統計的機械学習 （応用計量分析２）第４回

平均因果効果の推定法（参考pdf 5章）
条件付き平均因果効果の推定法（参考pdf 6章）にも少し触れる

出欠コード
XXXX

因果推論の意義と状況設定

- 因果推論の精度は意思決定の良さの保証を与える
- 潜在結果モデルはデータの表を拡張して反事実的結果を欠損とみなす
  - こうすることで因果効果を推定対象として表現でき、それが識別可能性を満たせばデータが無限にあれば（分布がわかれば）真値を特定可能
- （平均）因果効果の識別可能性を保証するには3つの仮定があれば十分
  - SUTVA、無視可能性、正値性
- IPW法は損失関数を傾向スコアの逆数で重み付ける
- 検証実験方法
  - ノイズありの精度でもモデル間の比較はできる

本日の内容

小テスト・前回の内容で質問等あれば

- 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. CATEの推定法２：二重機械学習
- 7. CATEの推定法３：決定木と決定森
- 8. 構造方程式モデルとバックドア基準
- 9. 因果探索
- 10. 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
- 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

（再掲）平均因果効果と評価指標

基本的な推定対象：（曝露群/統制群における）平均因果効果

- 推定対象
  - 平均因果効果（Average Treatment Effect; **ATE**）
    - $\tau = \mathbb{E}[y_1 - y_0]$
  - **曝露群**（サンプルのうち$a = 1$の群）における因果効果
    （Average Treatment Effect on the Treated; **ATT**）
    - $\tau_1 = \mathbb{E}[y_1 - y_0 | a = 1]$
  - 統制群（同$a = 0$の群）における因果効果
    （Average Treatment Effect on the Untreated; ATU）
    - $\tau_0 = \mathbb{E}[y_1 - y_0 | a = 0]$
- 評価指標
  - 推定対象が1次元なので単に数値どうしを比較する

←政策などを適用した効果を定量化したい場合

ATE推定の手法（1/4）

マッチング法：バランスされた標本を作って統計をとる

- 行動以外が似ているデータ点のペアを作る
  - 同じような背景 x をもつ人（データ点）を比較すると行動 a の効果だけに
  - 無視可能性 $(y_{a'})_{a'} \perp\!\!\!\perp a \mid x$ の仮定を満たす x が同じ

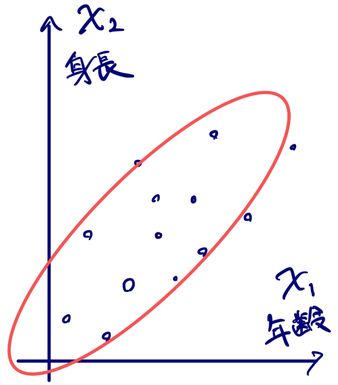
    $\Rightarrow$ 行動と潜在結果は独立（=RCTと同じ状況を再現）
- 曝露群のデータ点 i に対して距離 $d(x^i, x^j)$ が最小の統制群のデータ点を取得し、ペアとする
  - 曝露群の方がサンプルサイズが小さいことが多いため
  - 十分類似したペア得られれば（曝露群の分布$p(x \mid a=1)$が統制群の分布$p(x \mid a=0)$に覆われていれば）

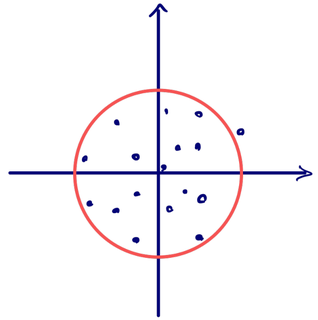
    曝露群における平均因果効果（ATT）の推定量となる
- そのペア間の結果の差を平均
  - $\hat{\tau} = \dfrac{1}{|M|} \displaystyle\sum_{(i,j)\in M} y_1^i - y_0^j$ ， Mはペアの集合

