Counterfactual Cross-Validation

Stable Model Selection Procedure for Causal Inference Models

Yuta Saito[1] and Shota Yasui[2]

[1]Tokyo Institute of Technology

[2]CyberAgent, Inc.

Accurate Causal Prediction = Accurate Decision Making

Causal prediction appears in all kinds of decision makings

- doctors want to decide whether they should administer a medication based on its causal effect on patients' survival rate

- advertisers want to decide whether they should advertise based on its causal effect on users' conversion rate

Optimal Decision Making Policy =
Treat when causal effect is larger than treatment cost

Rubin-Neyman Potential Outcome Framework

Basic notation

- **X**: Feature or Covariate Vector
- **T**: Binary Treatment Indicator
- **Y(1) / Y(0)**: Potential outcomes w/ or w/o a treatment
- **Y=TY(1)+(1-T)Y(0)**: Observed outcome,

Prediction Target: *Conditional Average Treatment Effect* (CATE)

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x],$$

Fundamental Problem in CATE Prediction

Counterfactual outcome makes it impossible to directly apply supervised machine learning to CATE predcition

Data	Feature	Treatment	Observed Outcome	Counterfactual Outcome	CATE
A	X_A	$T_A=1$	$Y_A(1)$	$Y_A(0)$	?
B	X_B	$T_B=0$	$Y_B(0)$	$Y_B(1)$	?
…	…	…	…	…	…

Recent advances in Treatment Effect Prediction

Broad applications progress the theoretical and empirical breakthroughs

- *Counterfactual Regression [Shalit et al. 2017]*
- *Propensity Dropout [Alaa et al. 2017]*
- *CEVAE [Louizos et al. 2017]*
- *CMGPs [Alaa&Van der Shaar. 2017]*
- *GAN-ITE [Yoon et al. 2018]*
- *SITE [Yao et al. 2018]*
- *ABCEI [Du et al. 2019]*
- *DragonNet [Shi et al. 2019]*

Is developing only prediction methods sufficient for applying CATE prediction to real-world?

Recent advances in Treatment Effect Prediction

Broad applications progress the theoretical and empirical breakthroughs

- *Counterfactual Regression [Shalit et al. 2017]*
- *Propensity Dropout [Alaa et al. 2017]*
- *CEVAE [Louizos et al. 2017]*
- *CMGPs [Alaa&Van der Shaar. 2017]*
- *GAN-ITE [Yoon et al. 2018]*
- *SITE [Yao et al. 2018]*
- *ABCEI [Du et al. 2019]*
- *DragonNet [Shi et al. 2019]*

Is developing only prediction methods sufficient for applying CATE pred[illegible]real-world?

Model Selection and Hyperparameter tuning are also essential

Our Focus: Model Selection and Hyperparamete Tuning

Model selection and hyperparameter tuning have not yet been fully investigated

Prior works focus on model evaluation, not model selection

- *Survey on heuristic metrics [Schuer et al. 2018]*
- *A meta-estimation method [Alaa & van der Schaar. 2019]*

We focus on developing model selection and hyperparameter tuning procedure used for CATE predictors

Goal in Model Selection and Hyperparameter Tuning

An observational validation set: $\mathcal{V} = \{X_i, T_i, Y_i\}_{i=1}^{n}$

A set of candidate CATE predictors: $\mathcal{M} = \{\widehat{\tau}_1, \ldots, \widehat{\tau}_{|\mathcal{M}|}\}$

We want to identify the best predictor among a set of candidates

$$\hat{\tau}_{best} = \arg\min_{\hat{\tau}\in\mathcal{M}} \mathcal{R}_{true}(\hat{\tau})$$

where $\mathcal{R}_{\text{true}}(\widehat{\tau}) = \mathbb{E}_X\left[(\tau(X) - \widehat{\tau}(X))^2\right]$ **expected MSE or PEHE**

Model Selection Approach for Building Evaluation Metric

We aimed to develop a metric that preserves the rank order of the ground-truth performance of candidate CATE predictors

$$\underbrace{\mathcal{R}_{\text{true}}(\hat{\tau}) \leq \mathcal{R}_{\text{true}}(\hat{\tau}')}_{\text{True Performance Ranking}} \Rightarrow \underbrace{\widehat{\mathcal{R}}(\hat{\tau}) \leq \widehat{\mathcal{R}}(\hat{\tau}')}_{\text{Ranking by Eval Metric}}, \ \forall \hat{\tau}, \hat{\tau}' \in \mathcal{M}.$$

