オフ方策評価の概要とそのリスクリターンに基づく有用性評価

齋藤 優太｜コーネル大学コンピュータサイエンス学部博士課程在籍

齋藤 優太

コーネル大学コンピュータサイエンス学部博士課程在籍。2021年3月、東京工業大学（当時）にて学士号を取得。主に機械学習分野やデータマイニング分野の主要国際会議に多数論文を出版。また、Netflix、Spotify、DMM、Sony、CyberAgent、LINEヤフーなど国内外のテック企業と連携し、検索・レコメンドの戦略策定やオフライン／オンライン評価手法の構築に従事。また、RecSys、KDD、IBISなどの国内外の会議にて、推薦システムや反実仮想学習に関するチュートリアルを実施。著書に『施策デザインのための機械学習入門』『反実仮想機械学習』がある。WSDM 2022 Best Paper Award Runner-Up、RecSys 2022 ~ 2024 Outstanding Reviewer Award、2021年イノベーション大賞内閣総理大臣賞、2022年 Forbes JAPAN「30 UNDER 30」、孫正義育英財団第5期生選出などの受賞歴。

要約

本稿では、過去のデータを活用して新たな意思決定方策の性能を推定するオフ方策評価と呼ばれる統計技術の基本概念、主要な推定量（DM、IPS、DR）、およびそれらの理論的性質や実験的比較を概説する。また、近年の研究動向として、推定量選択の自動化や大規模行動空間への対応といった課題についても取り上げる。さらに、リスク調整後の性能評価を可能にする最新の指標「SharpeRatio@k」について紹介し、オフ方策評価の実運用における信頼性や安全性の観点を強調する。金融・医療等のリスク感度の高い応用分野における有用性にも言及し、今後の応用・発展の方向性を示す。

1．はじめに

機械学習は、正確な予測を可能にする技術として注目を集めている。多くの論文は、解く意味があるとされているタスクにおいて、より良い予測精度を達成することを至上命題としている。もちろん（天気予報など）機械学習等によって得られる予測値をそのまま活用する場面もあり、学術研究の成果が実務に直結するケースも少なくない。しかし現実世界における応用場面に目を向けると、機械学習による予測値をそのまま用いるのではなく、予測値に基づいて何らかの意思決定を行っていることが多い（Swaminathan and Joachims,2015）。例えば、ECや音楽配信サービスの推薦システムでは、各ユーザーとアイテムのペアごとにクリック確率を予測し、その予測値に基づいてどのアイテムを推薦すべきか決定する。または、過去の株価の系列に基づいて最適なポートフォリオを決定する。あるいは、商品の購入確率予測に基づいてユーザーごとにどの商品の広告を提示するか決定する、などが典型例である（Saito and Joachims,2022）。実際、株式会社ZOZOが運営するファッション通販サイト「ZOZOTOWN」におけるファッションアイテム推薦枠（図表1）では、データ駆動のアルゴリズムに基づいて推薦の意思決定が自動化されていることが知られている（Saito et al.(2020)）。これらの例では、クリック確率や購入確率の予測精度よりも、それに基づく推薦や広告配信などの意思

決定の性能に関心がある。

本稿の主たる興味は、機械学習による予測値などに基づいて導かれる意思決定方策（decision-making policy）の性能を正確に評価することである。意思決定方策の性能評価の方法として理想的なのは、方策を実環境あるいはサービスに実装することでその成果を直接的に計測するオンライン実験（A/B テストと呼ばれることもある）だろう。しかし、オンライン実験には時間がかかることや、多大な実装労力が伴うという欠点がある（Levine et al.(2020)）。また性能の悪い意思決定方策を実験してしまったときに、実験期間においてユーザー体験を害してしまったり、収益を毀損してしまう恐れもある（Chandak et al.(2021)）。したがって、古い意思決定方策により既に蓄積されたログデータのみを用いて、新しい意思決定方策の性能を事前に見積もりたい、というモチベーションが自然に生じる。

このように、新たな意思決定方策の性能を蓄積データのみを用いて推定する問題のことをオフ方策評価（Off-Policy Evaluation; OPE）と呼ぶ。正確なオフ方策評価は、機械学習やデータ駆動の意思決定の実践において多くのメリットをもたらす。例えば、時間を要するオンライン実験を行うことなく方策の性能を正確に見積もることが可能ならば、オフ方策評価の結果をもとに方策の改善サイクルを素早く回すことができるだろう。また、ハイパーパラメータや機械学習アルゴリズムの組み合わせを変えることで多数生成される意思決定方策の候補のうち、どの候補を実装すべきか、あるいはどの候補についてオンライン実験を行うべきなのか事前に絞り込むことも可能になるだろう。オフ方策評価はこれらの実利から、産業および学術界において大きな注目を集めている（Uehara, et al.(2022)）。

以降では、オフ方策評価の標準的な定式化と推定量を紹介する。また、それらの基本的な理論性質や最近の研究動向、今後の課題についても解説する。最後に、最新の研究紹介として、金融におけるリスク・リターントレードオフの考え方を応用したオフ方策評価の推定量の精度評価フレームワークについて、簡単に紹介する。

図表1　ZOZOTOWNのファッションアイテム推薦枠

出所）Saito et al.(2020)、Figure 1

2．オフ方策評価の定式化

本節では、オフ方策評価の問題を具体的に定式化する。まず、意思決定問題における三大要素を導入する。具体的に、特徴量（feature）ベクトルを x、離散的な行動（action）を a、観測される報酬（reward）を r とする。これらに関する具体例を、図表2に示した。

図表2　各種応用における特徴量･行動･報酬の例

応用例	特徴量	行動	報酬
映画推薦	視聴履歴	映画	視聴時間
クーポン配布	購買履歴	クーポンの種類	収益
投薬	過去の検査結果	薬の種類	生存確率
金融	過去の株価推移	ポートフォリオ	投資損益