マッチングに用いる距離

ユークリッド距離・マハラノビス距離・傾向スコア差

- マッチングに使用する距離 d はいくつかの選択肢がある
  - ユークリッド距離
    - 標準化（xの各変数平均0分散1にする）しておく
  - マハラノビス距離
    - 説明変数の分散共分散行列 $\hat{\Sigma}$ を用いて
    - $d_M(x, x') = (x - x')^{\mathsf{T}} \hat{\Sigma}^{-1} (x - x')$
  - 傾向スコア差・傾向スコアの対数オッズ差（なぜこれで良いかは後述）
    - $d_p(x, x') = \left| \hat{\mu}(a = 1 \mid x) - \hat{\mu}(a = 1 \mid x') \right|$
    - $d_\ell(x, x') = \left| \log \frac{\hat{\mu}(a = 1 \mid x)}{\hat{\mu}(a = 0 \mid x)} - \log \frac{\hat{\mu}(a = 1 \mid x')}{\hat{\mu}(a = 0 \mid x')} \right|$
    - 1次元なので「近傍」を見つけやすい
- 傾向スコアマッチングは$\hat{\mu}$の推定にaを使うのでやや不安定
  - 曝露群/統制群に分ける際にも使うことになる（同じデータを二度使う）
  - 次元が高すぎなければマハラノビス距離を使っておくのが安全



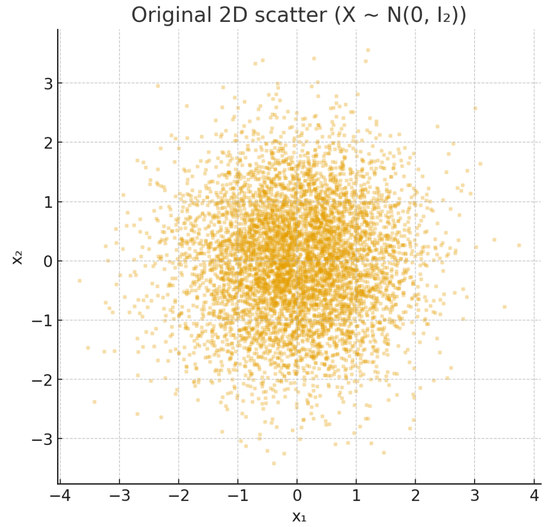


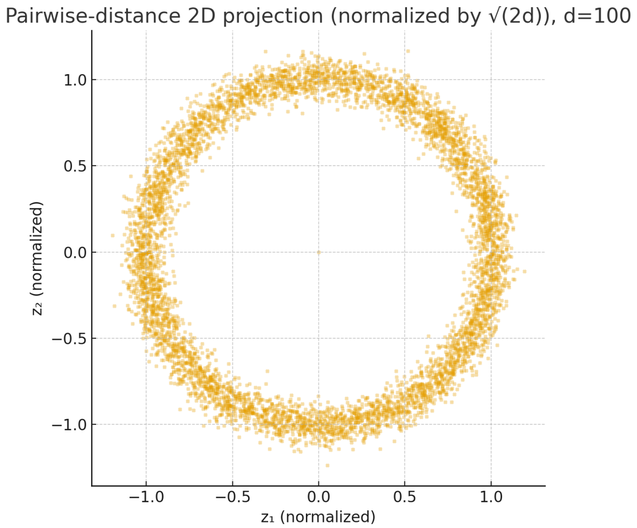
マハラノビス距離は説明変数間の相関を除去した空間でのユークリッド距離

出欠コード
XXXX

次元の呪い：100次元のジャガイモはほぼ皮

- 高次元では体積が距離に対して爆発的に増える
  - 次元 p において距離 r での微小体積は r^{p-1} に比例
- よって多次元正規分布でも原点からの距離が一定の層に集中
- 原点を多少取り替えても同じ
  →データ点間の距離も同じ
- 高次元では「全ての人どうしがほとんど同程度に違って見える」



