データ科学科目の総履修者数を増加する推薦モデルの提案
～反実仮想機械学習～

第６回早稲田大学データサイエンスコンペティション
チーム：フーヨン

教育学部数学科4年　岡村陸希
教育学部数学科4年　隆大志郎
教育学部数学科4年　南伊生太
理工学部経営システム工学科4年　駒崎智紀

目次

私達が学生生活を通して感じたこと

様々なGEC科目を受講し、自分の進路の新たな選択肢に気付いてほしい。

◇実際に私たちは、データ科学科目を履修したことが、進路に影響を与えた。

特に目標のない大学生

データ科学科目を履修

情報系の大学院に進学

データサイエンティストを
第一志望に就活

200以上のGEC科目の中に**自分の進路に影響を与える科目**があるかもしれない！ ！

私達が学生生活を通して感じたこと(結論)

様々なGEC科目を受講し、自分の進路の新たな選択肢に気付いてほしい。

◇実際に私たちは、データ科学科目を履修したことが、進路に影響を与えた。

特に目標のない大学生

データ科学科目を履修

情報系の大学院に進学

データサイエンティストを
第一志望に就活

現状はどうなっているのか？

200以上のGEC科目の中に**自分の進路に影響を与える科目**があるかもしれない！！

現状

履修率は高いとはいえず、単位のために楽そうなものを選んでいることが多い。

対面きつい、、
オンデマしか勝たん！

授業の形式

課題・試験の内容

試験なしでリアペのみ？
君に決めた！！

自分の興味からではなく、単位の取りやすさから科目を選んでいる。

課題

課題として認知率の低さや、科目数が多いが故の情報の複雑さがあげられる。

科目を知らない

■馬術？きもの学？なにそれ、、

科目にたどり着けない

■情報が多すぎて、興味のある科目を見逃している。

科目を取る意思がない

■わざわざ自分から探して授業取らなくてもいいかな

課題

課題として認知率の低さや、科目数が多いが故の情報の複雑さがあげられる。

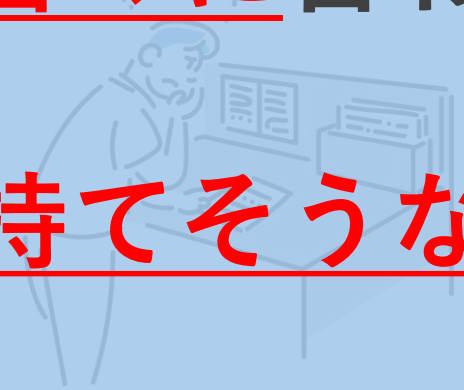
学生各々に合わせて

興味の持てそうな科目を

ダイレクトに紹介できる施策を考える

科目を知らない

■馬術？きもの学？なにそれ、、



■情報が多すぎて、興味のある科目を見逃している。

■わざわざ自分から探して授業取らなくてもいいかな

施策の概要と目標

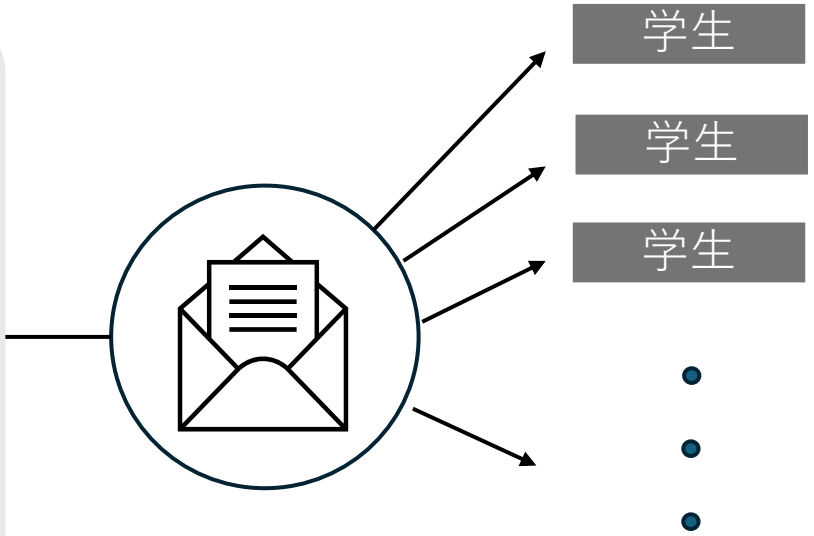
各学生に最適なデータ科学科目をWASEDAメールから推薦し、年間のデータ科学科目の総履修者数の増加を狙う。

メール配信で学生に授業を推薦

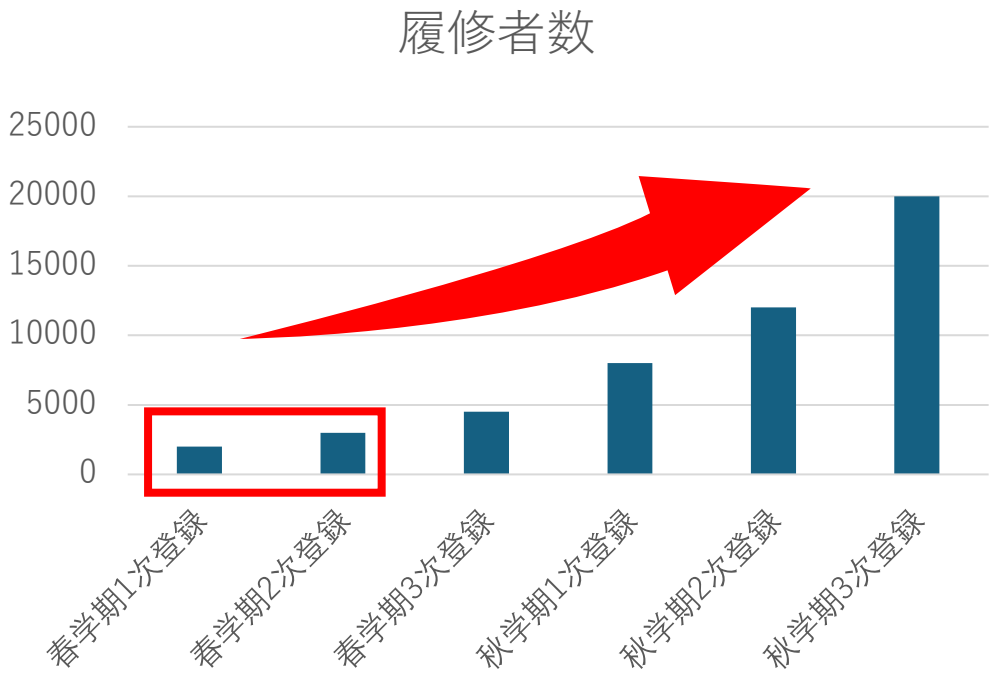
データ科学科目

・統計リテラシー

・データ科学実践



年間の総履修者数の増加

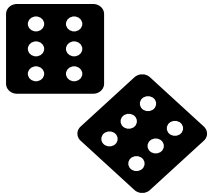


履修率を最大化する戦略

すべての学生にデータ科学の科目を1個推薦するモデルを構築する。今回はデータ収集後の2次科目登録時の総履修者数を最大化することを目標とする。

一次科目登録前に実施する施策

ランダムモデル



♦学生に授業を1つ推薦する

メール配信

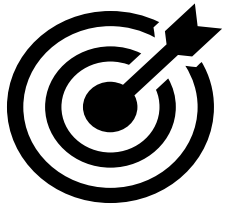
科目登録

♦推薦された授業を履修するか決める

データ収集

二次科目登録前に実施する施策

新推薦モデル



♦学生に授業を1つ推薦する

メール配信

科目登録

♦推薦された授業を履修するか決める

ログデータ

一次科目登録時に実施したランダムモデルによって観測されたログデータ。
このログデータを用いて、新推薦モデルの学習を行う。

学籍番号	性別	学年	学部	…	推薦する科目	履修したかどうか	推薦しなかったときの履修状況
1A000001	男性	1	政治		データ科学入門α	1（履修した）	データ科学入門β
1A000002	女性	1	政治		データ科学入門α	0（履修しない）	データ科学実践
1A000003	男性	2	政治		データ科学入門β	1	統計リテラシーα,統計リテラシーβ
1A000004	女性	2	政治		データ科学入門β	0	Null
1A000005	男性	3	政治		統計リテラシーα	1	統計リテラシーβ
1A000006	女性	3	政治		統計リテラシーα	0	Null
1A000007	男性	4	政治		Rによる統計解析	1	Null
1A000008	女性	4	政治		Rによる統計解析	0	Null

図1：春学期1次科目登録のログデータの例

理想的なデータセットと不完全なデータセットのギャップ

今回得られる不完全なログデータから意思決定モデルを学習・評価するためには、反事実的な状況を考慮する必要がある。

ある学生にデータ科学の1科目を推薦したときのログデータ（右が今回取得したログデータの例）

推薦する科目	履修したかどうか
データ科学入門α	1（履修した）
統計リテラシーα	0（履修しない）
Rによる統計解析	0

図2：理想的なログデータ

GAP

推薦する科目	履修したかどうか
データ科学入門α	未観測
統計リテラシーα	0
Rによる統計解析	未観測

図3：不完全なログデータ

- ある学生に「統計リテラシーα」を推薦する場合、この科目を履修したかどうかという情報だけが取得できる。つまり、他の科目を推薦するときの履修状況は取得不可能。
- 左図のように目的変数が既知の場合、正解データがあるため教師あり学習で解けるが、今回は右図のような不完全な（反事実的な）データが与えられているから教師あり学習で解けない。