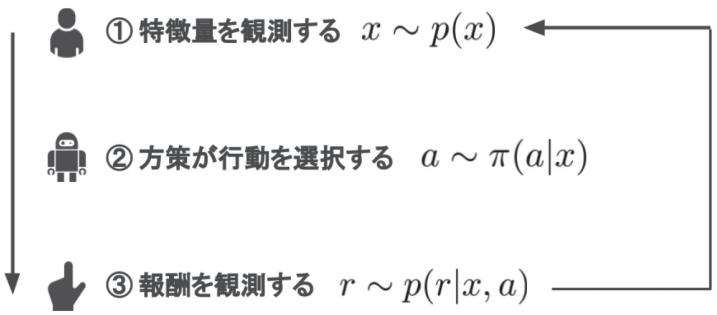
出所）筆者作成

さて意思決定方策とは、特徴量 x で条件付けられた行動空間上の確率分布である。つまり $\pi(a|x)$ は、あるベクトル x で特徴付けられるデータに対して、a という行動を選択する確率のことである。すなわち、$\pi(a|x) \geq 0, \forall(x,a)$ かつ $\sum_a \pi(a|x) = 1$、$\forall x$ が成り立つ。なお、本稿ではより一般的な確率的意思決定方策を主に考えるが、その特殊ケースとして、特徴量 x に対してある行動を確率

1 で選択する決定的な意思決定方策を考えることもできる。

図表３に、意思決定方策による一連の意思決定プロセスを示した。我々はまず、未知の確率分布 $p(x)$ に従う特徴量 x を観測する。これは、YouTube などのプラットフォームにユーザーが訪問したり、直近の株価の変動系列を観測することなどに対応する。次に観測した特徴 x に基づき、意思決定方策が行動を選択する。これは、x で特徴付けられるユーザーに対して推薦すべき動画 a を決めることや、直近の株価系列 x に基づきポートフォリオの構成に変更を加える a といった意思決定を行う場面に対応する。最後に特徴量 x と行動 a の両方に依存する形で報酬 r が未知の確率分布 $p(r|x,a)$ に従い観測される。この意思決定プロセスを大量に繰り返すことで、期待報酬の最大化を目指すのが意思決定方策の役割である。

図表3　方策による意思決定プロセス



出所）筆者作成

ここで、ある意思決定方策の性能（policy value とも呼ばれる）を、次のように定義する。

$$V(\pi) = E_{p(x)\pi(a|x)p(r|x,a)}[r] = E_{p(x)\pi(a|x)}[q(x,a)].$$

なお $q(x,a)=E[r|x,a]$ は、特徴量 x と行動 a を条件付けたときの報酬の期待値であり、期待報酬関数と呼ばれる。

$V(\pi)$ は、方策 π を環境あるいはサービスに実装した際に得られる報酬の期待値である。例えば、報酬がクリックの有無や運用損益で定義されるならば、$V(\pi)$ は方策が達成する期待クリック数や期待運用損益を表し、意思決定方策の性能の定義として妥当だろう。なお、報酬 r の定義は応用先やサービスの目的等によって異なり、実践者が適切に定める必要がある（図表２を参照）。

さて本稿の目的は、方策の性能、すなわち $V(\pi)$ を評価・推定することだった。仮に、方策を実システムに一定期間実装するオンライン実験が実行可能ならば、その期間に観測される報酬の経験平均を計算することで、方策の性能を正確に推定できる。しかし、冒頭で説明したように、オンライン実験はさまざまな理由で実施が難しいことが多い。また、仮にオンライン実験が可能だったとしても、ハイパーパラメータや最適化アルゴリズムなどを変更することで生

まれる大量の方策の候補すべての性能をオンライン実験で推定することは非現実的である。よってオンライン実験の代替として、あるいはオンライン実験を行うべき少数の方策を選択するためのスクリーニングの手段として、方策の性能をオフ方策評価したいというモチベーションが生じる。より具体的には、過去に運用されていたデータ収集方策（logging policy）と呼ばれる意思決定方策 π_0 によって既に収集された以下のログデータ D のみを用いて、π_0 とは異なる新たな方策 π の性能を正確に推定することを考えたい。

$$D = \{(x_i, a_i, r_i)\}_{i=1}^{n}$$

上記のログデータ D は、データ収集方策が環境で稼働していた際に観測された特徴量 x_i と選択された行動 a_i、そしてその結果として観測された報酬 r_i によって構成されている。

オフ方策評価の研究における主たる目標は、データ収集方策 $\pi_0\ (a|x)$ が収集したログデータ D のみを用いて、未だ実装したことのない新たな方策の性能 $V(\pi)$ をより正確に推定できる推定量 $\hat{V}$ を構築することである。ここで推定量 $\hat{V}$ とは、ログデータと評価対象の方策を入力として、真の性能を統計的に近似する次のような関数のことである（Dudik et al.(2014)）。

$$V(\pi) \approx \hat{V}(\pi ; D)$$

推定量 $\hat{V}$ の正確さ（推定精度）は、統計学における典型的な精度指標である平均二乗誤差（mean-squared-error; MSE）に基づき評価されることがほとんどである。

$$MSE\left(\hat{V}(\pi)\right) = E\left[\left(V(\pi) - \hat{V}(\pi; D)\right)^2\right] = Bias\left(\hat{V}(\pi)\right)^2 + Var\left(\hat{V}(\pi)\right)$$

なお、

$$Bias\left(\hat{V}(\pi)\right) = V(\pi) - E[\hat{V}(\pi; D)]$$

$$Var\left(\hat{V}(\pi)\right) = E\left[\left(\hat{V}(\pi; D) - E\left[\hat{V}(\pi; D)\right]\right)^2\right]$$

は、それぞれ推定量 $\hat{V}$ のバイアスとバリアンスである。ここで $E[\hat{V}(\pi;D)]$ は、推定量の期待値を表す。平均二乗誤差はバイアスの二乗とバリアンスの和に等しく、より良い平均二乗誤差を達成するためには、推定量のバイアスとバリアンスを抑えることが重要であることがわかる。

3．標準的な推定量とその性質

本節では、オフ方策評価における標準的な推定量として、Direct Method・Inverse Propensity Score・Doubly Robust という３つの推定量とそれらの基礎的な性質を紹介する。またシミュレーションデータを用いた簡単な実験を通して、これらの推定量の挙動を理解する。

