Label Shift Quantification

with Robustness Guarantees via Distribution Feature Matching

Dussap Bastien

Laboratoire de mathématiques d'Orsay
Université Paris-Saclay, Inria

June 22, 2023

Overview

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.

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.

A "source" distribution.

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

A "target" distribution.

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

Datasets

Datasets

- $\{(x_j, y_j)\}_{j\in[n]} \in (\mathcal{X} \times \mathcal{Y})^n$, a labelled dataset : "source".
- $\{x_{n+j}\}_{j\in[m]} \in \mathcal{X}^m$, an unlabelled dataset : "target".
- $\hat{\mathbb{P}}_i := \frac{1}{n_i} \sum_{j\in[n]:y_j=i} \delta_{x_j}(\cdot)$.
- $\hat{\mathbb{Q}} := \frac{1}{m} \sum_{j=1}^{m} \delta_{x_{n+j}}(\cdot)$.
- $\tilde{\beta}_i := \frac{n_i}{n}$.
- $\tilde{\alpha}_i := \frac{m_i}{m} \neq \alpha_i^*$.

Quantification

Quantification

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

Classify and Count

→ Use a classifier f.

Classifier

Count the number of times your classifier outputs each class.

Classify and Count (CC)

$$\hat{\alpha}_{cc} = \left(\frac{1}{m} \sum_{j=1}^{m} 1_{f(x_{n+j})=i} \right)_i$$

$$\alpha_{cc} = q(f(x) = i)_i$$

Black-Box Shift Estimator

$\rightarrow$ Use the confusion matrix of a classifier.

$$
\begin{aligned}
q(f(x) = i) &= \sum_{j=1}^{c} q(f(x) = i | y = j) q(y = j) \\
&= \sum_{j=1}^{c} p(f(x) = i | y = j) q(y = j) \\
\alpha_{cc} &= M_f \times \alpha^*
\end{aligned}
$$

Black-Box Shift Estimator

$$\hat{\alpha} = \hat{M_f}^{-1} \hat{\alpha}_{cc}$$

Black-Box Shift Estimator

Black-Box Shift Estimator

$$\hat{\alpha} = \hat{M}_f^{-1} \hat{\alpha}_{cc}$$

Alternative version: BBSE+

$$\hat{\alpha} = \underset{\alpha \in \Delta^c}{\arg\min} \; \|\alpha \hat{M}_f - \hat{\alpha}_{cc}\|_2$$

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

Kernel Mean Embedding

Kernel Methods

For a kernel k there exists a mapping $\Phi\colon \mathcal{X} \mapsto \mathcal{H}$ such as:

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$$

Kernel Mean Embedding

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

Kernel

Classical kernel

- $k(x, y) = x^T y$, linear.
- $k(x, y) = (\gamma x^T y + c_0)^d$, polynomial.
- $k(x, y) = \tanh(\gamma x^T y + c_0)$, sigmoid.
- $k(x, y) = \exp(\gamma \|x - y\|^2)$, gaussian.
- $k(x, y) = \exp(-\gamma \|x - y\|_1)$, laplacian.
- $k(x, y) = \|x\| + \|y\| - \|x - y\|$, energy.
- $k(x, y) = \left(1 + \frac{\|x-y\|^2}{\sigma^2}\right)^{-1}$, cauchy.

Maximum Mean Discrepancy

Maximum Mean Discrepancy

$$\begin{aligned}\mathrm{MMD}(\mathbb{P},\mathbb{Q}) &= \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|_{\mathcal{H}} \\ &= \mathbb{E}_{\mathbb{P},\mathbb{P}}[k(X,X)] + \mathbb{E}_{\mathbb{Q},\mathbb{Q}}[k(Y,Y)] - 2\mathbb{E}_{\mathbb{P},\mathbb{Q}}[k(X,Y)]\end{aligned}$$

Kernel Mean Matching

$$\arg\min_{\alpha \in \Delta^c} \mathrm{MMD}\left(\sum_{i=1}^{c} \alpha_i \mathbb{P}_i, \mathbb{Q}\right) = \arg\min_{\alpha \in \Delta^c} \mathrm{MMD}\left(\sum_{i=1}^{c} \alpha_i \mathbb{P}_i, \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i\right)$$
$$= \alpha^*$$

