Learning to Defer with an Uncertain Rejector via Conformal Prediction

Yizirui Fang, Eric Nalisnick
Johns Hopkins University
{yfang52, nalisnick}@jhu.edu
JOHNS HOPKINS UNIVERSITY

Project website

NEURAL INFORMATION PROCESSING SYSTEMS

Learning to defer with one expert

Learning to defer (L2D) is a framework for human-AI collaboration that divides responsibility between machine and human decision makers. For every test instance, a 'rejector' function decides if the case should be passed to either a human or model (but not both).

Learning in L2D requires we fit both the rejector and classifier. We assume that whoever makes the prediction - model or human - incurs a loss of zero (correct) or one (incorrect). To use the rejector to toggle between the human and model, function:

- the overall classifier-rejector loss

$$L_{0-1}(h, r) = \mathbb{E}_{\mathbf{x},\mathrm{y},\mathrm{m}}\left[(1 - r(\mathbf{x}))\, \mathbb{I}[h(\mathbf{x}) \neq \mathrm{y}] + r(\mathbf{x})\, \mathbb{I}[\mathrm{m} \neq \mathrm{y}]\right] \quad (1)$$

Bayes optimal by minimizing above loss function:

- classifier h*($\boldsymbol{x}$)

$$h^*(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}(\mathrm{y} = y|\boldsymbol{x})$$

- rejecter function r*($\boldsymbol{x}$)

$$r^*(\boldsymbol{x}) = \mathbb{I}\left[\mathbb{P}(\mathrm{m} = \mathrm{y}|\boldsymbol{x}) \geq \max_{y \in \mathcal{Y}} \mathbb{P}(\mathrm{y} = y|\boldsymbol{x})\right]$$

Surrogate losses

For classifier-rejector loss function as Eq 1

- One-over-All (OvA)

$$\begin{aligned}\psi_{\mathrm{OvA}}(g_1, \ldots, g_{K+1}; \boldsymbol{x}, y, m) = \\ \phi[g_y(\boldsymbol{x})] + \sum_{y' \in \mathcal{Y}, y' \neq y} \phi[-g_{y'}(\boldsymbol{x})] + \phi[-g_{K+1}(\boldsymbol{x})] \\ + \mathbb{I}[m = y]\left(\phi[g_{K+1}(\boldsymbol{x})] - \phi[-g_{K+1}(\boldsymbol{x})]\right)\end{aligned}$$

- Asymmetric Softmax (A-SM)

$$\begin{aligned}\psi_{\text{A-SM}}(g_1, \ldots, g_{K+1}; \boldsymbol{x}, y, m) = \\ -\log \phi_{\text{A-SM}}(g(\boldsymbol{x}), y) \\ -\mathbb{I}[m \neq y] \cdot \log\left(1 - \phi_{\text{A-SM}}(g(\boldsymbol{x}), K+1)\right) \\ -\mathbb{I}[m = y] \cdot \log \phi_{\text{A-SM}}(g(\boldsymbol{x}), K+1)\end{aligned}$$

where

$$\phi_{\text{A-SM}}(g(\boldsymbol{x}), y) = \begin{cases} \dfrac{\exp(g_y(\boldsymbol{x}))}{\sum_{y'=1}^{K} \exp(g_{y'}(\boldsymbol{x}))} & \text{if } y < K+1, \\ \dfrac{\exp(g_{K+1}(\boldsymbol{x}))}{\sum_{y'=1}^{K+1} \exp(g_{y'}(\boldsymbol{x})) - \max_{y' \in \mathcal{Y}} \exp(g_{y'}(\boldsymbol{x}))} & \text{otherwise.} \end{cases}$$

