To trust or not to trust: Assessing calibration error under covariate shift without labels

Teodora Popordanoska
ESAT-PSI, KU Leuven
teodora.popordanoska@kuleuven.be

Aleksei Tiulpin
HST Research Unit, University of Oulu
aleksei.tiulpin@oulu.fi

Matthew B. Blaschko
ESAT-PSI, KU Leuven
matthew.blaschko@esat.kuleuven.be

Abstract

In this work, we study the problem of estimating calibration error (CE) when facing covariate shift between the train (source) and test (target) distributions, while also not having access to labels from the target domain. We adapt an existing differentiable and consistent kernel density estimator of calibration error to account for the change in the marginal distribution. Our preliminary experiments on MNIST and SVHN datasets highlight the potential of the proposed method to address this challenge and successfully estimate CE.

1. Introduction

Reliable uncertainty estimation plays a pivotal role in predictive models, especially in safety-critical applications, where decisions based on neural network predictions can have significant consequences [5]. The *calibration error* (CE) [3, 12, 13] of a model provides insights into the reliability and trustworthiness of its predictions. Informally, CE is a measure of discrepancy between predicted probabilities and empirically observed class frequencies.

Many existing approaches for estimating CE and mitigating miscalibration rely on the availability of i.i.d. and labeled hold-out data. However, these assumptions are frequently violated in practical settings. On one hand, the training distribution may differ from the testing distribution, even though the conditional label distribution remains the same - a phenomenon known as *covariate shift* [11]. On the other hand, it is often the case that labels for the target data are not available or too costly to obtain.

In this work, we tackle the challenge of measuring calibration in the presence of covariate shift between the train and test distribution, while also assuming no labels for the target data. Previous works [9, 8] have studied this problem by building on ideas from domain adaptation and using importance weighting, whereas we derive a CE estimator directly from the properties of the covariate shift assumption. Empirically, we show the promising capability of our approach to effectively assess calibration in this context.

2. Calibration error under covariate shift

We consider a supervised classification setting under covariate shift [11]. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, ..., K\}$ be the set of labels. Let $f : \mathcal{X} \to \triangle^K$ be a classification model (e.g. a neural network) where $\triangle^K$ denotes the $K-1$ dimensional simplex. We denote the train (source) distribution as $\pi_{tr}(x, y) = \pi_{tr}(y \mid x)\pi_{tr}(x)$ and $\pi_{te}(x, y)$ as the test (target) distribution and assume covariate shift, i.e. $\pi_{tr}(x) \neq \pi_{te}(x)$ and $\pi_{tr}(y \mid x) = \pi_{te}(y \mid x)$.

A standard assumption in this setting is that $\pi_{te}(x, y)$ is absolutely continuous w.r.t. $\pi_{tr}(x, y)$, i.e., $\pi_{tr}(x, y) = 0$ implies $\pi_{te}(x, y) = 0$ [2]. In other words, the support of the target distribution is contained within the support of the source distribution. Our goal is to *assess the CE of the target distribition*, given labeled examples from the source and unlabeled examples from the target distribution. Assuming that $\pi_{te}(y \mid x) = \pi_{tr}(y \mid x)$ and f is constant, the L_p CE of f under covariate shift becomes:

$$\mathrm{CE}_p(f) = \left(\mathbb{E}_{\pi_{te}} \left[\left\| \mathbb{E}_{\pi_{tr}}[y \mid f(x)] - f(x) \right\|_p^p \right] \right)^{\frac{1}{p}}. \quad (1)$$

To estimate this quantity using a finite sample, we modify a recently proposed consistent and differentiable kernel density estimator (KDE) [10, Equation (9)]. Considering n labeled train examples and m unlabeled test examples, our new estimator is given by:

$$\widehat{\mathrm{CE}_p(f)}^p = \frac{1}{m} \sum_{j=n+1}^{m+n} \left\| \frac{\sum_{i=1}^{n} k(f(x_j); f(x_i)) y_i}{\sum_{i=1}^{n} k(f(x_j); f(x_i))} - f(x_j) \right\|_p^p, \quad (2)$$

Table 1. Performance of different model architectures trained on source (SVHN) dataset and evaluated using target test set (MNIST). We report mean and standard error across 5 different seeds. Target-gt is obtained using the labels of the target test set.

