統計的機械学習 （応用計量分析２）第7回

条件付き平均因果効果の推定法（参考pdf 5.3節、6章）

本日の内容

CATE推定法の中でも決定木系の手法について学ぶ
ベイズ推論についてもとりあげる

- 1. ガイダンス・因果推論と機械学習の概論
- 2. 意思決定理論（期待効用理論）の復習、因果推論との関係
- 3. 潜在結果モデルに基づく因果推論の枠組み
- 4. 平均因果効果の推定法
- 5. 条件付き平均因果効果（CATE）の推定法１：メタ学習器
- 6. メタ学習器の続きと二重機械学習
- **7. CATEの推定法３：決定木と決定森**
  - **深層学習に基づく方法**
- 8. 構造方程式モデルとバックドア基準
- 8. 因果探索
- 9. 発展的な因果推論手法：フロントドア調整、操作変数法、回帰不連続デザイン、代理変数法
- 10. 続き
- 11. 発展的な意思決定理論
- 12. 強化学習
- 13. オフライン強化学習
- 14. バンディット
- 15. まとめ

CATE推定のためのメタ学習とDML

- メタ学習器
  - S-Learnerは行動を入力変数に含めてモデル化
  - T-Learnerは行動ごとに分けてモデル化
  - X-LearnerはT-Learnerをベースに曝露/対照群間の偏りによって結果モデルの推定精度がネックにならないよう群ごとに$\hat{\tau}$を学習して重みつき平均
  - DR-LearnerはAIPW法の擬似結果を回帰する方法で二重頑健性を保証
- 二重機械学習（DML）
  - 行動が二値$a \in \{0,1\}$の場合に加えて連続値の場合に対して線形モデルとした

    セミパラメトリックモデル（a以外に関しては非線形を許容）
  - 結果yを説明変数（a以外）に回帰したモデルの残差を、行動を説明変数に回帰したモデルの残差に回帰
  - 残差が過小推定とならないようにデータ分割、クロスフィットを行う

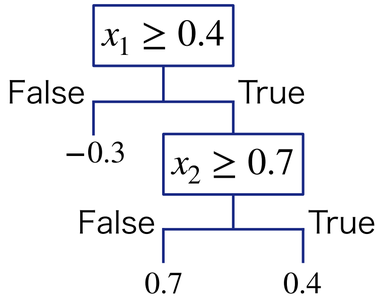
出欠コード
XXXX

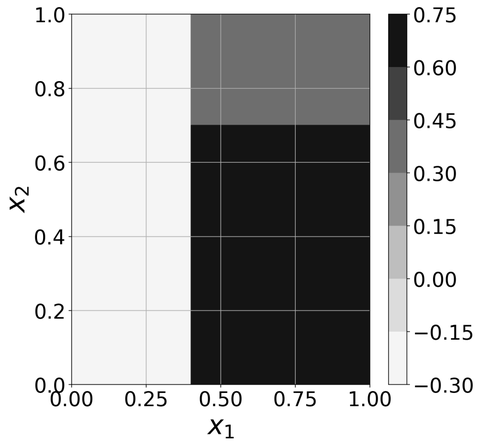
決定木は各次元で領域分割した上で定数で予測するモデル

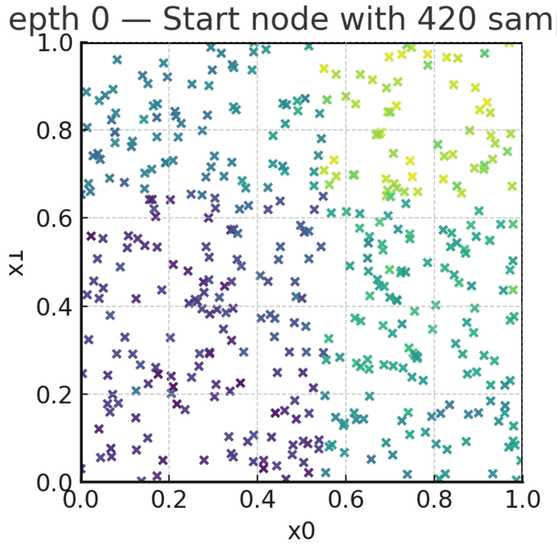
- 決定木は説明変数の各次元の値の大小によって場合分けした領域ごとに定数を予測値として出力
  - 分割ごとの変数としきい値、領域ごとの予測値が学習可能なパラメタ
- 決定木の基本的な学習法

