統計的機械学習　（応用計量分析２）第14回

バンディット（参考pdf 19章）

振り返り

出欠コード
XXXX

悲観的な価値推定によって曖昧さ回避するオフライン強化学習
因果推論的なバイアスのない効用推定を経ずデータが少ない領域を避ける

- 因果推論は期待効用を精緻に推定することで良い意思決定につなげる技術（講義第２回）
  - しかしデータが少ない行動がある場合は推定精度はどうしても上がらない
- 結果（の効用）が不確実な状況下で人間は曖昧さ回避する（と考えられている）
  - 曖昧さ回避は想定確率分布集合C内での期待効用の最悪値を基準に意思決定することで達成
- これを再現するのがオフライン強化学習
  - その一手法として保守的Q学習
- 悲観（評価値の）は最悪ケースの保証として証明可能な有効性をもつ
- データの行動(s, a)を単に教師あり学習する模倣学習も
  - 近年はtransformer等の系列学習により精緻に再現できるようになっている

本日の内容

データ収集も含めた意思決定（＝オンライン設定）と全体のまとめ

- 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
  - CATEの推定法２：二重機械学習
- 6. CATEの推定法３：決定木と決定森
  - 深層学習に基づく方法
- 7. 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
- 10. 発展的な因果推論手法：代理変数法、回帰不連続デザイン
- 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**

バンディットとは

報酬の履歴をもとに行動選択
学習データの取得のための行動も考慮する必要

- 以下を逐次的に繰り返す状況を考える
  - 行動 $a_t \in \mathcal{A}$ を選択（$|\mathcal{A}| < \infty$と仮定）
    - アルゴリズムとデータ$(a_1, r_1), \ldots, (a_{t-1}, r_{t-1})$とに従って選択
  - 対応する報酬 $r_t = r_{t,a_t} \in \mathbb{R}$ を受けとる
    - 報酬が固定の確率分布に従う（$p(r_{t,a}) = p(r_a)$）場合を**確率的バンディット**（本講義で扱うのはこちら）、
      従わない場合を**敵対的バンディット**と呼ぶ
  - 既知のラウンド数T回終わったら終了
    - ラウンド数が未知の場合、既知のTを想定した手法を応用可能
- （状況 x_t が与えられ、行動と報酬はxに依存する設定もある（文脈付きバンディット）が、本講義では割愛）
- 評価指標は典型的には単一のベストな行動a^*を取り続けた場合との総報酬の差（**累積リグレット**）
  - 各行動の期待報酬を$\mu_a = \mathbb{E}[r_a]$、最適行動の期待報酬を$\mu^* = \max_{a \in \mathcal{A}} \mu_a$ として、
  - $R_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} r_{t,a_t}\right]$
- 適当な戦略をとると（たとえば固定の確率分布$p(a)$に従って選択）Tに対して線形オーダー $R_T = O(T^1)$ となる
  - これを下げたい（劣線形$o(T)$にしたい）