3.1 Direct Method（DM）

DM 推定量ではまず、期待報酬関数 $q(x,a)$ に対する予測モデル $\hat{q}(x,a)$ を事前に得ておく。これは、例えば、ログデータを用いて、次のような回帰問題を解くことに対応する。

$$\hat{q}(x,a) = argmin_{q'} \sum_{i=1} \left(r_i - q'(x_i, a_i) \right)^2$$

ここで $\hat{q}(x、a)$ は期待報酬関数を推定するための予測モデルであり、線形回帰やサポートベクトルマシン、ランダムフォレストなど、よく知られた機械学習の手法を用いて定義できる（Farajtabar et al.(2018)）。

DM 推定量は、事前に得た報酬の予測モデル $\hat{q}(x,a)$ を用いて、次のように定義される。

$$\hat{V}_{DM}(\pi; D) = \frac{1}{n} \sum_{i} E_{\pi(a|x_i)}[\hat{q}(x_i, a)]$$

つまり DM 推定量は、方策の真の性能 $V(\pi)$ の定義において未知である期待報酬関数 $q(x,a)$ を、データから推定された $\hat{q}(x,a)$ で置き換えるというアイデアに基づく。その定義から推察される通り、DM 推定量の推定精度（平均二乗誤差）は、期待報酬関数 $q(x,a)$ の推定精度に大きく依存することが知られている。つまり、予測モデル $\hat{q}(x,a)$ が真の報酬関数 $q(x,a)$ を正確に近似するものであれば、DM 推定量 $\hat{V}_{DM}(\pi; D)$ も方策の真の性能 $V(\pi)$ を正確に推定できる。逆に、$\hat{q}(x,a)$ の推定精度が悪ければ、DM 推定量の推定精度もそれに連動して悪くなる。観測ノイズ、高次元特徴量、特徴量の分布シフトなどさまざまな困難が存在する現実において、真の期待報酬関数 $q(x,a)$ を正確に推定することは往々にして難しい（Farajtabar et al.(2018)）。これらの点から、DM 推定量は平均二乗誤差の推定量のなかでもバイアスが大きくなりやすいという問題を抱えていることが知られている。

3.2 Inverse Propensity Score（IPS）

IPS 推定量は DM 推定量とは大きく異なり、ログデータに含まれる報酬 r_i の重み付け平均により、方策の真の性能 $V(\pi)$ を推定する。同様の推定量のことを、Importance Sampling（IS）や Inverse Probability Weighting（IPW）と呼ぶこともある。

$$\hat{V}_{IPS}(\pi; D) = \frac{1}{n} \sum_{i} w(x_i, a_i)\, r_i$$

ここで、$w(x,a)=\pi(a|x)/\pi_0(a|x)$ は重要度重み（importance weight）と呼ばれる重みであり、「ある特徴量 x に対して、評価対象の方策がある行動 a を選択する確率」と「同じ特徴量 x に対して、データ収集方策が同じ行動 a を選択する確率」の比によって定義される。この重みは、収集されたデータが本来評価したい方策にとってどれだけ「重要」であるか、すなわち評価方策の観点から見た代表性や寄与の度合いを補正するために用いられる。

IPS 推定量は、次の共通サポート（common support）と呼ばれる条件のもとで、不偏性（すなわち、バイアスがゼロ）を持つことが知られている。

データ収集方策 π_0 は、ある評価対象の方策について
$\pi(a|x)>0 \Rightarrow \pi_0(a|x), \forall (x,a)$ であるとき、共通サポートを持つという。

すなわち共通サポートの条件は、ある行動 a が評価対象の方策によってある正の確率で選択されるならば、過去に採られていたデータ収集方策 π_0 もその行動をある正の確率で選択していなければならないことを要求する。これは言い換えると、データ収集方策が評価方策のもとで選択される可能性のある（選択確率がゼロではない）すべての行動の情報を集め、それがログデータとして収集されていることを要求していると理解することができる。

この共通サポートの条件のもとで、IPS 推定量の不偏性（IPS 推定量の期待値が、方策の真の性能に一致すること）を次のように確かめることができる（以下の期待値計算については、読み飛ばしても内容の理解に大きな影響はない）。

$$
\begin{aligned}
&E\left[\hat{V}_{IPS}(\pi;D)\right] \\
&= \frac{1}{n}\sum_i E_{p(x)\pi_0(a|x)p(r|x,a)}\left[\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)} r_i\right] \\
&= E_{p(x)\pi_0(a|x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)} q(x,a)\right] \\
&= E_{p(x)}\left[\sum_a \pi_0(a|x) \frac{\pi(a|x)}{\pi_0(a|x)} q(x,a)\right] \\
&= E_{p(x)}\left[\sum_a \pi(a|x) q(x,a)\right] \\
&= V(\pi)
\end{aligned}
$$

すなわち、IPS 推定量は $V(\pi)$ に対する不偏推定量であり、これは平均二乗誤差の構成要素のうちバイアスが完全にゼロであることを意味する。なお共通サポートの条件が満たされないとき IPS 推定量はもはや不偏ではなく、評価対象の方策に依存する量のバイアスを持つことが知られている（Sachdeva et al.(2020)）。

一方で、IPS 推定量のバリアンスは、次のように与えられることが知られている（Dudik et al.(2014)）。

$$
\begin{aligned}
n\,Var\left(\hat{V}_{PS}(\pi;D)\right) &= E_{p(x)\pi_0(a|x)}\left[w(x,a)^2\,\sigma^2(x,a)\right] \\
&+ Var_{p(x)}\left[E_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]\right] + E_{p(x)}\left[Var_{p(x)}\left[w(x,a)q(x,a)\right]\right]
\end{aligned}
$$

ここで $\sigma^2(x,a)=Var[r|x,a]$ は、x と a で条件付けたときの報酬のバリアンス（ノイズ）である。すなわち、報酬のノイズが大きい（$\sigma^2(x,a)$ が大きい）、言い換えると同一の特徴量と方策の組み合わせに対して得られる報酬の振れ幅が

大きい場合や、データ収集方策と評価対象の方策が大きく異なる（重要度重み $w(x,a)$ の値が大きくなり得る）場合に、IPS 推定量のバリアンスは非常に大きくなってしまう可能性がある。