理想的なデータセットと不完全なデータセットのギャップ

今回得られる不完全なログデータから意思決定モデルを学習・評価するためには、反事実的な状況を考慮する必要がある。

ある学生にデータ科学の1科目を推薦したときのログデータ（右が今回取得したログデータの例）

推薦する科目	履修したかどうか
データ科学入門α	1（履修した）
統計リテラシーα	0（履修しない）
Rによる統計解析	0

図2：理想的なログデータ

推薦する科目	履修したかどうか
データ科学入門α	未観測
統計リテラシーα	0
Rによる統計解析	未観測

図3：不完全なログデータ

そこで反事実的な状況を考慮するために、OPE,OPLを導入する。

- ある学生に「統計リテラシーα」を推薦する場合、この科目を履修したかどうかという情報だけが取得できる。つまり、他の科目を推薦するときの履修状況は取得不可能。
- 左図のように目的変数が既知の場合、正解データがあるため教師あり学習で解けるが、今回は右図のような不完全な（反事実的な）データが与えられているから教師あり学習で解けない。

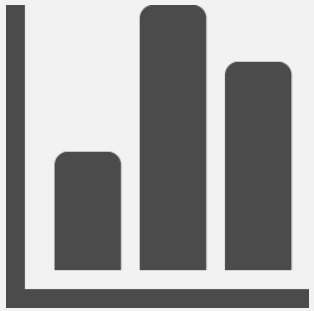
OPE・OPLとは (Off-Policy Evaluation, Off-Policy Learning)

OPE・OPLは、真の性能を上手に推定することが目的であり、教師あり学習では解けない「反実仮想的なデータセット」に取り組むための手法。

目的変数が未知

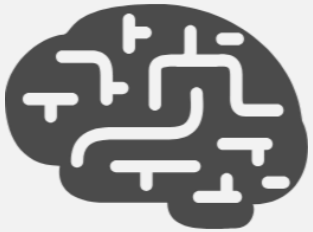
正解データが未知だから
教師あり学習は解けない

OPE



不完全なログデータから
新推薦モデルを評価する
ための評価指標を推定する

OPL



不完全なログデータから
OPEの推定量を用いて，最
適な推薦モデルを学習する

斎藤優太「反実仮想機械学習」

OPE・OPLの基本サイクル

OPE・OPLの最大のメリットは、オフラインで新推薦モデル$\hat{\pi}$を学習・評価できることだ。

1 ログデータの収集

旧推薦モデルπ_0がログデータを収集
（一般的にログデータは不完全）

学習のためのOPL 2

ログデータD_{train}から、
新推薦モデルπをオフライン学習
$\hat{\pi} = \mathrm{argmax}_{\pi} \hat{V}(\pi; \mathcal{D}_{train})$

$\hat{\pi} \rightarrow \pi_0$

4 実運用 or オンライン実験

オフライン実験の評価が高い
新推薦モデル$\hat{\pi}$を
実運用 or オンライン実験

評価のためのOPE 3

ログデータ$D_{validation}$から
新推薦モデルπのオフライン評価
$\hat{V}(\hat{\pi}; \mathcal{D}_{validation})$

OPE・OPLの基本サイクル

OPE・OPLの最大のメリットは、オフラインで新推薦モデル$\hat{\pi}$を学習・評価できることだ。

1 ログデータの収集

旧推薦モデルπ_0のログデータを収集
（一般的にログデータは不完全）

$\hat{\pi} \rightarrow \pi_0$

OPL 2

学習用ログデータ$\mathcal{D}_{train}$から、
新推薦モデルπを学習
$\hat{\pi} = \mathrm{argmax}_{\pi} \hat{V}(\pi; \mathcal{D}_{train})$

OPE 3

検証ログデータ$\mathcal{D}_{validation}$から新推薦モデルπのオフライン評価
$\hat{V}(\hat{\pi}; \mathcal{D}_{validation})$

4 実運用 or オンライン実験

オフライン実験の評価が高い
新推薦モデル$\hat{\pi}$を
実運用 or オンライン実験

新推薦モデルを学習するために、今回解くべき目的関数を考える。

目的関数の準備

正しい新推薦モデルを学習するには,メール配信の効果(増加量)を最大化するように目的関数を設定する。

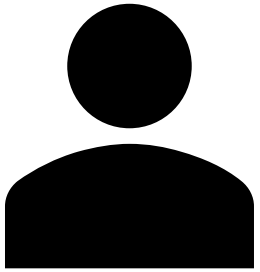
総履修者数の最大化問題

「科目A」を推薦するとき　：　(1) + 0 + 0 = 1
「科目B」を推薦するとき　：　1 + (1) + 0 = 2
「科目C」を推薦するとき　：　1 + 0 + (0) = 0

※( )は科目を推薦した時の報酬

注目するポイントは増加量

フーヨン君に科目Bを推薦することで，元々履修するつもりのなかった科目Bを履修した。
つまり，メール配信の効果の最大化に注目したい。



フーヨン君

科目名	推薦するとき	推薦しないとき	増加量
科目A	1	1	0
科目B	1（履修した）	0（履修しない）	+1
科目C	0	0	0

図4：ある学生に1個の科目を推薦するとき・しないときのデータの例

目的関数を定式化するプロセス

学生に対して、メール配信効果の高い科目を高確率で推薦する新推薦モデルの学習ができるように定式化を行う。

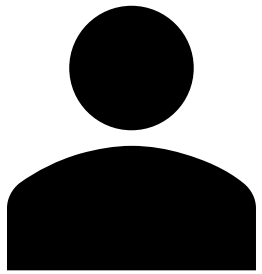
新施策πが学習する真の目的関数

※ $\max_{\theta} \sum_{u \in \mathcal{U}} \mathbb{E}_{\pi_\theta(a|x_u)}[q_1(x_u, a) - q_0(x_u, a)]$

$q_1(x, a) = \mathbb{E}_{p(r(a,1)|x)}[r(a, 1)]$
$q_0(x, a) = \mathbb{E}_{p(r(a,0)|x)}[r(a, 0)]$

目的関数の意味

21科目の内、メール配信の効果（増加量）が大きい科目を高確率で優先的に選択する新推薦モデルを学習する。



フーヨン君

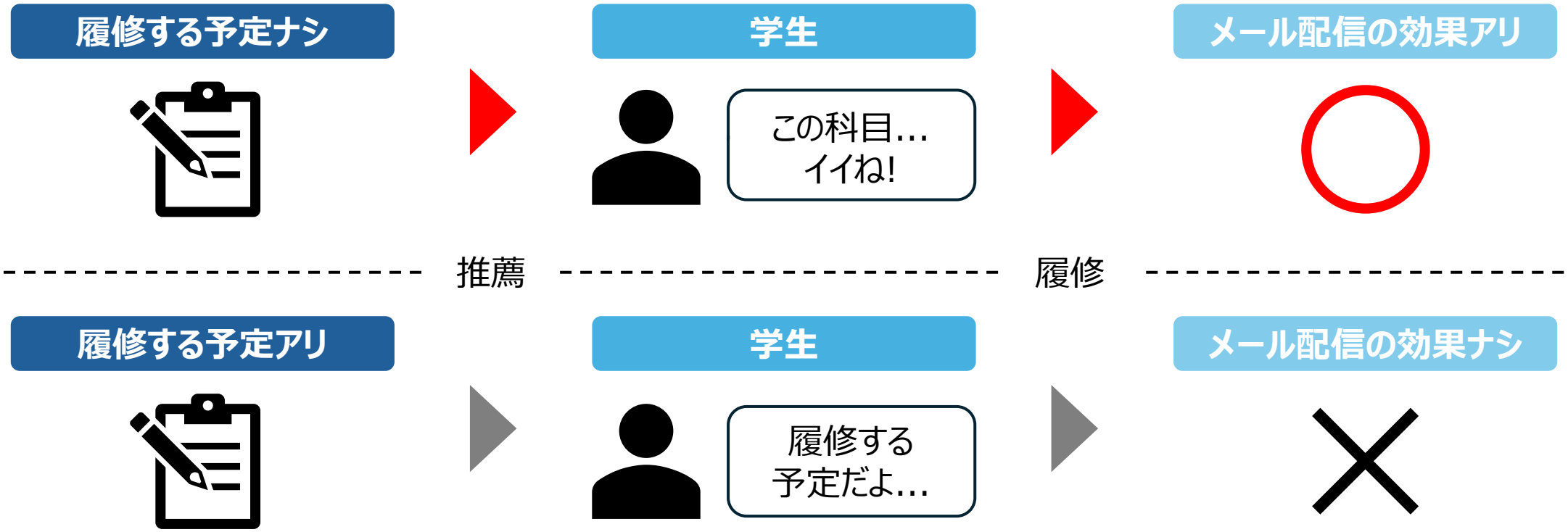
科目名	推薦するとき	推薦しないとき	増加量
科目A	1	1	0
科目B	1（履修した）	0（履修しない）	+1
科目C	0	0	0