Upper Confidence Bound（UCB）法

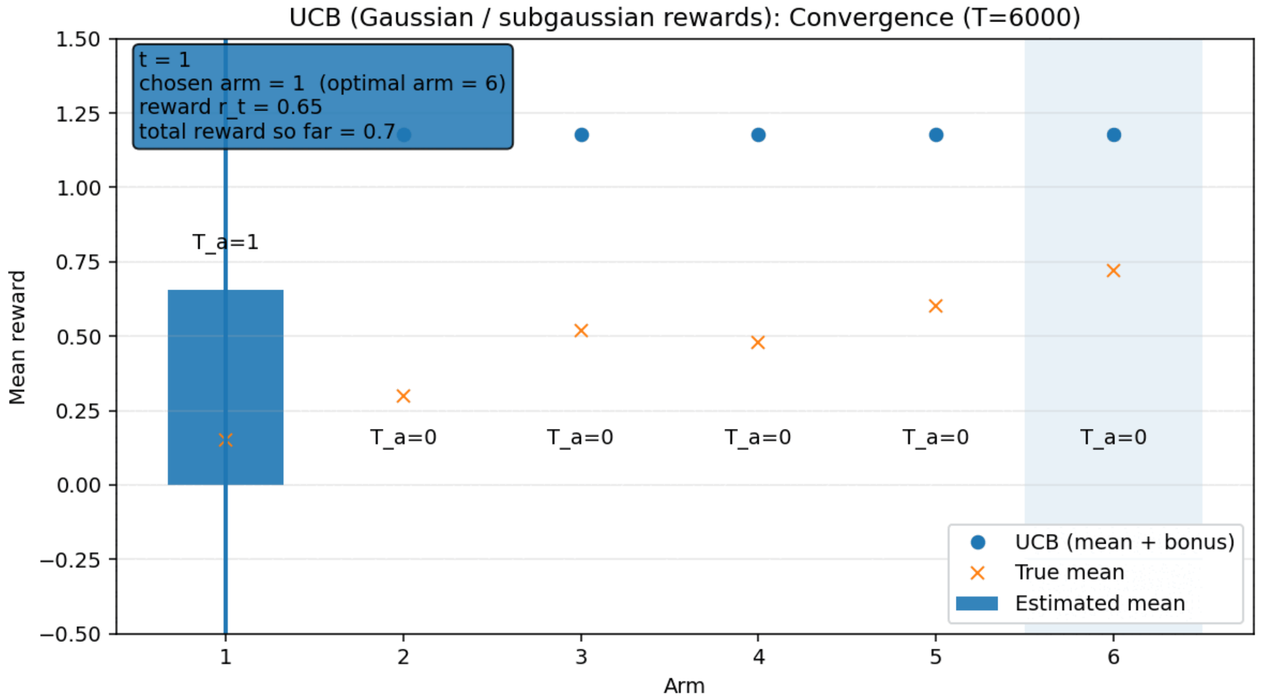
出欠コード
XXXX

期待報酬の推定の信頼区間の上限＝楽観的な推定を基準に

- 確率的バンディットに対する代表的なアルゴリズム：

  Upper confidence bound（UCB）法

  - 各行動aについて**期待報酬の信頼区間の上限** UCB_a を計算
  - 期待報酬の楽観的見積もりを最大化する行動を選択
    - $a_t = \arg\max_a \mathrm{UCB}_a$
  - 具体的なUCBの計算方法は報酬分布に関する仮定による（後述）



UCB-subgauss (colab)

正規分布と同等かそれより分布の裾が軽い分布を仮定

- 報酬分布$p(r_a)$がσ-**劣ガウス性**を持つと仮定できる場合を考える
  - 平均との差$r = r_a - \mathbb{E}[r_a]$が、$\forall \lambda \in \mathbb{R}$ に対し $\mathbb{E}[\exp(\lambda r)] \leq \exp\left(\lambda^2 \frac{\sigma^2}{2}\right)$
  - 正規分布 $\mathcal{N}(0, \sigma^2)$ はσ-劣ガウス
  - 以後、$\sigma = 1$とする（σがわかっていれば、r/σを改めてrとすれば1-劣ガウスになる）
- 平均との差が劣ガウスとなる確率分布からサンプリングしたT_a回の報酬の平均$\hat{\mu}$は、高い確率（$1 - \delta$以上）で真の期待値μから大きく外れない
  - $\mu \leq \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{T_a}}$ $\left(=: \mathrm{UCB}_a(\delta)\text{とする} \right)$

劣ガウス・確率的バンディットにおけるUCB(δ)のリグレット

劣線形オーダー（$O\left(\sqrt{T\log T}\right)$）＝長期的にはほぼ最適な行動だけ

- 報酬分布が1-劣ガウス性をもつとき、$\delta = 1/T^2$としたUCBアルゴリズムのリグレットは次のように抑えられる
  - $R_T \leq 8\sqrt{|\mathcal{A}|T\log(T)} + 3\sum_{a\in\mathcal{A}}\Delta_a$
  - ただしΔ_a は選択肢$a \in \mathcal{A}$の非最適性：$\Delta_a = \max_{a'\in\mathcal{A}} \mathbb{E}[r_{a'} - r_a]$
- Tが伸びるほどリグレットの平均は0に近づく
- $\mathrm{UCB}_a(\delta) := \hat{\mu} + \sqrt{\dfrac{2\log(1/\delta)}{T_a}}$ として選択回数T_aが少ない行動は高く見積もる（楽観的評価）ことで、
  - 効用が不確実な行動を積極的に選択し、報酬を観測しに行っているので、真に平均が高い行動が過小評価され永遠に選択されない確率は低い ⇒ 終盤はほとんど真に最適な行動が選択されがち

