統計的機械学習 （応用計量分析２）第13回

オフライン強化学習（参考pdf 18章）

振り返り

出欠コード
XXXX

曖昧さを避ける意思決定者は、期待効用の（想定確率集合での）最悪値を基準にする
強化学習の典型的な定式化MDP、その学習法Q学習

- マキシミン期待効用理論
  - 確率の凸集合を信念として持ち、その中の最悪期待効用を最大化する意思決定をする
- 強化学習の定式化MDPは状態、行動、遷移確率、報酬、割引率から定義される
- 各時刻の状態と行動に対する価値の評価値であるQ関数を定義
  - Q関数は再帰性があり、次の時刻の報酬とQ関数だけを用いて（2ステップ先以降を用いずに）書ける
- 再帰性の1時間ステップ差の整合性（TD誤差）から損失関数を定義
- TD誤差最小化としてQ学習を構成

振り返り：意思決定理論と因果推論の接続

目的（意思決定）を達成する経路の一つに因果推論

- 意思決定は（vNM公理系のもとで）期待効用がわかれば十分
- 期待効用の最大化は因果的推定対象と
  - 因果的推定対象の推定精度が高ければ良い

意思決定者の
思い

そのためには

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

そのためには

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

そもそも潜在結果を精緻に
推定できない場合、
人間はどうしているのか
⇒ 期待効用の最小値（マキシミン）
基準により不確実性を回避
⇒ これを再現する学習法？

を精緻に推定

$\arg\min_{f\in\mathcal{F}} L(f;D)$

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

無視可能性
$\mu(a\,|\,x) > 0$

本日の内容

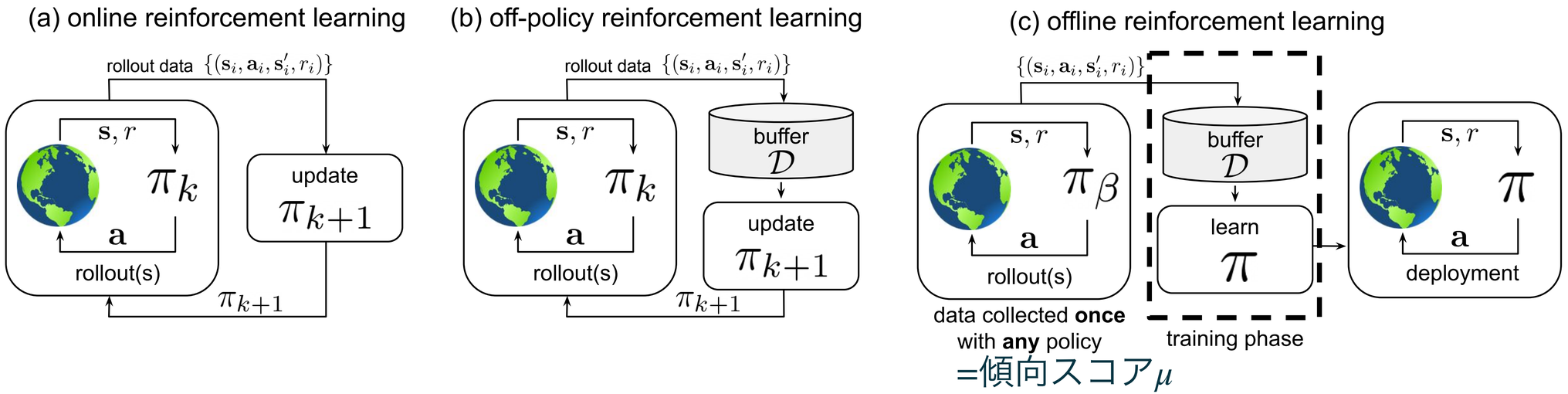
取得済みのデータのみから強化学習するオフライン強化学習

- 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

出欠コード
XXXX

一度取得したデータのみから学習する問題設定

- 強化学習は通常はオンライン設定
  - 学習した方策を実環境で実行↔データ取得し方策を更新
    - 方策オフ（off-policy）型のQ学習は必ずしもオンライン想定ではないが、
      実際にはオンラインでないと難しく、再生バッファDを持ちつつも新規データを取得する設定が多い
- 明確に取得済みデータのみから学習→オフライン強化学習



=傾向スコアμ

Levine, Sergey, et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems." *arXiv preprint arXiv:2005.01643* (2020).

