Trustworthy Machine Learning

Adaptive Conformal Prediction

Sangdon Park

POSTECH

Motivation: Distribution Shift

- The main assumption of conformal prediction: exchangeability (or i.i.d.)
- In practice, this is fragile due to distribution shifts.
- Type of distribution shifts
  - Covariate shift
  - Label shift
  - ...
  - Adversarial shift

Covariate Shift

- Setup: follows domain adaptation, *i.e.,*
  - There is only one shift
  - $p(x, y)$: a source distribution
  - $q(x, y)$: a target distribution
  - $S \sim p^m(x, y)$: i.i.d. label examples from source
  - $T \sim q^n(x)$: i.i.d. unlabeled examples from target

Covariate Shift

- Setup: follows domain adaptation, *i.e.*,
  - There is only one shift
  - $p(x, y)$: a source distribution
  - $q(x, y)$: a target distribution
  - $S \sim p^m(x, y)$: i.i.d. label examples from source
  - $T \sim q^n(x)$: i.i.d. unlabeled examples from target
- Assumption:

$$p(y|x) = q(y|x) \quad \text{but possibly} \quad p(x) \neq q(x)$$

Covariate Shift

- Setup: follows domain adaptation, *i.e.,*
  - There is only one shift
  - $p(x, y)$: a source distribution
  - $q(x, y)$: a target distribution
  - $S \sim p^m(x, y)$: i.i.d. label examples from source
  - $T \sim q^n(x)$: i.i.d. unlabeled examples from target
- Assumption:

$$p(y|x) = q(y|x) \quad \text{but possibly} \quad p(x) \neq q(x)$$

- Conformal prediction under covariate shift
  - Tibshirani et al. [2019]: provides the coverage guarantee
  - Park et al. [2022]: provides the PAC guarantee

Label Shift

- Setup: follows domain adaptation, *i.e.,*
  - There is only one shift
  - $p(x, y)$: a source distribution
  - $q(x, y)$: a target distribution
  - $S \sim p^m(x, y)$: i.i.d. label examples from source
  - $T \sim q^n(x)$: i.i.d. unlabeled examples from target

Label Shift

- Setup: follows domain adaptation, *i.e.*,
  - There is only one shift
  - $p(x, y)$: a source distribution
  - $q(x, y)$: a target distribution
  - $S \sim p^m(x, y)$: i.i.d. label examples from source
  - $T \sim q^n(x)$: i.i.d. unlabeled examples from target
- Assumption:

$$p(x|y) = q(x|y) \quad \text{but possibly} \quad p(y) \neq q(y)$$

Label Shift

- Setup: follows domain adaptation, *i.e.*,
  - There is only one shift
  - $p(x, y)$: a source distribution
  - $q(x, y)$: a target distribution
  - $S \sim p^m(x, y)$: i.i.d. label examples from source
  - $T \sim q^n(x)$: i.i.d. unlabeled examples from target
- Assumption:

$$p(x|y) = q(x|y) \quad \text{but possibly} \quad p(y) \neq q(y)$$

- Conformal prediction under label shift
  - Podkopaev and Ramdas [2021]: provides the coverage guarantee

Adversarial Shift

- Setup: follows an online learning setup, *i.e.,*
  - there are multiple shifts over time
  - $p_t(x, y)$: a distribution at time t
  - $(x_t, y_t) \sim p_t(x, y)$: a labeled example sampled at time t

Adversarial Shift

- Setup: follows an online learning setup, *i.e.,*
  - there are multiple shifts over time
  - $p_t(x, y)$: a distribution at time t
  - $(x_t, y_t) \sim p_t(x, y)$: a labeled example sampled at time t
- Assumption: no restriction on shifts

Adversarial Shift

- Setup: follows an online learning setup, *i.e.,*
  - there are multiple shifts over time
  - $p_t(x, y)$: a distribution at time t
  - $(x_t, y_t) \sim p_t(x, y)$: a labeled example sampled at time t
- Assumption: no restriction on shifts
- Conformal prediction under distribution shift
  - Gibbs and Candès [2021]: provides the coverage guarantee

Adaptive Conformal Prediction

Can we learn conformal sets under distribution shift?

