Distribution-free uncertainty quantification for classification under label shift

Aleksandr Podkopaev and Aaditya Ramdas

Kiho Park
parkkiho@uchicago.edu

1 Introduction

We focus on two categories of post-processing procedures: calibration via post-hoc binning and conformal prediction. In the context of distribution-free uncertainty quantification, we produce prediction sets and calibrated probabilities, which are complementary approaches for classifier UQ. Neither of two notions provide a complete answer to the question of UQ for classification on their own, but together they represent two of the more principled distribution-free approaches towards UQ that are practically efficient and theoretically grounded.

Consider labeled data $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$ where $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y} = \{1, \cdots, K\}$ for classification. In real-world applications, the target distribution might not be the same as the source distribution. Let P, Q stand for the source and target distributions defined on $\mathcal{X} \times \mathcal{Y}$ with p, q being the pdfs associated with P and Q. Two common assumptions about the type of shift include covariate shift: $q(x) \neq p(x)$ but $q(y \mid x) = p(y \mid x)$, and label shift: $q(y) \neq p(y)$ but $q(x \mid y) = p(x \mid y)$. For example, label shift is a reasonable assumption in medical applications where diseases (Y) cause symptoms (X).

For regression, Tibshirani et al. (2019) generalize construction of conformal prediction intervals to handle the case of known covariate likelihood ratio. For binary classification, Gupta et al. (2020) propose a way of calibrating probabilistic predictors under covariate shift. In this work, we close an existing gap for quantifying predictive uncertainty under label shift by adapting distribution-free calibration and conformal prediction.

2 Conformal classification

We wish to construct an uncertainty set function $C : \mathcal{X} \to 2^{\mathcal{Y}}$, such that for a new data point we can guarantee that

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$$

Conformal prediction does not require making any distributional assumptions. Also, it allows to proceed with both probabilistic and scoring classifiers.

2.1 Exchangeable conformal

Consider a sequence of candidate nested prediction sets $\{\mathcal{F}_\tau(x)\}_{\tau \in \mathcal{T}} : \mathcal{F}_{\tau_1}(x) \subseteq \mathcal{F}_{\tau_2}(x) \subseteq \mathcal{Y}$ for any $\tau_1 \leq \tau_2 \in \mathcal{T}$, with $\mathcal{F}_{\inf \mathcal{T}} = \emptyset$ and $\mathcal{F}_{\sup \mathcal{T}} = \mathcal{Y}$. For any point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we define

$$r(x, y) := \inf\{\tau \in \mathcal{T} : y \in \mathcal{F}_\tau(x)\}$$

Within split-conformal framework, the first dataset is used to construct a nested sequence and the second is used to select the smallest τ^* that guarantees validity.

If the true class-posterior distribution $\pi_y(x) = \mathbb{P}[Y = y \mid X = x]$ is known, we can construct the optimal prediction set

$$C_\alpha^{\text{oracle}}(x) := \{y \in \mathcal{Y} : \rho_y(x; \pi) < 1 - \alpha\}$$

where $\rho_y(x; \pi) := \sum_{y'=1}^{K} \pi_{y'}(x) \mathbb{1}\{\pi_{y'}(x) > \pi_y(x)\}$ is the total probability mass of labels that are more likely than $y \in \mathcal{Y}$. Naive conformalization of the nested sequence yields prediction sets with correct marginal coverage but typically inferior conditional coverage in practice. Thus, we instead use a randomized version defined as

$$\tilde{C}_\alpha^{\text{oracle}}(x) := \{y : \rho_y(x; \pi) + u \cdot \pi_y(x) \leq 1 - \alpha\}$$

where u is a realization of Unif($[0, 1]$). This form suggests to consider the following nested sequence:

$$\mathcal{F}_\tau(x, u; \widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y\left(x; \widehat{\pi}\right) + u \cdot \widehat{\pi}_y(x) \leq \tau\}$$

for $\tau \in \mathcal{T} = [0, 1]$. Then for any triple (X, Y, U), the corresponding radius (score) is given by

$$\begin{aligned} r(X, Y, U, \widehat{\pi}) &= \inf\left\{\tau \in \mathcal{T} : \rho_Y\left(X; \widehat{\pi}\right) + U \cdot \widehat{\pi}_Y(X) \leq \tau\right\} \\ &= \rho_Y(X; \widehat{\pi}) + U \cdot \widehat{\pi}_Y(X) \end{aligned}$$

