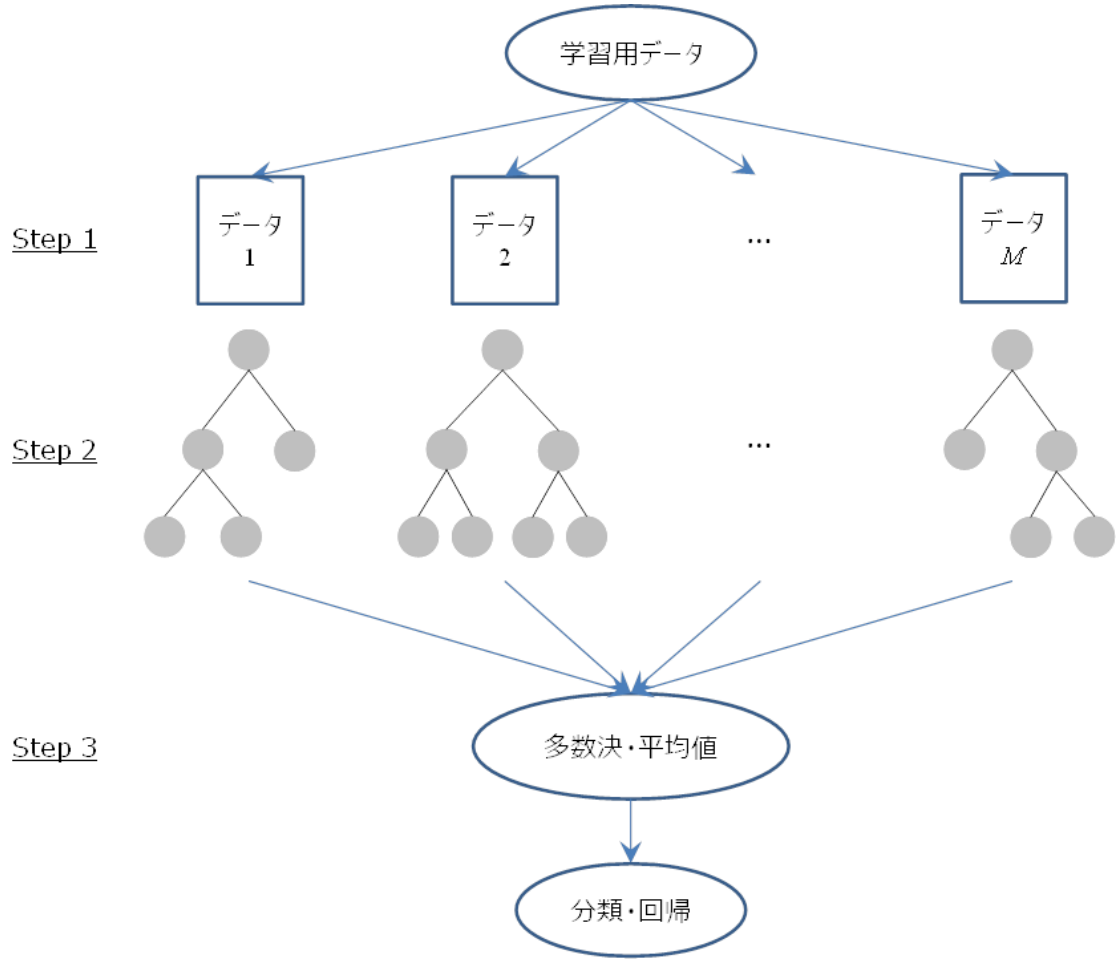
Introduction<1/3>

- 因果推論の応用先⇒医療、マーケティング、政策評価など様々
- これらの分野ではA/Bテスト(RCT)が行われる
- 医療や経済の分野ではRCTを行うための手法が決まっているが、heterogeneousな因果効果を評価するのは難しい
- 本論文ではランダムフォレストをベースにしたheterogeneousな因果効果を求める手法(causal forest)が紹介されている

＜補足＞ランダムフォレスト(ざっくりと)

- 観測データからランダムサンプリングされたトレーニングデータとランダムに選択された説明変数を用いることにより、相関の低い決定木群を多数作成する。
- 各決定木の出力を多数決、平均などで1つの値に集約する

$$\hat{\mu}(x) = \frac{1}{|\{i: X_i \in L(x)\}|} \sum_{\{i:X_i \in L(x)\}} Y_i$$



Introduction<2/2>

- ランダムフォレストを因果推論の枠組みで実用化するには

の漸近正規性と一致性の結果が必要である

　→Causal Forestはこの二つの性質を持ち、信頼区間の構築と検定ができる

Causal Forests<1/7>Notation

- サンプル数　$n(i = 1,2, \dots, n)$
- 共変量ベクトル　$X_i \in [0,1]^d$
- 応答変数(結果変数)　$Y_i \in \mathbb{R}$
- 割当変数　$W_i \in \{0,1\}$
- 潜在結果　Y_i^1, Y_i^0

Causal Forests<2/7>処置効果の推定

- 通常の処置効果$\tau(x)$の推定

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

無交絡性$Y_i^1, Y_i^0 \perp W_i | X_i$を仮定すれば、処置効果は例えば

$$E\left[Y_i\left(\frac{W_i}{e(x)} - \frac{1 - W_i}{e(x)}\right)\middle| X_i = x\right], e(x) = p[W_i | X_i = x]$$

となる(IPW推定量)