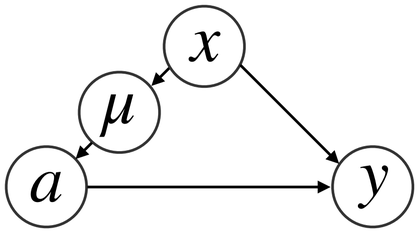


傾向スコアのバランシング特性

出欠コード
XXXX

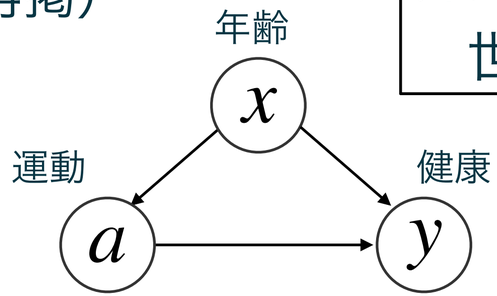
バランシングスコアが同じペアで十分、傾向スコアはその１つ

- バランシングスコアbは、xの関数であって、以下を満たすもの
  - $a \perp\!\!\!\perp x \mid b(x)$
- バランシングスコアをxの代わりに用いても無視可能性を満たせる
  - $(y_{a'})_{a'} \perp\!\!\!\perp a \mid b(x)$
- 定理：傾向スコア$\mu(x) = \mu(a = 1 \mid x)$ はバランシングスコアである
  - 証明略（pdf参照）



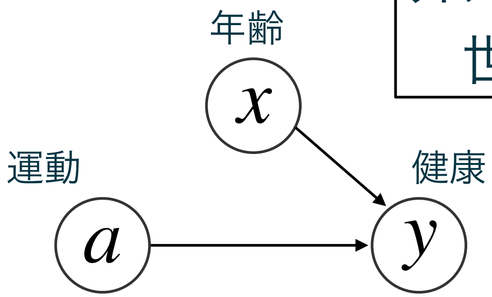
（再掲）

介入前の世界



年齢が高い人が運動する

介入後の世界



年齢が若くても運動する

傾向スコアの推定

プロパーな損失を用いて学習・クリッピングも

- 傾向スコアの学習は漸近的に真値に一致することが推定バイアスの観点では望ましい
  - $\hat{\mu} = \arg\min_{\mu \in \mathscr{F}} \frac{1}{N} \sum_i \ell(a^i, \mu(a \mid x^i))$
  - 関数クラス $\mathscr{F}$ が真の傾向スコアを含む
  - 損失 ℓ として、真の確率値においてのみ最小値をとる損失（**プロパーな損失**）を用いる
    - 真の傾向スコア値$\bar{a} = \mu(a = 1 \mid x)$に対して以下となる
    - $\bar{a} = \arg\min_{a' \in [0,1]} \mathbb{E}_{a \sim \mu(a|x)} \ell(a, a')$
    - 交差エントロピーや二乗損失はプロパーである
- ただし傾向スコアは逆数で使うことが多いため、小さい値は後段の推定分散の観点から避けたい
  - データ点の重み付け $1/\hat{\mu}(a^i \mid x^i)$ が1点だけ100などになると$N = 100$でもほぼその1点だけを重視した学習になる
  - 20以上は20に切り捨てるクリッピングなどを行うことが多い

ATE推定の手法（2/4）

IPW推定の応用：ホーヴィッツ＝トンプソン推定量

- 行動と潜在結果の積は観測できることを使う
  - $\tau = \mathbb{E}_x \mathbb{E}_{a\sim\mu(a|x)} \mathbb{E}_y \left[\left(\frac{a}{\mu(x)} - \frac{1-a}{1-\mu(x)} \right) y \right]$
    - 第3回でも既出の変換。最後の期待値は訓練データの分布 $p(y,a,x) = p(y|a,x)\mu(a|x)p(x)$ になっていることに注意

$$\leftarrow \mathbb{E}_{y_1,y_0,x}\left[y_1 - y_0\right] = \mathbb{E}_x \mathbb{E}_{a\sim\mu(a|x)} \mathbb{E}_{y_1,y_0} \left[\frac{a}{\mu(x)} y_1 - \frac{1-a}{1-\mu(x)} y_0 \right]$$
$$= \mathbb{E}_x \mathbb{E}_{a\sim\mu(a|x)} \mathbb{E}_y \left[\left(\frac{a}{\mu(x)} - \frac{1-a}{1-\mu(x)} \right) y \right]$$

