Exploiting Human-AI Dependency for Learning to Defer

Anonymous Authors[1]

In our paper, we proposed a surrogate loss called DCE for the vanilla Learning to Defer (L2D) setting. Here, we further provide an extension for the proposed DCE loss to the setting of L2D with deferral costs and L2D with multiple experts.

1. L2D with Deferral Costs

In this part, we would like to introduce the extension of DCE with deferral **c**ost (DCEc). In L2D with deferral cost $c(\boldsymbol{x})$. The performance of an L2D system could be formulated as follows:

$$L_{01c}^{\perp}(f(\boldsymbol{x}), y, m) = \mathbb{I}_{f(\boldsymbol{x})\neq y}\mathbb{I}_{f(\boldsymbol{x})\neq\perp} + (\mathbb{I}_{m\neq y} + c(\boldsymbol{x}))\mathbb{I}_{f(\boldsymbol{x})=\perp},$$

Previous studies showed that the Bayes optimal classifier for L2D with deferral cost $c(\boldsymbol{x})$ can be formulated as follows:

$$f^*(x) = \begin{cases} \perp, & \mathbb{P}(Y = M|\boldsymbol{x}) > c(\boldsymbol{x}) + \max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), \\ \arg\max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), & \text{otherwise.} \end{cases}$$

Then, the generalized version of DCE for L2D with deferral costs could be defined as:

$$\begin{aligned} L_{\mathrm{DCEc}}^{\perp}(g(\boldsymbol{x}), y, m) = &-\mathbb{I}_{m\neq y}\log(\psi_{\mathcal{Y}^{\perp}}^{y}(g(\boldsymbol{x}))) \\ &- \mathbb{I}_{m=y}\big(\log(\psi_{\mathcal{Y}}^{y}(g(\boldsymbol{x}))) + \log(\psi_{\mathcal{Y}^{\perp}/q}^{\perp}(g(\boldsymbol{x})))\big) \\ &- c(\boldsymbol{x})\log(\psi_{\mathcal{Y}^{\perp}}^{q}(g(\boldsymbol{x}))), \end{aligned}$$

Where $q = \arg\max_{y\in\mathcal{Y}} g_y(\boldsymbol{x})$ denotes the predicted class of the model.

2. L2D with Multiple Experts

In this part, we show how to extend the DCE loss to L2D with multiple experts. We call this extended version DCEm. L2D with multiple experts could be formulated as minimizing the following loss function

$$\begin{aligned} L_{01m}^{\perp}(f(\boldsymbol{x}), y, m_1, m_2, \ldots, m_J) = &\ \mathbb{I}_{f(\boldsymbol{x})\neq y}\mathbb{I}_{f(\boldsymbol{x})\in\mathcal{Y}} \\ &+ \sum\nolimits_{j=1}^{J}(\mathbb{I}_{m_j\neq y})\mathbb{I}_{f(\boldsymbol{x})=\perp_j}, \end{aligned}$$

Where J is used to denote the number of experts, and m_j denotes the prediction made by the j-th expert. When $f(\boldsymbol{x}) = \perp_j$, the L2D system would defer the prediction to the j-th expert. Previous studies have shown that the Bayes optimal classifier for L2D with multiple experts should meet the following condition:

$$f^*(x) = \begin{cases} \perp_t, & \mathbb{P}(Y = M_t|\boldsymbol{x}) > \max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), \\ \arg\max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), & \text{otherwise,} \end{cases}$$

where $t = \arg\max_{0\leq j\leq J} \mathbb{P}(Y = M_j|\boldsymbol{x})$.

Then the extended DCE for L2D with multiple experts could be formulated as:

$$\begin{aligned} L_{\mathrm{DCEm}}^{\perp}(g(\boldsymbol{x}), y, m_1, m_2, \ldots, m_J) = &\\ &- \log(\psi_{\mathcal{Y}}^{y}(g(\boldsymbol{x}))) - \log(\psi_{\mathcal{Y}\cup\{j|\mathbb{I}[m_j\neq y]\}}^{\mathcal{Y}}(g(\boldsymbol{x}))) \\ &- \sum\nolimits_{j=1}^{J}\mathbb{I}[m_j = y]\log(\psi_{\mathcal{Y}\cup\perp,j/q}^{\perp_j} g(\boldsymbol{x})), \end{aligned}$$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.