準備：ソフトなQ学習

Q関数に対する方策最適化をソフト化（cf 講義第３回）

- 学習のk反復目における方策をハードな最適行動選択 $\arg\max_a Q(s,a)$ ではなく正則化Rによってソフトな方策にする
  - $\pi_k = \arg\max_\pi \ \mathbb{E}_{\mathbf{a}\sim\pi(\mathbf{a}|\mathbf{s})}[\hat{Q}^k(\mathbf{s},\mathbf{a})] - R(\pi)$
- 正則化項Rとして事前分布ρとのKL距離を用いると
  - $R(\pi) = \mathrm{KL}(\pi||\rho) = \mathbb{E}_{a\sim\pi(a|s)}\left[\log\dfrac{\pi(a|s)}{\rho(a|s)}\right]$
  - 方策の最適化は陽に解けて（※次ページ）以下となる
  - $\pi_k(a|s) = \dfrac{1}{Z}\rho(a|s)\exp(\hat{Q}^k(a,s))$　　（Zは定数）
    - ρを一様分布とするとπはソフトマックス関数と呼ばれるものになる
- Fitted Q学習と同様、次ステップの行動価値$Q(s',a')$は固定（k反復目の値で代替）して次のように学習を行う
  - $\hat{Q}^{k+1} \leftarrow \arg\min_Q \dfrac{1}{2}\mathbb{E}_{s,a,s'\sim D}\left[\left(R(s,a) + \gamma\mathbb{E}_{a'\sim\pi_k(a'|s')}\left[\hat{Q}^k(s',a')\right] - Q(s,a)\right)^2\right]$

KL距離正則化付き方策最適化の導出

- 各状態 s ごとに、$\pi(\cdot\,|\,s)$ について
  - $$\pi_k(\cdot\,|\,s) = \arg\max_{\pi(\cdot|s)} \left\{ \sum_a \pi(a\,|\,s)\,\hat{Q}^k(s,a) - \tau \sum_a \pi(a\,|\,s) \log \frac{\pi(a\,|\,s)}{\rho(a\,|\,s)} \right\} \quad \text{s.t.} \ \sum_a \pi(a\,|\,s) = 1$$
- ラグランジュの未定乗数λを導入して制約を除去
  - $$L(\pi, \lambda) = \sum_a \pi(a\,|\,s) \hat{Q}^k(s,a) - \tau \sum_a \pi(a\,|\,s) \log \frac{\pi(a\,|\,s)}{\rho(a\,|\,s)} + \lambda \Big(\sum_a \pi(a\,|\,s) - 1 \Big)$$
- π で偏微分して 0 とおく：
  - $$\frac{\partial L}{\partial \pi(a\,|\,s)} = \hat{Q}^k(s,a) - \tau \Big(\log \frac{\pi(a\,|\,s)}{\rho(a\,|\,s)} + 1 \Big) + \lambda = 0$$
  - $$\Rightarrow \log \frac{\pi(a\,|\,s)}{\rho(a\,|\,s)} = \frac{\hat{Q}^k(s,a) + \lambda}{\tau} - 1$$
  - $$\Rightarrow \pi(a\,|\,s) = \rho(a\,|\,s) \exp\Big(\frac{\hat{Q}^k(s,a)}{\tau} \Big) \cdot \exp\Big(\frac{\lambda}{\tau} - 1 \Big)$$
- 定数 $\exp(\lambda/\tau - 1)$ は a によらないので正規化定数 $Z(s)$ と書いて
  - $$\pi_k(a\,|\,s) = \frac{1}{Z(s)}\, \rho(a\,|\,s) \exp\Big(\frac{\hat{Q}^k(s,a)}{\tau} \Big)$$