以上の内容から、DM 推定量と IPS 推定量の間にはバイアスとバリアンスについての明確なトレードオフが存在することがわかる。IPS 推定量は不偏推定量であるため、ログデータに含まれるデータ数 n が大きければ IPS 推定量がより正確であることが多いが、逆にデータが少ないときには DM 推定量の方が正確なことがある（Su et al.(2019)）。

3.3 Doubly Robust（DR）

これまでに紹介した DM 推定量と IPS 推定量には、バイアスとバリアンスについて明確なトレードオフが存在していた。より良い推定量を構築するための自然なアイデアは、異なる利点を持つ DM 推定量と IPS 推定量をうまく組み合わせることだろう。DR 推定量はそれを体現した推定量であり、次のように定義される（Dudik et al.(2014), Jiang and Li(2016)）。

$$\hat{V}_{DR}(\pi; D) = \frac{1}{n}\sum_{i} E_{\pi(a|x_i)}[\hat{q}(x_i, a)] + w(x_i, a_i)\big(r_i - \hat{q}(x_i, a_i)\big)$$

つまり、DR 推定量は期待報酬関数に対する予測モデルを用いる DM 推定量をベースラインとしつつも、第二項において IPS 推定量と類似する方法により期待報酬関数の予測モデル $\hat{q}(x_i, a_i)$、の推定誤差を補正している。このように 2 つの推定量を組み合わせることで、DR 推定量は IPS 推定量の不偏性を保ちつつ、多くの場合 IPS 推定量よりも小さいバリアンスを達成することが知られている。この DR 推定量の利点は、そのバリアンスを実際に計算することで、より正確に理解できる。

$$n\,Var\left(\hat{V}^{DR}(\pi; D)\right) = E_{p(x)\pi_0(a|x)}\left[w(x,a)^2\,\sigma^2(x,a)\right]$$
$$+ Var_{p(x)}\left[E_{\pi_0(a|x)}\left[w(x,a)q(x,a)\right]\right] + E_{p(x)}\left[Var_{p(x)}\left[w(x,a)\,\Delta(x,a)\right]\right]$$

ここで、$\Delta(x,a) = q(x,a) - \hat{q}(x,a)$ は、期待報酬関数の推定バイアスである。ここで導入した DR 推定量のバリアンス表現は、全分散の法則（law of total variance）を 2 度用いることで得ることができる。また、DR 推定量のバリアンスにおいて $\hat{q}(x,a)=0 \Rightarrow \Delta(x,a)=q(x,a)$ とすることで、前節において紹介した IPS 推定量のバリアンスを得ることができる（$\hat{q}(x,a)=0$ としたとき、DR 推定量の定義が IPS 推定量の定義に一致するため）。

さて、IPS 推定量と DR 推定量のバリアンスを比較すると、第三項のみが異なることがわかる。すなわち、期待報酬関数の予測モデル $\hat{q}(x,a)$ が正確であるほど、DR 推定量のバリアンスは小さくなる。一方で、IPS 推定量は $\hat{q}(x,a)$ に非依存であり、そのバリアンスは評価対象の方策および期待報酬や報酬のノイズの大きさのみから決定される。これらが方策の性能 $V(\pi)$ の推定量により大きな悪影響を及ぼすこともありうる。

ここで行った IPS 推定量と DR 推定量のバリアンスの比較は、DR 推定量

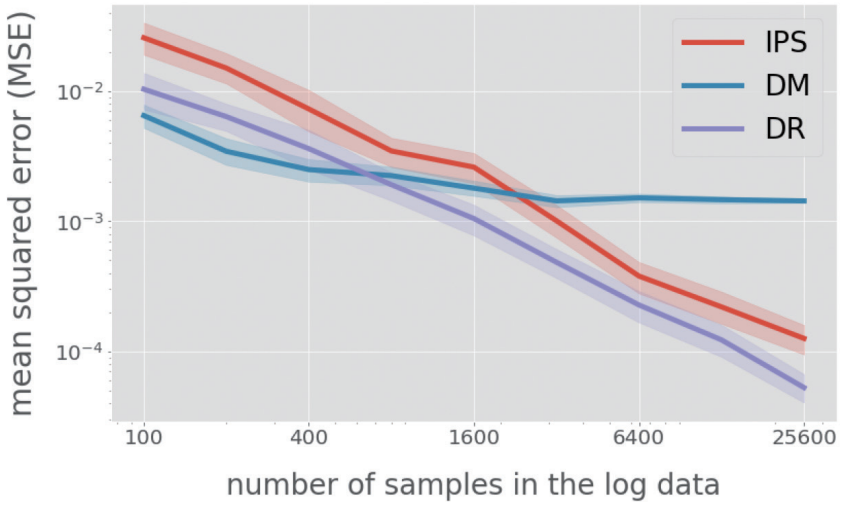
の利点を把握するためのごく初歩的なものである。実際には、先に導入した共通サポートの条件が満たされない状況におけるバイアスなど、より多角的な推定量の比較が可能であり、多くの場合、DR 推定量がほか 2 つの推定量よりも望ましい理論性質を持つことが知られている（齋藤 (2024)）。

3.4 推定量の性能の比較

本小節では、シミュレーションデータを用いて、オフ方策評価における標準的な推定量である DM 推定量・IPS 推定量・DR 推定量の挙動を比較する。実験の実装には、3.6 節で後述する Open Bandit Pipeline を用いた。方策の行動数を 20 で固定し、データ数を 100、200、400、800、…、25,600 と変化させ、事前に用意した評価対象の方策の性能を、DM 推定量・IPS 推定量・DR 推定量によりオフ方策評価を行う。これをデータ数ごとに 100 回繰り返すことで各推定量の平均二乗誤差を算出し、結果を図表 4 に示した（図表 4 は両対数グラフであることに注意されたい）。

