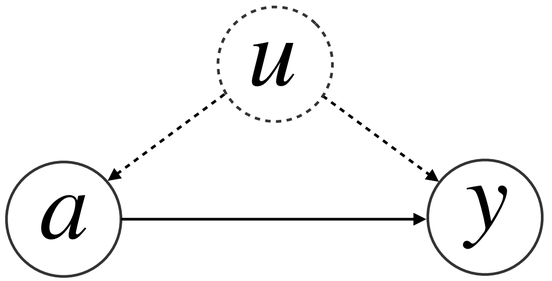
統計的機械学習 （応用計量分析２）第３回

因果推論（因果機械学習）の問題設定（参考pdf 4章）
　必要な仮定、実験・評価方法、公開データ

注意）機械学習分野における典型的な問題設定について解説する
**　　統計分野では若干異なる**

記法一覧

- x ・・・ 背景因子、確率変数一般
- x^n ・・・ n番目のデータ点のx。nはサンプルインスタンス番号
- a ・・・ 行動（意思決定変数）
- $\mathcal{A}$ ・・・ (花文字のA) 行動の候補の集合 $a \in \mathcal{A}$
- $|\mathcal{A}|$ ・・・ 集合$\mathcal{A}$のサイズ（有限離散集合を仮定）
- y ・・・ 結果、予測対象
  - y_a ・・・ 行動aに対する潜在結果
- $\hat{y}$ ・・・ y の推定値
- ℓ ・・・ 損失、損失関数、負の効用（嬉しくなさ、最小化したいもの）
- $a \perp\!\!\!\perp b | S$ ・・・ 確率変数 a と　bが変数集合 S の条件付きのもとで独立　（Sは空集合でもよい）
- $(a_i)_i = (\ldots, a_i, \ldots)$：ベクトルの添字表記
- $\prod_i a_i := a_1 \times a_2 \times \cdots$
- $\varnothing$ ・・・ 空集合
- $a \propto b$ ・・・ aはbに比例する
- グラフィカルモデル
  - ノードは確率変数、破線ノードは未観測変数
  - 矢印は因果関係、破線矢印は因果関係がある可能性、矢無し辺は関連性
  - pa_i ・・・ i 番目の変数の親変数群