証明 1/4

- 方針：以下のように選択肢 a を非最適性が大きい $\Delta_a > \Delta$ 選択肢とそれ以外に分けて示す
  - (1) 非最適性 $\Delta_a = \max_{a' \in \mathcal{A}} \mathbb{E}[r_{a'} - r_a]$ が大きい選択肢は高確率で少数回しか選ばれない
    - 後述
  - (2) 非最適性 $\Delta_a = \max_{a' \in \mathcal{A}} \mathbb{E}[r_{a'} - r_a]$ が小さい選択肢は選ばれ続けてもリグレットへの影響は小さい
    - 非最適性が小さい選択肢が非最適であることを学習するにはサンプルサイズが多数必要なので諦める
    - 最悪でもT回しか選ばれないのでリグレットへの影響は $T\Delta_a \leq T\Delta$ 以下
      - あとで $\Delta = O(1/\sqrt{T})$ くらいのオーダーでΔを選べば $O(\sqrt{T})$ で抑えられそう

証明 2/4

- (1) 非最適性が大きい $\Delta_a = \max_{a' \in \mathcal{A}} \mathbb{E}[r_{a'} - r_a] > \Delta$ なら、高確率で少ない回数（u_a回）以下しか選ばれないはず
  - u_a回選ばれた後は、他の行動（最適な行動a^*を含む）の方が選ばれる＝UCBが大きい、と良い
  - ⇒最適な行動のUCB_{a^*}が真の期待値$\mu_{a^*} =: \mu_*$より大きくて

    かつ非最適行動のu_a回選ばれた後のUCB_aがμ_*より小さい　と良い：

    - $$G_a = \left\{ \mu_* < \min_{t \in [T]} \mathrm{UCB}_{a^*}(t, \delta) \right\} \cap \left\{ \hat{\mu}_{a,u_a} + \sqrt{\frac{2\log(1/\delta)}{u_a}} < \mu_* \right\}$$
  - 良い事象G_aが起きた場合はaはu_a回しか選択されないのでリグレットへの影響は最悪でも$u_a\Delta_a$
- 良い事象G_aが起きないと最悪リグレットが$T\Delta_a$かかってしまうが、その確率は低くなるようにδとu_aを設定する
  - G_aが起きない確率 $1 - P(G_a)$ は、G_aの前半が起きないか、G_aの後半が起きない確率。
  - G_aの前半は、最適な行動の評価値UCBがT回中一度でも違反する確率。1回違反する確率は定義からδなのでT回では$T\delta$
  - G_aの後半は、劣ガウス分布の平均が真の平均から離れる確率の評価を再度用いる（次ページ）

証明 3/4

- G_aの後半は、劣ガウス分布の平均が真の平均から離れる確率の評価を再度用いる（次ページ）
- ここで$c \in (0,1)$をとり、 $u_a = \left\lceil \frac{2\log(1/\delta)}{(1-c)^2\Delta_a^2} \right\rceil$ と設定する
  - すると、G_aの後半が成立しない確率 $1 - p\left(\hat{\mu}_{a,u_a} + \sqrt{\frac{2\log(1/\delta)}{u_a}} < \mu_*\right)$は
  - 非最適性の定義 $\mu_* = \mu_a + \Delta_a$ と劣ガウス分布の平均の評価式から

$$p\left(\hat{\mu}_{a,u_a} + \sqrt{\frac{2\log(1/\delta)}{u_a}} \geq \mu_*\right) \leq p(\hat{\mu}_{a,u_a} - \mu_a \geq c\Delta_a) \qquad \leftarrow \text{こうなるように} u_a \text{を設定した}$$

$$\leq \exp\left(-\frac{u_a c^2 \Delta_a^2}{2}\right) \qquad \leftarrow \text{劣ガウス分布の平均の性質}$$

- aが選ばれる回数の期待値は、

$$\mathbb{E}[T_a] \leq \underbrace{u_a}_{G_a\text{が起きた場合}} + \underbrace{T\left(T\delta + \exp\left(-\frac{u_a c^2 \Delta_a^2}{2}\right)\right)}_{G_a\text{が起きない確率}}$$

  - 定理通り、$\delta = 1/T^2$と設定、u_aを代入し、$c = 1/2$（理由の詳細は省略）として整理すると、T回までにaが選択される回数は
  - $\mathbb{E}[T_a(T)] \leq 3 + \frac{16\log(T)}{\Delta_a^2}$