1. 現状の領域分割に追加して最も目的関数（損失）を改善する分割を逐次的に探す
   - 二乗誤差による回帰では、領域の予測値は平均値とする
2. 停止条件に合致したら停止
   - 木の深さが一定になるか、領域内のサンプルサイズが一定以下になったらその領域は分割終了

- 各分割は説明変数軸に沿って行われる
  - 「斜め切り」はしない。ただ、表データは各軸が意味的に独立していることが多く、現実には有益な制約であることが多いとされる







因果推論向けの決定木

因果木：データ分割の工夫と因果効果の直接推定

- 以下は割付確率pが一定のランダム化試験（0<p<1）
  - 決定木＝領域内は同一の値で近似
    - xの属する領域を$\ell(x)$として、　（※葉ノード＝リーフのL。損失関数のℓとは全く別）

      $\tau(x) = \mathbb{E}[y_1 \mid x] - \mathbb{E}[y_0 \mid x] \approx \mathbb{E}[y_1 \mid \ell(x)] - \mathbb{E}[y_0 \mid \ell(x)]$
  - 領域内の曝露群と対照群の平均差を使う（観察データでは交絡調整が必要）
- 領域分割の最適化に用いる訓練データ$\mathcal{D}^{\mathrm{tr}}$、予測値の計算に用いる推定用データを分ける：**Honestアプローチ**
  - 説明変数の空間$\mathcal{X}$の領域$\ell_k \subset \mathcal{X}$によるある分割 $\Pi = \{\ell_k\}_k$ $(\cup_k \ell_k = \mathcal{X})$ に対して推定用データ$\mathcal{D}^{\mathrm{est}}$を用いて
  - $\hat{\tau}(x; \mathcal{D}^{\mathrm{est}}, \Pi) = \dfrac{1}{|\mathcal{D}_1^{\ell(x)}|} \sum_{i \in \mathcal{D}_1^{\ell(x)}} y^i - \dfrac{1}{|\mathcal{D}_0^{\ell(x)}|} \sum_{i' \in \mathcal{D}_0^{\ell(x)}} y^{i'}$　と推定する。ここで$\mathcal{D}_a^{\ell(x)}$は$\mathcal{D}^{\mathrm{est}}$のうち$\ell(x)$に属し行動が$a$のサンプル
- 目的関数も推定分散を考慮して修正
  - 以下の目的関数を**最大化**するように領域Πを学習
  - $-\mathrm{EMSE}_\tau(\Pi) = \mathbb{E}_x[\tau^2(x;\Pi)] - \mathbb{E}_x\left[\mathbb{V}_{\mathcal{D}^{\mathrm{est}}}\left[\hat{\tau}(x; \mathcal{D}^{\mathrm{est}}, \Pi) \mid x\right]\right]$　の推定として以下の目的関数を得る
  - $-\widehat{\mathrm{EMSE}}_\tau\left(\mathcal{D}^{\mathrm{tr}}, \Pi\right) := \dfrac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{D}^{\mathrm{tr}}} \hat{\tau}^2\left(x^i; \mathcal{D}^{\mathrm{tr}}, \Pi\right) - \left(\dfrac{1}{N^{\mathrm{tr}}} + \dfrac{1}{N^{\mathrm{est}}}\right) \cdot \sum_{\ell \in \Pi} \left(\dfrac{(S_1^{\mathrm{tr}}(\ell))^2}{p} + \dfrac{(S_0^{\mathrm{tr}}(\ell))^2}{1-p} \right)$　ただし$(S_a^{\mathrm{tr}}(\ell))^2$は領域内の各群の結果の経験分散
  - $\hat{\tau}$は平均が入るので、それが領域ごとに異なっているほどxごとの異質性を捉えており近似精度が高い（第１項）
    かつ領域内の分散は小さいほど推定分散が小さい（第２項）