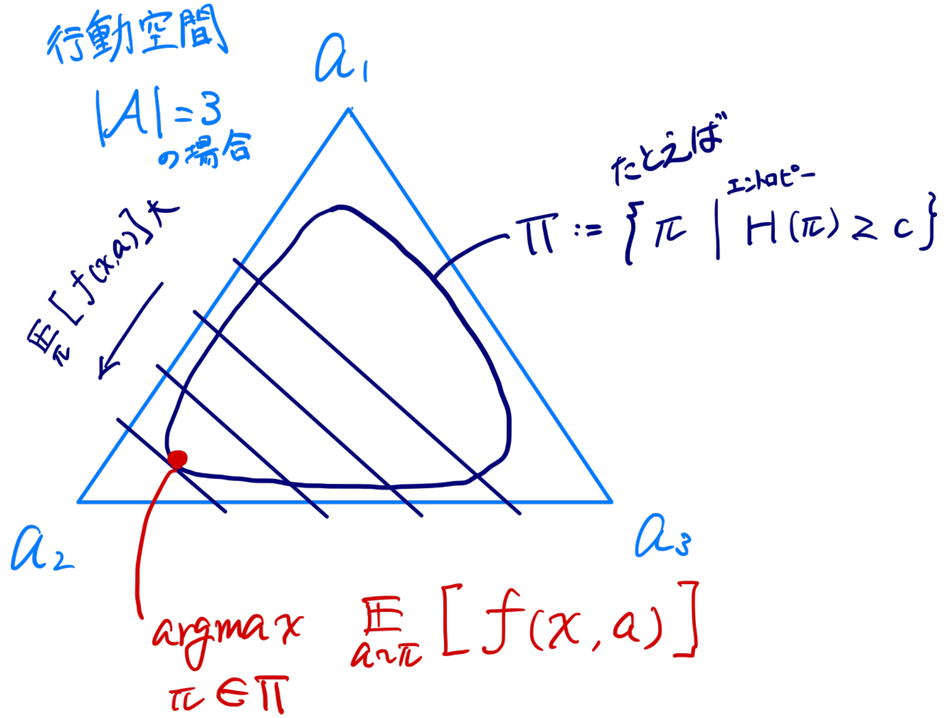
（参考）ソフトな意思決定方策を選ぶイメージ（参考pdf 5.2節）

出欠コード
XXXX

（過適合を防ぐために）方策空間を狭めると確率的な方策に

- プラグイン方策は制約無しだと決定論的になる（いずれかの行動を確率1で選択）
  - $\pi(a \mid x) = \arg\max_a f(x, a)$
- 方策空間を絞った中での最適化も考えておく
  - $\pi(a \mid x) = \arg\max_{\pi \in \Pi} \mathbb{E}_{a \sim \pi(a \mid x)} \left[f(x, a) \right]$
    - 確率分布を選ぶことは、行動の選択肢が3つの場合を考えると
      三角形の内部から1点を選ぶことと対応付けられる（右図）
- 例えば Π がエントロピー制約の場合
  - $\Pi := \{\pi \mid H(\pi) \geq c\}$
    - エントロピー $H(p) := \mathbb{E}_p \left[-\log(p) \right]$
  - プラグイン方策は（実は）Softmaxになる

    $\pi(a \mid x) \propto \exp\left(\beta_c f(x, a)\right)$



行動が3つなら確率単体Πは3次元ベクトル
（各要素≥0、合計=1）なので
(1,0,0),(0,1,0),(0,0,1)の三角形の内部

出欠コード
XXXX

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

（補足）因果推論の重要性

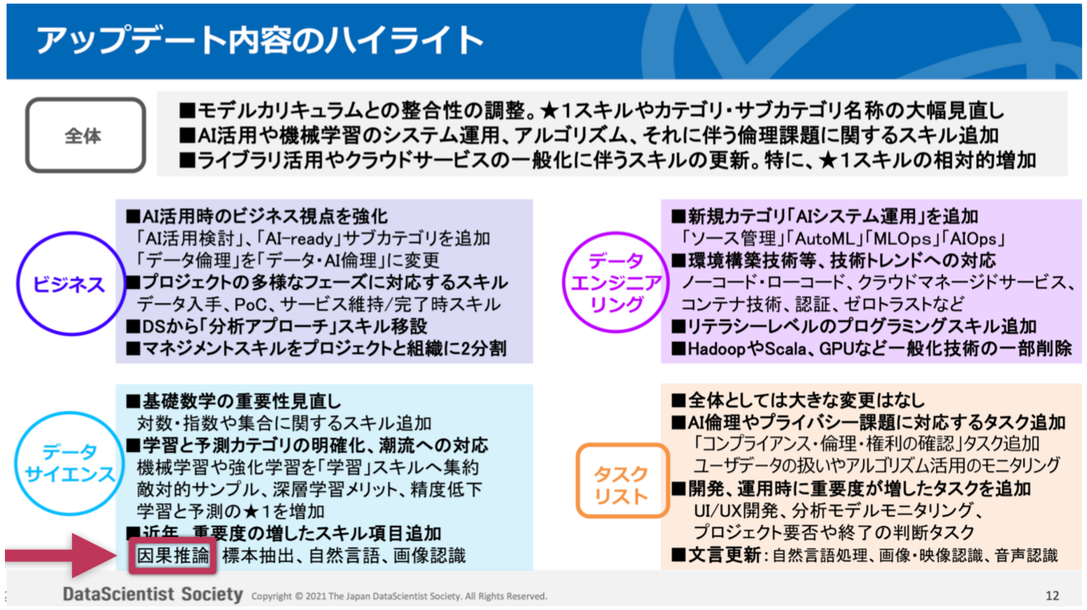
出欠コード
XXXX

2020年頃から因果推論への注目が急速に高まっている

“Causal inference” の検索量は増加傾向

2022/7

データサイエンティスト協会のスキル定義に追加 (’21/11)



関連書籍も
2020年頃
から急増

2021年
ノーベル
経済学賞が
因果推論関係

Ill. Niklas Elmehed © Nobel Prize Outreach.

