[Review] Explainability as statistical inference

ICML 23 Poster accepted

Chanmoo Park

December 12, 2023

Seoul National University

Notation

- $\mathcal{X} = \prod_{d=1}^{D} \mathcal{X}_i$ be a $D-$dimensional feature space : Image
- $\mathcal{Y}$ be the target space : Label
- Consider two random variables, $\mathbf{X} = (X_1, \ldots, X_D)$ and $Y \in \mathcal{Y}$, following the true data generating distribution $p_{\text{data}}(x, y)$.
- We have N i.i.d realizations, $x^1, \ldots, x^N \in \mathcal{X}$ and labels $y^1, \ldots, y^N \in \mathcal{Y}$
- As a statistician, we mostly like to figure out

$$p_{\text{data}}(y \mid x),$$

which is called ”Statistical Inference”.

LEX : Latent Variable as Explanation

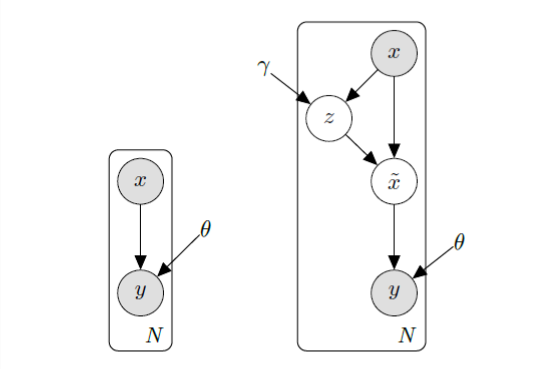
- In the standard predictive model, we usually approximate $p_{\text{data}}(y \mid x)$ using the parametric predictive model $p_\theta(y \mid x) = \Phi(y \mid f_\theta(x))$, maximizing

$$\max_\theta \mathcal{L}(\theta) = \max_\theta \sum_{n=1}^{N} \log p_\theta(y_n \mid x_n)$$

where $(\Phi(. \mid \eta))_{\eta \in H_1}$ is a parametric family.
- In this article, the latent variable Z is induced and maximize

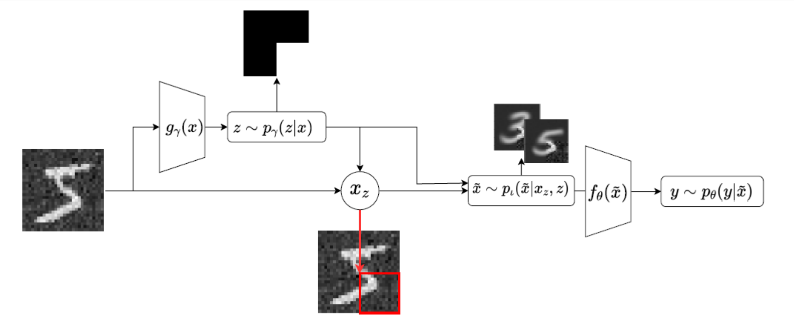
$$\max_{\theta,\gamma} \mathcal{L}(\theta,\gamma) = \max_{\theta,\gamma} \sum_{n=1}^{N} \log \left[\mathbb{E}_{z \sim p_\gamma(.\mid x_n)} p_\theta(y_n \mid x_n, z)\right]$$

LEX : Latent Variable as Explanation



- What is Z?
  $Z \in \{0,1\}^D$ corresponds to a subset of selected features. If $Z_d = 1$, then feature d is used by the predictor, and otherwise is not used by the predictor. We can call it as a "Mask" for the image.

LEX : Latent Variable as Explanation



- Neural net $g_\gamma : \mathcal{X} \to [0,1]^D$ is called "selector" with weight $\gamma \in \Gamma$
- $p_\gamma(z \mid x)$ is parametrized by g_γ (e.g. $p_\gamma(z \mid x) = \prod_{d=1}^{D} \mathcal{B}\left(z_d \mid g_\gamma(x)_d\right)$)
- $\tilde{X}$ is "masked" and "imputed" vector.

LEX is modular

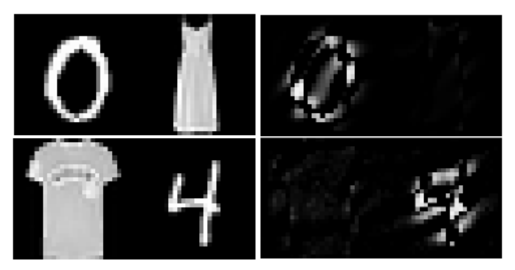
- Considering LEX framework, maximum likelihood problem is

$$\max_{\theta,\gamma} \sum_{n=1}^{N} \left[\log \mathbb{E}_{z\sim p_\gamma(.|x_n)} \mathbb{E}_{\tilde{x}\sim p_\iota(.|x_n,z)} p_\theta\left(y_n \mid \tilde{x}\right) - \lambda \mathrm{R}(z)\right].$$

Model	Sampling (p_γ) & Regularization (R)	Imputation (p_ι)	Training regime
L2X [1]	Subset Sampling & Implicit	0 imputation	Surrogate PostHoc
Invase [4]	Bernoulli & L1	0 imputation	Surrogate PostHoc
REAL-X [2]	Bernoulli & L1	Surrogate 0 imputation	Fixed θ In-Situ / Surrogate PostHoc
Rationale Selection [3]	Bernoulli, L1 & continuity regularization	Removed or imputed with padding value	Free In-Situ

Experiments : SP MNIST

- Switching Panels MNIST
  - Synthetic data using MNIST and FashionMNIST
  - Randomly sample a single image both from two datasets.
  - Arrange them in random order.
  - Target is the label of MNIST (number).



Experiments : CelebA

- CelebA dataset
  - Target : Smile or not