（参考）因果木の目的関数の導出

推定用データと訓練データの独立性を利用して訓練の楽観バイアスと推定分散を推定できる

- 固定した分割Πに関して、$\mathscr{D}^{\mathrm{est}}$を用いて前ページのように$\hat{\tau}$を定義し、これを$\mathscr{D}^{\mathrm{te}}$で評価した場合のMSEを考える
  - $$\mathrm{MSE}_{\tau}(\mathscr{D}^{\mathrm{te}}, \mathscr{D}^{\mathrm{est}}, \Pi) := \frac{1}{N^{\mathrm{te}}} \sum_{i \in \mathscr{D}^{\mathrm{te}}} \left\{ \left(\tau^i - \hat{\tau}(x^i; \mathscr{D}^{\mathrm{est}}, \Pi)\right)^2 - (\tau^i)^2 \right\}$$
  - ここで最後の $-(\tau^i)^2$ は解析の都合で付与した、モデルによらない定数
- $\mathscr{D}^{\mathrm{est}}$、$\mathscr{D}^{\mathrm{te}}$に関して期待値をとる
  - $$\mathrm{EMSE}_{\tau}(\Pi) := \mathbb{E}_{\mathscr{D}^{\mathrm{te}}, \mathscr{D}^{\mathrm{est}}} \left[\mathrm{MSE}_{\tau}(\mathscr{D}^{\mathrm{te}}, \mathscr{D}^{\mathrm{est}}, \Pi) \right]$$
- 負号をつけて式展開する。領域$\mathscr{\ell}(x)$における真の平均因果効果を$\tau(x^i; \Pi)$として分解して

