Trustworthy Machine Learning

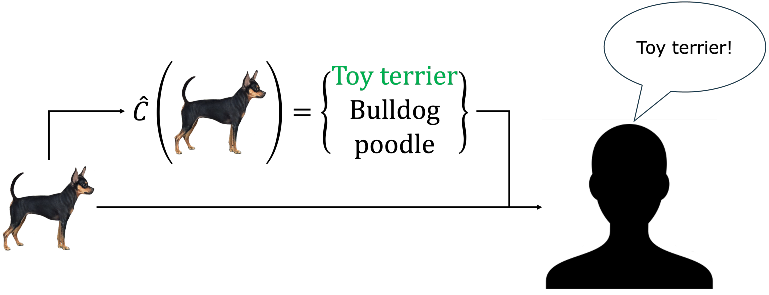
Selective Prediction

Sangdon Park

POSTECH

Motivation

- Conformal prediction is fine but requires post-processing (*i.e.*, human-in-the-loop)



- Can we use this in fully automated systems (*i.e.*, without-human-in-the-loop)?

Learning Setup

A standard supervised learning setup:

- $\mathcal{X}$: an example space
- $\mathcal{Y}$: an example space
- $\mathcal{D}$: a distribution over $\mathcal{X} \times \mathcal{Y}$
- $Z \sim \mathcal{D}^n$: a calibration set
- $\mathcal{F}$: a set of selective predictors (will introduce soon)
- loss: a false discovery rate (will introduce soon)

A Selective Classifier

Definition (a selective classifier)

$$\hat{S}(x) = \begin{cases} \hat{y}(x) & \text{if } f(x, \hat{y}(x)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

- $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$: a scoring function
- $\hat{y} : \mathcal{X} \to \mathcal{Y}$: a classifier, *i.e.*,

$$\hat{y}(x) \coloneqq \arg\max_{y \in \mathcal{Y}} \; f(x, y)$$

- $\tau \in \mathbb{R}_{\geq 0}$: a parameter
- $\texttt{IDK}$: "I don't know"
- $\hat{S} : \mathcal{X} \to \mathcal{Y} \cup \{\texttt{IDK}\}$: a selective classifier

A Goodness Metric: False Discovery Rate

Definition (false discovery rate (FDR))

$$\mathbb{P}\left\{y \neq \hat{S}(x) \,\middle|\, \hat{S}(x) \neq \texttt{IDK}\right\}$$

- the FDR is equivalent to the precision
- The following equivalence may be useful:

$$\begin{aligned}\mathbb{P}\left\{y \neq \hat{S}(x) \,\middle|\, \hat{S}(x) \neq \texttt{IDK}\right\} &= \mathbb{P}\left\{y \neq \hat{y}(x) \,\middle|\, \hat{S}(x) \neq \texttt{IDK}\right\} \\ &= \mathbb{P}\left\{y \neq \hat{y}(x) \,\middle|\, f(x, \hat{y}(x)) \geq \tau\right\}\end{aligned}$$

- uncomfortable fact: the FDR is not monotonic in τ (you will see why)

Goal: Achieving a PAC-Style Guarantee

Goal

Find a PAC algorithm that returns $\hat{S}$, *i.e.*,

$$\mathbb{P}\left\{\mathbb{P}\left\{y \neq \hat{S}(x) \,\middle|\, \hat{S}(x) \neq \texttt{IDK}\right\} \leq \varepsilon\right\} \geq 1 - \delta$$

- This is an ideal goal.
- This PAC guarantee for any ε can be achievable under some condition.
  - This is related to the monotonicity of the FDR.

Assumption

Assumption (i.i.d.)

Labeled examples are independently drawn from the same (and unknown) distribution $\mathcal{D}$ over labeled examples $\mathcal{X} \times \mathcal{Y}$.

- Same as the assumption for PAC algorithms

Motivation: Direct Comparison to Conformal Prediction

Conformal Prediction

- Predictor form:

$$\hat{C}(x) = \left\{ y \in \mathcal{Y} \;\middle|\; f(x, y) \geq \tau \right\}$$

- Guarantee: a coverage guarantee

$$\mathbb{P}\left\{ y \notin \hat{C}(x) \right\} \leq \varepsilon$$

- Assumption: exchangeable or i.i.d.