図4：ある学生に1個の科目を推薦するとき・しないときのデータの例

※Appendix①,⑨参照

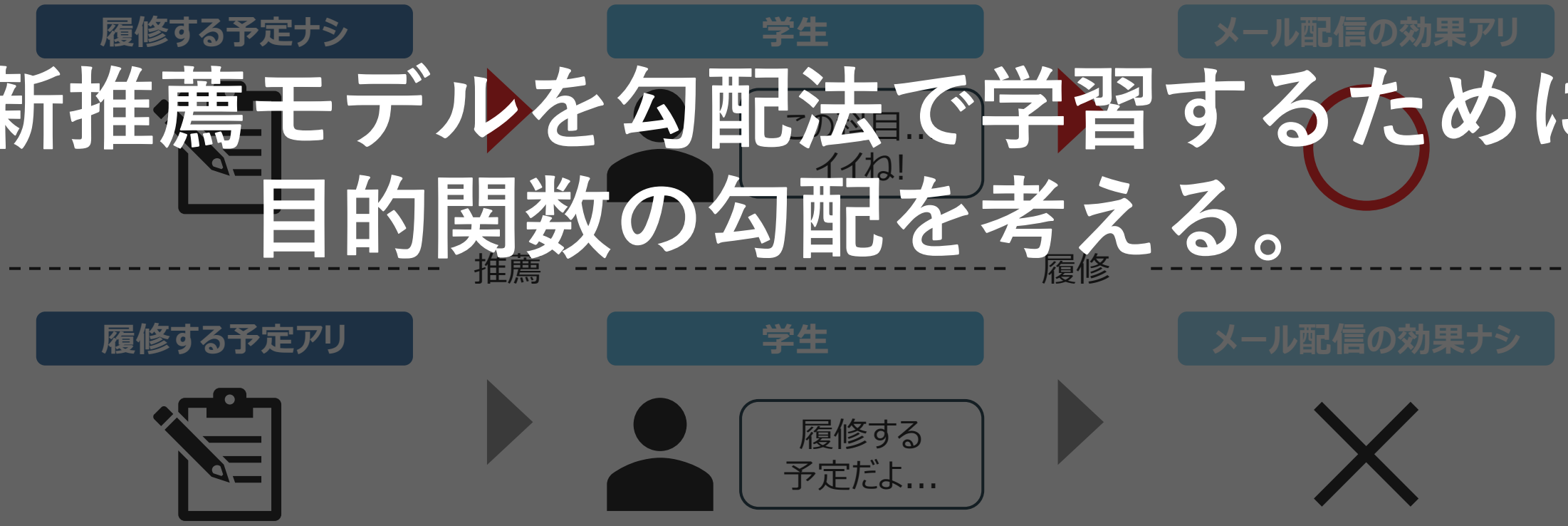
目的関数を定式化するプロセス

意思決定モデルπの目的は総履修者数を最大化すること。メール配信による推薦によって、その科目の履修者数を最大化することを考えたい。



目的関数の再定式化

意思決定モデルπの目的は総履修者数を最大化すること。メール配信による推薦によって、その科目の履修者数を最大化することを考えたい。



新推薦モデルを勾配法で学習するために目的関数の勾配を考える。

OPL ~新しい推定量~

既存のAdaptiveIPS推定量は分散が大きく上手く学習できないため、分散を抑えて安定した推定が可能な“NewDR推定量”を新しく開発した。

NewIPS推定量　既存　斎藤優太.反実仮想機械学習.技術評論社,2024,[336]

$$\widehat{\nabla_\theta V}_{\mathrm{newIPS}}(\pi_\theta; D) = \sum_{u\in\mathcal{U}} \left\{ \frac{\pi_\theta(a_u|x_u)}{\pi_0(a_u|x_u)} r_u(a,1) - \sum_{a\in\mathcal{A}\setminus\{a_u\}} \frac{\pi_\theta(a_u|x_u)}{1-\pi_0(a_u|x_u)} r_u(a,0) \right\} \nabla_\theta \log \pi_\theta(a|x_u)$$

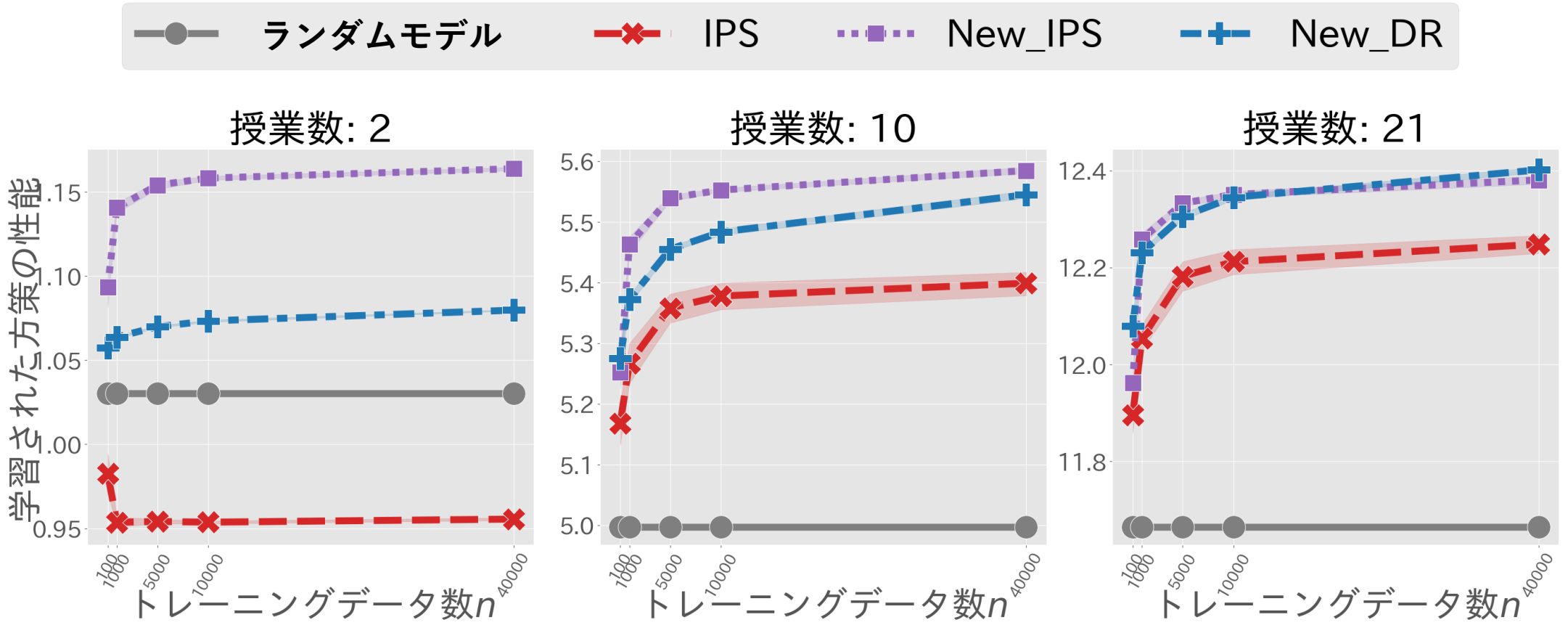
NewDR推定量　**New!!**

$$\begin{aligned}
&\widehat{\nabla_\theta V}_{\mathrm{newDR}}(\pi_\theta; D, q_1, q_0) \\
&= \sum_{u\in\mathcal{U}} \left\{ \frac{\pi_\theta(a_u|x_u)}{\pi_0(a_u|x_u)} \left(r_u(a,1) - \widehat{q_1}(x_u,a_u)\right) - \sum_{a\in\mathcal{A}\setminus\{a_u\}} \frac{\pi_\theta(a_u|x_u)}{1-\pi_0(a_u|x_u)} \left(r_u(a,0) - \widehat{q_0}(x_u,a_u)\right) \right\} \nabla_\theta \log \pi_\theta(a|x_u) \\
&+ \sum_{u\in\mathcal{U}} \sum_{a\in\mathcal{A}} \{\pi_\theta(a|x_u)\hat{q}_1(x_u,a) + \pi_\theta(a|x_u)\hat{q}_0(x_u,a)\} \nabla_\theta \log \pi_\theta(a|x_u)
\end{aligned}$$

※Appendix①,⑤,⑪,⑫参照

OPLの結果

最適な新推薦モデルの学習が進み，旧推薦モデルと比較して，観測される性能(※$_1$)が向上している。



※$_1$Appendix⑬,⑭,⑮,⑯参照　　※$_2$OPEの新規推定量：Appendix③,④,⑪,⑫参照

問題設定における課題と改善点

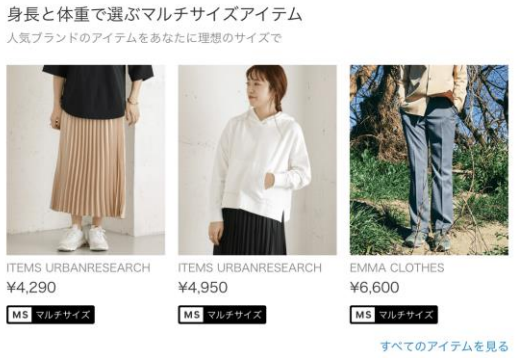
今回の定式化では、問題設定において考慮しきれなかった課題や改善点が挙げられる。

- シミュレーションデータで学習・評価したが、実際のデータセットでこのようなことは可能なのか？
- 授業の情報(特徴量)を考量していない。
- 一つの授業を推薦するのではなく，ランキング形式にして複数の授業を推薦することで学生の選択の幅が広がる。
- 今回提案する新推薦モデルを実運用して得たログデータを用いて，3次科目登録の最大化を目的とする推薦モデルを同様に考えることができる。

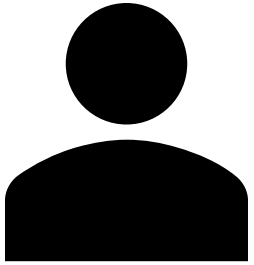
オープンデータを用いた検証

ZOZOTOWNが提供しているオープンデータセットから、推薦された商品のクリック率を最大化する推薦モデルを考える。

ECサイト ▶ **顧客** ▶ **クリック**



3種類の商品を
特定の位置に推薦



ZOZOTOWNのECに
訪れるすべての顧客

あるアイテム広告をクリック
すると報酬を与える

オープンデータとシミュレーションデータの類似性

観測された実際のログデータは，旧推薦モデルによって商品が推薦され，顧客のクリックの有無が記録されている。

ZOZOTOWNのオープンデータ

顧客	性別	推薦する商品	クリックしたか	…
1	男性	アイテムC	1	
2	男性	アイテムF	0	
3	女性	アイテムR	0	
4	男性	アイテムQ	0	
5	女性	アイテムB	1	
6	女性	アイテムQ	1	
7	女性	アイテムE	0	

シミュレーションデータ

学生	性別	推薦する科目	履修したか	…
1	男性	科目A	1	
2	女性	科目C	0	
3	男性	科目F	1	
4	女性	科目B	0	
5	男性	科目C	1	
6	女性	科目D	0	
7	男性	科目K	1	