証明 4/4

- aが選ばれる回数の期待値は（前ページから）
  - $\mathbb{E}[T_a(T)] \leq 3 + \dfrac{16 \log(T)}{\Delta_a^2}$
- リグレットは非最適性×選択回数の期待値と書けるので、最終的に

$$
\begin{aligned}
R_T &= \sum_{a:\Delta_a<\Delta} \Delta_a \mathbb{E}[T_a(T)] + \sum_{a:\Delta_a\geq\Delta} \Delta_a \mathbb{E}[T_a(T)] \\
&\leq T\Delta + \sum_{a:\Delta_a\geq\Delta} \Delta_a \left(3 + \frac{16\log(T)}{\Delta_a^2}\right) \\
&\leq T\Delta + \frac{16\,|\mathscr{A}|\log(T)}{\Delta} + 3\sum_a \Delta_a
\end{aligned}
$$

  - 1項めと2項めを均衡化するよう非最適性の分け目Δを調整し相加相乗平均不等式により定理の式を得る

バンディット（オンライン設定）のまとめ

出欠コード
XXXX

UCB＝不確実な選択肢は楽観的に評価することで探索（積極的にデータ取得）し長期的に最適な選択肢を選べるように

- オンライン＝逐次的にデータを取得し学習しながら意思決定する設定
  - 実行した行動に対応する報酬が与えられる問題設定をバンディット、中でも
    各行動の報酬の確率分布は一定している状況を確率的バンディットという
- 1-劣ガウスの報酬分布の確率的バンディットに対し、$\delta = 1/T^2$としたUCBアルゴリズム
  - $\mathrm{UCB}_a(\delta) := \hat{\mu} + \sqrt{\dfrac{2\log(T^2)}{T_a}}$ が最大となる行動を選択する
- そのリグレットは以下
  - $R_T \leq 8\sqrt{|\mathcal{A}|T\log(T)} + 3\sum_{a\in\mathcal{A}}\Delta_a$　ただし$\Delta_a = \max_{a'\in\mathcal{A}} \mathbb{E}[r_{a'} - r_a]$
- UCBの定義の2項めは平均に用いたサンプルサイズT_aが小さいほど大きい＝**不確実なときは楽観的に**見積もる
  - 効用が不確実な行動を積極的に選択し、報酬を観測しに行っているので、真に平均が高い行動が過小評価され
    永遠に選択されない確率は低い ⇒ 終盤はほとんど真に最適な行動が選択されがち

全体まとめ

データに基づく意思決定の手法を体系的に学習し、特に因果推論を中心に扱った

- 因果推論はドメイン知識に基づく識別可能性と推定法により期待効用を精緻に推定
- データからの推定に限界がある場合は実験可能ならオンライン、不可能ならオフラインRL等

意思決定者の思い

期待効用 $\mathbb{E}\left[u(y_a)\,|\,x\right]$ の最大化

因果的推定対象 $\mathbb{E}\left[y_a\,|\,x\right]$ を精緻に推定

統計的推定対象 $\mathbb{E}\left[y\,|\,x,a\right]$ を精緻に推定

推定量 $\hat{f} = \arg\min_{f\in\mathcal{F}} L(f;D)$

期待効用理論
vNM公理系

推定精度と
意思決定の関係

3つの仮定
識別可能性

（因果）機械学習
分布シフト対処

実験可能・長期 → オンライン設定（バンディット・強化学習）＝楽観戦略で探索と活用

実験不可能・短期 → オフライン設定（オフライン強化学習）＝悲観戦略で過去の性能を再現

到達度評価の問題例

出欠コード
XXXX

細かい定義の暗記は不要（ただしCATEなど重要かつ単純なものは既知と仮定します）、重要な概念・考え方を説明可能なレベルで理解しているか

- 例：以下の定義を参考に問いに解答せよ
  - 因果ダイアグラムGにおいて、aからyへの有向パスがあるとする。次の２条件を満たすとき、**変数集合 Z は順序対 (a, y) についてバックドア基準を満たす**という
    - (B1) aからZへの有向パスがない（行動より下流の変数を含まない）
    - (B2) aに入るパスを含む、aとy を結ぶパス（**バックドアパス**）において、ZがaとYを**有向分離** （ブロック）する
  - ただし、a-y間の全てのパス p に対してZが以下の条件のいずれかを満たすとき、Zはaとyを**有向分離する**という
    - 鎖 i→m→j またはフォーク i←m→j を含み、m は Z に含まれる
    - 合流点 i→m←j を含み、m及びその子孫は Z に含まれない
  - **問：右のDAGが想定されるとき調整すべき変数集合Zとして適切なものの例を挙げよ**