Selective Prediction

- Predictor form:

$$\hat{S}(x) = \begin{cases} \hat{y}(x) & \text{if } f(x, \hat{y}(x)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

- Guarantee: a false-discovery rate guarantee

$$\mathbb{P}\left\{ y \neq \hat{S}(x) \;\middle|\; \hat{S}(x) \neq \texttt{IDK} \right\} \leq \varepsilon$$

- Assumption: i.i.d.

Algorithm

Idea

- Enumerate all candidate hypotheses (*i.e.*, a set of τs), *i.e.*,

$$f(x_1, \hat{y}(x_1)), \dots, f(x_i, \hat{y}(x_i)), \dots, f(x_n, \hat{y}(x_n)) \quad \text{for } (x_i, \cdot) \in Z$$

- For each hypothesis, compute a binomial tail bound, *i.e.*,

$$\mathbb{P}\left\{ y \neq \hat{y}(x) \;\middle|\; f(x, \hat{y}(x)) \geq \tau_i \right\} \leq U_{\mathsf{Binomial}}(\tau_i, \dots) \quad \text{for } \tau_i = f(x_i, \hat{y}(x_i))$$

- Choose a hypothesis that has the binomial tail bound smaller than ε, *i.e.*,

$$U_{\mathsf{Binomial}}(\tau_i, \dots) \leq \varepsilon$$

- Minimize τ to maximize efficiency, *i.e.*, recall the definition of the selective classifier:

$$\hat{S}(x) = \begin{cases} \hat{y}(x) & \text{if } f(x, \hat{y}(x)) \geq \tau \\ \texttt{IDK} & \text{otherwise} \end{cases}$$

Algorithm

Algorithm 1 Selective Classifier Learning Algorithm $\mathcal{A}$ [Geifman and El-Yaniv, 2017]

procedure $\mathrm{SC}(f,\ \hat{y},\ Z,\ \varepsilon,\ \delta)$
    $Z \leftarrow \textsc{Sort}_f(Z)$ ($\triangleright$) Sort Z in an increasing order of $f(x_i, \hat{y}(x_i))$
    $(\underline{i}, \bar{i}) \leftarrow (1, |Z|)$
    for $i = 1$ **to** $\lceil \log_2 |Z| \rceil$ **do**
        $\tau^{(i)} \leftarrow f(x_{\lceil (\underline{i}+\bar{i})/2 \rceil}, \hat{y}(x_{\lceil (\underline{i}+\bar{i})/2 \rceil}))$
        $Z^{(i)} \leftarrow \{(x, y) \in Z \mid f(x, \hat{y}(x)) \geq \tau^{(i)}\}$
        $k^{(i)} \leftarrow \sum_{(x,y) \in Z^{(i)}} \mathbb{1}(y \neq \hat{y}(x))$
        $U^{(i)} \leftarrow U_{\mathsf{Binom}}(k^{(i)}; |Z^{(i)}|, \delta / \lceil \log_2 |Z| \rceil)$
        if $U^{(i)} \leq \varepsilon$ **then**
            $\bar{i} \leftarrow \lceil (\underline{i} + \bar{i})/2 \rceil$
        else
            $\underline{i} \leftarrow \lceil (\underline{i} + \bar{i})/2 \rceil$
        end if
    end for
    return $\tau^{(i)}, U^{(i)}$
end procedure

FDR Guarantee

Theorem

For any f and $\mathcal{D}$, we have

$$\mathbb{P}\left\{ y \neq \hat{y}(x) \;\middle|\; f(x, \hat{y}(x)) \geq \hat{\tau} \right\} \leq \hat{U}$$

with probability at least $1 - \delta$, where the probability is taken over $Z \sim \mathcal{D}^n$ and $(\hat{\tau}, \hat{U}) = \mathcal{A}(Z)$.