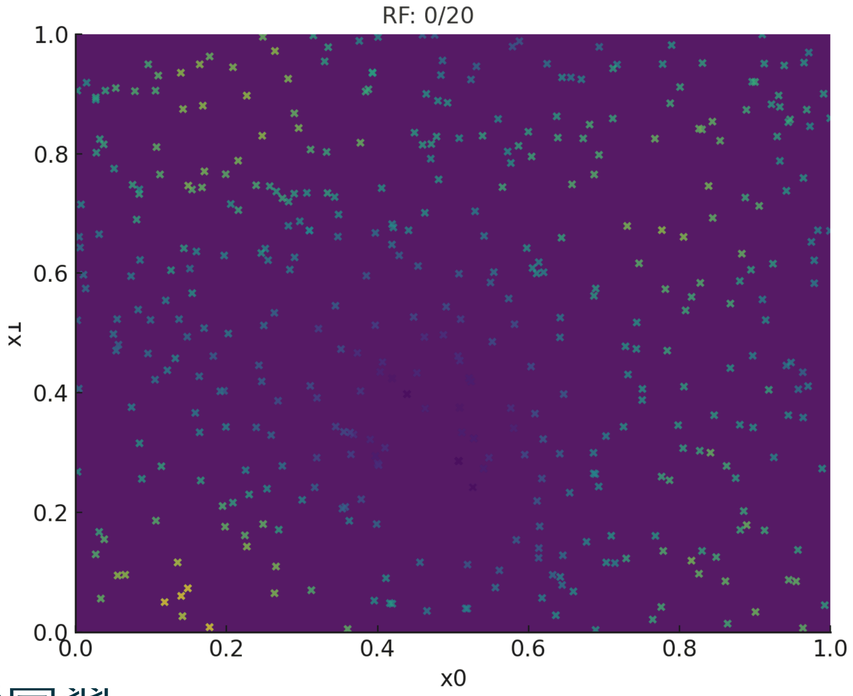
$$
\begin{aligned}
-\mathrm{EMSE}_{\tau}(\Pi) &= -\mathbb{E}_{(x^i, \tau^i), \mathscr{D}^{\mathrm{est}}} \left[(\tau^i - \hat{\tau}(x^i; \mathscr{D}^{\mathrm{est}}, \Pi))^2 - (\tau^i)^2 \right] \\
&= -\mathbb{E}_{(x^i, \tau^i)} \left[(\tau^i - \tau(x^i; \Pi))^2 - (\tau^i)^2 \right] \\
&\quad -\mathbb{E}_{x^i, \mathscr{D}^{\mathrm{est}}} \left[\left(\hat{\tau}(x^i; \mathscr{D}^{\mathrm{est}}, \Pi) - \tau(x^i; \Pi) \right)^2 \right] \\
&= -\mathbb{E}_{x^i} \left[-2\mathbb{E}[\tau^i \mid x^i] \tau(x^i; \Pi) + (\tau(x^i; \Pi))^2 \right] \\
&\quad -\mathbb{E}_{x^i} \left[\mathbb{V}_{\mathscr{D}^{\mathrm{est}}} \left[\hat{\tau}(x^i; \mathscr{D}^{\mathrm{est}}, \Pi) \mid x^i \right] \right] \\
&= \mathbb{E}_{x^i} \left[\tau^2 \left(x^i; \Pi \right) \right] - \mathbb{E}_{x^i} \left[\mathbb{V}_{\mathscr{D}^{\mathrm{est}}} \left[\hat{\tau}(x^i; \mathscr{D}^{\mathrm{est}}, \Pi) \mid x^i \right] \right]
\end{aligned}
$$

$\leftarrow \mathbb{E}\left[\tau^i \mid \mathscr{\ell}(x^i)\right] = \tau(x^i; \Pi)$

出欠コード
XXXX

因果フォレスト：因果木のアンサンブル

- ランダム森（教師あり学習法）
  - 決定木をベース学習器とするアンサンブル法
    1. データを重複ありで再抽出し、
       分割候補の変数もランダムに選んで決定木を学習する
       - 木の多様性を高める（木どうしの相関を下げる）ため
    2. これを繰り返し、出力は各木の出力の平均とする
- 因果フォレスト（Causal Forest）
  - 因果木をベース学習器$\hat{\tau}_j(x)$とし、重複なしのサブサンプルで学習
  - $$\hat{\tau}(x) = \frac{1}{m}\sum_{j=1}^{m}\hat{\tau}_j(x)$$