Setup:

- $\mathcal{X}$: example space
- $\mathcal{Y}$: label space
- $C_t : \mathcal{X} \to 2^{\mathcal{Y}}$: a conformal set
- A learning game between a learner and nature

for $t = 1, \ldots, T$ **do**
  Learner receives an example $x_t \in \mathcal{X}$
  Learner outputs a *conformal set* $C_t(x_t) \in 2^{\mathcal{Y}}$
  Learner receives a true label $y_t \in \mathcal{Y}$
  Learner suffers loss $\mathbb{1}(y_t \notin C_t(x_t))$
  Learner update a parameter of a conformal set
end for

A Goodness Metric: "Empirical" Coverage Guarantee

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \varepsilon$$

- $1 - \alpha$: a desired coverage rate
- T: a time horizon
- $\hat{C}_t$: a conformal set at time t constructed by an algorithm
- It is similar to the regret definition (but not exactly the same).
- We wish to bound this quantity.

A Goodness Metric: "Empirical" Coverage Guarantee

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \varepsilon$$

- $1 - \alpha$: a desired coverage rate
- T: a time horizon
- $\hat{C}_t$: a conformal set at time t constructed by an algorithm
- It is similar to the regret definition (but not exactly the same).
- We wish to bound this quantity.
- Why not use the PAC guarantee?

A Goodness Metric: "Empirical" Coverage Guarantee

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \varepsilon$$

- $1 - \alpha$: a desired coverage rate
- T: a time horizon
- $\hat{C}_t$: a conformal set at time t constructed by an algorithm
- It is similar to the regret definition (but not exactly the same).
- We wish to bound this quantity.
- Why not use the PAC guarantee?
  - the PAC guarantee is for the batch learning.

Algorithm

Main Ideas

- Run the batch conformal prediction (CP) for each time
- But adjust the coverage α for the batch CP to satisfy the empirical coverage guarantee.

Algorithm

Algorithm 1 A standard version of Adaptive Conformal Inference [Gibbs and Candès, 2021]

$t_1 \in \{1, \dots, T\}$
$\alpha_{t_1} \in [0, 1]$
for $t = t_1, \dots, T$ **do**
  $(\mathcal{D}_{\text{train}^{(t)}}, \mathcal{D}_{\text{cal}}^{(t)}) \leftarrow$ Randomly split the data $\{(x_i, y_i)\}_{i=1}^{t-1}$ and obtain non-conformity scores
  $S_t \leftarrow$ Update using $\mathcal{D}_{\text{train}}^{(t)}$
  $q_t \leftarrow \text{Quantile}(1 - \alpha_t, \mathcal{D}_{\text{cal}}^{(t)} \cup \{\infty\})$
  Observe x_t
  Predict $\hat{C}_t(x_t)$
  Observe y_t
  Update $\alpha_{t+1} \leftarrow \alpha_t + \gamma \left(\alpha - \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right)\right)$
end for

- A conformal set: $\hat{C}_t(x_t) := \{y \in \mathcal{Y} \mid S_t(x_t, y) \leq q_t\}$
- Until t_1, the algorithm simply collects data.
- The algorithm is not randomized.

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0, 1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1 - \alpha_1\} + \gamma}{T\gamma}$$

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0,1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1} \left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1 - \alpha_1\} + \gamma}{T\gamma}$$

- The coverage decreases by $\mathcal{O}\left(\frac{1}{T}\right)$

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0,1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1 - \alpha_1\} + \gamma}{T\gamma}$$

- The coverage decreases by $\mathcal{O}\left(\frac{1}{T}\right)$
- This holds for any sequence $((x_1, y_t), \ldots, (x_T, y_T))$!

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0,1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}$$

- The coverage decreases by $\mathcal{O}\left(\frac{1}{T}\right)$
- This holds for any sequence $((x_1, y_t), \ldots, (x_T, y_T))$!
  - If $\hat{C}_t(x_t) = \mathcal{Y}$, the adversary will never win without randomization.

Coverage Bound

Theorem

For all $T \in \mathbb{N}$, $\alpha \in (0,1)$, *and* $\gamma > 0$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) - \alpha \right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}$$

- The coverage decreases by $\mathcal{O}\left(\frac{1}{T}\right)$
- This holds for any sequence $((x_1, y_t), \ldots, (x_T, y_T))$!
  - If $\hat{C}_t(x_t) = \mathcal{Y}$, the adversary will never win without randomization.
- Suppose $\alpha_1 = 0$, $\gamma = 0.01$, and $\varepsilon = 0.01$. Then, we $T = 10,100$ observations to make the empirical coverage close to a desired coverage.