Kernel Mean Matching

$$\hat{\alpha} = \arg\min_{\alpha \in \Delta^c} \mathrm{MMD}\left(\sum_{i=1}^{c} \alpha_i \hat{\mathbb{P}}_i, \hat{\mathbb{Q}}\right)$$
$$= \arg\min_{\alpha \in \Delta^c} \|\sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}})\|_{\mathcal{H}}$$

A general Approach

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

Embedding

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

Distribution Feature Matching

Distribution Feature Matching (DFM)

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

BBSE+

$$\rightarrow \phi(x) = (1\{f(x) = i\})_{i=1,\ldots,c} \in \mathbb{R}^c$$

BBSE+

$$\rightarrow \phi(x) = (1\{f(x) = i\})_{i=1,\ldots,c} \in \mathbb{R}^c$$

BBSE+ as DFM

- $\phi(\mathbb{Q})_i = q(f(x) = i) = \alpha_{cc}$
- $\phi(\mathbb{P}_j)_i = p(f(x) = i | y = j) = (M_f)_{i,j}$
- $\phi(\hat{\mathbb{P}}_j)_i = (\hat{M}_f)_i$
- $\phi(\hat{\mathbb{Q}})_i = \hat{\alpha}_{cc}$

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 $1-\delta$:

$$\|\hat{\alpha} - \alpha^*\|_2 \leq \frac{2CR_{(\delta+\log c)}}{\sqrt{\Delta_{\min}}}\left(\frac{\|w\|_2}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)$$

where $R_x = 2 + \sqrt{2\log(2/x)}$, $w_i = \frac{\alpha_i^*}{\tilde{\beta}_i}$, 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.

$\Delta_{\min}$

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 $\displaystyle\min_{\substack{\|u\|=1 \\ \mathbf{1}^T u=0}} u^T M u$.

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

Properties of $\Delta_{\min}$

Theorem

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

Robustness to contamination

Robustness to contamination

Contaminated Label Shift

$$\mathbb{Q} = \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i + \alpha_0^* \mathbb{Q}_0 \qquad (\mathcal{CLS})$$

Robustness to contamination

Contaminated Label Shift

$$\mathbb{Q} = \sum_{i=1}^{c} \alpha_i^* \mathbb{P}_i + \alpha_0^* \mathbb{Q}_0 \qquad (\mathcal{CLS})$$

Soft Distribution Feature Matching (soft-DFM)

$$\hat{\alpha}_{\mathrm{soft}} = \operatorname*{arg\,min}_{\alpha \in \mathrm{int}(\Delta^c)} \left\| \sum_{i=1}^{c} \alpha_i \phi(\hat{\mathbb{P}}_i) - \phi(\hat{\mathbb{Q}}) \right\|_{\mathcal{F}}^2 ,$$

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

Theorem for $\mathcal{CLS}$

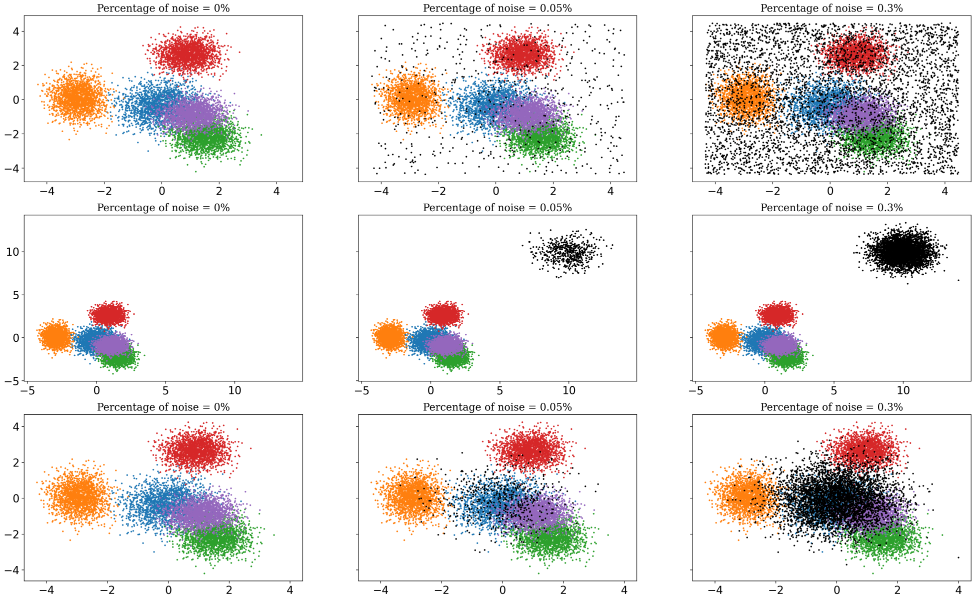
Introduce $\bar{V} := \mathrm{Span}\{\Phi(\mathbb{P}_i), i \in [c]\}$ and let $\Pi_{\bar{V}}$ be the orthogonal projection on $\bar{V}$.

