Out-of-Distribution Generalization of Uplift Models

Leo Guelman

Head Statistician | Royal Bank of Canada

ECML 2022
Uplift Modeling Workshop

Introduction

- Most uplift modeling methods assume i.i.d. data.
- In most practical settings this assumption is violated:
  - Data changes in space (e.g., different cities, different labs).
  - Data changes in time (e.g., same space at different points in time).
- These changes might induce non-robust uplift models: models that fail to generalize under changing data conditions.
- Active research area in out-of-distribution (OOD) generalization for supervised learning, but not much attention in uplift modeling literature.
- Goal of this talk is to provide an overview of the problem, and sketch a proposed approach (in progress).

Motivating Example

Healthcare: Uplift model developed based on a randomized controlled trial composed of lung cancer patients from different medical clinics/locations in Hospital 1 identifies for which patients a treatment is more effective.

Can this model be safely used to predict treatment effectiveness on patients in Hospital 2? (patients coming from a different set of clinics/locations)

Motivating Example

Healthcare: Uplift model developed based on a randomized controlled trial composed of lung cancer patients from different medical clinics/locations in Hospital 1 identifies for which patients a treatment is more effective.
Can this model be safely used to predict treatment effectiveness on patients in Hospital 2? (patients coming from a different set of clinics/locations)

Program Evaluation: Uplift model developed based on experimental data from schools in California identifies for which students a new educational program is more effective (improvement in grades).

Can this model be reliably used in other US states?

Motivating Example

Healthcare: Uplift model developed based on a randomized controlled trial composed of lung cancer patients from different medical clinics/locations in Hospital 1 identifies for which patients a treatment is more effective.
Can this model be safely used to predict treatment effectiveness on patients in Hospital 2? (patients coming from a different set of clinics/locations)

Program Evaluation: Uplift model developed based on experimental data from schools in California identifies for which students a new educational program is more effective (improvement in grades).

Can this model be reliably used in other US states?

Marketing: A company develops an uplift model to estimate which customers will most likely respond to a price incentive based on experimental data.

Can we rely on this model to predict price-elasticities on a population of clients whose characteristics differ from the study distribution?

Conditional Average Treatment Effect (CATE)

- We refer to Uplift models as CATE models.
- Let T be a binary variable representing treatment, $Y \in \mathbb{R}$ be the observed outcome, and $X \in \mathbb{R}^p$ a vector of covariates.
- Using the Neyman/Rubin Potential Outcome notation, the CATE $\tau(x)$ is defined as the following estimand:

$$\tau(x) \triangleq \mathbb{E}[Y_i(1) - Y_i(0)|X = x].$$

- Alternatively, using Pearl's *do*-operator, we can equivalently define the CATE as

$$\tau(x) \triangleq \mathbb{E}[Y_i|do(T = 1), X = x] - \mathbb{E}[Y_i|do(T = 0), X = x].$$

- We will occasionally work with the full distribution rather than just the means:

$$P(y|do(t), x).$$

The Setting

- Let $\Psi := \langle G, P_{(Y,X,T)} \rangle$ be *probabilistic causal model* (PCM), where G is a causal graph, and $P(Y, X, T)$ a joint distribution over the variables in G.

The Setting

- Let $\Psi := \langle G, P_{(Y,X,T)} \rangle$ be *probabilistic causal model* (PCM), where G is a causal graph, and $P(Y, X, T)$ a joint distribution over the variables in G.
- Assume $P(Y, X, T)$ satisfies the *causal Markov assumption*

$$P(Y, X, T) = P(Y|\mathrm{PA}_Y)P(T|\mathrm{PA}_T)\prod_{j=1}^{p} P(X_j|\mathrm{PA}_{X_j}),$$

where $P(V|\mathrm{PA}_V)$ represents the *causal mechanism* for variable V, and we assume it remains invariant to interventions in variables other than V (a.k.a. *modularity assumption*).

The Setting

- Let $\Psi := \langle G, P_{(Y,X,T)} \rangle$ be *probabilistic causal model* (PCM), where G is a causal graph, and $P(Y, X, T)$ a joint distribution over the variables in G.
- Assume $P(Y, X, T)$ satisfies the *causal Markov assumption*

$$P(Y, X, T) = P(Y|\text{PA}_Y)P(T|\text{PA}_T)\prod_{j=1}^{p} P(X_j|\text{PA}_{X_j}),$$

  where $P(V|\text{PA}_V)$ represents the *causal mechanism* for variable V, and we assume it remains invariant to interventions in variables other than V (a.k.a. *modularity assumption*).
- Let Π^{tot} be a collection of 'environments' (e.g., different cities, labs, perturbations, etc.) such that for each environment

