Incorporating treatment-outcome conditional mutual-information in estimating individual treatment effect

Ron Teichner

097400 - Introduction to causal inference
Final project

Project's objective

- Estimating ITE based on observational data
  (assumptions and notations as in [Shalit et al., 2017])

Project's objective

- Estimating ITE based on observational data (assumptions and notations as in [Shalit et al., 2017])
- Learn a model $\hat{y} \sim p_\theta(\hat{y} \mid t, x)$

Project's objective

- Estimating ITE based on observational data (assumptions and notations as in [Shalit et al., 2017])
- Learn a model $\hat{y} \sim p_\theta(\hat{y} \mid t, x)$
- Prediction errors might arise in regions where $\{x, t\}$ is sparse

Project's objective

- Estimating ITE based on observational data (assumptions and notations as in [Shalit et al., 2017])
- Learn a model $\hat{y} \sim p_\theta(\hat{y} \mid t, x)$
- Prediction errors might arise in regions where $\{x, t\}$ is sparse
- Resulting in ITE estimating errors

Project's objective

- Estimating ITE based on observational data (assumptions and notations as in [Shalit et al., 2017])
- Learn a model $\hat{y} \sim p_\theta(\hat{y} \mid t, x)$
- Prediction errors might arise in regions where $\{x, t\}$ is sparse
- Resulting in ITE estimating errors
- [Shalit et al., 2017] introduce a representation function $\Phi : X \to R$ and a balancing-regulation loss

Project's objective

- Estimating ITE based on observational data (assumptions and notations as in [Shalit et al., 2017])
- Learn a model $\hat{y} \sim p_\theta(\hat{y} \mid t, x)$
- Prediction errors might arise in regions where $\{x, t\}$ is sparse
- Resulting in ITE estimating errors
- [Shalit et al., 2017] introduce a representation function $\Phi : X \to R$ and a balancing-regulation loss
- We suggest the conditional mutual-information $I(\hat{Y}, T \mid X)$ as an alternative regulation

Loss function

The loss function incorporates negative log-likelihood and mutual-information:

$$L = L_o(\hat{y}, y) - \gamma I(\hat{y}; t \mid X) \tag{1}$$

The mutual-information:

$$\begin{aligned} I(\hat{y}; t \mid X) &= H(\hat{y} \mid X) - H(\hat{y} \mid t, X) \\ &= \mathrm{E}_{x,t \sim p_{data}(x,t)} \mathrm{E}_{\hat{y} \sim p_\theta(\hat{y} \mid x,t)} \left[\log p_\theta(\hat{y} \mid t, x)\right] \\ &- \mathrm{E}_{x \sim p_{data}(x)} \mathrm{E}_{t \sim p_{data}(t)} \mathrm{E}_{\hat{y} \sim p_\theta(\hat{y} \mid x,t)} \left[\log \sum_t p_\theta(\hat{y} \mid x, t) p(t)\right] \end{aligned} \tag{2}$$

Intuition and drawbacks

- Prediction error in y:

Intuition and drawbacks

- Prediction error in y:
  - The model ignores t: The Loss L_o is amplified by the low value of $I(\hat{y}; t \mid X)$

Intuition and drawbacks

- Prediction error in y:
  - The model ignores t: The Loss L_o is amplified by the low value of $I(\hat{y}; t \mid X)$
  - $I(\hat{y}; t \mid X)$ has a high value that masks a high value of L_o

Intuition and drawbacks

- Prediction error in y:
  - The model ignores t: The Loss L_o is amplified by the low value of $I(\hat{y}; t \mid X)$
  - $I(\hat{y}; t \mid X)$ has a high value that masks a high value of L_o
- Correct prediction of y:
  - Although the Neural-Net correctly predicts y it is motivated in increasing I which will result in wrong-prediction. To avoid this we need L_o to rise quicker than $I(\hat{y}; t \mid X)$.

Preliminary results

Consider the next example:

Idx	x	t	y	nRepetitions
1	0	1	1	3
2	1	1	0	3
3	0	0	0	3
4	1	0	1	1

Preliminary results

Consider the next example:

Idx	x	t	y	nRepetitions
1	0	1	1	3
2	1	1	0	3
3	0	0	0	3
4	1	0	1	1

We set $p_\theta(\hat{y} \mid t, x)$ as a Bernoulli distribution and we choose a model with parameters $\{a, b_x\}$:

$$\begin{aligned} p_\theta(\hat{y} \mid t, x) &= \theta\hat{y} + (1-\theta)(1-\hat{y}) \\ \theta &= \sigma(a + \mathrm{ReLU}(f_{b_x}(x,t)) + \mathrm{ReLU}(g_{b_x}(x,t))) \end{aligned} \tag{3}$$

Preliminary results

Model predictions:

Idx	x	t	y	$p_\theta(\hat{y} = 1 \mid x, t); \gamma = 0$	$p_\theta(\hat{y} = 1 \mid x, t); \gamma = 0.55$
1	0	1	1	0.9499	0.8656
2	1	1	0	0.1391	0.1715
3	0	0	0	0.1391	0.1715
4	1	0	1	0.3155	0.4692

ITE errors:

x	ITE(x=0) error; $\gamma = 0$	ITE(x=1) error; $\gamma = 0.55$
0	19%	30%
1	82%	70%

Project goals

- Develop a theoretical basis for the suggested approach **(enabling or disabling)**
- On a nominal dataset identify inputs x for which the model ignores t
- Run the proposed method and analyze the obtained results

References I

Shalit, U., Johansson, F. D., and Sontag, D. (2017).
Estimating individual treatment effect: generalization bounds and algorithms.
In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085, International Convention Centre, Sydney, Australia. PMLR.