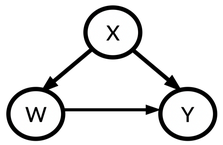
Empirical Analysis of Model Selection for Heterogeneous Causal Effect Estimation

Divyat Mahajan, Ioannis Mitliagkas, Brady Neal, & Vasilis Syrgkanis

Contributions

- We perform a comprehensive empirical study over **78 datasets** to benchmark **34 surrogate metrics** for conditional average treatment effect (CATE) model selection, where the model selection task is made challenging by training **415 CATE estimators** per dataset.
- We introduce novel surrogate metrics as well as novel strategies of **two-level model selection** and **causal ensembling** for CATE model selection.

Background: CATE Estimation



X : Covariates
W : Binary Treatments
$Y(w)$: Potential Outcomes

- **CATE:** $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$
- **Meta-Learners** estimate $\tau(x)$ as a function of nuisance models $\hat{\eta} = (\hat{\mu}, \hat{\pi})$
  - Potential Outcome Model: $\hat{\mu}_w(x) = \mathbb{E}[Y|W = w, X = x]$
  - Propensity Model: $\hat{\pi}_w(x) = \mathbb{P}(W = w|X = x)$
- Indirect Meta-Learner:
  - **T-Learner:** $\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$
- Direct Meta-Learner:
  - **DR-Learner:** $\hat{\tau}_{DR} := \hat{f}_{DR} = \arg\min_{f \in F} \sum_{\{x,w,y\}} \left(y^{DR}(\hat{\eta}) - f(x)\right)^2$

Motivation: CATE Model Selection

Precision of Heterogeneous Effects (PEHE): $L(\hat{\tau}) = \mathbb{E}_X[(\hat{\tau}(X) - \tau(X))^2]$

Input CATE Estimate — True CATE

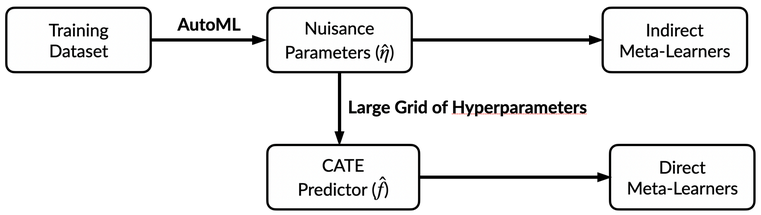
- True CATE ($\tau(X)$) is unknown as we don't observe both potential outcomes
- Cannot perform cross-validation unlike machine learning!

Surrogate PEHE: $L(\hat{\tau}) = \mathbb{E}_X[(\hat{\tau}(X) - \tilde{\tau}(X))^2]$

Input CATE Estimate — Proxy CATE

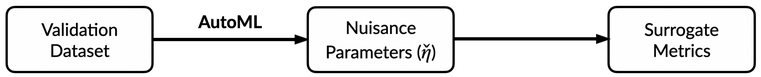
- **Surrogate Metrics**: Estimate true CATE on the validation set ($\tilde{\tau}(X)$)
- Different strategies for estimating $\tilde{\tau}(X)$ lead to different surrogate metrics
  - **T Score:** $\tilde{\tau}_T(x) = \check{\mu}_1(x) - \check{\mu}_0(x)$
  - **DR Score:** $\tilde{\tau}_{DR} = y^{DR}(\check{\eta}) = y_1^{DR}(\check{\eta}) - y_0^{DR}(\check{\eta})$ where $y_w^{DR}(\check{\eta}) = \check{\mu}(x, w) + \frac{y - \check{\mu}(x,w)}{\check{\pi}_w(x)}$
- We have a poor understanding about the relative advantages/disadvantages of surrogate metrics!

CATE Estimators in our study



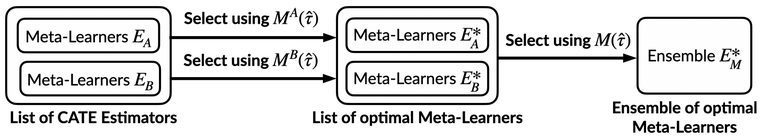
- We use AutoML to tune nuisance parameters ($\hat{\eta}$) of meta-learners.
- We allow for diverse collection of estimators for each direct meta-learner to make the task of CATE model selection more challenging.

Surrogate Metrics in our study



- Prior works estimate the nuisance parameters ($\check{\eta}$) of surrogate metrics using small grid search for hyperparameters.
- We use AutoML to have low bias in estimating the nuisance parameters ($\check{\eta}$) of surrogate metrics, which enhances their model selection ability.