- $\hat{U} \leq \varepsilon$ may not be true

FDR Guarantee: A Proof Sketch I

Let

- $\mathcal{R}(\tau) := \mathbb{P}\left\{y \neq \hat{y}(x) \mid f(x, \hat{y}(x)) \geq \tau\right\}$
- $\mathcal{H}$ is a data-dependent set of hypotheses with a fixed size, *i.e.*, $|\mathcal{H}| = m = \lceil \log_2 n \rceil$
- $Z_\tau := \left\{(x, y) \in Z \mid f(x, \hat{y}(x)) \geq \tau\right\}$

FDR Guarantee: A Proof Sketch II

Then, we have

$$
\begin{aligned}
\mathbb{P}\left\{\mathcal{R}(\hat{\tau}) > \hat{U}\right\} &\le \mathbb{P}\left\{\exists \tau \in \mathcal{H}, \mathcal{R}(\tau) > U_{\mathsf{FDR}}(\tau, \delta/m, Z_\tau)\right\} \\
&\le \sum_{j=1}^{m} \mathbb{P}\left\{\mathcal{R}(\tau_j) > U_{\mathsf{FDR}}(\tau_j, \delta/m, Z_{\tau_j})\right\} \\
&= \sum_{j=1}^{m} \sum_{i_j=0}^{n} \mathbb{P}\left\{\mathcal{R}(\tau_j) > U_{\mathsf{FDR}}(\tau_j, \delta/m, Z_{\tau_j}), |Z_{\tau_j}| = i_j\right\} \\
&= \sum_{j=1}^{m} \sum_{i_j=0}^{n} \mathbb{P}\left\{\mathcal{R}(\tau_j) > U_{\mathsf{FDR}}(\tau_j, \delta/m, Z_{\tau_j}) \mid |Z_{\tau_j}| = i_j\right\} \mathbb{P}\left\{|Z_{\tau_j}| = i_j\right\} \\
&\le \sum_{j=1}^{m} \sum_{i_j=0}^{n} \frac{\delta}{m} \mathbb{P}\left\{|Z_{\tau_j}| = i_j\right\} \\
&= \frac{\delta}{m} \sum_{j=1}^{m} 1 \\
&= \delta.
\end{aligned}
$$

Ideal Case (when our life gets easy)

Definition (prefect calibration)

We say that a scoring function f is perfectly calibrated with respect to $\mathcal{D}$ and $\hat{y}$ if

$$\mathbb{P}\left\{y = \hat{y}(x) \mid f(x, \hat{y}(x)) = t\right\} = t, \forall t \in [0, 1]$$

- Recall the definition of precision, *i.e.*,

$$\mathbb{P}\left\{y = \hat{y}(x) \mid f(x, \hat{y}(x)) \geq \tau\right\}$$

- If f is perfectly calibrated, we have

$$\begin{aligned}
\frac{d}{d\tau}\mathbb{P}\left\{y = \hat{y}(x) \mid f(x, \hat{y}(x)) \geq \tau\right\} &= \frac{d}{d\tau}\frac{\mathbb{P}\left\{y = \hat{y}(x) \mid f(x, \hat{y}(x)) \geq \tau\right\}\mathbb{P}\left\{f(x, \hat{y}(x)) \geq \tau\right\}}{\mathbb{P}\left\{f(x, \hat{y}(x)) \geq \tau\right\}} \\
&= \frac{d}{d\tau}\frac{\int_{\tau}^{1}\mathbb{P}\left\{y = \hat{y}(x) \mid f(x, \hat{y}(x)) = t\right\}\mathbb{P}\{f(x, \hat{y}(x)) = t\}\mathrm{d}t}{\int_{\tau}^{1}\mathbb{P}\{f(x, \hat{y}(x)) = t\}\mathrm{d}t} \\
&= \frac{d}{d\tau}\frac{\int_{\tau}^{1}t\mathbb{P}\{f(x, \hat{y}(x)) = t\}\mathrm{d}t}{\int_{\tau}^{1}\mathbb{P}\{f(x, \hat{y}(x)) = t\}\mathrm{d}t} = \cdots \geq 0
\end{aligned}$$

- This means precision is monotonically non-decreasing in τ [Lee et al., 2024] – this is a good property!

Conclusion

- Selective prediction could be a good alternative for conformal prediction.
- Selective prediction may not satisfy the PAC guarantee.

Reference I

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

M. Lee, K. Kim, T. Kim, and S. Park. Selective generation for controllable language models. 2024.