保守的Q学習

データへの適合（TD誤差）に加えて（ソフトな）最大値の最小化により不確実な行動の価値を悲観的に（保守的、低めに）見積もる

- ソフトなQ学習のTD誤差に加えて、**Q関数の値そのもの**を最小化する
  - ただし、過去の方策μ（データ分布）に関しては
  - $$\min_{Q}\max_{\pi}\frac{1}{2}\mathbb{E}_{s,a,s'\sim D}\left[\left(R(s,a)+\gamma\mathbb{E}_{a'\sim\pi_k(a'|s')}\left[\hat{Q}^k(s',a')\right]-Q(s,a)\right)^2\right]+\alpha\left(\underbrace{\mathbb{E}_{s\sim D,a\sim\pi(a|s)}[Q(s,a)]}_{\text{(A)}}-\underbrace{\hat{\mathbb{E}}_{s\sim D,a\sim\mu(a|s)}[Q(s,a)]}_{\text{(B)}}\right)-R(\pi)$$
    - $\hat{\mathbb{E}}$は実データの(s,a)を用いることを表す
  - 方策最適化$\max_{\pi}$に関与するのは(A)と正則化項$R(\pi)$でソフトなQ学習と同じ → RをKL距離とすると方策は陽に書ける
- $\alpha\{(\text{A})-(\text{B})\}$の項を$Q(s,a)$で微分すると $\alpha\{\pi(a|s)-\mu(a|s)\}$
  - → μに対してπが大きい(s,a)の価値$Q(s,a)$を下げる方向に最適化が働く（逆に$\pi<\mu$の場合はQ関数値を上げる）
- データへの適合（第１項）が悪くなりすぎない範囲で、方策πの重みが（μより）大きいQ関数値を最小化

  →データが少ない(s,a)におけるQ関数値は相対的に低めに見積もられる
- **（推定の）不確実性が高い選択肢は避ける、曖昧さ回避を再現** cf) マキシミン基準：$\min_{p\in C}\int_{\Omega}\mathbb{E}_{f(\omega)}[u]\mathrm{d}p(\omega)$

悲観の利点

楽観側誤差は1つでも高いとそれが最適化で選択される
→ 不確実なときは悲観的に見積もるのがオフラインにおける王道

- なんらかの推定値（評価値）$\hat{J}(\pi)$ を最適化する一般的な状況を考える
  - $\hat{\pi} = \arg\max_{\pi \in \Pi} \hat{J}(\pi)$
- このとき、選ばれた方策の真の評価値$J(\hat{\pi})$は次のように、その”楽観側”評価誤差（過大評価）$\hat{J}(\pi) - J(\pi)$と”悲観側”評価誤差（過小評価）$J(\pi) - \hat{J}(\pi)$を用いて分解できる
  - $$J(\pi^*) - J(\hat{\pi}) \leq \inf_{\pi} \{ \underbrace{\left(J(\pi^*) - J(\pi)\right)}_{\text{各方策}\pi\text{の非最適性}} + \underbrace{(J(\pi) - \hat{J}(\pi))}_{\text{各方策}\pi\text{の過小評価（悲観側推定誤差）}} \}$$
  - $$+ \sup_{\pi} \{ \underbrace{\hat{J}(\pi) - J(\pi)}_{\text{各方策}\pi\text{の過大評価（楽観側推定誤差）}} \}$$
- 過小評価は$\inf_{\pi}$をとるのでΠの中で1つでも小さければ良いが、
  過大評価は$\sup_{\pi}$をとるのでΠの中で1つでも大きければ影響大 → **不確実な時は悲観的に**

