The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

作用によって生じる共変量シフトバイアスの補正

Covariate shift bias correction induced by action

比嘉　亮太
Ryota Higa

NEC データサイエンス研究所
NEC Data Science Research Laboratories

In this study, we consider the covariates shift caused by the action on the data problem. Also, consider the method of bias correction using the density ratio estimation.In the numerical experiments, the action actually cause unsteady/extrapolation problem of distribution. It shows that have a significant bias to the expected value evaluation.Further illustrates that the method using the density ratio even in the case of changing the data size and dimensions are effective.

1. はじめに

人工知能/機械学習における研究やビジネスにおいて、データの蓄積と活用手法の確立がますます重要になっている。解くべき課題の明確化と十分なデータをそろえることが、機械学習を用いた予測モデルを構築するためのスタートラインである。しかしながら、これまでにとったことの無い行動を含むリーンスタートアップやキャンペーン最適化、新素材探索などは、本質的にデータ取得が部分的になる。その様な場合は、機械学習の手法が適用困難であり、適用できたとしても適切な結果を導けない。

データ取得が部分的である場合に生じる問題として、共変量シフトによって生じるバイアスの問題が指摘されている。これは、予測と評価時の入力分布が異なる、つまり非定常な分布によって生じる期待値評価のズレである。密度比推定を用いたアプローチは、非定常な分布によって引き起こされる回帰や分類において有効であることが知られている [Shimodaira 00, Sugiyama 07]。

また、我々が知りたい量としては、データのミクロな分布そのものではない。これまでと異なったアプローチを行った場合の費用対効果や売り上げなどの期待値であるマクロな量である。期待値は、上昇、一定、減少など現状との相対的な比較によって評価できる。

本研究において、データに対する作用によって生じる共変量シフトの問題と密度比推定を用いたバイアス補正の方法について考察する。数値実験において、最適化の作用が実際に分布の非定常/外挿問題を引き起こし、期待値評価に大きなバイアスを与えることを示す。さらに、データサイズや次元を変えた場合においても密度比を用いた方法が有効であることを示す。

2. 問題設定

具体例として、以下の目的関数の期待値評価を考える。

$$l(a, x, y) := xy - ca \tag{1}$$

ここで、a は操作変数、x は入力、y は出力であり、c は定数とする。例えば、a はキャンペーン、x は収入、y は解約であり、$l(a, x, y)$ はキャンペーンの費用対効果の指標と成る。または、a は電圧、 x は電流、y は抵抗とする。操作変数 a に対して、あるアルゴリズムを適用した後の、l の期待値を

$$\mathbb{E}_F[l(a, x, y)] := \sum_{a,y} \int \mathrm{d}x p_F(a, x, y) l(a, x, y) \tag{2}$$

と定める。本問題は、上記の期待値を評価することが目的である。しかし、a に対するアルゴリズム適用前のイニシャルデータセット $\mathcal{D}_I = \{a_n, x_n, y_n\}$ は得られているが、操作後のデータセットは部分的 $\mathcal{D}_F = \{a_n, x_n\}$ であるとする。

2.1 仮定

出力 y の条件付き分布において、以下の仮定を置く。

$$p_I(y|x, a) = p_F(y|x, a) \tag{3}$$

ここで、添え字の I は a の操作前、F は a の操作後とする。つまり、出力結果においては、保存則が成立りつしているとする。1 つの方法としては、過去のデータから $p(y|x, a)$ を推定して、それをデータを生成するシミュレータとみなして a 操作後の期待値を求める方法がある。しかしながら、推定したモデルに対して操作変数を変えて期待値評価を行うと、データの無い部分の外挿問題に対応してしまい、過大、過小評価の結果となる。以下では、密度比を用いたアプローチを示す。

3. 密度比を用いた定式化

操作変数 a によって、$p(a, x, y)$ に対して定常な分布は仮定できない。以下のような状況を考える。

$$p_I(a, x, y) \neq p_F(a, x, y) \tag{4}$$

ここで、両者の比を γ とすると、y の条件付分布の仮定より

$$\gamma(a, x) := \frac{p_F(y|a, x) p_F(a, x)}{p_I(y|a, x) p_I(a, x)} \tag{5}$$

$$= \frac{p_F(a, x)}{p_I(a, x)} \tag{6}$$

となり、y の情報が不要になることがわかる。

連絡先: 比嘉　亮太、NEC データサイエンス研究所、神奈川県川崎市中原区下沼部 1753、email:r-higa@ct.jp.nec.com

密度比 $\gamma(a,x)$ を用いると、(2) は

$$\begin{aligned}\mathbb{E}_F[l(a,x,y)] &= \sum_{a,y}\int \mathrm{d}x p_F(a,x,y)l(a,x,y)\\ &= \sum_{a,y}\int \mathrm{d}x \frac{p_F(a,x,y)}{p_I(a,x,y)} p_I(a,x,y)l(a,x,y)\\ &= \sum_{a,y}\int \mathrm{d}x \gamma(a,x)p_I(a,x,y)l(a,x,y) \quad (7)\\ &= \mathbb{E}_I[\gamma(a,x)l(a,x,y)] \quad (8)\end{aligned}$$