オープンデータとシミュレーションデータの類似性

観測された実際のログデータは，旧推薦モデルによって商品が推薦され，顧客のクリックの有無が記録されている。

ZOZOTOWNのオープンデータ

顧客ID	性別	推薦する商品	クリックしたか	…
1	男性	アイテムC	1	
2	男性	アイテムF	0	
3	女性	アイテムR	0	
4	男性	アイテムQ	0	
5	女性	アイテムB	1	
6	女性	アイテムQ	1	
7	女性	アイテムE	0	

シミュレーションデータ

学生	性別	推薦する科目	履修したか	…
1	男性	科目A	1	
2	女性	科目C	0	
3	男性	科目F	1	
4	女性	科目B	0	
5	男性	科目C	1	
6	女性	科目D	0	
7	男性	科目K	1	

オープンデータを用いたOPE, OPLで上手く学習・評価できた。

※Appendix⑱, ⑲参照

まとめ

実験

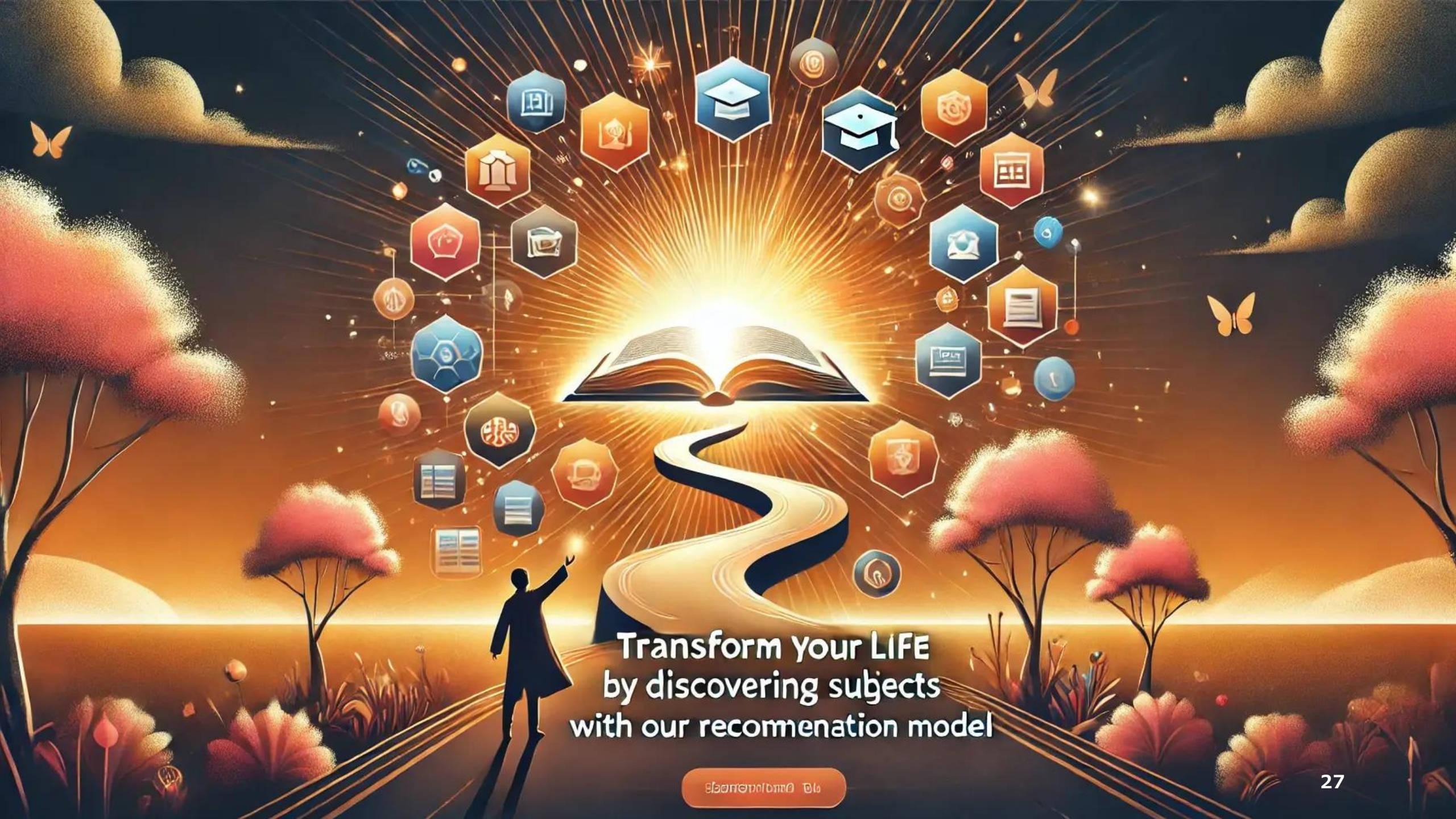
シミュレーションデータ（現実を模倣したデータセット）から、OPE・OPLで新推薦モデルが学習・評価できるか試した.
結果、新推定量を用いて推薦モデルの性能の向上に成功した。

検証

シミュレーションデータと似た形式のオープンデータで異なる問題設定におけるOPE・OPLが上手くいくかどうか検証し、性能の良い推薦モデルの開発に成功した。

結論

オープンデータで上手くいったことから、シミュレーションデータで学習・評価した新しい施策も、実データがあれば実際に運用できる可能性がある。


Transform your LIFE
by discovering subjects
with our recomenation model

参考文献

◆ 斎藤優太.反実仮想機械学習.技術評論社,2024,[336]
◆ 斎藤優太.施策デザインのための機械学習入門,2021,[336]
◆ https://research.zozo.com/data.html 「Open Bandit Dataset」
◆ https://github.com/st-tech/zr-obp 「Open Bandit Pipelineのgithub」
◆ 齋藤優太."Off-Policy Evaluationの基礎とZOZOTOWN大規模公開実データおよびパッケージ紹介". 2020-09-03.https://techblog.zozo.com/entry/openbanditproject .
◆ https://zr-obp.readthedocs.io/en/latest/ 「Open Bandit Pipelineのドキュメント」

論文
◆ Yuta Saito, Shunsuke Aihara, Megumi Matsutani, Yusuke Narita . "Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation". Arxiv.2020.
◆ Noveen Sachdeva, Yi Su, and Thorsten Joachims. "Off-policy Bandits with Deficient Support. ". In Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2020
◆ Haruka Kiyohara, Masahiro Nomura, and Yuta Saito. Off-Policy Evaluation of Slate Bandit Policies via Optimizing Abstraction. arXiv preprint arXiv:2402.02171.2024

コード
◆ 斎藤優太.反実仮想機械学習.https://github.com/ghmagazine/cfml_book
◆ 書いたコード(随時更新). https://github.com/taishirooo/cfml-WASEDA-datascience-competition

ご清聴ありがとうございました。

Appendix⓪ ~ログデータの定式化~

データ科学科目	推薦するとき：$r(a,1)$	推薦しないとき：$r(a,0)$	増加量
データ科学入門α	1	1	0
統計リテラシーα	1	0	+1
Rによる統計解析	0	1	-1

図5：ある学生に1個の科目を推薦するとき・しないときのデータの例

$$\mathcal{D} = \left\{(x_u, a_u, r_u(a_u,1), \{r_u(a,0)\}_{a\in A\setminus\{a_u\}})\right\}_{u\in\mathcal{U}} \sim \prod_{u\in\mathcal{U}} \pi_0(a_u|x_u)p(r_u(a_u,1)|x_u) \prod_{a\in \mathrm{A}\setminus\{a_u\}} p(r_u(a,0)|x_u)$$

$u \in \mathcal{U} = \{1,2,\dots,50000\}$	:	学生
$x \in \mathcal{X}\ (\in \mathbb{R}^{\text{students}\times\text{features}})$	:	学生の特徴量
$a \in \mathcal{A} = \{1,2,\dots,21\}$	:	データ科学科目
$r(a,1) = \{0,1\}$	:	科目を推薦したときの報酬
$r(a,0) = \{0,1\}$	:	科目を推薦しないときの報酬
$\pi_0(a\|x)$	:	データ収集方策（ランダムモデル）
$p(r(a,1)\|x)$	:	ある学生に対して、科目を推薦したときの報酬分布
$p(r(a,0)\|x)$	:	ある学生に対して、科目を推薦しないときの報酬分布

Appendix① ～定義について～

定義0.1

今回の問題設定（データ科学の科目推薦問題）について、意思決定モデルπの真の勾配を定義する。

$$\nabla_\theta V(\pi_\theta) = \nabla_\theta \sum_{u \in \mathcal{U}} \mathbb{E}_{\pi_\theta(a|x_u)}[q_1(x_u, a) - q_0(x_u, a)] = \sum_{u \in \mathcal{U}} \mathbb{E}_{\pi_\theta(a|x_u)}[\{q_1(x_u, a) - q_0(x_u, a)\}\nabla_\theta \log \pi_\theta(a|x_u)]$$

定義0.2

今回の問題設定（データ科学の科目推薦問題）について、意思決定モデルπの真の性能を定義する。

$$V(\pi) = \frac{1}{u}\sum_{u \in \mathcal{U}} \sum_{a \in \mathcal{A}} \{\pi(a|x_u) q_1(x_u, a) + \big(1 - \pi(a|x_u)\big) q_0(x_u, a)\}$$

注意

今回は、OPLとOPEについて解く問題が異なることに注意してください。OPLが解く問題は、期待報酬関数の差を最大化する問題に帰着したから、期待報酬関数の和を考えるOPEの問題とは別の問題です。

Appendix② ～新しい定義について～

仮定0

すべての$x \in \mathcal{X}$および$a \in \mathcal{A}$について、以下の式を満たすとき、データ収集方策π_0は意思決定モデルπに対して共通サポートを持つという。

$$\forall x \in \mathcal{X}, \forall a \in \mathcal{A}, \qquad \pi(a|x) > 0 \Rightarrow \pi_0(a|x) > 0$$