Proposed Two-Level Model Selection Strategy



- Prior works selected over the entire population of CATE estimators in a single step using a surrogate metric.
- We propose a novel two-step approach that carefully tunes the hyperparameters of Meta-Learners to aid the surrogate metics in model selection.
  - **Step 1:** Perform intra Meta-Learner selection using surrogate metric based on the respective Meta-Learner.
  - **Step 2:** Select across optimal Meta-Learners from the first step using the input surrogate metric.

Emperiment Setup

- **Datasets:** 75 synthetic (ACIC 2016 benchmark) and 3 realistic datasets
- **CATE Estimators:** Large collection of both direct and indirect meta-learners trained for each dataset.
- **Surrogate Metrics.** Comprehensive collection of prior metrics as well as novel metrics like adaptive propensity adjustment, TMLE, etc.

Results: Single-Level Strategy

Metric	ACIC 2016	LaLonde CPS	LaLonde PSID	TWINS
Value Score	1.05e+7 (4.31e+6)	6.63 (5.52)	**0.48 (0.06)**	0.57 (0.15)
S Score	0.95 (0.02)	0.90 (0.04)	0.74 (0.04)	**0.29 (0.05)**
T Score	**0.56 (0.02)**	**0.16 (0.03)**	**0.42 (0.03)**	**0.31 (0.05)**
X Score	**0.56 (0.02)**	**0.16 (0.03)**	**0.41 (0.03)**	0.35 (0.06)
R Score	4.0 (0.11)	0.83 (0.04)	0.67 (0.03)	0.60 (0.11)
Influence Score	1455.75 (1439.46)	0.95 (0.04)	0.80 (0.02)	1.08 (0.1)
DR T Score	**0.56 (0.02)**	**0.16 (0.02)**	**0.41 (0.03)**	**0.32 (0.07)**
DR Switch T Score	**0.56 (0.02)**	**0.16 (0.03)**	**0.41 (0.03)**	**0.28 (0.05)**
TMLE T Score	**0.64 (0.03)**	**0.16 (0.03)**	**0.42 (0.03)**	**0.31 (0.05)**
Cal DR T Score	3.45 (0.11)	**0.17 (0.03)**	**0.42 (0.03)**	**0.21 (0.03)**
Qini DR T Score	1.32 (0.07)	2.87 (1.53)	0.57 (0.05)	2.08e+7 (1.90e+7)

Table 1: Normalized PEHE of the **best estimators** chosen by each metric with the **single-level model selection strategy**; results report the mean (standard error) across 20 seeds and also across datasets for the ACIC 2016 benchmark. **Lower value is better.**

- Plug-in surrogate metrics (T/X Score) are optimal (Thanks to AutoML!)

Results: Two-Level Strategy

Metric	ACIC 2016	LaLonde CPS	LaLonde PSID	TWINS
Value Score	3.97 (1.98)	**0.34 (0.09)**	**0.43 (0.03)**	**0.21 (0.03)**
S Score	0.93 (0.02)	0.90 (0.04)	0.75 (0.04)	**0.21 (0.03)**
T Score	**0.56 (0.02)**	**0.16 (0.03)**	**0.41 (0.03)**	**0.21 (0.03)**
X Score	**0.56 (0.02)**	**0.16 (0.03)**	**0.41 (0.03)**	**0.21 (0.03)**
R Score	3.88 (0.11)	0.86 (0.03)	0.62 (0.03)	**0.21 (0.03)**
Influence Score	3.26 (0.1)	0.93 (0.04)	0.77 (0.03)	**0.16 (0.02)**
DR T Score	**0.56 (0.02)**	**0.16 (0.02)**	**0.41 (0.03)**	**0.21 (0.03)**
DR Switch T Score	**0.56 (0.02)**	**0.16 (0.03)**	**0.41 (0.03)**	**0.21 (0.03)**
TMLE T Score	**0.61 (0.03)**	**0.16 (0.03)**	**0.42 (0.03)**	**0.21 (0.03)**
Cal DR T Score	**0.62 (0.02)**	**0.19 (0.04)**	**0.42 (0.03)**	**0.22 (0.03)**
Qini DR T Score	**0.58 (0.02)**	**0.14 (0.03)**	0.52 (0.03)	**0.24 (0.04)**

Table 2: Normalized PEHE of the **best estimators** chosen by each metric with the **two-level model selection strategy**; results report the mean (standard error) across 20 seeds and also across datasets for the ACIC 2016 benchmark. **Lower value is better.**

- Strict improvement over single-level selection strategy! Better performance in **28.7 %** cases, otherwise statistically indistinguishable.