となる。ただし、$\gamma(x,a) < \infty$ とする。つまり、本来知りたい操作後の期待値は、操作前の期待値と γ によって評価できることを示している。これは、y に対して手元に得られているデータの活用方法やシミュレータなどのドメイン知識などの助けを借りることによって、マクロな期待値の評価は可能であることを意味している。

仮定を検証する方法としては、$r(x,a) = 1$ となる場合において、操作前後の期待値の差であるバイアス

$$B = \mathbb{E}_F[l(a,x,y)] - \mathbb{E}_I[\gamma(a,x)l(a,x,y)] \quad (9)$$

がゼロに成ることを確かめればよい。これは、y を生成するモデルに対する改良にも使える。例えば、仮説の再検討において、時間 t を与えた $p_I(y|x,a) = p_F(y|x,a)\exp(-\alpha t)$ として、パラメータ α を決定する問題とする。ミュレータと現実の環境の変化への適用といった、シミュレータ開発者にとっても、データ主導による改良できる可能性を示している。

4. 数値実験

密度比を用いた方法と、仮説を用いて y のモデルを推定する方法の比較を行い、分布のズレによるバイアス補正の効果を実験的に確かめる。

4.1 設定

検証の実験として、以下の様な設定を行う。

$$a \sim Bernoulli(a|\mu_a) \quad (10)$$

$$x|a \sim \mathcal{N}(x|\mu_a, \sigma_a^2) \quad (11)$$

$$y|a,x \sim Bernoulli(y|\mu_a, x) \quad (12)$$

$$(13)$$

操作変数 a に対しては、x 全体の上位 K[%] にたいして $a = 1$、それ以外を $a = 0$ とする操作アルゴリズム (以下、Top of K と呼ぶ) を用いる。図 1 は、操作前後においての $p(a,x)$ 分布の変化である。Top of K の適用後の y は得られないとする。

4.2 結果

設定した人工モデルに対して、データサイズを変更しながら、モンテカルロにより 1000 回の試行によって目的関数の期待値を評価した。共変量シフトを用いたアプローチとの比較手法としては、$p(y|x,a)$ の保存則を用いて、捜査前のデータより、ロジスティック回帰、SVM(RBF), Deep(4 層) により推定した。

結果は、2 で与えられている。黒が操作前、赤が操作後の期待値である。青が共変量シフトを用いた期待値評価であり、データサイズ 100 で Top of K 適用後の期待値評価が出来ている。一方、ロジスティック (緑)、SVM(茶), Deep(橙) は過大、過小に期待値を見積もる傾向にあることが確かめられた。

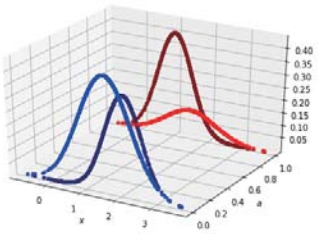

図 1: Top of K による分布の変化。$a = 0$ の x の分布は左側へシフトし、$a = 1$ の x の分布は右側へシフトしている。

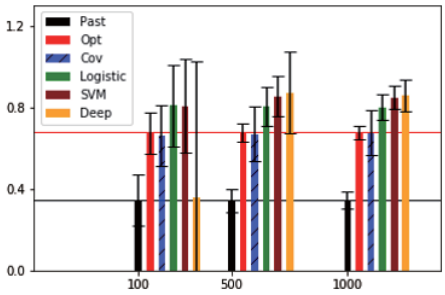

図 2: サンプル数を 100, 500, 1000 と変化させた場合での共変量シフトとロジスティック、SVM, Deep との比較。赤が正しい結果であり、共変量シフトが最もバイアスが補正できている結果である。

4.3 考察

直接 y を推定することは、分布の非定常性に加えデータの偏りなどで生じる外挿問題にダイレクトに挑戦する難しい問題である。一方、密度比を用いた期待値評価は、観測されたデータの内挿問題であり、分布の詳細な情報をあきらめる代わりに予測の難易度は下がる。マクロな量である期待値の評価を行う場合は、分布の普遍的な部分やドメイン知識で補える部分できりわけるアプローチは、既存のデータを活用する上で有効な手法であることが経験的に確かめられた。

5. おわりに

本研究では、密度比を用いたデータの活用が、将来の期待値評価に有効であることを示した。今後は、本研究の一般的な定式化を含め、逐次的な操作を含む期待値評価や、目的関数の推定を含む問題へ拡張する予定である。

参考文献

[Shimodaira 00] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” J. Stat. Plan. Inference, vol. 90, no. 2, pp. 227244, 2000.

[Sugiyama 07] M. Sugiyama, M. Krauledat, and K.-R. Mller, “Covariate shift adaptation by importance weighted cross validation,” J. Mach. Learn. Res., vol. 8, pp. 9851005, 2007.