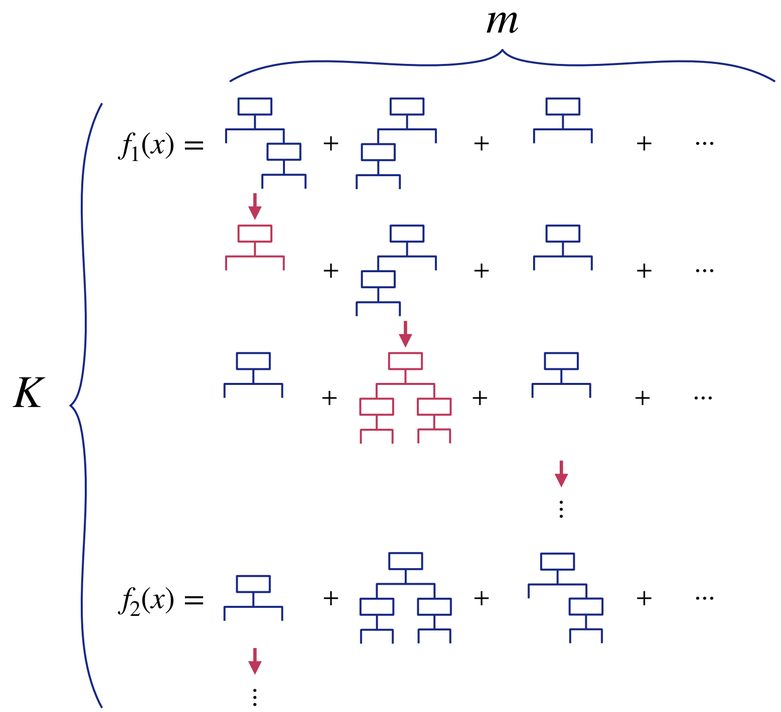
ベイズ的決定森

BARTはベイズ推論（MCMC）による決定森
もともとは因果推論用ではないが、因果推論にもよく使われる

- ベイズ推論
  - ベイズの定理を用いてデータからパラメタの分布を推論する
    - $p(\theta \mid D) = \dfrac{p(D \mid \theta)}{\int p(D \mid \theta')p(\theta')\mathrm{d}\theta'} p(\theta) \quad \cdots (\bigstar)$
    - 予測時はパラメタに関して期待値をとる： $\hat{y} = \mathbb{E}_{\theta \sim p(\theta \mid D)} \mathbb{E}[y \mid x, \theta]$
      - 実際には$p(\theta \mid D)$からサンプリングしたパラメタ集合$\{\theta_1, \ldots, \theta_K\}$の各モデルが返す予測値を平均する
  - BARTでは(★)式の分母の積分を直接計算するのが困難
    → マルコフ連鎖モンテカルロ（MCMC）法
    - あるパラメタθ_{t-1}から対称な提案分布で候補θ^*に遷移するか止まるかを

      受容確率 $\min\left\{1, \dfrac{p(\theta^* \mid D)}{p(\theta_{t-1} \mid D)}\right\}$に従って決める（対称提案のメトロポリス法）。適切な条件のもとで$\{\theta_t\}_t$は

      $p(\theta \mid D)$を定常分布とする連鎖になる
    - MCMC Interactive Gallery
- BART（Bayesian Additive Regression Trees）
  - 複数の木の和を1つのモデルとし、そのパラメタ θ をMCMCで推定する回帰モデル
  - これをS-Learnerとして因果推論に用いることが提案されている [Hill 2011]

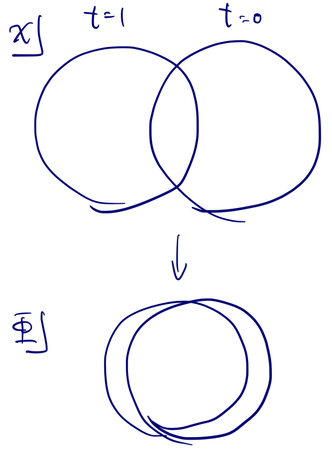


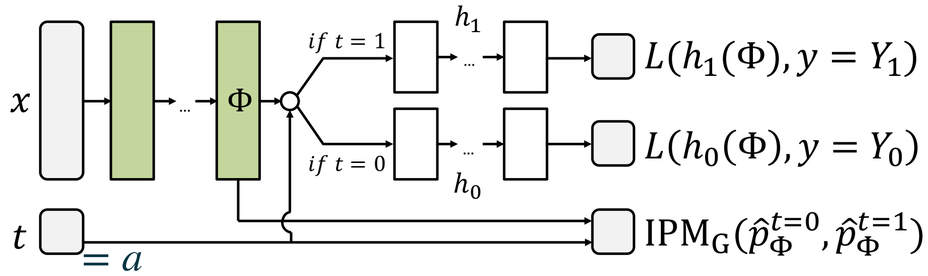
木を1つずつ更新する
（backfitting MCMC）

Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics* 20.1 (2011): 217-240.

