Label Shift Quantification

with Robustness Guarantees via Distribution Feature Matching

Dussap Bastien[†], Gilles Blanchard[†], Badr-Eddine Chérief-Abdellatif[§]

[†]Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay
[§]CNRS, LPSM, Sorbonne Université, Université Paris Cité

July 3, 2023

Introduction

Label Shift

Model

- $\mathcal{X}$: the data space.
- $\mathcal{Y}$: the label space, $\{1, \cdots, c\}$.
- $\mathbb{P}_1, \cdots, \mathbb{P}_c$: A list of c distributions, one for each class.
- $\mathbb{P}_i = p(X|Y = i)$, conditional distribution.

Label Shift

A "source" distribution.

$$\mathbb{P} = \sum_{i=1}^{c} \beta_i \mathbb{P}_i$$

A "target" distribution.

$$\mathbb{Q} = \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i$$

Training data.

$$\{(x_j, y_j)\}_{j \in [n]} \in (\mathcal{X} \times \mathcal{Y})^n$$

Testing data.

$$\{x_{n+j}\}_{j \in [m]} \in \mathcal{X}^m$$

Notations

Notations

- $\hat{\mathbb{P}}_i := \frac{1}{n_i} \sum_{j \in [n]: y_j = i} \delta_{x_j}(\cdot)$.
- $\hat{\mathbb{Q}} := \frac{1}{m} \sum_{j=1}^{m} \delta_{x_{n+j}}(\cdot)$.
- $\tilde{\beta}_i := \frac{n_i}{n}$.
- $\tilde{\alpha}_i := \frac{m_i}{m} \neq \alpha_i^*$.

Quantification

Quantification

- Using $\hat{\mathbb{P}}_i$ and $\hat{\mathbb{Q}}$, estimate α^*.
- Using $\hat{\mathbb{P}}_i$ and $\hat{\mathbb{Q}}$, estimate $\tilde{\alpha}$.

Distribution Feature Matching

A general Embedding

$\rightarrow$ Let $\phi : \mathcal{X} \rightarrow \mathcal{F}$ be a fixed feature mapping from $\mathcal{X}$ into a Hilbert space $\mathcal{F}$ (possibly $\mathcal{F} = \mathbb{R}^D$).

A general Embedding

→ Let $\phi : \mathcal{X} \to \mathcal{F}$ be a fixed feature mapping from $\mathcal{X}$ into a Hilbert space $\mathcal{F}$ (possibly $\mathcal{F} = \mathbb{R}^D$).

Embedding

$$\begin{aligned} \phi \colon \mathcal{M}_1^+(\mathcal{X}) &\to \mathcal{F} \\ \mathbb{P} &\mapsto \mathbb{E}_{X \sim \mathbb{P}}[\phi(X)] = \phi(\mathbb{P}) \end{aligned}$$

Pseudo-Distance

Examples

- $\phi(x) = (1\{\hat{f}(x) = i\})_{i=1,\dots,c} \in \mathbb{R}^c$.
- $\phi(x) = (\hat{f}(i))_{i=1,\dots,c} \in \mathbb{R}^c$.
- $\phi(x)$ of a neural networks : $\hat{f}(x) = \sigma(w^T \phi(x) + b) \in \mathbb{R}^c$
- $\phi(x) = (1\{x \in C_i\})_i \in \mathbb{R}^M$. Histogram.
- $\phi(x) = (y \mapsto k(x, y)) \in \mathcal{H}_k$

Pseudo-Distance

Examples

- $\phi(x) = (1\{\hat{f}(x) = i\})_{i=1,\dots,c} \in \mathbb{R}^c$.
- $\phi(x) = (\hat{f}(i))_{i=1,\dots,c} \in \mathbb{R}^c$.
- $\phi(x)$ of a neural networks : $\hat{f}(x) = \sigma(w^T\phi(x) + b) \in \mathbb{R}^c$
- $\phi(x) = (1\{x \in C_i\})_i \in \mathbb{R}^M$. Histogram.
- $\phi(x) = (y \mapsto k(x,y)) \in \mathcal{H}_k$

Pseudo-Distance

$$D_\phi(\mathbb{P}, \mathbb{Q}) = \|\phi(\mathbb{P}) - \phi(\mathbb{Q})\|$$

Distribution Feature Matching

Idea

$$\alpha^* \in \operatorname*{arg\,min}_{\alpha \in \Delta^c} D_\phi\left(\sum_{i=1}^{c} \alpha_i \mathbb{P}_i, \mathbb{Q}\right)$$
$$= \operatorname*{arg\,min}_{\alpha \in \Delta^c} D_\phi\left(\sum_{i=1}^{c} \alpha_i \mathbb{P}_i, \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i\right)$$

where $\Delta^c := \{x \in \mathbb{R}^c \colon x \geq 0,\ \sum x_i = 1\}$.