Corollary

With probability greater than $1 - \delta$:

$$\|\hat{\alpha}_{\text{soft}} - \alpha^*\|_2 \leq \frac{1}{\sqrt{\Delta_{\min}}}\Big(3\epsilon_n + \varepsilon_m + \sqrt{2\alpha_0^* \, \epsilon_n \, \|\phi(\mathbb{Q}_0)\|} + \|\Pi_{\bar{V}}(\phi(\mathbb{Q}_0))\|_{\mathcal{F}}\Big),$$

$$\epsilon_n = C\frac{R_{\delta+\log c}}{\sqrt{\min_i n_i}}; \qquad \varepsilon_m = C\frac{R_\delta}{\sqrt{m}};$$


Percentage of noise = 0%
Percentage of noise = 0.05%
Percentage of noise = 0.3%
Percentage of noise = 0%
Percentage of noise = 0.05%
Percentage of noise = 0.3%
Percentage of noise = 0%
Percentage of noise = 0.05%
Percentage of noise = 0.3%

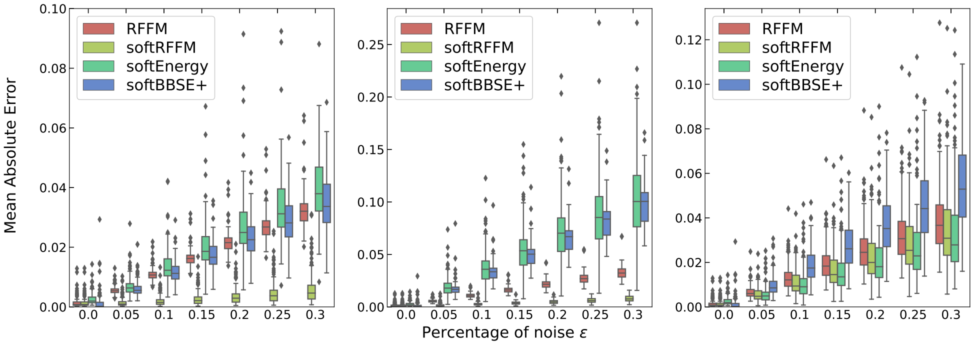


Figure: RFFM is Kernel Mean Matching with Random Fourier Features to speed up the computation.

Conclusion

- We introduced a general approach for Label Shift Quantification.
- We proposed to use Random Fourier Features to speed up the computation.
- We provided a general theoretical analysis of DFM, improving over previously known bounds derived for specific instantiations only.
- We analysed theoretically the behavior of DFM under a new setting and put into light a robustness against certain types of perturbations, depending on the feature mapping used.

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*.

González, Pablo et al. (2017). "A review on quantification learning". In: *ACM Computing Surveys (CSUR)* 50.5, pp. 1–40.

Gretton, Arthur et al. (2006). "A kernel method for the two-sample-problem". In: *Advances in neural information processing systems* 19.

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

Kawakubo, Hideko, Marthinus Christoffel Du Plessis, and Masashi Sugiyama (2016). "Computationally efficient class-prior estimation under class balance change using energy distance". In: *IEICE Transactions on Information and Systems* 99.1, pp. 176–186.

Lipton, Zachary, Yu-Xiang Wang, and Alexander Smola (2018). "Detecting and correcting for label shift with black box predictors". In: *International conference on machine learning*. PMLR, pp. 3122–3130.

Rahimi, Ali and Benjamin Recht (2007). "Random features for large-scale kernel machines". In: *Advances in neural information processing systems* 20.