デイビッド・カード教授

Ill. Niklas Elmehed © Nobel Prize Outreach.

ヨシュア・アングリスト教授

Ill. Niklas Elmehed © Nobel Prize Outreach.

グイド・インベンス教授

（補足）因果推論の応用状況

出欠コード
XXXX

各科学の領域から企業へも活用が広がっているが、ドメイン知識も重要でハードルが高い状況

元々は経済学、疫学、医学、社会学等の各領域で専門家が行ってきた。
近年では、テクノロジー企業を中心に企業でも活用が広がっている。
ただし、専門の高度な分析部隊を抱えていない企業への浸透＝民主化は
これから進む段階と思われる。
“

（５）科学技術的課題

❷自動化による分析者の負担の軽減

因果推論を行うためには、データだけでなく領域知識が必要である。（中略）個々の技術は各分野で行われているが、それを効果的に結び付け、一つの因果分析システムとして、分析者が比較的手軽にアクセス可能な状態にすること肝要であろう。

各種AI技術の研究及び活用動向をまとめたJSTの報告書

⇨データ分析者にとって今後重要

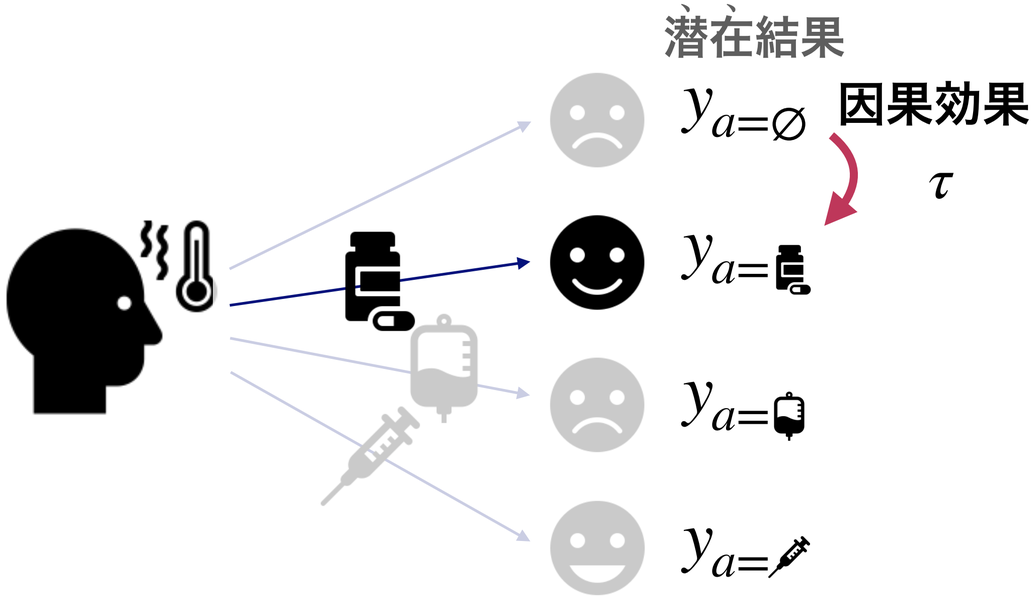
出欠コード
xxxx

実際に取らなかった行動にも潜在的な結果を想定し変数を拡張することで確率モデルとして意思決定問題を記述できる

- 介入は通常の確率の言葉では表現できない
- 各行動をとった世界線ごとの結果＝潜在結果に変数を拡張
  - 実際取らなかった（反事実的）行動の結果は**欠損**しているとみなす → **欠損データからの学習**の問題
- **推定対象（Estimand）** $\mathbb{E}[y_a|x]$ や $\tau(x) = \mathbb{E}[y_1 - y_0|x]$ がデータが無限にあれば特定できる条件を**識別可能性**という
  - =推定対象が観察データ分布から一意に計算可能