Split Conformal Predictor

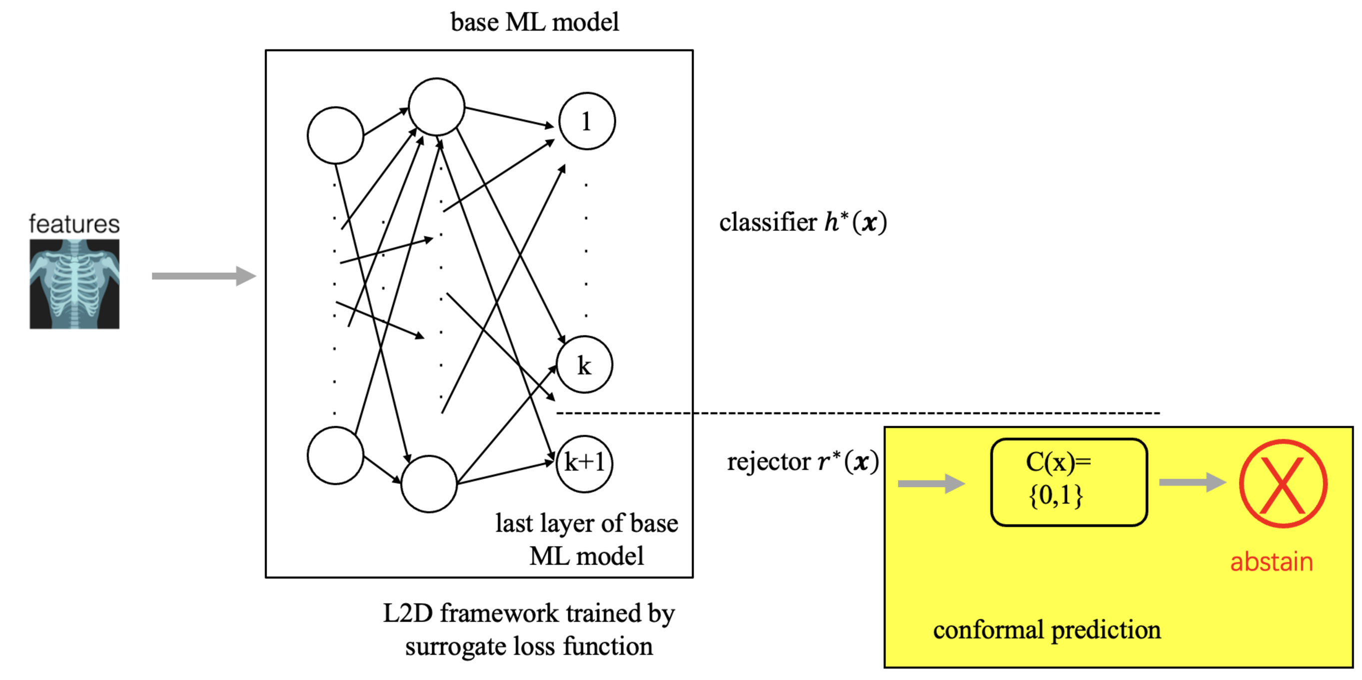
- distribution-free and finite sample guarantees
- $\hat{\tau}$ is computed as $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of calibration scores
- At test-time, given a feature vector $\boldsymbol{x}_{N+1}$, **marginal guarantee** is

$$\mathbb{P}\left(\mathrm{y}_{N+1} \in C\left(\mathbf{x}_{N+1}; \tau\right)\right) \geq 1 - \alpha, \text{ for } \alpha \in [0, 1]$$

- Prediction set constructed as

$$C(\mathbf{x}_{N+1}) = \{j | f_j(\mathbf{x}_{N+1}) > 1 - \hat{\tau}\}$$

- desired coverage is achieved in practice while also having efficient set sizes



Abstention L2D Decision Making Workflow

Uncertain Deferral Via Conformal Prediction

- CP framework to quantify the uncertainty in the rejector sub-component of an L2D system
- Conformal set $C_r(\mathbf{x}; \tau)$ is $\{\{0\}, \{1\}, \{0, 1\}\}$
- Ideal construction **marginal guarantee**

$$\mathbb{P}\left(r^*(\mathbf{x}_{N+1}) \in C_r(\mathbf{x}_{N+1}; \tau)\right) \geq 1 - \alpha$$

- Practical construction **marginal guarantee**

$$\mathbb{P}\left(\mathbb{I}\left[\mathrm{m}_{N+1} = \mathrm{y}_{N+1}\right] \in C_r(\mathbf{x}_{N+1}; \tau)\right) \geq 1 - \alpha$$

- Probability parameterization that the expert will be correct

  OvA

$$\hat{p}(\mathrm{m} = \mathrm{y}|\mathbf{x}) = \phi\left[g_{K+1}(\mathbf{x})\right] = \left(1 + \exp\{-g_{K+1}(\boldsymbol{x})\}\right)^{-1}$$

  A-SM

$$\hat{p}(\mathrm{m} = \mathrm{y}|\mathbf{x}) = \phi_{\text{A-SM}}(g(\boldsymbol{x}), K+1) = \frac{\exp(g_{K+1}(\boldsymbol{x}))}{\sum_{y'=1}^{K+1} \exp(g_{y'}(\boldsymbol{x})) - \max_{y' \in \mathcal{Y}} \exp(g_{y'}(\boldsymbol{x}))}$$