まず、データ数が小さいときは DM 推定量が最も正確であることがわかる。これはデータ数が小さい状況では、平均二乗誤差のうちバリアンスが支配的であり、バリアンスが比較的小さい DM 推定量が IPS 推定量や DR 推定量よりも小さい平均二乗誤差を達成するからである。一方で、データ数が大きくなるにつれ、IPS 推定量と DR 推定量はバリアンスを軽減し、徐々に正確さを増すことがわかる。これらに対し、不偏推定量ではない DM 推定量は、データ数に関わらず一定のバイアスを発生してしまうため、平均二乗誤差はほとんど改善していない。このように DM 推定量と IPS 推定量・DR 推定量は、データ数の変化に対して、全く異なる挙動を持つことがわかる。また DR 推定量は IPS 推定量と同様の挙動を示しつつも、そのバリアンス軽減の効果により、IPS 推定量よりも正確なオフ方策評価を達成していることもわかる。

図表4　データ数nが変化したときの推定量の平均二乗誤差の比較



出所）筆者作成

3.5 その他の推定量

本節では、DM 推定量・IPS 推定量・DR 推定量というオフ方策評価において標準的な推定量の定義およびそれらの基礎的な性質について紹介した。また推定量の挙動の違いを簡易的な実験を行うことで理解した。本節で紹介した推定量の定義や性質を把握しておけば、他の最新でより発展的な推定量の構造や理論性質についても困難なく理解できるはずである。より発展的な内容に興味を持った読者は、Switch（Wang et al.(2017)）、More Robust DR（Farajtabar et al.(2018)、Continuous Adaptive Blending（Su et al.(2019)）、DR with Optimistic Shrinkage（Su et al.(2019)）などの推定量について調べてみると、オフ方策評価の実践における選択肢が増えるだろう。

3.6 オフ方策評価の実装

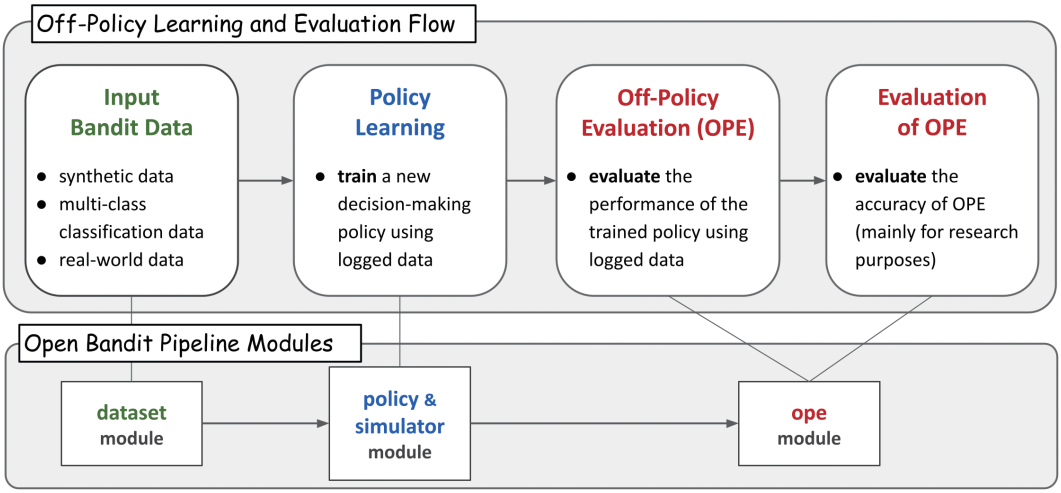
オフ方策評価は非常に実践的なモチベーションに端を発する分野であるが、実際はやや理論偏重的な論文が多数を占め、特に企業の実務家など実践者にとって参入障壁が高い現状がある。誰でも簡単にオフ方策評価の恩恵を受けることができる状況を作り、かつ学術研究における実験評価の再現性も担保すべく、著者らの研究グループは、オフ方策評価のための Python パッケージ「Open Bandit Pipeline（OBP）」を GitHub（https://github.com/st-tech/zr-obp）にて公開している（Saito et al.(2020)）。OBP を活用することで、研究者はオフ方策評価の実装に集中しつつ、再現性のある手順で推定量の精度比較を行うことができる。また現場の機械学習エンジニアやデータサイエンティストは、既に実装されている幅広いかつ最新の推定量を用いて、容易にオフ方策評価を実践できる。OBP は、機械学習において世界的に著名な学会に採択されたオフ方策評価に関する多数の論文の実験で活用されるなど、学界においても一定の地位を得ている。

OBP は、主に以下のモジュールで構成される。

- dataset モジュール：オフ方策推定量の評価に適したシミュレーションデータの生成や多クラス分類データをオフ方策評価で用いられるログデータに変換することで論文執筆用の実験を円滑に行うための機能を提供。
- policy モジュール：標準的なバンディットアルゴリズムやオフ方策学習アルゴリズムを提供。
- ope モジュール： 本節で解説した 3 つの推定量を含む幅広いオフ方策推定量の実装を提供。また、新たにオフ方策推定量を実装するための柔軟なインターフェースも提供。

図表 5 に、OBP の全体像を示した。なお本稿にて行ったいくつかの簡易実験は OBP で実装されており、詳細は Google Colaboratory にて公開している（https://drive.google.com/drive/folders/1YcT7Y0xkVmX_WOyyER3G32bw7DfpEEwA?usp=sharing）。オフ方策評価に関する研究開発や応用に関心を持った読者は、そちらも参考にされたい。

図表5　Open Bandit Pipelineの概要



出所）Saito et al.(2020)、Figure 3

4．近年の研究動向と課題

本節では、オフ方策評価に関連する注目すべき研究動向やビジネスへの実応用を見据える上での課題を紹介する。なお、本節で取り上げる内容は著者の興味・関心に大きく依存していることには注意されたい。

4.1 付随するタスクの自動化

オフ方策評価の推定量には、ハイパーパラメータを持つものも多い。例えば、3節で紹介したDM推定量やDR推定量を用いるには、事前に期待報酬関数を予測する必要がある。ここで、予測モデル $\hat{q}(x,a)$ を得るために用いる機械学習手法やその機械学習手法にまつわるハイパーパラメータは、推定量のハイパーパラメータと見なすことができる。また、より発展的な推定量の中には、バイアスとバリアンスのトレードオフを大きく左右する推定量独自のハイパーパラメータを有しているものも多い。これらの推定量のハイパーパラメータの選択は、推定量の精度に大きな影響を与えることが報告されている（Su et al.(2019)）。しかし、特徴量の分布など取り巻く環境が変化する中で、推定量のハイパーパラメータを実践者の目算やセンスに基づいて適切に定めることは容易ではない。そこで、観測可能なログデータのみに基づいて推定量のハイパーパラメータを選択する方法の研究が徐々に行われるようになっている（Tucker and Lee(2021)）。

