CORE: Self- and Semi-supervised Tabular Learning with COnditional REgularizations

Xintian Han [*,1]   Rajesh Ranganath[1]

[1]New York University

Self- and Semi-supervised Learning

Setup:

- A small labeled dataset $D_l = \{x_i, y_i\}_{i=1}^{N_l}$
- A larger unlabeled dataset $D_u = \{x_i\}_{i=N_l+1}^{N_l+N_u}$
- Utilize unlabeled data to help label data training.

Self-Supervised Learning:

- Create pretext tasks with target t
- Use an encoder enc to generate representation
- Use a predictor g to learn the target t from the representation
- With loss l_{self}, we minimize

$$\mathbb{E}_{x,t\sim p(x,t)} l_{self}(g(enc(x)), t).$$

Semi-Supervised Learning:

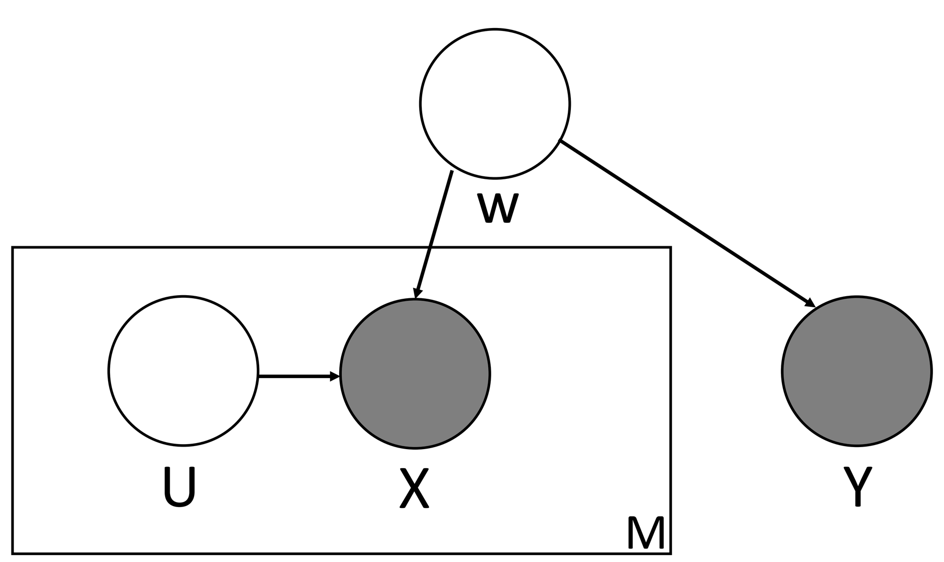
- Use $f(x)$ to model y
- With supervised loss l_{sup} and consistency loss l_c we minimize

$$\mathbb{E}_{x,y\sim p(x,y)} l_{sup}(f(x), y) + \beta \mathbb{E}_{x,x'\sim p(x'|x)} l_c(f(x), f(x'))$$

Our Assumptions

We assume the following data generation process:

- $W \sim p(W)$
- Individual noises U^j are independently drawn from $p(U^j)$
- Inputs $X^j \sim p(X^j|W, U^j)$
- $Y \sim P(Y|W)$



CORE

Knockoff Generator:

- For any index set S, the knockoff $\tilde{X}$ satisfies

$$(\tilde{X}, X) \overset{d}{=} (\tilde{X}, X)_{swap[S]}$$

- For $j \in S$, the swapping operation exchanges $\tilde{X}^j$ and X^j
- We use DDLK to generate knockoffs.

With knockoffs, CORE creates $\hat{X}$

$$\hat{X}(j)^j = \tilde{X}^j; \quad \hat{X}(j)^{-j} = X^{-j}$$

with j-th dimension replaced by samples from $p(X^j|X^{-j})$

Self-supervised CORE:

Denote dec as a decoder and ng as no gradient.
Self-supervised CORE minimizes

$$\mathbb{E}_X \|dec(enc(X)) - X\|_2^2 + \alpha \cdot \sum_{j=1}^{M} \mathbb{E}_{X,\hat{X}(j)} \|ng(dec)(enc(X)) - ng(dec)(enc(\hat{X}(j)))\|_2^2$$

Semi-supervised CORE:

Semi-supervised CORE minimizes

$$\mathbb{E}_{X,Y} l_{sup}(f(enc(X)), Y) + \beta \cdot \sum_{j=1}^{M} \mathbb{E}_{X,\hat{X}(j)} l_c(f(enc(X)), f(enc(\hat{X}(j))))$$

Why CORE does not memorize the noise?

- Conditional regularization

$$\sum_{i=1}^{M} \|ng(dec)(enc(X)) - ng(dec)(enc(\hat{X}(i)))\|_2^2$$

  can help us avoid memorizing the individual noise
- Individual noise is resampled in $\hat{X}$
- Memorize the individual noise, the conditional regularization term is large
- The conditional distribution still have information about W
- Memorizing W, the conditional regularization is not large

Experiments

Linear Simulation	Method	MSE
Supervised	Supervised Linear Regression	9444.25
	PCA	11.75
Self-Supervised	CORE	**1.17 ± 0.05**
	Denoising Auto-encoder	108.93 ± 6.80
	Context Encoder	1.49 ± 0.05
	VIME	104.07 ± 3.00

Higgs	Method	Accuracy
Supervised	4-layer perceptron	0.6055 ± 0.0041
	2-layer perceptron	0.6101 ± 0.0032
Self-supervised	CORE	**0.6692 ± 0.0055**
	Denoising Auto-encoder	0.6088 ± 0.0055
	Context Encoder	0.6096 ± 0.0154
	VIME	0.6675 ± 0.0056
Semi-supervised	CORE	0.6189 ± 0.0078
	VIME	0.6115 ± 0.0118
Self + Semi-supervised	CORE	0.6667 ± 0.0058
	VIME	0.6595 ± 0.0048

Mortality Prediction	Method	AUC
Supervised	4-layer perceptron	0.7837 ± 0.0029
	2-layer perceptron	0.7790 ± 0.0021
Self-supervised	CORE	0.7941 ± 0.0051
	Denoising Auto-encoder	0.7918 ± 0.0053
	Context Encoder	0.7806 ± 0.0042
	VIME	0.7914 ± 0.0028
Semi-supervised	VIME	0.7994 ± 0.0037
	CORE	0.7992 ± 0.0048
Self+Semi-supervised	VIME	0.7889 ± 0.0037
	CORE	0.7930 ± 0.0027