$$\pi \in \Pi^{\text{tot}}, (Y^\pi, X^\pi, T^\pi) \sim P^\pi.$$

The Setting

- Let $\Psi := \langle G, P_{(Y,X,T)} \rangle$ be *probabilistic causal model* (PCM), where G is a causal graph, and $P(Y, X, T)$ a joint distribution over the variables in G.
- Assume $P(Y, X, T)$ satisfies the *causal Markov assumption*

$$P(Y, X, T) = P(Y|\mathrm{PA}_Y)P(T|\mathrm{PA}_T)\prod_{j=1}^{p} P(X_j|\mathrm{PA}_{X_j}),$$

  where $P(V|\mathrm{PA}_V)$ represents the *causal mechanism* for variable V, and we assume it remains invariant to interventions in variables other than V (a.k.a. *modularity assumption*).
- Let Π^{tot} be a collection of 'environments' (e.g., different cities, labs, perturbations, etc.) such that for each environment

$$\pi \in \Pi^{\text{tot}}, (Y^{\pi}, X^{\pi}, T^{\pi}) \sim P^{\pi}.$$

- The causal mechanisms $P(X_j|\mathrm{PA}_{X_j})$ and $P(T|\mathrm{PA}_T)$ are allowed to change between environments, but assume no changes in $P(Y|\mathrm{PA}_Y)$ or the graph G.

The Problem

- At training time, we observe n_k samples

$$(Y_i^{\pi_k}, X_i^{\pi_k}, T_i^{\pi_k})_{i=1}^{n_k} \sim \; P^{\pi_k}\left(Y^{\pi_k}, X^{\pi_k} | do(T := \text{Bernoulli}(0.5)\right),$$

from a subset $\{\pi_{k=1}, \ldots, \pi_K\} = \Pi^{\text{obs}} \subseteq \Pi^{\text{tot}}$ of the environments, where $P^{\pi_k}\left(Y^{\pi_k}, X^{\pi_k} | do(T := \text{Bernoulli}(0.5)\right)$ represents an *interventional distribution* obtained by randomizing T.

The Problem

- At training time, we observe n_k samples

$$(Y_i^{\pi_k}, X_i^{\pi_k}, T_i^{\pi_k})_{i=1}^{n_k} \sim \; P^{\pi_k}\left(Y^{\pi_k}, X^{\pi_k} | do(T := \text{Bernoulli}(0.5)\right),$$

  from a subset $\{\pi_{k=1}, \ldots, \pi_K\} = \Pi^{\text{obs}} \subseteq \Pi^{\text{tot}}$ of the environments, where $P^{\pi_k}\left(Y^{\pi_k}, X^{\pi_k} | do(T := \text{Bernoulli}(0.5)\right)$ represents an *interventional distribution* obtained by randomizing T.

- At test time, we want to predict the CATE $\tau(x)$ from a potentially unseen environment $\pi_* \in \Pi^{\text{tot}} \setminus \Pi^{\text{obs}}$, from samples drawn from $\sim \; P^{\pi_*}(Y^{\pi_*}, X^{\pi_*}, T^{\pi_*})$.

The Problem

- At training time, we observe n_k samples

$$(Y_i^{\pi_k}, X_i^{\pi_k}, T_i^{\pi_k})_{i=1}^{n_k} \sim \; P^{\pi_k}\left(Y^{\pi_k}, X^{\pi_k} | do(T := \text{Bernoulli}(0.5)\right),$$

  from a subset $\{\pi_{k=1}, \dots, \pi_K\} = \Pi^{\text{obs}} \subseteq \Pi^{\text{tot}}$ of the environments, where $P^{\pi_k}\left(Y^{\pi_k}, X^{\pi_k} | do(T := \text{Bernoulli}(0.5)\right)$ represents an *interventional distribution* obtained by randomizing T.

- At test time, we want to predict the CATE $\tau(x)$ from a potentially unseen environment $\pi_* \in \Pi^{\text{tot}} \setminus \Pi^{\text{obs}}$, from samples drawn from $\sim \; P^{\pi_*}(Y^{\pi_*}, X^{\pi_*}, T^{\pi_*})$.

- The goal is to build a CATE estimator $\hat{\tau}(x)$ that minimizes the expected loss

$$\mathbb{E}_{(Y^{\pi_*}, X^{\pi_*}, T^{\pi_*}) \sim P^{\pi_*}} \ell(\hat{\tau}(x), \tau),$$

  based on experimental data from the source environments Π^{obs}.