A Lemma for the Coverage Bound: A Proof Sketch

Lemma

For all $t \in \mathbb{N}$, we have

$$\alpha_t \in [-\gamma, 1 + \gamma].$$

- Recall our update rule:

$$\alpha_{t+1} \leftarrow \alpha_t + \gamma \left(\alpha - \mathbb{1} \left(y_t \notin \hat{C}_t(x_t) \right) \right)$$

A Lemma for the Coverage Bound: A Proof Sketch

Lemma

For all $t \in \mathbb{N}$, we have

$$\alpha_t \in [-\gamma, 1+\gamma].$$

- Recall our update rule:

$$\alpha_{t+1} \leftarrow \alpha_t + \gamma \left(\alpha - \mathbb{1} \left(y_t \notin \hat{C}_t(x_t) \right) \right)$$

- Observe that the update cannot be larger than (and equal to) γ, *i.e.*,

$$\sup_t |\alpha_{t+1} - \alpha_t| = \sup_t \left| \gamma \left(\alpha - \mathbb{1} \left(y_t \notin \hat{C}_t(x_t) \right) \right) \right| < \gamma$$

A Lemma for the Coverage Bound: A Proof Sketch

Lemma

For all $t \in \mathbb{N}$, we have

$$\alpha_t \in [-\gamma, 1+\gamma].$$

- Recall our update rule:

$$\alpha_{t+1} \leftarrow \alpha_t + \gamma \left(\alpha - \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) \right)$$

- Observe that the update cannot be larger than (and equal to) γ, *i.e.*,

$$\sup_t |\alpha_{t+1} - \alpha_t| = \sup_t \left| \gamma \left(\alpha - \mathbb{1}\left(y_t \notin \hat{C}_t(x_t) \right) \right) \right| < \gamma$$

  - Thus, the claim intuitively make sense.

A Lemma for the Coverage Bound: A Proof Sketch

- Suppose that there is $\{\alpha_t\}_{t\in\mathbb{N}}$ such that $\inf_t \alpha_t < -\gamma$.
- Due to the update, we have positive probability to have $\alpha_t < 0$ and $a_{t+1} < \alpha_t$ for some t.
- Contradiction:

$$
\begin{aligned}
\alpha_t < 0 \quad & \implies \quad q_t := \text{Quantile}(1-\alpha_t, \mathcal{D}_{\text{cal}}^{(t)} \cup \{\infty\}) = \infty \\
& \implies \quad \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right) = 0 \\
& \implies \quad \alpha_{t+1} = \alpha_t + \gamma\left(\alpha - \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right)\right) = \alpha_t + \gamma\alpha \geq \alpha_t
\end{aligned}
$$

Coverage Bound: A Proof Sketch

- Let $e_t := \mathbb{1}\left(y_t \notin \hat{C}_t(x_t)\right)$
- Recall the recursive update rule, *i.e.*,

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - e_t)$$

- Due to the recursive update rule,

$$\alpha_{T+1} = \alpha_1 + \sum_{t=1}^{T} \gamma(\alpha - e_t)$$

- Due to the previous lemma,

$$-\gamma \leq \alpha_1 + \sum_{t=1}^{T} \gamma(\alpha - e_t) \leq 1 + \gamma.$$

- This implies

$$\frac{\alpha_1 - (1+\gamma)}{T\gamma} \leq \frac{1}{T}\sum_{t=1}^{T}(e_t - \alpha) \leq \frac{\alpha_1 + \gamma}{T\gamma}$$

Conclusion

- Adaptive Conformal Inference [Gibbs and Candès, 2021] is the first approach to learn a conformal set under distribution shift.
- Running a batch algorithm within an online algorithm.
  - The time and memory complexity is linear in T.
  - See a more efficient (and general) approach [Bastani et al., 2022]

Reference I

O. Bastani, V. Gupta, C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35: 29362–29373, 2022.

I. Gibbs and E. Candès. Adaptive conformal inference under distribution shift, 2021.

S. Park, E. Dobriban, I. Lee, and O. Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=DhP9L8vIyLc`.

A. Podkopaev and A. Ramdas. Distribution-free uncertainty quantification for classification under label shift. *arXiv preprint arXiv:2103.03323*, 2021.

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32:2530–2540, 2019.