また、問題設定に応じて適切な推定量を自動で選択する問題も、オフ方策評価の実応用上重要である（Udagawa et al.(2022)）。すなわち、本稿で紹介したDM推定量・IPS推定量・DR推定量に留まらず多数の推定量が提案されている現状において、実践者は所与の問題設定ごとにどの推定量を用いるべきか？という難しい判断に迫られることになる。推定量の精度は問題設定に依存して大きく変化するため、どの推定量を用いるべきかという判断を適切に下せ

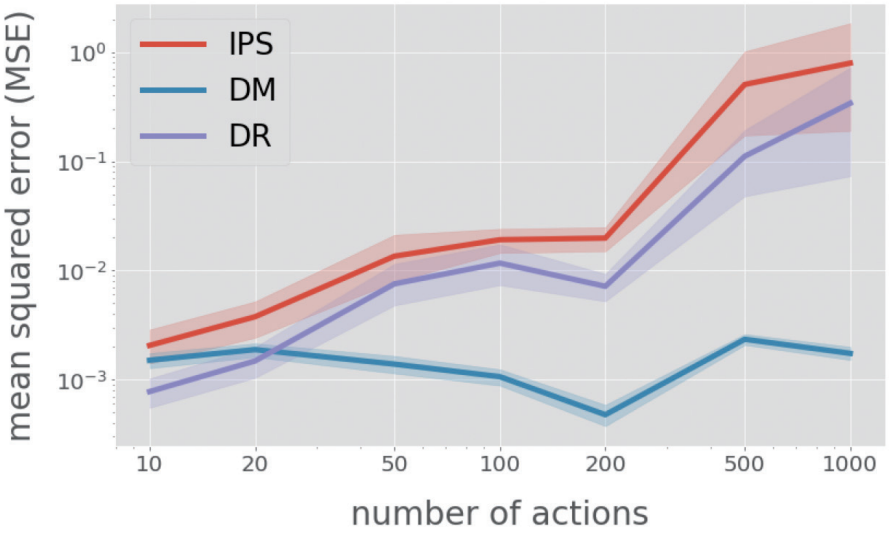
ることは、オフ方策評価の応用上極めて重要である。

この事実は、３節で行った実験の結果（図表４）からも読み取ることができる。データ数 n が小さいときは DM 推定量が最も正確であるが、n が大きくなるにつれ IPS 推定量や DR 推定量の正確さが増している。ここで問題になるのは、現実において図表４に描かれているような平均二乗誤差の値にはアクセスできないという事実である。すなわち図表４を見ると、n=200 以下においては DM 推定量を用い、それ以外の場合は DR 推定量を用いるという推定量の選択が適切であることが一目瞭然なわけだが、推定量の真の平均二乗誤差が未知の現実世界において、このような理想的な推定量選択を実行することは不可能である。また、推定量の正確さはデータ数 n のみならず、行動数や報酬のノイズなど多くの要因の複雑な組み合わせによって決定する。このいわゆる推定量選択の問題は先述のハイパーパラメータチューニングと類似した問題であるように思えるが、ハイパーパラメータチューニングのための既存の手法は、ある特定の推定量が与えられたときにその推定量のハイパーパラメータを選択するタスクにしか適用できない（という仮定に基づいている）(Su et al.(2020))。観測可能なログデータのみに基づいて適切な推定量を選択する問題に関しては、その必要性に反して十分な研究蓄積がないのが現状である。

４.２ 大規模な問題への対応

オフ方策評価の主たる応用に推薦システムや広告配信があるが、これらの問題は往々にして非常に大規模である。すなわち、データ数 n に加えて行動数（施策の選択数）が数千から数百万といったスケールになることが珍しくない。例えば、大規模な顧客にどの商品を推奨するかといった実務上のプラクティスを考えてみると、問題の規模の大きさが判る。図表４を見ると、データ数 n が大きくなるにつれ IPS 推定量と DR 推定量はより正確になっており、大規模な問題におけるオフ方策評価は簡単に思えるかもしれない。しかし、図表４の結果はあくまで行動数を 20 に固定した上での結果である。では、データ数 n を固定した状態で行動数を増やすと、推定量の精度はどのように変化するだろうか。図表６に、データ数を n=3,000 で固定し、行動数を 10、20、50、100、…、2,000 と変化させたときの DM 推定量・IPS 推定量・DR 推定量の平均二乗誤差の変化を示した。この結果から、特に IPS 推定量と DR 推定量は、行動の数が多くなるにつれ推定精度が著しく悪化することがわかる。具体的には、IPS 推定量と DR 推定量の平均二乗誤差は、行動数が 10 から 1,000 に増えたときに 100 倍以上も悪化している。DM 推定量は行動数の増加に対して比較的頑健であるが、図表４で既に確認した通り、データ数 n が増加したときにその恩恵を受けることができない。大規模な問題に対応するためには、データ数 n が増加するにつれて精度が改善し、かつ行動数の増加に頑健な推定量が望まれる。

図表6　行動数が変化したときの推定量の平均二乗誤差の比較



出所）筆者作成

大規模な問題に対処するためのアプローチはいくつか存在する。なかでもユーザーの行動パターンに仮定を課すことで、IPS 推定量や DR 推定量のバリアンスを軽減するアプローチが主流である（McInerney et al.(2020)）。このアプローチは、情報検索の分野で発展してきたユーザー行動モデリングの知見を、オフ方策評価におけるバリアンス軽減に活かしていると理解することができる。例えば、Kiyohara et al.(2022) では、情報検索の分野ではよく知られたカスケードモデルをユーザー行動に仮定することで、DR 推定量を推薦システムのオフ方策評価に活用する際のバリアンスを軽減している。類似のアイデアは、音楽配信サービス Spotify のプレイリスト推薦におけるオフ方策評価に活用されるなど、応用事例が存在する（McInerney et al.(2020)）。