Assume that the dataset is split at random into training $\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}$ and calibration $\{(X_i, Y_i)\}_{i \in \mathcal{I}_2}$ where $\mathcal{I}_2 = \{1, \cdots, n\}$. Let $\tau^\star = Q_{1-\alpha}(\{r_i\}_{i \in \mathcal{I}_2} \cup \{1\})$ for the non-conformaity scores $r_i = r(X_i, Y_i, U_i; \widehat{\pi}) \in [0, 1]$, $i \in \mathcal{I}_2 \cup \{n + 1\}$.

Theorem 1. *If $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, then*

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{F}_{\tau^\star}(X_{n+1}, U_{n+1}; \widehat{\pi})\right) \mid \{(X_i, Y_i)\}_{i \in \mathcal{I}_1}) \geq 1 - \alpha$$

Further, if the non-conformity scores are almost surely distinct, then the above probability is upper bounded by $1 - \alpha + 1/(n + 1)$.

2.2 Label-shifted conformal

Assume that the true likelihood ratios $w(y) = q(y)/p(y)$ are known for all $y \in \mathcal{Y}$. Let

$$\mathcal{F}_{\tau^\star}^{(w)}(x, u; \widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leq \tau_w^\star(y)\}$$

for $\tau_w^\star = Q_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{p}_i^w(y)\delta_{r_i} + \tilde{p}_{n+1}^w(y)\delta_1\right)$ where $\tilde{p}_i^w(y) = w(Y_i)/(\sum_{j=1}^{n} w(Y_j) + w(y))$ for $i = 1, \cdots, n$ and $\tilde{p}_{n+1}^w(y) = w(y)/(\sum_{j=1}^{n} w(Y_j) + w(y))$.

Theorem 2. *For any $\alpha \in (0, 1)$, if the true likelihood ratios $w(y) = q(y)/p(y)$ are known for all $y \in \mathcal{Y}$, it holds that*

$$\mathbb{P}(Y_{n+1} \in \mathcal{F}_{\tau^*}^{(w)}(X_{n+1}, U_{n+1}; \widehat{\pi}) \mid \{(X_i, Y_i)\}_{i \in \mathcal{I}_1}) \geq 1 - \alpha$$

A toy classification task demonstrates the necessity of correcting for label shift as the classic conformal prediction sets fail to achieve the correct marginal coverage. In contrast, it illustrates that calibrating the threshold $\tau_w^\star$ with either oracle or importance weights estimated by BBSE allows to achieve the target marginal coverage. The conformal sets that proceed with estimated ones have a slightly downgraded performance.

Corollary 1. *Fix $\alpha \in (0,1)$. Assume that $\widehat{w}_k$ is a consistent estimator of w where $k = |\mathcal{D}_{\text{est}}|$ denotes the total size of the datasets used for construction. Further, assume that for the true w and all for $y \in \mathcal{Y}$, the discrete distribution $\sum_{i=1}^{n} \tilde{p}_i^w(y)\delta_{r_i} + \tilde{p}_{n+1}^w(y)\delta_1$ does not have a jump at level $1-\alpha$. Then,*

$$\lim_{k\to\infty} \mathbb{P}\left(Y_{n+1} \in \mathcal{F}_{\tau^*}^{(\widehat{w}_k)}(X_{n+1}, U_{n+1}; \widehat{\pi})\right) \geq 1-\alpha$$

Label-conditional conformal prediction. Assume that a set of significance levels for each class $\{\alpha_y\}_{y\in\mathcal{Y}}$ has been chosen. By further splitting the calibration set $\mathcal{I}_2$ into $|\mathcal{Y}| = K$ groups depending on the corresponding labels, $\mathcal{I}_{2,y} := \{i \in \mathcal{I}_2 : Y_i = y\}$, one can consider prediction sets

$$\mathcal{F}_{\tau_c^\star}^c(x, u; \widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leq \tau_c^\star(y)\}$$

where $\tau_c^\star(y) = Q_{1-\alpha_y}\left(\{r_i\}_{i\in\mathcal{I}_{2,y}} \cup \{1\}\right)$. For a test point (X_{n+1}, Y_{n+1}), the corresponding non-conformity score $r(X_{n+1}, Y_{n+1}, U_{n+1}; \widehat{\pi})$ together with $\{r_i\}_{i\in\mathcal{I}_2, Y n+1}$ forms a collection of exchangeable random variables, and it implies label-conditional validity:

$$\mathbb{P}\left(Y_{n+1} \notin \mathcal{F}_{\tau_c^\star}^c(X_{n+1}, U_{n+1}; \widehat{\pi}) \mid Y_{n+1} = y\right) \leq \alpha_y \text{ for all } y \in \mathcal{Y}$$