- **Non-conformity score** for binary classification

$$s(\mathbf{x}, \mathrm{y}, \mathrm{m}; \hat{p}) = \begin{cases} 1 - \hat{p}(\mathrm{m} = \mathrm{y}|\mathbf{x}) & \text{if } \mathrm{m} = \mathrm{y} \\ \hat{p}(\mathrm{m} = \mathrm{y}|\mathbf{x}) & \text{if } \mathrm{m} \neq \mathrm{y} \end{cases}$$

- Given the empirical threshold $\hat{\tau}$, the **deferral set** can be constructed

$$C_r(\mathbf{x}; \hat{\tau}) = \begin{cases} \{0\} & \text{if } 1 - \hat{p}(\mathrm{m} = \mathrm{y}|\mathbf{x}) \geq 1 - \hat{\tau} \\ \{1\} & \text{if } \hat{p}(\mathrm{m} = \mathrm{y}|\mathbf{x}) \geq 1 - \hat{\tau} \\ \{0, 1\} & \text{otherwise} \end{cases}$$

Experiments

- Both OvA and A-SM improve upon the accuracy
- Coverage reduction is variable
- No clear superiority between the parameterizations

Coverage and efficiency of conformal prediction given confidence level $1 - \alpha = 90\%$

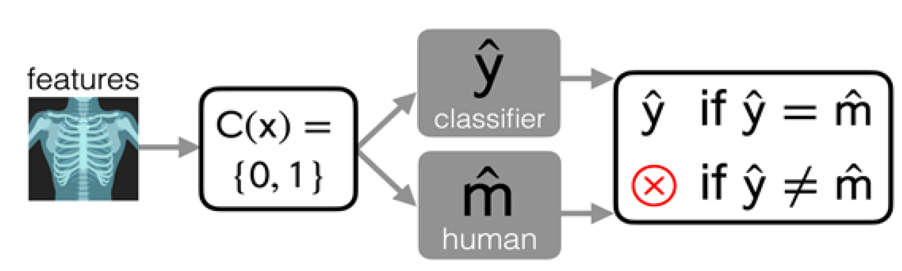
Dataset	Param.	Coverage (%)	Avg. Size
CIFAR-10	OvA	86.94 ± 0.86	1.07 ± 0.03
	A-SM	90.53 ± 0.56	1.37 ± 0.01
HAM10k	OvA	90.65 ± 0.63	1.25 ± 0.01
	A-SM	91.13 ± 0.58	1.28 ± 0.03
HateSpeech	OvA	90.35 ± 0.53	1.03 ± 0.03
	A-SM	90.67 ± 0.52	1.01 ± 0.01

L2D with Abstention and Consensus Prediction

	Param.	Method	Sys. Acc.	Ratio Deferred	Sys. Cov.
CIFAR-10	OvA	Base Model	84.71 ± 0.46	55.26 ± 1.76	100
		Abstention	86.72 ± 1.02	56.41 ± 2.30	92.14 ± 0.48
		Consensus	**86.79** ± 1.07	56.38 ± 2.31	93.32 ± 0.52
	A-SM	Base Model	84.01 ± 0.45	56.63 ± 3.73	100
		Abstention	87.05 ± 0.76	84.13 ± 4.56	62.53 ± 0.75
		Consensus	**87.58** ± 0.61	79.62 ± 4.31	67.57 ± 0.75
HAM10k	OvA	Base Model	82.1 ± 0.49	33.71 ± 2.39	100
		Abstention	**87.48** ± 0.51	35.91 ± 2.84	75.23 ± 1.40
		Consensus	85.72 ± 0.63	34.27 ± 2.52	88.39 ± 1.85
	A-SM	Base Model	78.92 ± 0.29	26.68 ± 3.07	100
		Abstention	**87.05** ± 0.87	28.11 ± 3.45	72.82 ± 1.19
		Consensus	84.76 ± 0.44	27.49 ± 3.16	84.48 ± 0.95
Hate Speech	OvA	Base Model	92.09 ± 0.07	42.41 ± 0.99	100
		Abstention	**92.28** ± 0.14	42.48 ± 0.96	99.38 ± 0.43
		Consensus	92.25 ± 0.13	42.42 ± 0.96	99.78 ± 0.22
	A-SM	Base Model	91.82 ± 0.32	67.91 ± 1.76	100
		Abstention	**91.88** ± 0.15	67.79 ± 1.74	99.16 ± 0.75
		Consensus	**91.88** ± 0.12	67.81 ± 1.73	99.65 ± 0.28

Conclusions

- The uncertainty in the rejector translates to safer decisions via two forms of selective prediction
- Conformal scoring functions shall be carefully parameterized



Consensus Prediction L2D Decision Making Workflow