定義1 **New!**

あるデータ収集方策π_0が収集したログデータDが与えられている。意思決定モデルπ_θの真の勾配$\nabla_\theta V(\pi_\theta)$に対するnewDR推定量$\widehat{\nabla_\theta V}_{\mathrm{newDR}}(\pi_\theta; D, \hat{q}_1, \hat{q}_1)$を定義する。

$$\begin{aligned}
&\widehat{\nabla_\theta V}_{\mathrm{newDR}}(\pi_\theta; D, q_1, q_0) \\
&= \sum_{u \in \mathcal{U}} \left\{ \frac{\pi_\theta(a_u|x_u)}{\pi_0(a_u|x_u)} \left(r_u(a,1) - \widehat{q_1}(x_u, a_u)\right) - \sum_{a \in \mathcal{A} \setminus \{a_u\}} \frac{\pi_\theta(a_u|x_u)}{1 - \pi_0(a_u|x_u)} \left(r_u(a,0) - \widehat{q_0}(x_u, a_u)\right) \right\} \nabla_\theta \log \pi_\theta(a|x_u) \\
&+ \sum_{u \in \mathcal{U}} \sum_{a \in \mathcal{A}} \{\pi_\theta(a|x_u)\hat{q}_1(x_u, a) + \pi_\theta(a|x_u)\hat{q}_0(x_u, a)\} \nabla_\theta \log \pi_\theta(a|x_u)
\end{aligned}$$

Appendix③ ~新しい定義について~

定義2 **New!**

意思決定モデルπがオンライン実験から収集したログデータ$\mathcal{D}_{\text{online}}$が与えられている。意思決定モデル$\pi$の真の性能$V(\pi)$に対するnewAVG推定量$\hat{V}_{\text{newAVG}}(\pi; D_{\text{online}})$を定義する。

$$\hat{V}_{\text{newAVG}}(\pi; D_{\text{online}}) = \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{r(a_u, 1) + \sum_{a\in\mathcal{A}\setminus\{a_u\}} r(a, 0)\right\}$$

定義3 **New!**

あるデータ収集方策π_0が収集したログデータ$\mathcal{D}$が与えられている。意思決定モデルπ_θの真の性能$V(\pi)$に対するnewIPS推定量$\hat{V}_{\text{newIPS}}(\pi; \mathcal{D})$を定義する。

$$\hat{V}_{\text{newIPS}}(\pi; \mathcal{D}) = \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)} r(a_u, 1) + \sum_{a\in\mathcal{A}\setminus\{a_u\}} \frac{1-\pi(a_u|x_u)}{1-\pi_0(a_u|x_u)}\big(r(a, 0)\big)\right\}$$

Appendix④ ～新しい定義について～

定義4 **New!**

あるデータ収集方策π_0が収集したログデータ$\mathcal{D}$が与えられている。意思決定モデルπ_θの真の性能$V(\pi)$に対するnewDR推定量$\hat{V}_{\mathrm{newIPS}}(\pi;\mathcal{D},\hat{q}_1,\hat{q}_0)$を定義する。

$$\hat{V}_{\mathrm{newDR}}\left(\pi;\mathcal{D}_{\log},\widehat{q_1},\widehat{q_0}\right)=\frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}\left(r(a_u,1)-\hat{q}_1(x_u,a_u)\right)+\sum_{a\in\mathcal{A}\setminus\{a_u\}}\frac{1-\pi(a|x_u)}{1-\pi_0(a|x_u)}\left(r(a,0)-\hat{q}_0(x_u,a)\right)\right\}$$
$$+\frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\pi(a_u|x_u)\hat{q}_1(x_u,a_u)+\left(1-\pi(a_u|x_u)\right)\hat{q}_0(x_u,a_u)\right\}$$

注意
今回は、OPLとOPEについて解く問題が異なることに注意してください。OPLが解く問題は、期待報酬関数の差を最大化する問題に帰着したから、期待報酬関数の和を考えるOPEの問題とは別の問題です。

Appendix⑤ ~新しい定理について~

定理1.1 **New!**

データ収集方策π_0が収集したログデータ$\mathcal{D}$が与えられている。NewDR推定量$\widehat{\nabla_\theta V}_{\mathrm{newDR}}(\pi_\theta; D, \hat{q}_1, \hat{q}_1)$は、共通サポートの仮定（仮定1）の下で、意思決定モデルπの真の勾配$\nabla_\theta V(\pi_\theta)$に対する不偏推定量である。

$$
\begin{aligned}
&\mathbb{E}\left[\hat{V}_{\mathrm{newDR}}(\pi_\theta; \mathcal{D}, \hat{q}_1, \hat{q}_0)\right] \\
&= \mathbb{E}_{p(\mathcal{D})}\left[\sum_{u\in\mathcal{U}}\left\{\frac{\pi_\theta(a_u|x_u)}{\pi_0(a_u|x_u)}\left(r(a_u,1)-\hat{q}_1(x_u,a)\right)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{\pi_\theta(a_u|x_u)}{1-\pi_0(a'|x_u)}\left(r(a',0)-\hat{q}_0(x_u,a')\right)\right\}+\sum_{u\in\mathcal{U}}\sum_{a\in\mathcal{A}}\{\pi_\theta(a|x_u)\hat{q}_1(x_u,a)+\pi_\theta(a|x_u)\hat{q}_0(x_u,a)\}\right] \\
&= \sum_{u\in\mathcal{U}}\mathbb{E}_{\pi_0(a|x_u)p(r(a,1),\{r(a',0)\}_{a'\in\mathcal{A}\setminus\{a_u\}}|x_u)}\left[\left\{\frac{\pi_\theta(a_u|x_u)}{\pi_0(a_u|x_u)}\left(r(a_u,1)-\hat{q}_1(x_u,a)\right)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{\pi_\theta(a_u|x_u)}{1-\pi_0(a'|x_u)}\left(r(a',0)-\hat{q}_0(x_u,a')\right)\right\}+\sum_{a\in\mathcal{A}}\{\pi_\theta(a|x_u)\hat{q}_1(x_u,a)+\pi_\theta(a|x_u)\hat{q}_0(x_u,a)\}\right] \\
&= \sum_{u\in\mathcal{U}}\mathbb{E}_{\pi_0(a|x_u)}\left[\left\{\frac{\pi_\theta(a_u|x_u)}{\pi_0(a_u|x_u)}\left(q_1(x_u,a)-\hat{q}_1(x_u,a)\right)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{\pi_\theta(a_u|x_u)}{1-\pi_0(a'|x_u)}\left(q_0(x_u,a')-\hat{q}_0(x_u,a')\right)\right\}+\sum_{a\in\mathcal{A}}\{\pi_\theta(a|x_u)\hat{q}_1(x_u,a)+\pi_\theta(a|x_u)\hat{q}_0(x_u,a)\}\right] \\
&= \sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\pi_0(a_u|x_u)\frac{\pi_\theta(a_u|x_u)}{\pi_0(a_u|x_u)}\left(q_1(x_u,a)-\hat{q}_1(x_u,a)\right)+\sum_{a\in\mathcal{A}}\pi_0(a_u|x_u)\sum_{a'\in\mathcal{A}}\mathbb{I}\{a'\neq a\}\frac{\pi_\theta(a_u|x_u)}{1-\pi_0(a'|x_u)}\left(q_0(x_u,a')-\hat{q}_0(x_u,a')\right)+\sum_{a\in\mathcal{A}}\{\pi_\theta(a|x_u)\hat{q}_1(x_u,a)+\pi_\theta(a|x_u)\hat{q}_0(x_u,a)\}\right\} \\
&= \sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\pi_\theta(a_u|x_u)\left(q_1(x_u,a)-\hat{q}_1(x_u,a)\right)+\sum_{a'\in\mathcal{A}}\mathbb{I}\{a'\neq a\}\pi(a|x_u)\sum_{a'\in\mathcal{A}}\frac{\pi_\theta(a_u|x_u)}{1-\pi_0(a'|x_u)}\left(q_0(x_u,a')-\hat{q}_0(x_u,a')\right)+\sum_{a\in\mathcal{A}}\{\pi_\theta(a|x_u)\hat{q}_1(x_u,a)+\pi_\theta(a|x_u)\hat{q}_0(x_u,a)\}\right\} \\
&= \sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\left\{\pi_\theta(a_u|x_u)\left(q(x_u,a)-\hat{q}_1(x_u,a)\right)+\pi_\theta(a_u|x_u)\left(q_0(x_u,a')-\hat{q}_0(x_u,a')\right)\right\}+\sum_{a\in\mathcal{A}}\{\pi_\theta(a|x_u)\hat{q}_1(x_u,a)+\pi_\theta(a|x_u)\hat{q}_0(x_u,a)\}\right\} \\
&= \sum_{u\in\mathcal{U}}\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)q_1(x_u,a)+\left(1-\pi(a|x_u)\right)q_0(x_u,a)\right\}=V(\pi)
\end{aligned}
$$

Appendix⑥ ~新しい定理について~

定理2.1 **NEW!**

意思決定モデルπがオンライン実験から収集したログデータ$\mathcal{D}_{\text{online}}$が与えられている。このとき、定義2で定義されるnewAVG推定量$\hat{V}_{\text{newAVG}}(\pi; D_{\text{online}})$は、意思決定モデル$\pi$の真の性能$V(\pi)$に対する不偏推定量である。

$$
\begin{aligned}
\mathbb{E}\left[\hat{V}_{\text{newAVG}}(\pi; D_{\text{online}})\right] &= \mathbb{E}_{p(\mathcal{D}_{\text{online}})}\left[\frac{1}{u}\sum_{u\in\mathcal{U}}\left\{r(a_u,1)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}} r(a',0)\right\}\right] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}} \mathbb{E}_{\pi(a|x_u)p(r(a,1),\{r(a',0)\}_{a'\in\mathcal{A}\setminus\{a_u\}}|x_u)}\left[r(a_u,1)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}} r(a',0)\right] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}} \mathbb{E}_{\pi(a|x_u)}[q_1(x_u,a)+\sum_{a'\in\mathcal{A}} \mathbb{I}\{a'\neq a\}q_0(x_u,a')] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\pi(a|x_u)q_1(x_u,a)+\sum_{a\in\mathcal{A}}\pi(a|x_u)\sum_{a'\in\mathcal{A}}\mathbb{I}\{a'\neq a\}q_0(x_u,a')\right\} \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\pi(a|x_u)q_1(x_u,a)+\sum_{a'\in\mathcal{A}}\mathbb{I}\{a'\neq a\}\,\pi(a|x_u)\sum_{a'\in\mathcal{A}}q_0(x_u,a')\right\} \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)q_1(x_u,a)+\left(1-\pi(a|x_u)\right)q_0(x_u,a)\right\} = V(\pi)
\end{aligned}
$$