The label-conditional conformal framework yields a stronger guarantee than the standard marginal conformal and it is automatically robust to changes in class proportions. The price to pay for the stronger conditional guarantee is larger prediction sets. When the number of classes K is large, we can use the label-conditional conformal framework as a complementary approach.

3 Calibration

While prediction sets describe a construction on top of the output of a predictor, calibration quantifies whether the output itself admits a rigorous frequentist interpretation.

Definition 1 (Canonical Calibration)**.** *Let Δ_K denote the probability simplex in $\mathbb{R}^K$. A probabilistic predictor $f : \mathcal{X} \to \Delta_K$ is said to be calibrated if*

$$p(Y = y \mid f(X)) = f_y(X), \quad y \in \mathcal{Y}$$

where $f_y(x)$ denotes the y-th coordinate of $f(x)$.

Model miscalibration is usually assessed using either reliability curves or related one-dimensional summary statistics. It is known that popular metrics, such as Expected Calibration Error (ECE), can be biased if binning or discretization of the output of the resulting model is not performed. Binning represents coarsening of the sample space and is defined as the partition: $\Delta_K = B_1 \cup \cdots \cup B_M$, $B_i \cap B_j = \emptyset$ for $i \neq j$. Then a predictor f induces a partition of the sample space:

$$\mathcal{X}_m := \{x \in \mathcal{X} : f(x) \in B_m\}, \quad m \in \mathcal{M} := \{1, \cdots, M\}$$

3.1 Calibration for i.i.d. data

Assume that the binning scheme has been chosen and use $g : \mathcal{X} \to \mathcal{M}$ to denote the bin-mapping function: $g(x) = m \iff f(x) \in B_m$. The calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ is used for estimating

$$\pi_{y,m}^P := \mathbb{P}(Y = y \mid f(X) \in B_m), \quad y \in \mathcal{Y}$$

for all bins $m \in \mathcal{M}$. Let $N_m = |\{(X_i, Y_i) \in \mathcal{D}_{\text{cal}} : f(X_i) \in B_m\}|$ denote the number of calibration points that fall into bin $m \in \mathcal{M}$. Empirical frequencies of class labels $y \in \mathcal{Y}$ in each bin $m \in \mathcal{M}$ are

$$\widehat{\pi}^P_{y,m} := \frac{1}{N_m} \sum_{i=1}^{n} \mathbb{1}\{Y_i = y, f(X_i) \in B_m\}$$

Let $\pi^P_m := (\pi^P_{1,m}, \cdots, \pi^P_{K,m})^T$ and $h : \mathcal{X} \to \Delta_K$ such that $h(x) = \widehat{\pi}_{g(x)}$.

Theorem 3. *Fix $\alpha \in (0,1)$. With probability at least $1-\alpha$, $||\widehat{\pi}^P_m - \pi^P_m||_1 \leq \varepsilon_m$, simultaneously for all $m \in \mathcal{M}$, where*

$$\epsilon_m := \frac{2}{\sqrt{N_m}} \sqrt{\frac{1}{2} \ln\left(\frac{M 2^K}{\alpha}\right)}$$

As a consequence, with probability at least $1-\alpha$,

$$\sum_{y=1}^{K} |\mathbb{P}(Y = y \mid h(X) = z) - z_y| \leq \max_{m \in \mathcal{M}} \varepsilon_m$$

simultaneously for all z in the range of h.

Theorem 3 states that as long as the least-populated bin contains sufficiently many points, the output of the recalibrated predictor will approximately satisfy condition $\mathbb{P}\left(Y = y \mid \widehat{\pi}^P_{y,g(X)}\right) \approx \widehat{\pi}^P_{y,g(X)}$.

