Label Shift Quantification with Robustness Guarantees via Distribution Feature Matching

Bastien Dussap ✉, Gilles Blanchard, Badr-Eddine Chérief-Abdellatif
bastien.dussap@inria.fr, gilles.blanchard@univerite-paris-saclay.fr, badr-eddine.cherief-abdellatif@cnrs.fr

Label Shift Quantification

Consider a covariate space $\mathcal{X} \subset \mathbb{R}^d$, a label space $\mathcal{Y} := [c]$. Consider the Label Shift Hypothesis, where the test distribution $\mathbb{Q}$ verified:

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

With $\mathbb{P}_i = p(X|Y=i)$. We have access to samples: $\hat{\mathbb{P}}_1, \cdots, \hat{\mathbb{P}}_c$ and $\hat{\mathbb{Q}}$.
We also consider a new setting, Contaminated Label Shift defined as :

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

The distribution $\mathbb{Q}_0$ is seen as a contamination, for which we have no prior knowledge nor sample.

Goal : Estimate the proportions α^*. This is called Quantification [1].

Distribution Feature Matching

Let $\Phi : \mathcal{X} \to \mathcal{F}$ be a fixed feature map from $\mathcal{X}$ into a Hilbert space $\mathcal{F}$. We extend the mapping to probability distributions on $\mathcal{X}$:

$$\Phi \colon \mathbb{P} \mapsto \Phi(\mathbb{P}) := \mathbb{E}_{X\sim\mathbb{P}}[\Phi(X)] \in \mathcal{F}.$$

We call *Distribution Feature Matching* (DFM) any estimation procedure that can be formulated as the minimiser of the following problem:

$$\hat{\alpha} = \underset{\alpha\in\Delta^c}{\arg\min} \left\| \sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}}) \right\|_{\mathcal{F}}^2 \qquad (\mathcal{P})$$

$\Delta^c := \{x \in \mathbb{R}_+^c \colon \sum_{i=1}^{c} x_i = 1\}$ is the $(c-1)$ dimensional simplex.

Related literature

Kernel Mean Matching (KMM) [2]:

$$\Phi(x) = (y \mapsto k(x,y)) \in \mathcal{H}_k$$

Black-Box Shift Estimation (BBSE) [3]:

$$\Phi(x) = (1\{\hat{f}(x) = i\})_{i=1,\dots,c} \in \mathbb{R}^c$$

Definitions

$$\hat{\boldsymbol{G}}_{ij} = \langle \Phi(\hat{\mathbb{P}}_i), \Phi(\hat{\mathbb{P}}_j) \rangle$$
$$\hat{\boldsymbol{M}}_{ij} = \langle \Phi(\hat{\mathbb{P}}_i) - \overline{\Phi}, \Phi(\hat{\mathbb{P}}_j) - \overline{\Phi} \rangle$$

$\Delta_{\min}$ is the second smallest eivenvalue of $\hat{\boldsymbol{M}}$ and $\lambda_{\min}$ the smallest eigenvalue of $\hat{\boldsymbol{G}}$. In particular, it holds:

$$\Delta_{\min} \geq \lambda_{\min}.$$

References

[1] González, P., Castaño, A., Chawla, N. V., and Coz, J. J. D. A review on quantification learning. *ACM Computing Surveys (CSUR)* (2017).

[2] Iyer, A., Nath, S., and Sarawagi, S. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning* (2014).

[3] Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning* (2018).

Consistency of Distribution Feature Matching

We make the following identifiability hypothesis on the mapping Φ:

$$\sum_{i=1}^{c} \lambda_i \Phi(\mathbb{P}_i) = 0 \iff \lambda = 0 \qquad (\mathcal{A}_1)$$
$$\exists C > 0 : \ \|\Phi(x)\|_{\mathcal{F}} \leq C \text{ for all } x. \qquad (\mathcal{A}_2)$$

Theorem 1 *If the Label Shift hypothesis $(\mathcal{LS})$ holds, and if the mapping Φ verifies Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$, then for any $\delta \in (0,1)$, with probability greater than $1-\delta$, the solution $\hat{\alpha}$ of $(\mathcal{P})$ satisfies:*

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

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

- The bound (1) improves upon existing bounds in the literature ([2, 3]).
- The (empirical) quantity $\Delta_{\min}$ provides a natural criterion for the choice of the feature map hyperparameter.

Robustness to contamination

In the Contaminated Label Shift setting, we aim at finding the proportions of the non-noise classes of the target. As these proportions don't sum to one, the "hard" condition $\sum_i \alpha_i = 1$ is replaced by the "soft" condition $\sum_i \alpha_i \leq 1$.

$$\hat{\alpha}_{\text{soft}} = \underset{\alpha\in\text{int}(\Delta^c)}{\arg\min} \left\| \sum_{i=1}^{c} \alpha_i \Phi(\hat{\mathbb{P}}_i) - \Phi(\hat{\mathbb{Q}}) \right\|_{\mathcal{F}}^2, \qquad (\mathcal{P}_2)$$

If $\alpha_0^* = 0$, then $\|\hat{\alpha}_{\text{soft}} - \alpha^*\|_2$ is bounded by (1) and (2) with $\Delta_{\min}$ replaced by $\lambda_{\min}$.