Appendix⑦ ~新しい定理について~

定理3.1 **NEW!**

あるデータ収集方策π_0が収集したログデータ$\mathcal{D}$が与えられている。このとき、定義4で定義されるNewIPS推定量$\hat{V}_{\mathrm{newIPS}}(\pi;\mathcal{D})$は、共通サポートの仮定（仮定1）の下で、意思決定モデルπの真の性能$V(\pi)$に対する不偏推定量である。

$$
\begin{aligned}
&\mathbb{E}\left[\hat{V}_{\mathrm{newIPS}}(\pi;\mathcal{D})\right] \\
&= \mathbb{E}_{p(\mathcal{D})}\left[\frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}r(a_u,1)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}r(a',0)\right\}\right] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\mathbb{E}_{\pi_0(a|x_u)p(r(a,1),\{r(a',0)\}_{a'\in\mathcal{A}\setminus\{a_u\}}|x_u)}\left[\left\{\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}r(a_u,1)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}r(a',0)\right\}\right] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\mathbb{E}_{\pi_0(a|x_u)}\left[\left\{\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}q_1(x_u,a)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}q_0(x_u,a')\right\}\right] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\pi_0(a_u|x_u)\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}q_1(x_u,a)+\sum_{a\in\mathcal{A}}\pi_0(a_u|x_u)\sum_{a'\in\mathcal{A}}\mathbb{I}\{a'\neq a\}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}q_0(x_u,a')\right\} \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\pi(a|x_u)q_1(x_u,a)+\sum_{a'\in\mathcal{A}}\mathbb{I}\{a'\neq a\}\pi(a|x_u)\sum_{a'\in\mathcal{A}}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}q_0(x_u,a')\right\} \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)q_1(x_u,a)+\left(1-\pi(a'|x_u)\right)q_0(x_u,a')\right\}\right\} \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)q_1(x_u,a)+\left(1-\pi(a|x_u)\right)q_0(x_u,a)\right\} = V(\pi)
\end{aligned}
$$

Appendix⑧ ~新しい定理について~

定理4.1 **NEW!**

あるデータ収集方策π_0が収集したログデータ$\mathcal{D}$が与えられている。このとき、定義4で定義されるNewDR推定量$\hat{V}_{\mathrm{newDR}}(\pi;\mathcal{D},\hat{q}_1,\hat{q}_0)$は、共通サポートの仮定（仮定1）の下で、意思決定モデルπの真の性能$V(\pi)$に対する不偏推定量である。

$$
\begin{aligned}
&\mathbb{E}\left[\hat{V}_{\mathrm{newDR}}(\pi;\mathcal{D},\hat{q}_1,\hat{q}_0)\right] \\
&= \mathbb{E}_{p(\mathcal{D})}\left[\frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}\left(r(a_u,1)-\hat{q}_1(x_u,a)\right)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}\left(r(a',0)-\hat{q}_0(x_u,a')\right)\right\}+\frac{1}{u}\sum_{u\in\mathcal{U}}\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)\hat{q}_1(x_u,a)+\left(1-\pi(a|x_u)\right)\hat{q}_0(x_u,a)\right\}\right] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\mathbb{E}_{\pi_0(a|x_u)p(r(a,1),\{r(a',0)\}_{a'\in\mathcal{A}\setminus\{a_u\}}|x_u)}\left[\left\{\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}\left(r(a_u,1)-\hat{q}_1(x_u,a)\right)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}\left(r(a',0)-\hat{q}_0(x_u,a')\right)\right\}+\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)\hat{q}_1(x_u,a)+\left(1-\pi(a|x_u)\right)\hat{q}_0(x_u,a)\right\}\right] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\mathbb{E}_{\pi_0(a|x_u)}\left[\left\{\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}\left(q_1(x_u,a)-\hat{q}_1(x_u,a)\right)+\sum_{a'\in\mathcal{A}\setminus\{a_u\}}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}\left(q_0(x_u,a')-\hat{q}_0(x_u,a')\right)\right\}+\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)\hat{q}_1(x_u,a)+\left(1-\pi(a|x_u)\right)\hat{q}_0(x_u,a)\right\}\right] \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\pi_0(a_u|x_u)\frac{\pi(a_u|x_u)}{\pi_0(a_u|x_u)}\left(q_1(x_u,a)-\hat{q}_1(x_u,a)\right)+\sum_{a\in\mathcal{A}}\pi_0(a_u|x_u)\sum_{a'\in\mathcal{A}}\mathbb{I}\{a'\neq a\}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}\left(q_0(x_u,a')-\hat{q}_0(x_u,a')\right)+\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)\hat{q}_1(x_u,a)+\left(1-\pi(a|x_u)\right)\hat{q}_0(x_u,a)\right\}\right\} \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\pi(a|x_u)\left(q_1(x_u,a)-\hat{q}_1(x_u,a)\right)+\sum_{a'\in\mathcal{A}}\mathbb{I}\{a'\neq a\}\pi(a|x_u)\sum_{a'\in\mathcal{A}}\frac{1-\pi(a'|x_u)}{1-\pi_0(a'|x_u)}\left(q_0(x_u,a')-\hat{q}_0(x_u,a')\right)+\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)\hat{q}_1(x_u,a)+\left(1-\pi(a|x_u)\right)\hat{q}_0(x_u,a)\right\}\right\} \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)\left(q_1(x_u,a)-\hat{q}_1(x_u,a)\right)+\left(1-\pi(a'|x_u)\right)\left(q_0(x_u,a')-\hat{q}_0(x_u,a')\right)\right\}+\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)\hat{q}_1(x_u,a)+\left(1-\pi(a|x_u)\right)\hat{q}_0(x_u,a)\right\}\right\} \\
&= \frac{1}{u}\sum_{u\in\mathcal{U}}\sum_{a\in\mathcal{A}}\left\{\pi(a|x_u)q_1(x_u,a)+\left(1-\pi(a|x_u)\right)q_0(x_u,a)\right\}=V(\pi)
\end{aligned}
$$

Appendix⑨ ～目的関数の設定の証明～

定理5

今回の問題設定において、新施策π_θを学習するときの目的関数は次の式である。

$$\max_\theta V(\pi_\theta) \Rightarrow \max_\theta \sum_{u \in \mathcal{U}} \mathbb{E}_{\pi_\theta(a|x_u)}[q_1(x_u, a) - q_0(x_u, a)]$$

今、目的関数は、推薦する場合と推薦しない場合を考慮する、総履修者数を最大化することを考える。
以下の目的関数を計算すると、

$$V(\pi) = \sum_{u \in U} \sum_{a \in \mathcal{A}} \{\pi(a|x_u) q_1(x_u, a) + \big(1 - \pi(a|x_u)\big) q_0(x_u, a)\}$$

$$= \sum_{u \in \mathcal{U}} \left\{ \mathbb{E}_{\pi(a|x_u)}[q_1(x_u, a) - q_0(x_u, a)] + \underbrace{\sum_{a \in \mathcal{A}} q_0(x_u, a)}_{①} \right\}$$

ここで,下線部①は新施策πに依存しないから、

$$\max_\theta V(\pi_\theta) \Rightarrow \max_\theta \sum_{u \in \mathcal{U}} \mathbb{E}_{\pi_\theta(a|x_u)}[q_1(x_u, a) - q_0(x_u, a)]$$

Appendix⑩ ～新しい推定量の良さ～

OPEの問題について、真の性能$V(\pi)$の推定量$\hat{V}(\pi; D)$の“良さ”を平均二乗誤差で出力する。つまり、平均二乗誤差が小さい場合、推定量$\hat{V}(\pi; D)$は真の性能$V(\pi)$に対する“良い”推定量ということだ。

$$\mathrm{MSE}\left[\hat{V}\right] = \mathbb{E}_{p(\mathcal{D})}\left[\left(V - \hat{V}\right)^2\right]$$
$$\left(V = V(\pi), \hat{V} = \hat{V}(\pi; \mathcal{D})\right)$$

