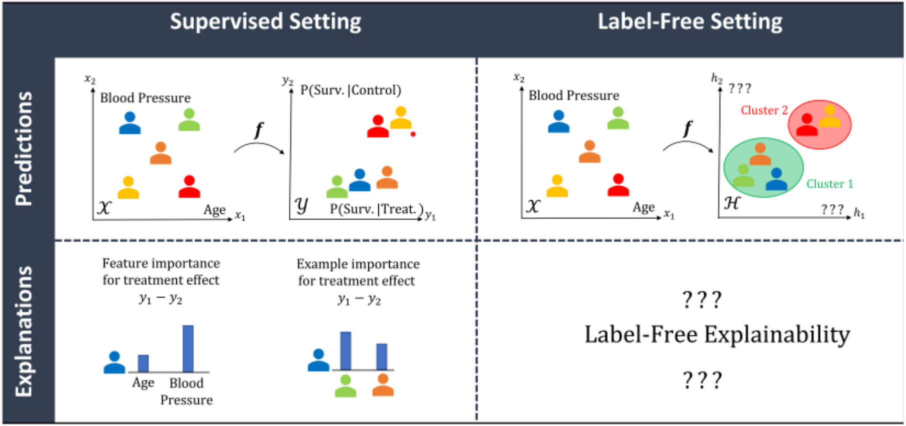
Label free explainability for Unsupervised Model

December 19, 2023

Seoul National University

Label-Free Explainability



Label-free Importance

1 Feature Importance

2 Example Importance

- Feature Importance With Labels.

$$b_i(\boldsymbol{f},\boldsymbol{x}) \equiv \sum_{j=1}^{d_y} f_j(\boldsymbol{x}) \cdot a_i(f_j,\boldsymbol{x}).$$

- Feature Importance With Label-Free

$$b_i(\boldsymbol{f},\boldsymbol{x}) \equiv a_i\left(g_{\boldsymbol{x}},\boldsymbol{x}\right)$$
$$g_{\boldsymbol{x}} : \mathcal{X} \to \mathbb{R} \text{ such that for all } \tilde{\boldsymbol{x}} \in \mathcal{X} :$$
$$g_{\boldsymbol{x}}(\tilde{\boldsymbol{x}}) = \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\tilde{\boldsymbol{x}}) \rangle_{\mathcal{H}} \, ,$$

- Label-Free Completeness.

$$\sum_{i=1}^{d_X} b_i(\boldsymbol{f}, \boldsymbol{x}) = \|\boldsymbol{f}(\boldsymbol{x})\|_{\mathcal{H}}^2 - b_0.$$

Label-free importance scores = sum to the black-box norm

Example Importance

- Loss-Based Example Importance

(Supervised setting)

In a supervised setting, this typically correspond to a couple z = (x, y) with an input $x \in X$ and a label $y \in Y$.

$$\boldsymbol{\delta}_{\boldsymbol{\theta}}^{n} L(\boldsymbol{z}, \boldsymbol{\theta}_*) \equiv L(\boldsymbol{z}, \boldsymbol{\theta}_*^{-n}) - L(\boldsymbol{z}, \boldsymbol{\theta}_*).$$

- Loss-Based Example Importance

  (Label-free setting)

Is it enough to drop the label and fix z = x in all the above expressions? No. -> Loss function can be different!

Example Importance

- Representation-Based Example Importance

(Supervised setting)

$$\boldsymbol{f}_l \circ \boldsymbol{f}_e : \mathcal{X} \to \mathcal{Y},$$

$\boldsymbol{f}_e : \mathcal{X} \to \mathcal{H}$ Inputs -> representations

$\boldsymbol{f}_l : \mathcal{H} \to \mathcal{Y}$ representations -> labels

$$\boldsymbol{f}_e(\mathcal{D}_{\text{train}}): \ \boldsymbol{f}_e(\boldsymbol{x}) \approx \sum_{n=1}^{N} w^n(\boldsymbol{x}) \cdot \boldsymbol{f}_e(\boldsymbol{x}^n).$$

$$w^n(\boldsymbol{x}) = \mathbf{1}\left[n \in \mathrm{KNN}(\boldsymbol{x})\right] \cdot \kappa\left[\boldsymbol{f}_e(\boldsymbol{x}^n), \boldsymbol{f}_e(\boldsymbol{x})\right]$$