- μを$\hat{\mu}$で、期待値を平均で置き換えたものがホーヴィッツ・トンプソン推定量
  - $\hat{\tau}_{\mathrm{HT}} := \frac{1}{N} \sum_i \left(\frac{a^i}{\hat{\mu}(x^i)} - \frac{1-a^i}{1-\hat{\mu}(x^i)} \right) y^i$
- 上記は、結果の期待値 $\bar{y} = \mathbb{E}[y]$ が非ゼロであった場合、 ()内の重み部分のばらつきにより$\bar{y}$の影響が出て推定精度が悪化する
  - 因果推論では通常τよりも$\bar{y}$の方がスケールが大きいので重みの合計が曝露群と統制群で同じでないと$\bar{y}$の影響が大きく出る
    - yが一定値で$\tau = 0$でも重みの合計が違うと$\hat{\tau}$は非ゼロ
  - よって重みを正規化した修正版を用いることが多い（ハイェク推定量ともいう）
  - $\hat{\tau}_{\mathrm{Hájek}} := \frac{\sum_i \frac{a^i y^i}{\hat{\mu}(x^i)}}{\sum_i \frac{a^i}{\hat{\mu}(x^i)}} - \frac{\sum_i \frac{(1-a^i) y^i}{1-\hat{\mu}(x^i)}}{\sum_i \frac{(1-a^i)}{1-\hat{\mu}(x^i)}}$

G計算：CATE推定量を利用する

- ATEはCATEの平均
  - $\tau = \mathbb{E}\left[y_1 - y_0\right]$
  - $= \mathbb{E}_x[\mathbb{E}[y \mid x, a = 1] - \mathbb{E}[y \mid x, a = 0]]$
- CATE推定量 $\hat{f}(x, a)$ を（なんらか）学習して代入
  - $\hat{\tau}_{\mathrm{gc}} = \frac{1}{N} \sum_i \hat{f}(x^i, a = 1) - \hat{f}(x^i, a = 0)$
    - $f(x, a) = af_1(x) + (1 - a)f_0(x)$ として曝露群の関数f_1と統制群の関数f_0に分ける場合も
      - T(wo)-Learnerという（次回内容）

ATE推定の手法（4/4）

IPWとG計算のハイブリッド：二重に頑健な推定量（Doubly robust estimator、またはAugmented IPW）

- 以下の２つはそれぞれ$\mathbb{E}[a \mid x]$、$\mathbb{E}[y \mid a, x]$の推定精度に依存
  - IPW $\hat{\tau}_{\mathrm{HT}} := \frac{1}{N} \sum_i \left(\frac{a^i}{\hat{\mu}(x^i)} - \frac{1 - a^i}{1 - \hat{\mu}(x^i)} \right) y^i$
  - G計算 $\hat{\tau}_{\mathrm{gc}} = \frac{1}{N} \sum_i \hat{f}(x^i, a = 1) - \hat{f}(x^i, a = 0)$
- これら2つの推定量を組み合わせる
  - $\hat{\tau}_{\mathrm{dr}} = \frac{1}{N} \sum_i \hat{f}(x^i, a = 1) - \hat{f}(x^i, a = 0) + \left(\frac{a^i}{\hat{\mu}(x^i)} - \frac{1 - a^i}{1 - \hat{\mu}(x^i)} \right) (y^i - \hat{f}(x^i, a^i))$
- $\hat{f}$ **または** $\hat{\mu}$ が正確であれば $\hat{\tau}_{\mathrm{dr}}$ も正確（**二重の頑健性**）
  - 実務的には、両方がそこそこ正確であれば $\hat{\tau}_{\mathrm{dr}}$ はかなり正確
  - 理論的には、それぞれの収束率の積が効率的であれば $\hat{\tau}_{\mathrm{dr}}$ の推定も効率的

ATE（ATT）推定法

- 4つの推定法
  - マッチング法
    - 距離の定義によってマッチの取り方がいくつかある
    - 高次元では距離は難しいので1次元の傾向スコアを使う方法も
      - 傾向スコアはバランシングスコアなのでそれが類似ならペアとしてよい
    - 傾向スコアの推定にはプロパーな損失を用いる
  - IPW法（ホーヴィッツ＝トンプソン推定量）
    - 行動と結果の積は観測可能、行動をかけて歪んだ分布を傾向スコアで重み付けて戻す
  - G計算（CATE推定量を用いる）
  - 二重に頑健な推定量（AIPW）
    - $\hat{\mu}$ と $\hat{f}$ のどちらかが正しければ正しい推定