Dataset	Model	Acc (source)	Acc (target)	CE (source)	CE (target)	CE (target-gt)
SVHN → MNIST	LeNet	$88.63_{\pm 0.08}$	$57.54_{\pm 1.03}$	$0.23_{\pm 0.0021}$	$0.20_{\pm 0.0044}$	$0.69_{\pm 0.0211}$
	PreResNet20	$93.46_{\pm 0.11}$	$54.48_{\pm 3.51}$	$0.12_{\pm 0.0023}$	$0.44_{\pm 0.0300}$	$0.74_{\pm 0.0604}$
	PreResNet56	$93.98_{\pm 0.13}$	$59.35_{\pm 2.58}$	$0.13_{\pm 0.0032}$	$0.49_{\pm 0.0057}$	$0.79_{\pm 0.0452}$

where k is a Dirichlet kernel, as defined in [10, Equation (10)]. We select the bandwidth that maximizes the likelihood of the target points under the density estimated from the source points.

3. Experiments

Datasets We evaluate our method under covariate shift between two digit datasets: SVHN [7] and MNIST [1]. The **SVHN** dataset contains color images of size 32×32, divided into 73257 training and 26032 test examples. The **MNIST** consists of hand-written, 28×28 grayscale images, split into 50000 training examples and 10000 test examples. For both datasets, we further partition the training examples to create a separate validation set of 10000 points, which we use for model and hyperparameter selection. To ensure compatibility between MNIST and SVHN images, we resize SVHN to 28×28 pixels and convert them to grayscale.

Setup The classifiers are trained using only the labeled source data. We report L_1 canonical calibration error on the source and target domain, using the estimator in Equation (2). We compare to the calibration error obtained on the target data using labels from the target domain.

We consider two sets of experiments – one on simulated data, and one on covariate shift in real data. For the first case, we divide the MNIST test set in two equal parts. The first half represents the i.i.d. source test set and the second half corresponds to the target test set, which is rotated with various degrees. For the second case, we train models (LeNet [6], PreResNet20, PreResNet56 [4]) on SVNH, and evaluate CE on MNIST.

Results In Figure 1 we show the changes in CE with the ground labels and using our method on a rotated MNIST dataset. The performance of the model degrades with the increase of the rotation angle, both in terms of accuracy and CE. We observe that the estimator effectively captures canonical calibration error without using labels on the test set up to 30° of rotation.

In Table 1 we evaluate accuracy and CE on the test set from the source (SVNH) and target distribution (MNIST), using a neural network trained on the source dataset. The covariate shift is clearly seen through the change in accuracy and CE. The quantitative results suggest that substantial changes in CE can be observed without having any labels from the target distribution.

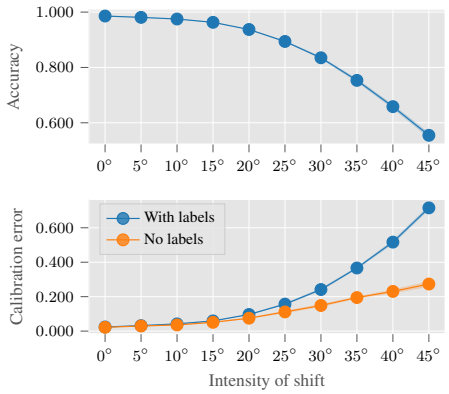

Figure 1. Accuracy and calibration error of LeNet architecture trained on MNIST as a function of dataset shift. The test set is split into source and target, with rotations applied to the target images at various degrees. The results are computed on target test set and the shaded regions represent standard error across five seeds.

4. Discussion and conclusion

In this work, we addressed the challenge of estimating the calibration error of an unlabeled target distribution under covariate shift. We proposed an estimator of CE that accounts for the change in the marginal distribution. The proposed estimator has advantages of being consistent and differentiable, making it suitable for integrating in both post-hoc and trainable calibration methods.

Our work has some limitations. Specifically, we focus solely on covariate shift, neglecting other types of dataset shifts, which we leave for future work. Moreover, we observed that our method underestimates the CE on the target domain, which also needs to be studied further.

To conclude, we proposed a variant of a KDE-based estimator, designed to assess CE in the presence of covariate shift, and empirically demonstrated that it allows us to effectively reveal a substantial increase in CE even when labels for the target data are unavailable. By accounting for covariate shift, we can conduct more comprehensive and realistic evaluation of the model's calibration performance.

Acknowledgements

This research received funding from the Flemish Government (AI Research Program), the Research Foundation - Flanders (FWO) through project number G0G2921N, and the Academy of Finland (Profi6 336449 funding program).

References

[1] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[2] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. *Advances in neural information processing systems*, 33:11996–12007, 2020.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[5] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[7] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[8] Anusri Pampari and Stefano Ermon. Unsupervised calibration under covariate shift, 2020.

[9] Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR, 2020.

[10] Teodora Popordanoska, Raphael Sayer, and Matthew B. Blaschko. A consistent and differentiable lp canonical calibration error estimator. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[11] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 10 2000.

[12] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.

[13] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems*, 32, 2019.