Our approach is specific to model selection and might be easier than directly estimating the ground-truth performance

Research Questions

$$\mathcal{R}_{\text{true}}(\hat{\tau}) = \mathbb{E}_X\left[(\tau(X) - \hat{\tau}(X))^2\right]$$

We use the following **flexible** and **feasible** class of evaluation metrics

$$\underbrace{\widehat{\mathcal{R}}(\hat{\tau})}_{\text{Resulting Metric}} := \frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{\tilde{\tau}(X_i, T_i, Y_i)}_{\text{plug-in tau}} - \underbrace{\hat{\tau}(X_i)}_{\text{A CATE predictor}}\Big)^2$$

Resulting Metric   ***plug-in tau***   **A CATE predictor**

Research Questions.

1. **What is the ideal plug-in tau to identify the performance ranking?**
2. **How can we obtain it from observable validation data?**

Technical Contributions

- **We identify *two* conditions that the ideal plug-in tau should satisfy**
  - plug-in tau should be unbiased and has a small expectation of conditional variance
- **We propose a method to obtain a plug-in tau that satisfies the conditions well**
  - combining doubly robust estimation and a modified version of CFR (shalit et al. 2017)

The first condition for building a plug-in tau

Condition 1: A plug-in tau should be an unbiased estimator for the CATE

Reason (cf. Proposition 1)

Suppose $\mathbb{E}\left[\tilde{\tau}\left(X, T, Y^{obs}\right)|X\right] = \tau(X)$

$$\mathcal{R}_{true}(\hat{\tau}) \leq \mathcal{R}_{true}\left(\hat{\tau}'\right) \Longrightarrow \mathbb{E}\left[\widehat{\mathcal{R}}(\hat{\tau})\right] \leq \left[\widehat{\mathcal{R}}\left(\hat{\tau}'\right)\right]$$

The resulting evaluation metric identifies the true performance

But, we cannot take the expectation in finite samples..

In reality, we cannot take

the expectation

this motivates us to investigate

finite sample error in ranking

Decomposition of Evaluation Metric

$$\widehat{\mathcal{R}}(\hat{\tau}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \hat{\tau}(X_i))^2}_{\textit{converges to } \mathcal{R}_{true}(\hat{\tau})} \underbrace{- \frac{2}{n}\sum_{i=1}^{n}\left(\hat{\tau}(X_i) - \tau(X_i)\right)\left(\tilde{\tau}(X_i, T_i, Y_i) - \tau(X_i)\right)}_{\mathcal{W}\text{:source of uncertainty}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \tilde{\tau}(X_i, T_i, Y_i))^2}_{\text{independent of } \hat{\tau}}. \quad (5)$$

The second condition for building a plug-in tau

Condition 2: A plug-in tau should have a small expectation of conditional variance

Reason (cf. Theorem 2)

Variance of a finite sample error term

$$\mathbb{V}(\mathcal{W}) \leq 4C_{\max}n^{-1}\mathbb{E}[\mathbb{V}(\tilde{\tau}(X,T,Y)|X)]$$

Expectation of conditional variance of a plug-in tau

$$C_{\max} = \max_{i\in[n]}\left(\tau\left(x_i\right) - \hat{\tau}\left(x_i\right)\right)^2$$

A guildeline for obtaining a good plug-in tau

A guideline to build an evaluation metric for CATE predictors

Condition 2
$$\min_{\tilde{\tau}} \mathbb{E}\left[\mathbb{V}\left(\tilde{\tau}(X,T,Y) \mid X\right)\right],$$

Condition 1
$$\text{s.t. } \mathbb{E}[\tilde{\tau}(X,T,Y) \mid X] = \tau(X).$$

- Condition 1 ensures the identification of the performance ranking
- Condition 2 minimizes the finite sample error in ranking CATE predictors