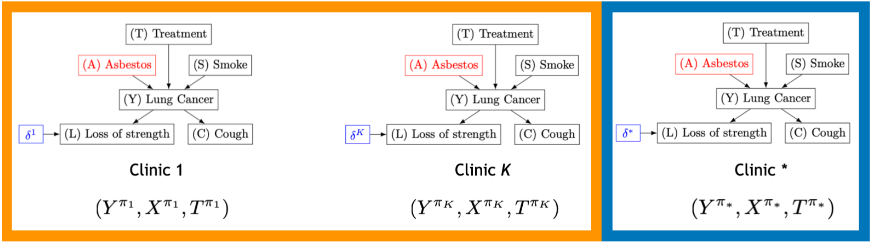
Illustration

Observed

$$\{\pi_1, \ldots, \pi_K\} = \Pi^{\text{obs}} \subseteq \Pi^{\text{tot}}$$

Unobserved

$$\pi_* \in \Pi^{\text{tot}} \setminus \Pi^{\text{obs}}$$



- We let $\delta^k, k = \{1, \ldots, K\}$, be a set of K auxiliary variables which turn G into an augmented graph G_δ.
- An edge $\delta^k \to X$ denotes a change in the causal mechanism that generates X.
- In this example, the causal mechanism for *Loss of Strength* changes between environments:

$$P^{\pi_i}(\text{Loss of strength}|\text{Cancer}) \neq P^{\pi_j}(\text{Loss of strength}|\text{Cancer}) \ \forall i \neq j \in K.$$

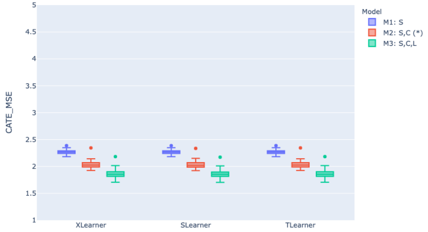
Illustration (cont'd)

Suppose the augmented Causal Graphs G_δ above are induced by the following Structural Causal Model (SCM):

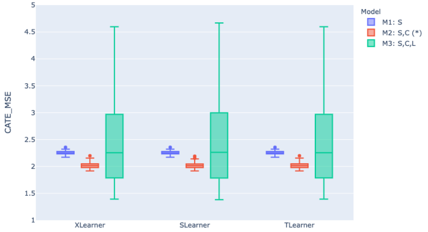
$$
\begin{aligned}
A &:= N_A \\
S &:= N_S \\
Y &:= A + S + T + 1.5 \times A \times T + 0.5 \times S \times T + N_Y \\
L &:= \delta \times Y + N_L \\
C &:= 0.3 \times Y + N_C \\
\delta^k,\ k = \{1, \ldots, K\} &\sim \mathrm{U}(0, 1) \\
\delta^* &\sim \mathrm{U}(-1, 1) \\
T &\sim \text{Bernoulli}(0.5) \\
N_j &\sim \mathcal{N}(0, 1)
\end{aligned}
$$

Illustration (cont'd)

In-Distribution (Π^{obs})



Out-of-Distribution ($\pi_* \in \Pi^{\text{tot}} \setminus \Pi^{\text{obs}}$)



- Inclusion of *Cough* (C) is beneficial for generalization performance.
- Inclusion of *Loss of Strength* (L) is harmful for OOD generalization performance.
- Our proposed approach selects *Smoke* (S) and *Cough* (C) as input features in CATE estimation (**Model 2**).

Key Challenges for Building Robust CATE Estimators

- Recall the goal is to build a CATE estimator $\hat{\tau}(x)$ that minimizes the expected loss

$$\mathbb{E}_{(Y^{\pi*}, X^{\pi*}, T^{\pi*}) \sim P^{\pi*}} \ell(\hat{\tau}(x), \tau), \tag{1}$$

based on experimental data from the source environments Π^{obs}.

- We have two problems with 1:
  1. No data from $P^{\pi*}$ are available at training time.
  2. τ_i is not observed for an individual (due to the fundamental problem of causal inference).

Invariant CATE Features

Definition (Invariance)

A set of features X^{I}, $\mathrm{I} \subseteq \{1, \ldots, p\}$, for estimating the CATE $P(y|do(t), x)$ from Π^{obs} is invariant if for all $\pi_i, \pi_j \in \Pi^{\text{obs}}$ and for all $x \in \mathbb{X}$

$$P^{\pi_i}(y|do(t), x^{\mathrm{I}}) = P^{\pi_j}(y|do(t), x^{\mathrm{I}}).$$

Invariant sets are not unique. We let Ω be the collection of invariant CATE feature sets.

Proposed CATE Estimator

Assumptions

A1. There exists an invariant set of CATE features X^{I} (i.e., satisfying the invariance property defined above).

A2. The invariance property also hold for unseen environments $\pi^* \in \Pi^{\text{tot}} \setminus \Pi^{\text{obs}}$.