証明

$$J(\pi^*) - J(\hat{\pi})$$

$$= \inf_{\pi \in \Pi} \left\{ J(\pi^*) - \hat{J}(\pi) + \underbrace{\hat{J}(\pi) - \hat{J}(\hat{\pi})}_{\leq 0 \ \forall \pi \in \Pi} + \hat{J}(\hat{\pi}) - J(\hat{\pi}) \right\}$$

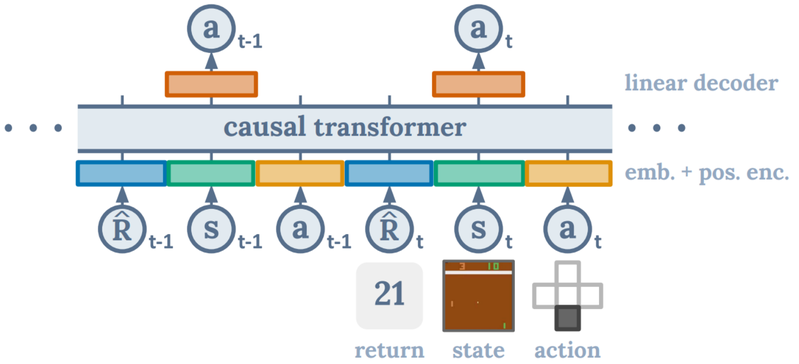
$$\leq \inf_{\pi \in \Pi} \left\{ J(\pi^*) - \hat{J}(\pi) \right\} + \hat{J}(\hat{\pi}) - J(\hat{\pi})$$

$$\leq \inf_{\pi \in \Pi} \left\{ J(\pi^*) - \hat{J}(\pi) \right\} + \sup_{\pi \in \Pi} \left\{ \hat{J}(\pi) - J(\pi) \right\}$$

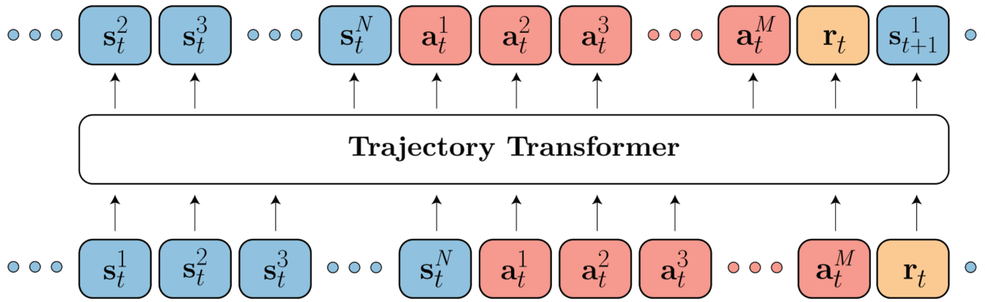
出欠コード
XXXX

行動選択を評価値の最適化ではなくただ真似ることにより学習

- 過去の方策μをただ教師あり学習する**行動クローニング**
  - $$\hat{\pi} = \arg\min_{\pi} \frac{1}{|D|} \sum_{(s,a)\in D} \ell(a, \pi(s))$$
  - 毎ステップ t において「正解」の行動が与えられるためデータが多い
  - 過去の方策μが最適ではない場合でも、それを超える意思決定は基本的にできない
- 近年はTransformerでより精緻な行動予測も
  - 方策が過去の系列全体$\tau_t = \left(s_0, \tilde{R}_0, a_0, \ldots, s_t, \tilde{R}_t\right)$を用いる
    - Decision transformerは「今後の報酬（推定値）」$\tilde{R}_t$ も用いる



Chen, Lili, et al. "Decision transformer: Reinforcement learning via sequence modeling." *NeurIPS 2021.*



Janner, Michael, Qiyang Li, and Sergey Levine. "Offline reinforcement learning as one big sequence modeling problem." *NeurIPS 2021.*

まとめ

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