深層学習による敵対的ドメイン適応として解く

- 実績分布が傾向スコア $\mu(a|x)$ によって偏ることが問題
- Counterfactual Regression（CFR）
  - 行動 a によって分布変化しない表現 $\phi_x = \phi(x)$ を抽出し、
    $p(\phi_x|a=0) \simeq p(\phi_x|a=1)$ になるようにして、
    ϕ_x 空間上で構築したモデルの誤差低減を目指す（均衡化だけでは保証されない）
- CFRは積分確率計量 Integral Probability Metric (IPM) によって
  分布間不一致を測り、これを同時最小化（損失に加える）
  - $$\mathrm{IPM}_G(p_1, p_2) := \sup_{g \in G} \left| \int_\Phi g(\phi)(p_1(\phi) - p_2(\phi))\mathrm{d}\phi \right| .$$
  - 関数クラスGの設定は1-リップシッツ関数などが用いられる
    - この場合のIPMはワッサースタイン計量とも呼ばれる
  - 群間の差を測る関数gをGの制約のもとで最大化する（敵対的学習）
    - （上記supの双対問題と呼ばれる別問題を毎回解く方法もある）





（参考）敵対的ドメイン適応の理論

群間のx 分布による損失の差を sup で上から抑える

- 理論（引用式のtは行動a、Φは表現抽出器φ）
  - 表現抽出器 ϕ が逆関数を持つなら、反事実損失 $\mathrm{MSE}_{\mathrm{CF}}$ が、
    各群の事実損失の重み付き和 R_F と分布間IPMを用いて、下記の条件のもとで上から抑えられる
  - → R_F + B_Φ IPM を最小化して $\mathrm{MSE}_{\mathrm{CF}}$ の低減を目指す

反事実（観測されなかった潜在結果）の誤差

$$\mathrm{MSE}_{CF}(h,\Phi) \le \underbrace{(1-u)\mathrm{MSE}_F^{t=1}(h,\Phi) + u\mathrm{MSE}_F^{t=0}(h,\Phi)}_{\text{各群の事実損失の重み付き和（通常の事実損失とは重みが逆）}} + B_\Phi \cdot \mathrm{IPM}_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right),$$

（$\downarrow p(a=1)$）

- Φ による表現上の予測モデル h と損失 L (MSE) に対する期待損失を正の定数 B_Φ で割った関数が、両行動でGに入ると仮定。

$$\ell_{h,\Phi}(x,t) = \int_{\mathcal{Y}} L(Y_t, h(\Phi(x),t)) p(Y_t|x) dY_t$$

- 未知の期待損失 $\ell_{h,\Phi}$ の群間の差を sup で抑える

（参考）敵対的ドメイン適応の理論

（観測できない）x による損失の差を sup で置き換える

- 理論
  - 表現抽出器 ϕ が逆関数を持つなら、一様分布上の損失 MSE^u が、データ分布上の損失 MSE と分布間IPMを用いて上から抑えられる
  - → MSE + IPM を最小化すれば MSE^u が抑えられる

CounterFactual error

$$\epsilon_{CF}(h,\Phi) \leq (1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi) + B_\Phi \cdot IPM_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right),$$

- Φ 上の真のモデル h と損失 L の合成
  - $$\ell_{h,\Phi}(x,t) = \int_{\mathcal{Y}} L(Y_t, h(\Phi(x),t)) p(Y_t|x) dY_t$$
  - が関数クラスGに入るように定数(パラメタ) B_Φ を決めれば、その未知の関数 $\ell_{h,\Phi}$ に関して最悪 sup をとれば上界になる