また、平均二乗誤差は二乗バイアスとバリアンスに分解可能である。

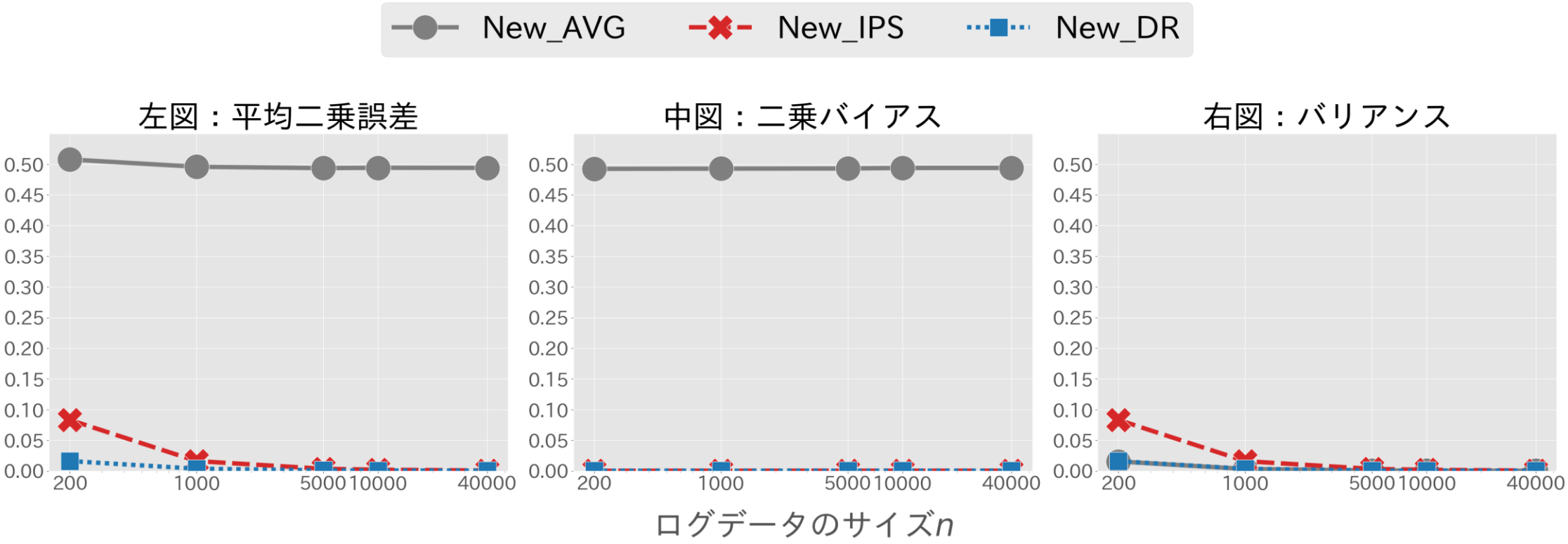
$$\begin{aligned}
\mathrm{MSE}\left[\hat{V}\right] &= \mathbb{E}_{p(\mathcal{D})}\left[\left(V - \hat{V}\right)^2\right] \\
&= \mathbb{E}\left[\left(V - \mathbb{E}\left[\hat{V}\right] + \mathbb{E}\left[\hat{V}\right] - \hat{V}\right)^2\right] \\
&= \mathbb{E}\left[\left(V - \mathbb{E}\left[\hat{V}\right]\right)^2 + 2\left(V - \mathbb{E}\left[\hat{V}\right]\right)\left(\mathbb{E}\left[\hat{V}\right] - \hat{V}\right) + \left(\hat{V} - \mathbb{E}\left[\hat{V}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(V - \mathbb{E}\left[\hat{V}\right]\right)^2 + \left(\hat{V} - \mathbb{E}\left[\hat{V}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(V - \mathbb{E}\left[\hat{V}\right]\right)^2\right] + \mathbb{E}\left[\left(\hat{V} - \mathbb{E}\left[\hat{V}\right]\right)^2\right] \\
&= \mathrm{Bias}\left[\hat{V}\right]^2 + \mathrm{Var}\left[\hat{V}\right]
\end{aligned}$$

Appendix⑪ ~新しい推定量の良さ~

考察1.1

新しく定義した３つの推定量についてMSE ($= \mathrm{Bias}^2 + \mathrm{Variance}$)を出力した結果、NewIPS推定量とNewDR推定量が真の性能$V(\pi)$を“良好”に推定したことが確認できた。

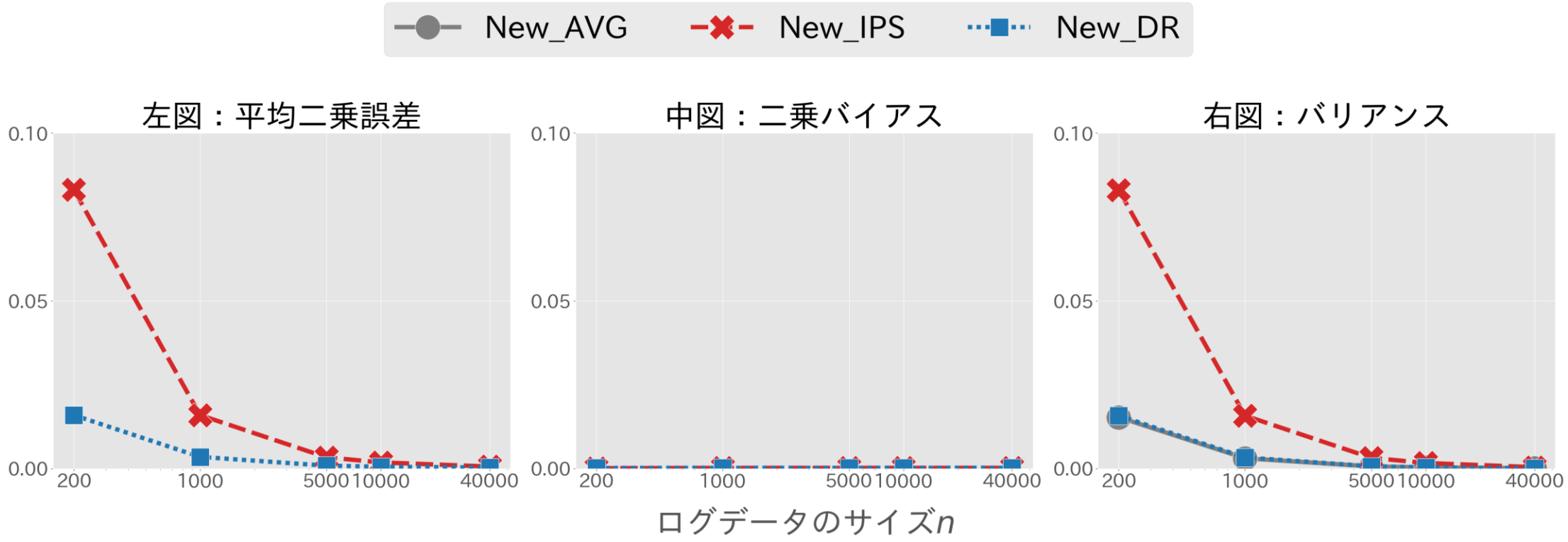
したがって、NewIPS推定量とNewDR推定量を、新施策の評価に適用できる根拠を得ることができた。



Appendix⑫ ~新しい推定量の良さ~

考察1.2

NewIPS推定量とNewDR推定量を比較した結果、NewDR推定量は真の性能$V(\pi)$を推定する際のバリアンスが小さいため、平均二乗誤差がNewIPS推定量よりも小さく抑えられていることが分かる。



Appendix⑬～OPLのコード上での学習の流れ～

目的関数の真の勾配を既存の推定量と新規推定量を用いて表現。その推定量の最大化を解くことで，最適な新推薦モデルを学習可能。

学習の流れ	対応した数式
目的関数の真の勾配を定義	$\nabla_\theta V(\pi_\theta) = \nabla_\theta \sum_{u \in \mathcal{U}} \mathbb{E}_{\pi_\theta(a\|x_u)}[q_1(x_u, a) - q_0(x_u, a)]$
真の勾配は未知であるため近似できる推定量を定義	$\widehat{\nabla_\theta V}_{\mathrm{NewIPS}}(\pi_\theta; D) \ / \ \widehat{\nabla_\theta V}_{newDR}(\pi_\theta; D, q_1, q_0)$
各推定量の最大化問題を解き，最適な$\hat{\pi}$を導く	$\max_\theta \widehat{\nabla_\theta V}_{\mathrm{NewIPS}}(\pi_\theta; D) \ / \ \max_\theta \widehat{\nabla_\theta V}_{newDR}(\pi_\theta; D, q_1, q_0)$
$\hat{\pi}$の性能を定義し，測定	$V(\hat{\pi}) = \sum_{u \in \mathcal{U}} \sum_{a \in \mathcal{A}} \{\hat{\pi}(a\|x_u) q_1(x_u, a) + \big(1 - \hat{\pi}(a\|x_u)\big) q_0(x_u, a)\}$

Appendix⑭～性能$V(\hat{\pi})$の意味（数式編）～

p.21の表の縦軸の「推薦モデルの性能$V(\hat{\pi})$」について説明する。また、次スライドのAppendix⑮では、具体例を用いて説明する。

新推薦モデル$\hat{\pi}$の性能$V(\hat{\pi})$

$$V(\hat{\pi}) = \sum_{u \in \mathcal{U}} \sum_{a \in \mathcal{A}} \{\hat{\pi}(a|x_u) q_1(x_u, a) + \big(1 - \hat{\pi}(a|x_u)\big) q_0(x_u, a)\}$$

展開した式

$$V(\hat{\pi}) = \sum_{u \in \mathcal{U}} \sum_{a \in \mathcal{A}} \left\{\hat{\pi}(a|x_u) \mathbb{E}_{p(r_1|x,a)}[r_1] + \big(1 - \hat{\pi}(a|x_u)\big) \mathbb{E}_{p(r_0|x,a)}[r_0]\right\}$$

意味

- 学生がその科目を履修する可能性が高いと期待される科目を正確に高確率で推薦できているかを評価する。
- さらに、推薦することによって初めて履修が促進される科目を効果的に推薦できる場合、モデルの性能がより優れていると判断されます。

Appendix⑮～性能$V(\hat{\pi})$の意味（数式編）～

p.21の表の縦軸の「推薦モデルの性能$V(\hat{\pi})$」について説明する。また、次スライドのAppendix⑮では、具体例を用いて説明する。

新推薦モデル$\hat{\pi}$の性能$V(\hat{\pi})$

$$V(\hat{\pi}) = \sum_{u \in \mathcal{U}} \sum_{a \in \mathcal{A}} \{\hat{\pi}(a|x_u) q_1(x_u, a) + (1 - \hat{\pi}(a|x_u)) q_0(x_u, a)\}$$