これらのアプローチは、ウェブ産業におけるオフ方策評価の活用において有効であるが、情報検索の知見が活用できない医療や金融、ロボティクスなどの分野においては効力を発揮できない。また、ウェブ応用においても、ユーザーの行動は多様であることが知られており、単一の仮定で全てのユーザーの行動を説明しようとすると、予期せぬバイアスが生じる可能性もある（Kiyohara et al.(2023)）。したがって、Saito and Joachims（2022）や Saito et al.(2023) は、大規模な問題に対応するためのより汎用的な枠組みとして、行動の特徴表現を自然に扱えるようデータ生成過程を一般化し、ユーザー行動モデリングなど特定の分野の知見に依存せずとも巨大な行動空間に対応できる推定量を提案している。

5．リスクリターン指標に基づくオフ方策評価の有用性評価

前節までに述べた通り、オフ方策評価の研究が進展し、現実の意思決定問題への応用が期待されている。その一方、多数の推定量が開発され、現場ごとに

適切な推定量を選択することの難易度が年々増している。既存の研究論文や実践では、主に前述の平均二乗誤差や、後続の方策選択における順位相関およびリグレットといった典型的な精度指標を用いて推定量の正確さの評価がなされてきた。しかし、これらの従来指標は、実運用時におけるオンライン実験を通じて選択された方策が引き起こす潜在的なリスクを捉えるには不十分である。

この問題を解決するために、Kiyohara et al.(2024) は、SharpeRatio@k というオフ方策評価の推定量に対する新たな精度指標を提案した。本指標は、金融におけるポートフォリオ理論に着想を得ており、推定量によって選ばれた上位 k 個の方策群（いわゆる方策のポートフォリオ）におけるリスクリターンのトレードオフを定量化することで、リスクを考慮に入れた推定量の有用性評価を可能にする。

推定量の有用性を測るための新指標である SharpeRatio@k は、以下の数式で定義される。

$$\mathrm{SharpeRatio@k}\left(\hat{V}\right) = \frac{\mathrm{best@k}(\hat{V}) - \mathrm{V}(\pi_0)}{\mathrm{std@k}(\hat{V})}$$

ここで、

- $\hat{V}(\pi; D)$：オフ方策評価における推定量、
- π_0：データ収集方策、
- best@k $(\hat{V})$：推定量 $\hat{V}(\pi; D)$ によって選ばれた上位 k 個の方策ポートフォリオのうち、真の性能が最大の方策の性能値、
- std@k $(\hat{V})$：それら k 個の方策ポートフォリオ内の性能の標準偏差、
- $V(\pi_0)$：データ収集方策の性能。

この指標は、オフ方策評価の「平均的な正確さ」と「ばらつき（リスク）」の比率を取ることで、より実践的かつ信頼性の高い推定量を特定する。たとえば、推定量が選択する方策ポートフォリオの平均性能が高くとも、性能のばらつきが大きければ SharpeRatio@k は低くなり、逆に一貫性が高く低リスクな方策ポートフォリオを形成できる推定量が高評価を得る。

SharpeRatio@k の実用性を示すために、Kiyohara et al.(2024）では、強化学習の分野でよく用いられるシミュレーション環境を通じたデモンストレーションを行なっている。例えば、MountainCar 環境（谷底をスタート地点とし、車を加速させて斜面を登らせるタスクで、物理的な運動量をうまく調整しないと目標に到達できないよう設計された強化学習における古典的ベンチマーク環境）における実験では、SharpeRatio@k がリスクを考慮しない平均二乗誤差やリグレットとは異なる視点から推定量の良さを順位付けする能力が示された。具体的には、平均的な推定精度が高く見えるが、誤って低性能な方策を上位に選出してしまう推定量は、SharpeRatio@k においては低評価となる。

この挙動は、平均二乗誤差などの従来指標が精度のみを基準にしていたのに対し、SharpeRatio@k がリスクという実運用上の視点を併せ持っていることに起因する。

また、SharpeRatio@k はオンライン評価予算（すなわち、方策ポートフォリオの大きさ）の違いにも柔軟に対応できる特性を持つ。ここでいう予算とは、オンラインで実際に評価可能な方策の数に制約があることを意味し、一般に方策数が多くなるほど、各方策を十分に評価するために必要な予算（試行回数や時間的資源、実装に要する手間など）も増加する傾向がある。オンライン評価予算が小さい場合、選ばれる方策ポートフォリオ内のばらつきが大きくなり、1 つの方策の誤選択が全体の成果に大きな影響を与える。一方、オンライン評価予算が大きくなるにつれ、ポートフォリオのリスクが平均化されやすくなり、より安定した評価が可能となる。この性質により、SharpeRatio@k は限られた予算下でのオフ方策評価においても、安定した評価を可能にする推定量を特定できる。

本指標の導入は、オフ方策評価の推定量選択における新たな評価基準の潮流を生み出し始めている。これまで、オフ方策評価に関する研究は主に意思決定アルゴリズムの性能に対する推定精度のみに着目してきたが、大失敗を避けたいことの多い意思決定最適化の現場では、安全性や信頼性も重要な観点である。SharpeRatio@k は、こうしたリスク評価の観点を織り込むことで、オフ方策評価の実用性向上に大きく寄与することが見込まれている。さらに、SharpeRatio@k は新たな推定量設計の基盤ともなり得る。従来の推定量は、報酬の期待値に対する推定精度向上や重要度重みの定義の工夫に注力してきたが、SharpeRatio@k による評価を意識することで、リスクに頑健な推定量の開発が進む可能性がある。例えば、特にリスクの高い方策の悪影響を抑えるような重み付け手法や、方策間の分散を抑制するようなバイアス補正技術が今後ますます台頭するかもしれない。

