推定された密度比を用いる
重要度重み付けによる共変量シフト適応の再考

加藤真大 [1]
松井孝太 [2]
井口亮 [1]

1：みずほ第一FT
2：名古屋大学

みずほ第一フィナンシャルテクノロジー

Double Debiased Covariate Shift Adaptation Robust to Density-Ratio Estimation
arXiv:2310.16638. URL: https://arxiv.org/abs/2310.16638.

重要度重み付けによる共変量シフト適応と密度比推定によって生じる問題

- 観察データ（ラベルなしテストデータを観測できる半教師あり学習設定）.
  - 訓練データ: $\{(X_i, Y_i)\}_{i=1}^n$. $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R} \sim P_0$.
  - テストデータ: $\{\tilde{X}_j\}_{j=1}^m$. $(\tilde{X}_j, \tilde{Y}_j) \in \mathbb{R}^d \times \mathbb{R} \sim Q_0$
    - $\tilde{X}_j$に対応するラベル$\tilde{Y}_j$ が未観測.
- 目標：テスト分布Q_0のもとでの期待リスクを最小化する予測器の学習.
  - 予測器: $f: \mathbb{R}^d \to \mathbb{R}$.
- テスト分布Q_0のもとでの期待リスク: $R(f) = \mathbb{E}_{Q_0}\left[\ell\left(\tilde{X}_j, \tilde{Y}_j\right)\right]$.

仮定：共変量シフト

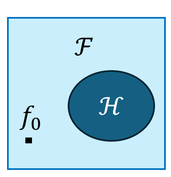
- P_0とQ_0がともに密度$p_0(x, y)$と$q_0(\tilde{x}, \tilde{y})$を持つとする.
- Y_iと$\tilde{Y}_j$条件付き確率は同じで，X_iと$\tilde{X}_j$の周辺密度だけ異なる：

$$p_0(x, y) = p_0(x)\zeta_0(y \mid x), \quad q_0(x, y) = q_0(x)\zeta_0(y \mid x).$$

仮定：共通サポート

- ある定数$C > 0$が存在し，すべてのxに対して，
  もし$q_0(x) > 0$であれば，$p_0(x) > 0$が成立.

➢ 訓練データとテストデータの分布が異なる場合の学習：



- モデル誤特定のもとで学習が頑健ではない.
- 可測関数全体$\mathcal{F}$と仮説集合$\mathcal{H} \subset \mathcal{F}$.
- 最適予測器$f_0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P_0}[\ell(X_i, Y_i)] = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{Q_0}\left[\ell\left(\tilde{X}_j, \tilde{Y}_j\right)\right]$.
- もし$f_0 \notin \mathcal{H}$であれば，最適予測器が訓練・テストで異なる可能性：

$$f^* = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{P_0}[\ell(X_i, Y_i)] \neq \tilde{f}^* = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{Q_0}\left[\ell\left(\tilde{X}_j, \tilde{Y}_j\right)\right]$$

- 重要度重み付けによる共変量シフト適応
  - 密度比$r_0(x) = \frac{p_0(x)}{q_0(x)}$.
  - 重要度重み付けのもとでの経験リスク: $\tilde{R}(f) = \frac{1}{n}\sum_{i=1}^n \ell\left(X_i, \tilde{Y}_i\right) r_0(X_i)$.

密度比$r_0(X_i)$が未知の場合；推定量$\hat{r}(x)$で置き換える.
（推定量の構成方法はSugiyama et al, (2012)など）

- 推定された密度比を用いる重要度重み付けのもとでの経験リスク:

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^n \ell\left(X_i, \tilde{Y}_i\right) \hat{r}(X_i).$$

- **リスクの近似誤差**: $R(f) - \hat{R}(f) = R(f) - \tilde{R}(f) + \tilde{R}(f) - \hat{R}(f)$
  - オラクル近似誤差項$R(f) - \tilde{R}(f)$: 通常の近似誤差解析（r_0既知）.
  - 密度比推定誤差項$\tilde{R}(f) - \hat{R}(f)$: 密度比の推定誤差.
- 回帰の設定（二乗損失 + 回帰関数$f_0(X) = \mathbb{E}[Y \mid X]$の推定・予測）

$$R(f) - \hat{R}(f) = R(f) - \tilde{R}(f) + \tilde{R}(f) - \hat{R}(f)$$

<u>オラクル近似誤差項$R(f) - \tilde{R}(f)$:</u>

- パラメトリックなら$O_p(1/\sqrt{n})$.
- ノンパラメトリックなら$O_p\left(n^{\frac{\beta}{2\beta+d}}\right)$.

βは回帰関数f_0の滑らかさ.

<u>密度比推定誤差項$R(f) - \tilde{R}(f)$:</u>

- パラメトリックなら$O_p(1/\sqrt{n})$.
- ノンパラメトリックなら $O_p\left(n^{\frac{\gamma}{2\gamma+d}}\right)$

γは密度比関数r_0の滑らかさ.