展開した式

$$V(\hat{\pi}) = \sum_{u \in \mathcal{U}} \sum_{a \in \mathcal{A}} \{\underbrace{\hat{\pi}(a|x_u) \mathbb{E}_{p(r_1|x,a)}[r_1]}_{①} + \underbrace{(1 - \hat{\pi}(a|x_u)) \mathbb{E}_{p(r_0|x,a)}[r_0]}_{②}\}$$

①最大値：1

②最大値：num(a) − 1

意味

- 学生がその科目を履修する可能性が高いと期待される科目を正確に高確率で推薦できているかを評価する。
- さらに、推薦することによって初めて履修が促進される科目を効果的に推薦できる場合、モデルの性能がより優れていると判断されます。

Appendix⑯～性能$V(\hat{\pi})$の意味（具体例編）　～

ある1人の学生に対して、それぞれ4つの科目を推薦する確率や、その学生が履修する確率などが与えられている状況を考える。

a	$\pi(a\|x)$	$p(r_1\|x,a)$	r_1	$p(r_0\|x,a)$	r_0
科目A	0.2	0.7	1	0.6	1
科目B	0.7	0.7	1	0.4	0
科目C	0.05	0.3	0	0.6	1
科目D	0.05	0.3	0	0.4	0

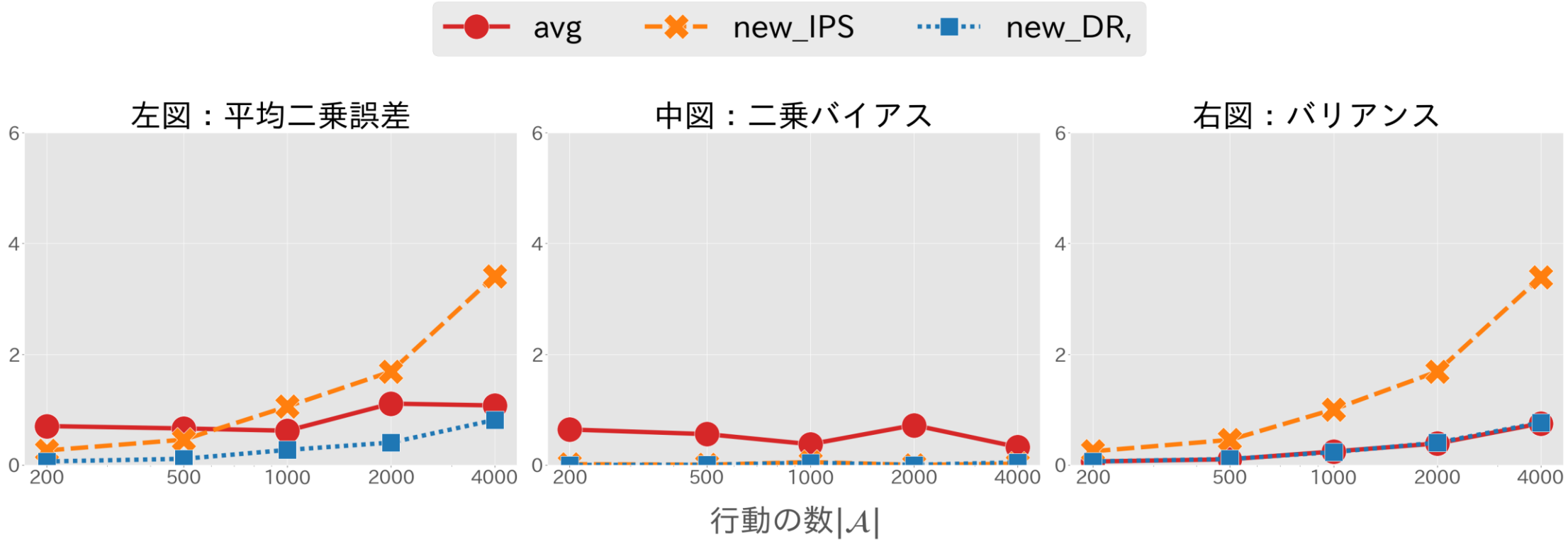
$$V(\hat{\pi}) = \sum_{u\in\mathcal{U}}\sum_{a\in\mathcal{A}}\left\{\hat{\pi}(a|x_u)\mathbb{E}_{p(r_1|x,a)}[r_1] + \left(1-\hat{\pi}(a|x_u)\right)\mathbb{E}_{p(r_0|x,a)}[r_0]\right\}$$

$$= (0.2\ 0.7\ 0.05\ 0.05)\begin{pmatrix}0.7\times 1\\ 0.7\times 1\\ 0.3\times 0\\ 0.3\times 0\end{pmatrix} + (0.8\ 0.3\ 0.95\ 0.95)\begin{pmatrix}06\times 1\\ 0.4\times 0\\ 0.6\times 1\\ 0.4\times 0\end{pmatrix}$$

$$= 0.63 + 1.05 = 1.68$$

Appendix⑰～授業数が増加した時のOPEの精度～

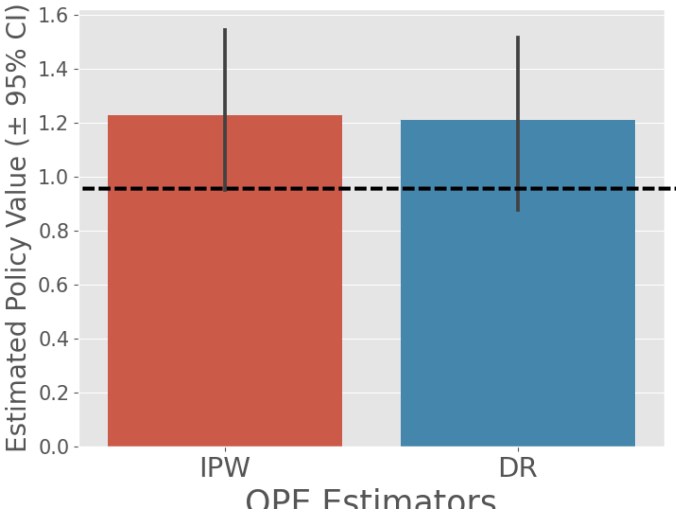
各推定量の平均二乗誤差の推移から，行動空間が500以内の場合は、性能高く推薦モデルを学習することができる。



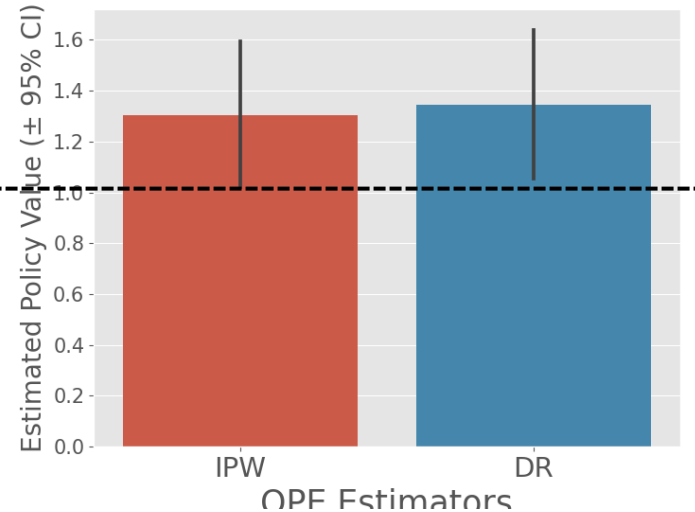
Appendix⑱～機械学習を用いた推薦モデルの構築～

顧客の特徴量から機械学習モデルを用いて$\pi(a|x)$を学習した。旧推薦モデルπ_0(BTSアルゴリズム)と比較して，オフライン上で優れた新推薦モデルだと分かった。

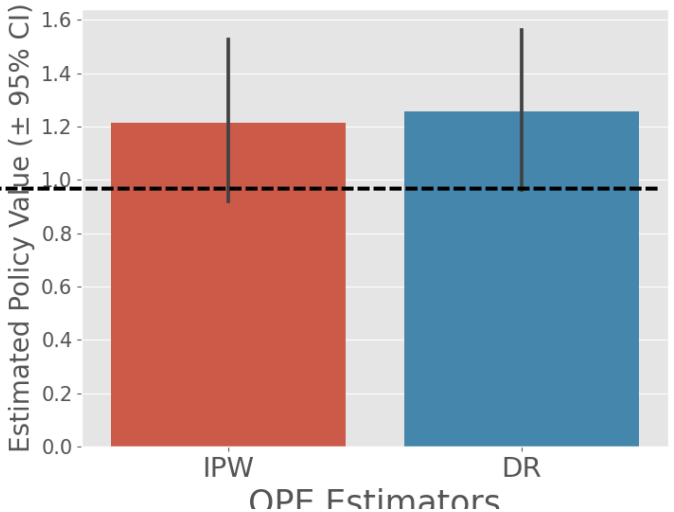
出力 $=\frac{\text{新推薦モデル}}{\text{旧推薦モデル}}$ なので、縦軸の値が1.0を超える場合、新推薦モデルは旧推薦モデルに比べて“良い“モデル



①ロジスティック回帰



②ランダムフォレスト



③勾配ブースティング

Appendix⑲～従来の推定量～

DM推定量

$$\hat{V}_{\mathrm{DM}}(\pi; D_{online}) = \frac{1}{\mathrm{n}}\sum_{i=1} q(x_i, \pi) = \frac{1}{n}\sum_{i=1}\sum_{a\in A} \pi(a|x_i) q^*(x_i, a)$$

IPS推定量

$$\hat{V}_{\mathrm{IPS}}(\pi; D) = \frac{1}{n}\sum_{i=1} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)} r_i = \frac{1}{n}\sum_{i=1} w(x_i, a_i) r_i$$

DR推定量

$$\hat{V}_{\mathrm{DR}}(\pi; D) = \frac{1}{n}\sum_{i=1} \{q^*(x_i, \pi) + w(x_i, a_i)(r_i - q^*(x_i, a_i))\}$$