SharpeRatio@k が提示するリスクとリターンの最適バランスという視点は、これまでより広範な分野への応用も期待される。医療現場では、新しい治療方針を導入する際、その有用性だけでなく副作用などのリスクも重要である。ファイナンス分野においては、投資方針の最適化に際して、リターンとリスクのバランスを定量的に捉えることが求められる。こうした背景から、SharpeRatio@k のようなリスクとリターンの両面に基づく推定量選択は、多彩な応用先が見込まれる意思決定方策の評価において、重要性が高まっていくことだろう。また、SharpeRatio@k の理論的な枠組みを拡張する試みも今後進む可能性がある。例えば、標準偏差の代わりに条件付き VaR（Conditional Value at Risk; CVaR、ファイナンスの先進的なリスク計量の手法の一つ）など、より厳密なリスク指標を用いた拡張を検討することが可能だろう。これにより、リスクに対してさらに慎重な評価が可能となり、安全性が重視される環境において特に有効な評価が期待できる。

6. 総括

本稿では、意思決定方策のデータに基づく性能評価のための統計技法として注目されるオフ方策評価（Off-Policy Evaluation; OPE）を取り上げ、典型的な定式化や推定量の導入、シミュレーションによる比較を行った。また、従来の精度中心の推定量評価が、実運用の観点からは不十分であるという問題を提起した。特に、医療や金融、教育といった領域では、方策が行う意思決定が生み出すリスクに対する定量的評価が不可欠なのである。

その解決策として最近提案されたのが、SharpeRatio@k である。本指標は、金融のポートフォリオ理論における Sharpe Ratio の概念に着想を得ており、リターンとリスクのバランスを考慮する形で推定量の有用性を定量評価する新たな枠組みを提供するものである。単なる推定精度の高さだけでなく、「安全に高性能な方策ポートフォリオを構成できるか」という実務上重要な視点を加える点が革新的である。理論的には、SharpeRatio@k は上位 k 個の推定方策のうち、最も高い実際性能（リターン）とその集合内の標準偏差（リスク）を用いて定義される。この定義は、意思決定に伴う不確実性を定量化する上で有用であり、オンライン実験のような本番環境への導入前段階において、信頼性の高い推定量の選定を可能にする。実験結果からも、SharpeRatio@k は従来の平均二乗誤差とリグレットいった指標と異なる傾向を示し、リスクの低い推定量を適切に評価できることが示された。

加えて、本稿では SharpeRatio@k の金融分野への応用可能性にも言及した。意思決定方策の集合を金融商品のポートフォリオに見立てることで、リスク調整後リターンを考慮するという視点は、投資戦略の評価と対応付けできる。今後の展望としては、SharpeRatio@k を最大化するような新たな推定量の開発、ならびに環境ごとに動的に指標を調整する自律的評価フレームワークの構築が考えられる。また、リスクの性質に応じた派生指標の導入なども考えられる。総じて、SharpeRatio@k は、データ駆動の意思決定最適化やオフ方策評価による性能評価に安全性という軸を加えるための重要な一歩となることが期待されている。

参考文献

Chandak, Y., Niekum, S., Bruno Castro da Silva, Learned-Miller, E., Brunskill, E., & Philip S Thomas. (2021). Universal off-policy evaluation. *Advances in Neural Information Processing Systems,* 34.

Dudik, M., Erhan, D., Langford, J., & Li, L. (2014). Doubly robust policy evaluation and optimization. *Statistical Science,* 29(4):485- 511.

Farajtabar, M., Chow, Y., & Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 1447-1456. PMLR.

Jiang, N., & Li, L.(2016). Doubly robust offpolicy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 48, 652–661. PMLR.

Kiyohara, H., Kishimoto, R., Kawakami, K., Kobayashi, K., Nakata, K., & Saito,

Y. (2024). Towards Assessing and Benchmarking Risk-Return Tradeoff of Off-Policy Evaluation. In *The Twelfth International Conference on Learning Representations*.

Kiyohara, H., Saito, Y., Matsuhiro, T., Narita, Y., Shimizu, N., & Yamamoto, Y. (2022). Doubly robust off-policy evaluation for ranking policies under the cascade behavior model. *arXiv preprint* arXiv:2202.01562.

Kiyohara, H., Uehara, M., Narita, Y., Shimizu, N., Yamamoto, Y., & Saito, Y. (2023). Off-Policy Evaluation of Ranking Policies under Diverse User Behavior. *arXiv preprint* arXiv:2306.15098.

Levine, S., Kumar, A., Tucker, G., & Fu, J.(2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint* arXiv:2005.01643.

McInerney, J., Brost, B., Chandar, P., Mehrotra, R., & Carterette, B. (2020). Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions. *arXiv preprint* arXiv:2007.12986.

Sachdeva, N., Su, Y., & Joachims, T. (2020). Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 965-975.

Saito, Y., Aihara, S., Matsutani, M., & Narita, Y.(2020). Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv preprint* arXiv:2008.07146.

Saito, Y., & Joachims, T. (2022). Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 828–830.

Saito, Y., Ren, Q., & Joachims, T. (2023). Off-policy evaluation for large action spaces via conjunct effect modeling. In *Proceedings of the 40th International Conference on Machine learning*. PMLR, 29734-29759.

Su, Y., Srinath, P., & Krishnamurthy, A. (2020). Adaptive estimator selection for off policy evaluation. In *International Conference on Machine Learning*, 9196-9205. PMLR.

Su, Y., Wang, L., Santacatterina, M., & Joachims, T. (2019). Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, vol. 84, 6005-6014.

Swaminathan, A., & Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research,* 16(1), 1731–1755.

Tucker, G., & Lee, J. (2021). Improved estimator selection for off-policy evaluation. Workshop on Reinforcement Learning Theory at *the 38th International Conference on Machine Learning*.

Wang, Y., Agarwal, A., & Dudík, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits" . In *International Conference on Machine Learning*, 3589-3597. PMLR.

Udagawa, T., Kiyohara, H., Narita, Y., Saito, Y., & Tateno, K. (2022). "Policy-Adaptive Estimator Selection for Off-Policy Evaluation." *arXiv preprint* arXiv:2211.13904.

Uehara, M., Shi, C., & Kallus, N. (2022). A review of off-policy evaluation in reinforcement learning. *arXiv preprint* arXiv:2212.06355.

齋藤優太 (2024)『反実仮想機械学習』技術評論社.