Doubly Robust class for plug-in tau

We use a class of doubly robust plug-in tau

$$\tilde{\tau}_{DR}\left(X, T, Y; f_t\right)$$

$$:= \frac{T - e(X)}{e(X)(1 - e(X))}\left(Y - f_T(X)\right) + f_1(X) - f_0(X)$$

e(X): propensity score, f: regression function

Doubly robust plug-in tau is unbiased

The DR plug-in tau is unbiased regardless of a regression function

$$\mathbb{E}\left[\tilde{\tau}_{DR}(X,T,Y)\,|X\right] = \tau(X)$$

Doubly robust plug-in tau is unbiased

The DR plug-in tau is unbiased regardless of a regression function

$$\mathbb{E}\left[\tilde{\tau}_{DR}\left(X, T, Y\right) \middle| X\right] = \tau(X)$$

It satisfies the condition 1, and we can focus on condition 2

when deriving a "regression function" f

Optimizing variance as a loss function of f

Condition 2 with DR plug-in tau

$$\min_{f \in \mathcal{F}} \mathbb{E}_X \left[\mathbb{V} \left(\tilde{\tau}_{DR}(X, T, Y; f) | X \right) \right]$$

minimization of the expectation of the conditional variance

= use it as a loss function when training a “regression function” f

The expectation of conditional variance is counterfactual

A problem is that the expectation of conditional variance of the doubly robust plug-in tau cannot be optimized directly...

$$\mathbb{E}\left[\mathbb{V}\left(\tilde{\tau}_{DR}\left(X, T, Y; f_t\right) | X\right)\right]$$
$$= \zeta + \mathbb{E}_X\left[\left\{\sum_{t \in \mathcal{T}} \sqrt{w_t(X)}\left(f_t(X) - m_t(X)\right)\right\}^2\right]$$

where $m_t(x) := \mathbb{E}_{Y(t)}[Y(t)|X = x], \forall t \in \{0, 1\}$

Factual upper bound of the expectation of conditional variance

We optimize the factual version of the upper bound

by a weighted version of CFR (shalit et al. 2017)

$$\mathbb{E}_X\left[\left\{\sum_{t\in\mathcal{T}}\sqrt{w_t(X)}\,(f_t(X)-m_t(X))\right\}^2\right]$$
$$\leq 2\left(\epsilon_{F_1}^{w_1}(h,\Phi)+\epsilon_{F_0}^{w_0}(h,\Phi)+B_\Phi\,\mathrm{IPM}_G\left(p_t^\Phi,p_{1-t}^\Phi\right)-2\sigma^2\right)$$

weighted factual losses

regularization
(integral probability metric)

Our proposed model selection procedure

1. Estimate the propensity score (if needed)
2. Train a regression function by “Weighted CFR”
3. Calculate the doubly robust plug-in tau for a given validation set
4. Calculate evaluation metric for every candidate CATE predictor
5. Deploy a CATE predictor having the best performance in our metric

IHDP Dataset

We use **IHDP dataset** [Hill 2011.]

- contains the ground-truth CATE, enabling the **evaluation of evaluation metrics**
- 747 samples with 25 features
- Used in many experiments on CATE prediction methods

$$\mathcal{M} = \{\widehat{\tau}_1, \ldots, \widehat{\tau}_{|\mathcal{M}|}\}$$

Experimental Procedure

1. Construct a set of candidate CATE predictors (|M|=25)
2. Split the IHDP data into training/validation/test sets
3. Train 25 candidate CATE predictors on the training set
4. **Evaluate predictors using the validation set and evaluation metrics**
5. Calculate the ground-truth performance using the test set
6. Evaluate evaluation metrics

Evaluation Metrics for Evaluation Metrics

1. **Spearman Rank Correlation**

   Rank correlation between the model ranking by the evaluation metric values and the ground-truth performance

2. **Regret in model selection**

   The performance of a CATE predictor selected by each metric

Experimental Results

Our procedure stably ranks the performance and selects the best one

	Methods	Rank Correlation		Regret	
		Larger value is better		Lower value is better	
		Mean ±StdErr	Worst-Case	Mean ±StdErr	Worst-Case
baselines	IPW	0.195 ±0.039	-0.749	1.032 ±0.100	6.779
	τ-risk	0.312 ±0.030	-0.553	1.392 ±0.130	7.884
	Plug-in	0.914 ±0.006	0.591	0.073 ±0.012	0.780
proposed	CF-CV (ours)	**0.921** ±0.005	**0.666**	**0.066** ±0.012	**0.562**

Thank you for Listening!

website: https://usaito.github.io/
email: saito.y.bj at m.titech.ac.jp