Causal Forest<4/7>アイデア

- s個(s/n<<1)のサブサンプルからB個の木を作成し、それぞれについて以下を推定する

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

- その後、推定した$\hat{\tau}(x)$を集約する

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

Causal Forests<5/7>使用条件

- $E[Y^0|X=x], E[Y^1|X=x]$がリプシッツ連続である

(https://ja.wikipedia.org/wiki/%E3%83%AA%E3%83%97%E3%82%B7%E3%83%83%E3%83%84%E9%80%A3%E7%B6%9A)

- 識別性条件$\varepsilon < P[W=1|X=x] < 1-\varepsilon$を満たす

- “honesty”条件を満たす

  Y_iはL内の処置効果の推定と Xの分割のどちらか一方にしか用いられないという条件

  本論文ではhonesty条件を満たす木を作るアルゴリズムを2つ紹介している

Causal Forests<6/7>double-sample tree

1. データ{1,2,…,n}からサブサンプルs個を重複なしで取り出し、2等分する $|\mathcal{I}| = \lfloor s/2 \rfloor \quad |\mathcal{J}| = \lceil s/2 \rceil$
2. 再帰的分割によって木を作成する。分割にはJサンプルのデータとIサンプルのX,Wを用いる
3. IサンプルだけのYを用いてL内の処置効果を推定する

   Lには最低k個の処置の異なるIサンプルのデータが含まれる

   必要がある

Causal Forests<7/7>Propensity tree

1. データからサブサンプルs個を重複なしで取り出す
2. サブサンプルのX,Wを使って、分類木を作る
3. 各Lについて処置効果を推定する
   L内には最低k個の処置の異なるデータが含まれる必要がある

Simulation Experiments

- 本論文では、Propensity treeを用いたCausal forestの性能をk最近傍法(kNN)の性能とシミュレーションによって比較している
- 設定

$$\text{main effect: } m(x) = 2^{-1}\,\mathbb{E}[Y^{(0)} + Y^{(1)} \mid X = x],$$

$$\text{treatment effect: } \tau(x) = \mathbb{E}[Y^{(1)} - Y^{(0)} \mid X = x],$$

$$\text{treatment propensity: } e(x) = \mathbb{P}[W = 1 \mid X = x].$$

$$Y^{(0/1)} \sim \mathcal{N}(\mathbb{E}[Y^{(0/1)} \mid X],\ 1)$$

Simulation Experiments<>

- 処理効果$\tau(x)$を0に固定し、e(x)とm(x)の間の相互作用によるバイアスによる推定誤差への強さを測る

➢設定

$$e(X) = \frac{1}{4}(1 + \beta_{2,4}(X_1)), \qquad m(X) = 2X_1 - 1,$$

➢500サンプル、2~30変数、B=1000、s=50のpropensity treeをもとに評価

Simulation Experiments

Table 1. Comparison of the performance of a causal forests (CF) with that of the k-nearest neighbors (k-NN) estimator with $k = 10, 100$, on the setup (27). The numbers in parentheses indicate the (rounded) standard sampling error for the last printed digit, obtained by aggregating performance over 500 simulation replications

	Mean-squared error			Coverage		
d	CF	10-NN	100-NN	CF	10-NN	100-NN
2	0.02 (0)	0.21 (0)	0.09 (0)	0.95 (0)	0.93 (0)	0.62 (1)
5	0.02 (0)	0.24 (0)	0.12 (0)	0.94 (1)	0.92 (0)	0.52 (1)
10	0.02 (0)	0.28 (0)	0.12 (0)	0.94 (1)	0.91 (0)	0.51 (1)
15	0.02 (0)	0.31 (0)	0.13 (0)	0.91 (1)	0.90 (0)	0.48 (1)
20	0.02 (0)	0.32 (0)	0.13 (0)	0.88 (1)	0.89 (0)	0.49 (1)
30	0.02 (0)	0.33 (0)	0.13 (0)	0.85 (1)	0.89 (0)	0.48 (1)

Simulation Experiments

- m(x)=0、e(x)=0.5で固定し、$\tau(x)$の不均一性に対する性能を測る

➢設定　$$\tau(x) = \varsigma(X_1)\,\varsigma(X_2), \quad \varsigma(x) = 1 + \frac{1}{1 + e^{-20(x-1/3)}}.$$

➢5000サンプル、2~8変数、B=2000、s=2500のdouble-sample treeをもとに評価

Simulation Experiments

Table 2. Comparison of the performance of a causal forests (CF) with that of the k-nearest neighbors (k-NN) estimator with $k = 7,\ 50$, on the setup (28). The numbers in parentheses indicate the (rounded) standard sampling error for the last printed digit, obtained by aggregating performance over 25 simulation replications

	Mean-squared error			Coverage		
d	CF	7-NN	50-NN	CF	7-NN	50-NN
2	0.04 (0)	0.29 (0)	0.04 (0)	0.97 (0)	0.93 (0)	0.94 (0)
3	0.03 (0)	0.29 (0)	0.05 (0)	0.96 (0)	0.93 (0)	0.92 (0)
4	0.03 (0)	0.30 (0)	0.08 (0)	0.94 (0)	0.93 (0)	0.86 (1)
5	0.03 (0)	0.31 (0)	0.11 (0)	0.93 (1)	0.92 (0)	0.77 (1)
6	0.02 (0)	0.34 (0)	0.15 (0)	0.93 (1)	0.91 (0)	0.68 (1)
8	0.03 (0)	0.38 (0)	0.21 (0)	0.90 (1)	0.90 (0)	0.57 (1)

Simulation Experiments