3.2 Label-shifted calibration

Consider $\mathcal{Y} = \{0,1\}$ and $X \mid Y = y \sim \mathcal{N}(\mu_y, \Sigma)$ with $p(y)$ and $q(y)$. For the source distribution, a perfect probabilistic predictor is given by

$$\pi^P_1(x) = \frac{p(1) \cdot \varphi(x; \mu_1, \Sigma)}{p(0) \cdot \varphi(x; \mu_0, \Sigma) + p(1) \cdot \varphi(x; \mu_1, \Sigma)}$$

where $\varphi(x; \mu_i, \Sigma)$, $i = 0, 1$ denotes the PDF of a Gaussian random vector. Even though the Bayes-optimal rule is calibrated on the source, a correction is required to obtain a calibrated classifier under label shift.

If both the true class-posterior distribution $\pi^P_y(x)$ and the true label likelihood ratios w are known, the form of the adjustment of the probabilistic classifier under label shift is

$$\pi^Q_y(x) = \frac{w(y) \cdot \pi^P_y(x)}{\sum_{k=1}^{K} w(k) \cdot \pi^P_k(x)}$$

In practice, the output of the original predictor has to be binned to obtain the distribution-free guarantees.

Proposition 1. *Under label shift, for any class label $y \in \mathcal{Y}$ and any bin B_m, $m \in \mathcal{M}$, it holds that*

$$\pi^Q_{y,m}(x) = \frac{w(y) \cdot \pi^P_{y,m}(x)}{\sum_{k=1}^{K} w(k) \cdot \pi^P_{k,m}(x)}$$

With an estimator $\widehat{w}$ and empirical frequencies of class labels $\{\widehat{\pi}^P_m\}_{m \in \mathcal{M}}$, Proposition 1 suggests an appropriate correction to provably obtain asymptotically calibrated predictors on the target:

$$\widehat{\pi}^{(\widehat{w})}_{y,m} = \frac{\widehat{w}(y) \cdot \widehat{\pi}^P_{y,m}(x)}{\sum_{k=1}^{K} \widehat{w}(k) \cdot \widehat{\pi}^P_{k,m}(x)}$$

for all bins $m \in \mathcal{M}$. Define the condition number:

$$\kappa := \frac{\sup_k w(k)}{\inf_{k:w(k)\neq 0} w(k)}$$

Now we quantify the miscalibration of the predictor.

Theorem 4. *Let $\widehat{w}$ be an estimator of w and let $\widehat{\pi}_{y,m}^{(\widehat{w})}$ denote the reweighted empirical frequencies for all labels $y \in \mathcal{Y}$ and bins $m \in \mathcal{M}$. For any bin $m \in \mathcal{M}$, it holds that:*

$$\left\| \widehat{\pi}_m^{(\widehat{w})} - \pi_m^Q \right\|_1 \leq 2\kappa \cdot \left\| \widehat{\pi}_m^P - \pi_m^P \right\|_1 + \frac{2\|\widehat{w} - w\|_\infty}{\inf_{l:w(l)\neq 0} w(l)}$$

The calibration error on the target decomposes into two terms: the first term is controlled by the calibration error on the source and the second term is controlled by the importance weights estimation error. Finite-sample guarantees for the miscalibration of the resulting predictor on the target domain follow by Theorem 4 via invoking simple probabilistic arguments. Figures with reliability curves indicate that shift-corrected binning with either true or estimated importance weights yields a calibrated predictor on the target domain while uncorrected fails to do so as expected.

4 Discussion

We analyzed distribution-free uncertainty quantification for classification via two complementary notions: prediction sets and calibration. We focused on a setting of label shift for which a correction is necessary. By adapting conformal prediction sets and calibration via binning to label shift, we close an existing gap for distribution-free uncertainty quantification without requirement of labeled target data. We note that handling label shift should be expected to be an easier task rather than handling covariate shift which typically involves estimating a high-dimensional and continuous likelihood ratio. Lastly, establishing meaningful guarantees for full canonical calibration in multiclass setting remains a future research direction.

References

Gupta, C., Podkopaev, A., & Ramdas, A. (2020). *Distribution-free binary classification: prediction sets, confidence intervals and calibration.* In Advances in Neural Information Processing Systems.

Tibshirani, R. J., Barber, R. F., Candès, E., & Ramdas, A. (2019). *Conformal prediction under covariate shift.* In Advances in Neural Information Processing Systems, 2019.