年齢 x	処方 a	潜在結果 y_a				治癒 y
		∅	💊	🩸	💉	
23	∅	1	–	–	–	1
57	💊	–	–	–	1	1
43	🩸	–	0	–	–	0
72	💉	–	–	–	0	0

拡張



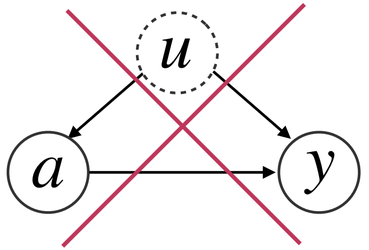
Rubin因果モデル (RCM) ともいう

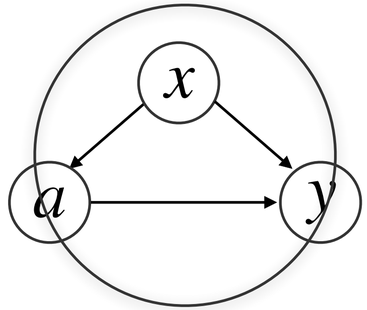
識別可能性を保証する典型的な仮定

出欠コード
XXXX

以下の3つの仮定のもとで因果効果は識別可能

1. **SUTVA** (Stable Unit Treatment Value Assumption) サトヴァ
   - 措置対象間の干渉がない（自身への介入行動からのみ影響を受ける）：$y^i_{a^1,\ldots,a^n} = y^i_{a^i} \quad \forall i, \forall (a^1, \ldots, a^n)$
     - NG) あるSNSユーザへの広告配信アルゴリズムの変更が、
       その人の反応を通してフォロワーに影響する → 漏出効果
   - 各行動は1種類の処置に対応
     - NG) 同じ「薬Aを投与する」行動でも、子どもには1錠、Bさんには2錠
2. **無視可能性** Ignorability / 未交絡因子の不存在: $(y_{a'})_{a'} \perp\!\!\!\perp a \mid x$
   - 観測されていない交絡因子の対処はできない
3. **正値性** Positivity / Overlap: $\mu(a \mid x) > 0 \quad \forall a, x$
   - 観測されない行動パターンの学習はできない





- 上記3仮定が満たされれば $\mathbb{E}[y_a \mid x]$ や $\tau(x)$ が識別可能
  - 満たされない場合の対処法は第5回以降。別のEstimandなら推定可能、といった場合もある

具体的な推定手法の一例

傾向スコア重み付け法（IPW）

- 教師あり学習問題（データの欠損がない問題）の損失関数を変形

$$L = \frac{1}{N}\sum_n^N \sum_{a\in\mathscr{A}} \ell_a^n \qquad \ell_a^n = -\,y_a^n \log \hat{y}_a^n \text{ (交差エントロピー) や } (y_a^n - \hat{y}_a^n)^2 \text{ など}$$

$$= \frac{|\mathscr{A}|}{N}\sum_n^N \mathbb{E}_{a\sim \mathrm{Unif}(\mathscr{A})}[\ell_a^n]$$

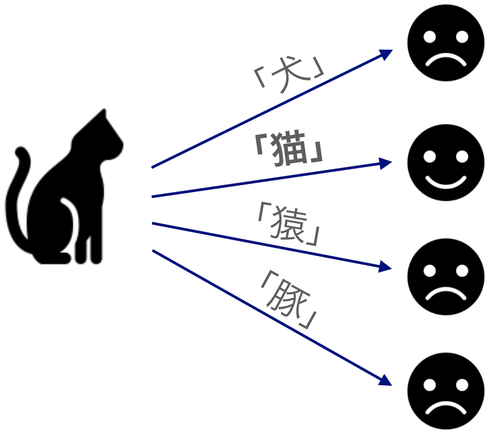
一様分布 ($=1/|\mathscr{A}|$)

$$= \frac{|\mathscr{A}|}{N}\sum_n^N \mathbb{E}_{a\sim\mu(a|x^n)}\left[\frac{\mathrm{Unif}(\mathscr{A})}{\mu(a\,|\,x^n)}\ell_a^n\right]$$