Distribution Feature Matching

Distribution Feature Matching (DFM)

$$\hat{\alpha} = \underset{\alpha \in \Delta^c}{\arg\min}\; D_\phi\Big(\sum_{i=1}^{c} \alpha_i \hat{\mathbb{P}}_i, \hat{\mathbb{Q}}\Big) = \underset{\alpha \in \Delta^c}{\arg\min} \Big\| \sum_{i=1}^{c} \alpha_i \phi(\hat{\mathbb{P}}_i) - \phi(\hat{\mathbb{Q}}) \Big\| \qquad (\mathcal{P})$$

Related Literature

Lipton, Zachary, Yu-Xiang Wang, and Alexander Smola. *"Detecting and correcting for label shift with black box predictor"*. In *International Conference on Machine Learning*, pages 3122–3130, 2018. Published by PMLR.

Saerens, Marco, Patrice Latinne, and Christine Decaestecker. *"Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure"*. In *Neural Computation*, 14(1):21–41, 2002. Published by MIT Press.

Iyer, Arun, Saketha Nath, and Sunita Sarawagi. *"Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection"*. In *International Conference on Machine Learning*, pages 530–538, 2014. Published by PMLR.

Kawakubo, Hideko, Marthinus Christoffel du Plessis, and Masashi Sugiyama. *"Computationally efficient class-prior estimation under class balance change using energy distance"*. In *IEICE Transactions on Information and Systems*, 99(1):176–186, 2016. Published by The Institute of Electronics, Information and Communication Engineers.

Theoretical guarantees

Assumption

$$\sum_{i=1}^{c} \beta_i \phi(\mathbb{P}_i) = 0 \iff \beta = 0 \tag{$\mathcal{A}_1$}$$

and

$$\exists C > 0 : \; \|\phi(x)\| \leq C \text{ for all } x. \tag{$\mathcal{A}_2$}$$

Main Theorem

Theorem

For any $\delta \in (0,1)$, with probability greater than $1-\delta$:

$$\begin{aligned} \|\hat{\alpha}-\alpha^*\|_2 &\leq \frac{2CR_{c/\delta}}{\sqrt{\Delta_{\min}}}\left(\frac{\|w\|_2}{\sqrt{n}}+\frac{1}{\sqrt{m}}\right) \\ &\leq \frac{2CR_{c/\delta}}{\sqrt{\Delta_{\min}}}\left(\frac{1}{\sqrt{\min_i n_i}}+\frac{1}{\sqrt{m}}\right), \end{aligned}$$

where $R_x = 2+\sqrt{2\log(2x)}$, $w_i = \frac{\alpha_i^*}{\tilde{\beta}_i}$, and $\Delta_{\min} := \Delta_{\min}(\phi(\hat{\mathbb{P}}_1), \cdots, \phi(\hat{\mathbb{P}}_c))$.
The same result holds when replacing α^* by the (unobserved) vector of empirical proportions $\tilde{\alpha}$ in the target sample, both on the left-hand side and in the definition of w.

Properties of $\Delta_{\min}$

Let $(b_1, \cdots, b_c)$ be a c-uple of vectors of $\mathbb{R}^D$ assumed to be linearly independant. We denote M the Gram matrix of those vectors, i.e. $M_{ij} = \langle b_i, b_j \rangle$.

Theorem

For any number of classes c, Δ_{min} is equal to $\min\limits_{\substack{\|u\|=1 \\ \mathbf{1}^T u = 0}} u^T M u$.

$\rightarrow$ $\Delta_{\min}(b_1, \cdots, b_c)$ is always greater than the smallest eigenvalue of the Gram matrix.

Properties of $\Delta_{\mathbf{min}}$

Theorem

In particular for two classes, $\Delta_{min}(b_1, b_2) = \frac{1}{2}\|b_1 - b_2\|^2$.

Properties of Δ_{min}

Theorem

In particular for two classes, $\Delta_{min}(b_1, b_2) = \frac{1}{2}\|b_1 - b_2\|^2$.

Theorem

For more than two classes,

$$\Delta_{min}(b_1, \cdots, b_c) \leq (1 - \frac{1}{c}) \inf_i \|b_i - P_{C_{-i}}(b_i)\|^2.$$

Conclusion

- We introduced a general approach for Label Shift Quantification.
- We provided a general theoretical analysis of DFM, improving over previously known bounds derived for specific instantiations only.

Thank you for your attention.

References

Dussap, Bastien, Gilles Blanchard, and Badr-Eddine Chérief-Abdellatif (2023). "Label Shift Quantification with Robustness Guarantees via Distribution Feature Matching". In: *arXiv preprint arXiv:2306.04376*.