真の密度比を知っている場合のオラクルな近似誤差 $R(f) - \tilde{R}(f)$ と比較して，
推定誤差項$\tilde{R}(f) - \hat{R}(f)$を無視できない（ノンパラの場合，$O_p\left(n^{\frac{\beta}{2\beta+d}}\right)$と$O_p\left(n^{\frac{\beta}{2\beta+d}} + n^{\frac{\gamma}{2\gamma+d}}\right)$）.
→ <u>密度比の知識に応じて，達成できる収束レートが異なる.</u>

提案法：二重に頑健な推定量

- 簡単化のために線形モデルを仮定する：$f_0(x) = x^\top \theta_0$.
  - 真の密度比r_0を知っている場合のθ_0の推定量の漸近分布と，
  - 真の密度比r_0を知らない場合のθ_0の推定量の<u>漸近分布</u>を調べる.
- 真の密度比のもとでの推定量：$\tilde{\theta} \coloneqq \left(\frac{1}{m}\sum_{j=1}^m \tilde{X}_j \tilde{X}_j^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i Y_i r_0(X_i)$.
  - 漸近分布：$\sqrt{n}\left(\tilde{\theta} - \theta_0\right) \to^d \mathcal{N}(0, \Omega)$.
    - $\Omega = \mathbb{E}_{Q_0}\left[\tilde{X}_j \tilde{X}_j^\top\right]^{-1} \mathbb{E}_{Q_0}\left[\sigma^2\left(\tilde{X}_j\right) r_0\left(\tilde{X}_j\right) \tilde{X}_j \tilde{X}_j^\top\right] \mathbb{E}_{Q_0}\left[\tilde{X}_j \tilde{X}_j^\top\right]^{-1}$
    - $\sigma^2\left(\tilde{X}_j\right) = \mathbb{E}\left[\left(Y - f_0(X)\right)^2 \,\middle|\, X\right]$.
- 素朴に$\tilde{\theta}$内の真の密度比r_0を推定量$\hat{r}$に置き換える.
  → 誤差を無視できず，同じ漸近分布を得られない.
- **提案推定量**：二重に頑健な推定量：

$$\hat{\theta} \coloneqq \left(\frac{1}{m}\sum_{j=1}^m \tilde{X}_j \tilde{X}_j^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n X_i \left(Y_i - \hat{f}(X_i)\right) \hat{r}(X_i) + \frac{1}{m}\sum_{j=1}^m \tilde{X}_j \hat{f}\left(\tilde{X}_j\right)\right).$$

  - $\hat{f}$は回帰関数f_0の一致推定量. $\hat{r}$は密度比r_0の一致推定量.
- この推定量は，密度比既知の場合の推定量と同じ漸近分布を持つ.

条件：局外母数の推定量

- 収束レート：
  $\hat{f}$と$\hat{r}$がそれぞれ一致推定量 and $\mathbb{E}\left[\left\|\hat{f} - f_0\right\|_2 \|\hat{r} - r_0\|\right] = o_p\left(\frac{1}{\sqrt{n}}\right)$.
- 推定量の構築方法：
  Donsker条件を満たしている or サンプル分割で構築されている.

二重に頑健な推定量$\hat{\theta}$の漸近分布：

- 局外母数の推定量が上記の条件を満たしているとき，以下が成立：

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to^d \mathcal{N}(0, \Omega). \quad (n \to \infty).$$

サンプル分割（交差適合）

- データセットを複数に分割していくつかのデータセットを作る.
- 推定量内の$X_i\left(Y_i - \hat{f}(X_i)\right)\hat{r}(X_i)$を計算する際に，$(X_i, Y_i)$を$\hat{f}(X_i)$や$\hat{r}(X_i)$に用いない ⟹ (X_i, Y_i)を含まないデータセットで$\hat{f}(X_i)$と$\hat{r}(X_i)$を推定.
- $\tilde{X}_j \hat{f}\left(\tilde{X}_j\right)$についても同様にする.

二重頑健性

- $\hat{f}$もしくは$\hat{r}$のどちらかが一致推定量であるならば，$\hat{\theta}$はθ_0の一致推定量.
  - $\hat{f}$はパラメトリックでもノンパラメトリックでも良い.

漸近有効性

- 提案推定量$\hat{\theta}$は漸近的に有効<u>ではない</u>.
  - 漸近分散Ωは（セミパラメトリック効率）下限より大きい.
- モデルを正しく特定していれば，密度比を使わない推定量の方が良い.

線形モデルからの拡張

- 論文中では単純な線形モデルではなく以下の拡張も提案：
  - 基底を入れた線形モデル.
  - 適当な条件を満たすパラメトリックモデル$f_0 = f_{\theta_0}$.

進行中の研究

- ノンパラメトリック回帰への拡張を研究中.
  - 真の回帰関数の推定誤差のミニマックス最適性で評価.
  - 密度比の滑らかさに依存するミニマックス下限を導出できると予想.

参考文献

- Shimodaira, H. (2000), ‘Improving predictive inference under covariate shift by weighting the log-likelihood function’, JSPI.
- Sugiyama, M., Suzuki, T. & Kanamori, T. (2012), Density Ratio Estimation in Machine Learning, 1st edn, Cambridge University Press.
- Uehara, M., Kato, M. & Yasui, S. (2020), “Off-policy evaluation and learning for external validity under a covariate shift” in NeurIPS.
- Kato, M. & Teshima, T. (2020), ‘Non-negative bregman divergence minimization for deep direct density ratio estimation’ in ICML