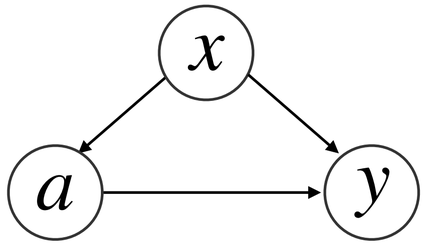
  - **傾向スコア** $\mu(a\,|\,x)$: 過去の意思決定者の選択分布（条件付き確率分布）
  - 最後の期待値はデータの従う分布と同じなので、データで近似できる
  - μ は結果 y 以外のデータ $\{(x_n, a_n)\}_n$ から教師あり学習で推定できる
- 推定値 $\hat{\mu}(a\,|\,x)$ を代入、期待値 $\mathbb{E}_{a\sim\mu(a|x^n)}$ データで置き換えて

$$L(f;D) = \frac{1}{N}\sum_n^N \frac{1}{\hat{\mu}(a^n\,|\,x^n)}\ell_a^n$$

- Inverse Probability Weighting using Propensity (傾向) Score (**IPW**法)



前提：



推定対象と評価指標

因果推論の典型的な精度は潜在結果や効果の精度

- 意思決定につながる推定対象（第2回内容）
  - **条件付き平均因果効果**（Conditional Average Treatment Effect; **CATE**） $\tau(x) = \mathbb{E}[y_1 - y_0 \mid x]$
    - ある程度個別化された効果。因果機械学習（機械学習分野における因果推論研究）でよく推定対象とされる
  - 条件付き平均潜在結果 $\mathbb{E}[y_a \mid x]$
- その他の典型的な推定対象
  - 平均因果効果（Average Treatment Effect; **ATE**） $\tau = \mathbb{E}[y_1 - y_0]$
  - 個別因果効果（Individual Treatment Effect; **ITE**） $\tau^i = y_1^i - y_0^i$
    - 究極的な目的だが、潜在結果が片方しか観測されないため識別のためには特別な追加仮定が必要（第8回以降）
- 評価指標
  - 因果効果の精度 $\mathrm{PEHE}(\hat{\tau}) = \mathbb{E}_x\left[\left(\tau(x) - \hat{\tau}(x)\right)^2\right]$
  - 潜在結果の精度 $\mathrm{MSE}^u(\hat{f}) := \mathbb{E}_x\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}\left(\mathbb{E}\left[y_a \mid x\right] - \hat{f}(x,a)\right)^2\right]$
  - 意思決定価値 $V(\pi) := \mathbb{E}_x\mathbb{E}_{a\sim\pi(a|x)}\mathbb{E}\left[y_a\right]$

性能評価自体も因果推論問題

- 評価自体も難しい
  - 評価が正しくできるならその評価指標を損失とする学習器が作れるはず
    - 損失は微分可能であることが通常なので若干異なるが、計算量の問題が異なるだけで統計的な問題は同じ
- いくつかの典型的な検証方法があり、一長一短ある
  - 擬似的に潜在結果が揃っているデータを使う（例：Twinsデータ）
    - 双子のうち体重の大小を行動とみなし、それぞれの死亡率への影響を調べる
      遺伝子が同じ＝ほぼ同一の個人とみなせる
    - そのようなデータが取れる状況は限定的。ほぼ一卵性双生児に限定
  - 潜在結果に人工データを用いる（例：乳幼児健康発達プログラム（IHDP）データセット）
    - 人工データであるため、実データの複雑さ（例えばノイズ分散のx依存性など）を再現できない可能性
  - RCTされたデータを用いる（例：職業訓練（Jobs）データセット）
    - PEHE等そのものは測れず、ノイズありの精度で測るか、意思決定性能で測る