Lemma A4 (Lemma 1, main text). *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be an invertible representation with Ψ its inverse. Let $p_\Phi^{t=1}, p_\Phi^{t=0}$ be defined as in Definition A3. Let $u = p(t=1)$. Let G be a family of functions $g : \mathcal{R} \to \mathbb{R}$, and denote by $IPM_{\mathrm{G}}(\cdot,\cdot)$ the integral probability metric induced by G. Let $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ be an hypothesis. Assume there exists a constant $B_\Phi > 0$, such that for $t = 0, 1$, the function $g_{\Phi,h}(r,t) := \frac{1}{B_\Phi} \cdot \ell_{h,\Phi}(\Psi(r),t) \in \mathrm{G}$. Then we have:*

$$\epsilon_{CF}(h,\Phi) \leq (1-u)\epsilon_F^{t=1}(h,\Phi) + u\epsilon_F^{t=0}(h,\Phi) + B_\Phi \cdot IPM_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right). \tag{6}$$

Proof.

$$\epsilon_{CF}(h,\Phi) - \left[(1-u)\cdot\epsilon_F^{t=1}(h,\Phi) + u\cdot\epsilon_F^{t=0}(h,\Phi)\right] =$$
$$\left[(1-u)\cdot\epsilon_{CF}^{t=1}(h,\Phi) + u\cdot\epsilon_{CF}^{t=0}(h,\Phi)\right] -$$
$$\left[(1-u)\cdot\epsilon_F^{t=1}(h,\Phi) + u\cdot\epsilon_F^{t=0}(h,\Phi)\right] =$$
$$(1-u)\cdot\left[\epsilon_{CF}^{t=1}(h,\Phi) - \epsilon_F^{t=1}(h,\Phi)\right] +$$
$$u\cdot\left[\epsilon_{CF}^{t=0}(h,\Phi) - \epsilon_F^{t=0}(h,\Phi)\right] = \tag{7}$$
$$(1-u)\int_{\mathcal{X}} \ell_{h,\Phi}(x,1)\left(p^{t=0}(x) - p^{t=1}(x)\right) dx +$$
$$u\int_{\mathcal{X}} \ell_{h,\Phi}(x,0)\left(p^{t=1}(x) - p^{t=0}(x)\right) dx = \tag{8}$$
$$(1-u)\int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(r),1)\left(p_\Phi^{t=0}(r) - p_\Phi^{t=1}(r)\right) dr +$$
$$u\int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(r),0)\left(p_\Phi^{t=1}(r) - p_\Phi^{t=0}(r)\right) dr =$$
$$B_\Phi\cdot(1-u)\int_{\mathcal{R}} \frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),1)\left(p_\Phi^{t=0}(r) - p_\Phi^{t=1}(r)\right) dr +$$
$$B_\Phi\cdot u\int_{\mathcal{R}} \frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(r),0)\left(p_\Phi^{t=1}(r) - p_\Phi^{t=0}(r)\right) dr \leq \tag{9}$$
$$B_\Phi\cdot(1-u)\sup_{g\in\mathrm{G}}\left|\int_{\mathcal{R}} g(r)\left(p_\Phi^{t=0}(r) - p_\Phi^{t=1}(r)\right) dr\right| +$$
$$B_\Phi\cdot u\sup_{g\in\mathrm{G}}\left|\int_{\mathcal{R}} g(r)\left(p_\Phi^{t=1}(r) - p_\Phi^{t=0}(r)\right) dr\right| = \tag{10}$$
$$B_\Phi\cdot\mathrm{IPM}_{\mathrm{G}}(p_\Phi^{t=0}, p_\Phi^{t=1}). \tag{11}$$

決定森と深層学習を用いた手法

- 因果木は出力をτとし分割学習用と予測値推定用にデータ分割した決定木
- 因果フォレストは因果木をベース学習器とするランダム森
  - ランダム森はデータをサブサンプリングして学習した決定木を平均したもの
- BARTはベイズ推論を複数の木の和のモデルに適用したもの
- 深層学習を用いるとIPMによる正則化を敵対的学習で実現可
  - 重み付けとは異なるアプローチ