Theorem 2 *Introduce $\bar{V} := \text{Span}\{\Phi(\mathbb{P}_i), i \in [c]\}$ and let $\Pi_{\bar{V}}$ be the orthogonal projection on $\bar{V}$. If the Contaminated Label Shift hypothesis $(\mathcal{CLS})$ holds, and if the mapping Φ verifies Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$. Then, with probability greater than $1-\delta$:*

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

with:

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

- Bound (3) shows the robustness of DFM against perturbations $\mathbb{Q}_0$ that are orthogonal to $\bar{V}$.
- For BBSE, the feature space is of the same dimension as the number of sources hence the orthogonal component will always be 0 and we expect no robustness property for BBSE.
- For KMM with a Gaussian kernel: $\Phi(\mathbb{P})$ and $\Phi(\mathbb{P}')$ will be close to orthogonal if $\mathbb{P}$ and $\mathbb{P}'$ are well-separated. We expect robustness property for KMM if the main mass of $\mathbb{Q}_0$ is far away from the source distributions.

Experiments

The source is a list of c Gaussian distributions. α_0^* ranges from from 0 to 0.3. We will test three kinds of noise $\mathbb{Q}_0$: uniform distribution over the data range, a new Gaussian with a mean distant from the other means and a new Gaussian with a similar mean to the source.

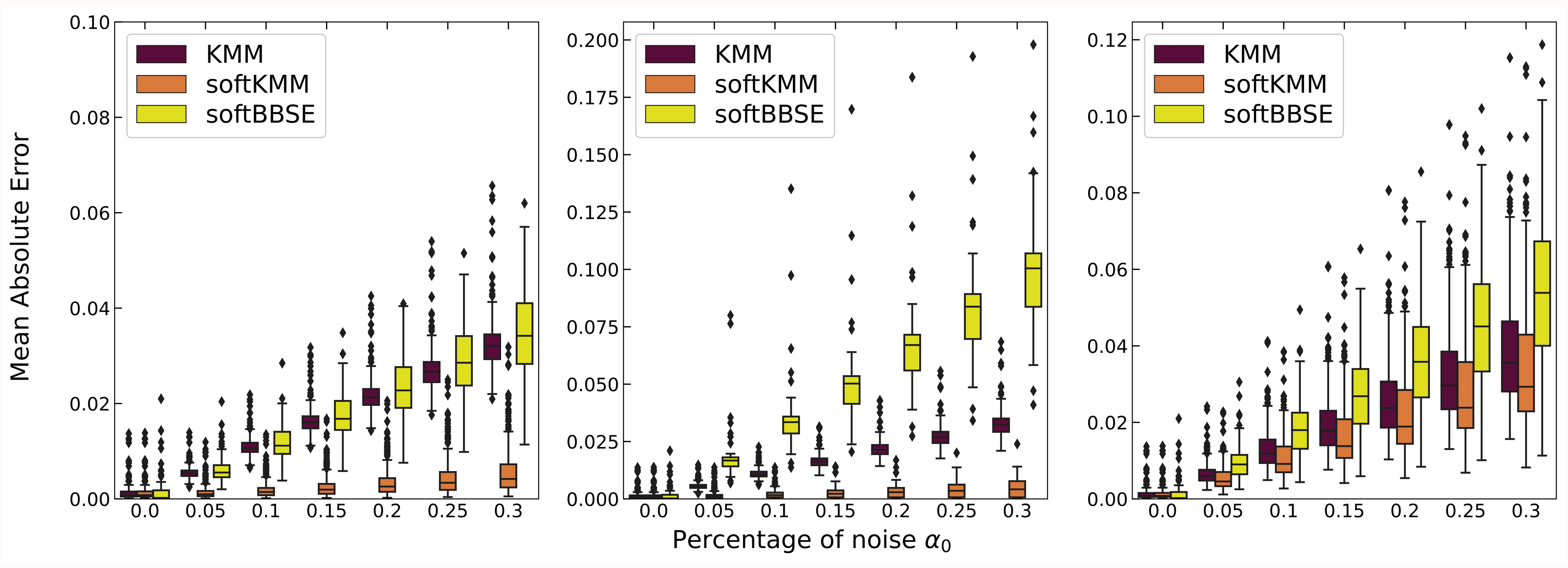

Figure 1: Robustness of the algorithms to three types of noise. Left: background noise; middle: noise is a new class far from the others; right: noise is a new class in the middle of the others.