ノイズあり/なしの推定精度

ノイズありの精度でもモデル間の比較には十分

- 実際に観測される（潜在）アウトカムは一般にノイズつき
  - $y_a = \mathbb{E}[y_a|x] + \varepsilon \quad (\mathbb{E}[\varepsilon|x] = 0)$
- y_aに対するMSEとCATEに対するMSEは期待値においてモデルによらない定数を除いて同等

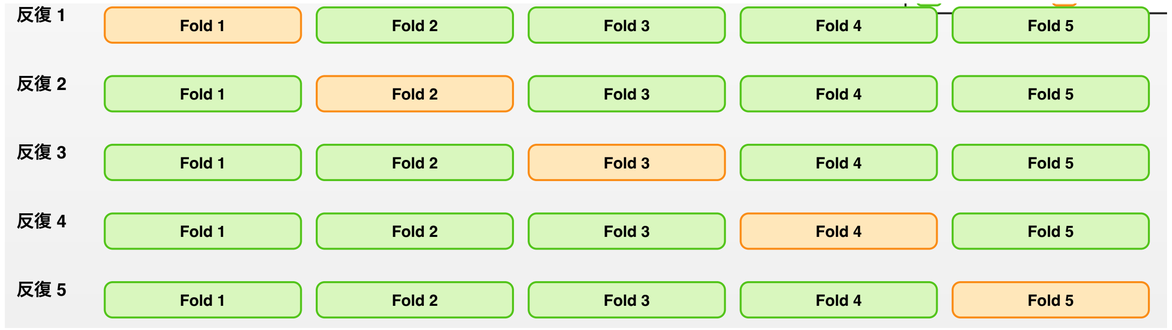
$$\mathbb{E}_\varepsilon\left[\left(\underbrace{\mathbb{E}\left[y_a|x\right]+\varepsilon}_{y_a} - \hat{f}(x,a)\right)^2\right] =$$

  - $\left(\mathbb{E}\left[y_a|x\right] - \hat{f}(x,a)\right)^2 - 2\underbrace{\mathbb{E}\left[\varepsilon\right]}_{=0}\left(\mathbb{E}\left[y_a|x\right] - \hat{f}(x,a)\right) + \underbrace{\mathbb{E}\left[\varepsilon^2\right]}_{\text{定数}}$
- → 学習やモデル選択（モデル f 間の相対評価）には使える
  - 意思決定性能の評価（第2回講義参照）に使うと過剰に悪い評価になるので注意
- 欠損データからノイズ付き因果効果を推定できる
  - $\mathbb{E}[\tau|x] = \mathbb{E}[y_1|x] - \mathbb{E}[y_0|x] = \dfrac{\mathbb{E}[ay_1|x]}{p(a=1|x)} - \dfrac{\mathbb{E}[(1-a)y_0|x]}{p(a=0|x)}$ ・・・観測不可能な反事実的結果にはゼロが掛かるので平均は計算可能

出欠コード
XXXX

検証用データでハイパーパラメタを調整する

- 多くの機械学習手法はハイパーパラメタが存在
  - モデルクラスの広さや正則化の強さなど、学習途中ではなく事前に設定するパラメタ
  - 複数のモデルを学習し、学習に用いずに取っておいた検証用データの精度に従い選択するのが標準的
    - 因果推論では、RCT（ランダム化実験）データを検証用に用いる場合も
  - データが少なくできるだけ全て使いたい場合は交差検証などの方法も



まとめ

意思決定とデータ分析をつなぐ各ステップで仮定や近似（上界）がある

- 統計的推定には誤差がつきもの
- 因果推論の評価指標の意味で高精度なモデルは意思決定に資する
- そのようなモデルを得るための手続きが因果推論