A3. The conditional distribution $P(y|do(t), x)$ is linear (this addresses potential issues with non-overlapping feature supports).

We propose linear CATE estimators $\hat{\tau}(x^{\mathrm{I}*}; \theta) = \theta' x^{\mathrm{I}*}$, using an invariant set of features $X^{\mathrm{I}*} \in \ \Omega$, identified from Π^{obs}.

Specifically, our proposed estimator with squared-error loss is given by

$$\theta^*(x^{\mathrm{I}*}) = \arg\min_{\theta} \frac{1}{|\mathcal{V}|} \sum_{i \in |\mathcal{V}|} \left(\hat{\tau}_i(x^{\mathrm{I}}; \theta) - \check{\tau}_i \right)^2 \ \ \forall \ x^{\mathrm{I}} \in \Omega,$$

where $\check{\tau}_i$ is a plug-in estimate of τ_i estimated using data from a validation set $\mathcal{V}$ (Shuler et al., 2018). This attempts to circumvent the issue of unobserved τ.

Robustness

Theorem (Adversarial[a])

[a] Adapted from Rojas-Carulla et al. (2018).

Consider $(Y^{\pi_1}, X^{\pi_1}, T^{\pi_1}) \sim P^{\pi_1}, \ldots, (Y^{\pi_K}, X^{\pi_K}, T^{\pi_K}) \sim P^{\pi_K}$ *and an invariant set of CATE features* $\mathrm{I}*$ *satisfying A1-A3. The proposed estimator satisfies the following optimality statement over a set of distributions:*

$$\theta^*(\mathrm{I}*) \in \arg\min_{\theta} \sup_{P^{\pi_*} \in \mathcal{P}} \mathbb{E}_{(Y^{\pi_*}, X^{\pi_*}, T^{\pi_*}) \sim P^{\pi_*}} \ell\Big(\hat{\tau}(x;\theta), \check{\tau}\Big).$$

Here $\mathcal{P}$ represents a family of distributions composed of all interventions on any subset of variables excluding Y.

Learning Invariant CATE Features

Algorithm 1 Invariant CATE Features

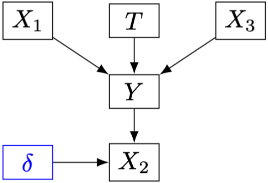
Inputs: Samples $(y_i^{\pi_k}, x_i^{\pi_k}, t_i^{\pi_k})_{i=1}^{n_k}$ from each environment $\pi_k, k \in \{1, \ldots, K\}$, and threshold α^c for independent test.
Output: Estimated invariant CATE feature set $X^{\mathrm{I}*}$.

Set MSE=[ ], $\mathrm{I} = [\ ]$.
Create pseudo-outcome $W = 2YT$ with $T = \pm 1$. (See Tian et al., 2014).
for $\mathrm{I} \subseteq \{1, \ldots, p\}$ **do**
    Linearly regress W on X^{I} and compute the residuals $R_\theta^k = W^k - \theta' X^k$, $k \in \{1, \ldots, K\}$ on a validation set $\mathcal{V}$.
    Test for equality in distributions of residuals across environments

$$H_0 = R_\theta^1 \stackrel{\mathrm{d}}{=} R_\theta^2 \stackrel{\mathrm{d}}{=} \ldots \stackrel{\mathrm{d}}{=} R_\theta^K,$$

    and the corresponding p-value α.
    if $\alpha > \alpha^c$ **then**
        Compute $\hat{\ell}_\theta = \frac{1}{|\mathcal{V}|} \sum_{i \in |\mathcal{V}|} \left(\hat{\tau}(x^{\mathrm{I}}; \theta) - \check{\tau}\right)^2$.
        I.append(X^{I}), MSE.append($\hat{\ell}_\theta$).
    end if
end for
Set $X^{\mathrm{I}*} = \mathrm{I}[\arg\min\,[\mathrm{MSE}]]$.

Numerical Experiments



$$
\begin{aligned}
X_1, X_3 \sim & N(0,1) \\
Y := & \alpha_1 X_1 + \alpha_2 X_3 + \alpha_3 T + \alpha_4 X_1 \times T + \\
& \alpha_5 X_3 \times T + N_Y \\
X_2 := & \delta_k Y + N_{X_2} \\
N_Y \sim & N(0, 1.5) \\
\delta^k, \delta* \sim & \mathrm{U}(0,1), \;\; k = \{1, \ldots, K\} \\
N_{X_2} \sim & N(0, 0.1) \\
\alpha_1, \ldots, \alpha_5 \sim & \mathrm{U}(-1, 2.5)
\end{aligned}
$$

Numerical Experiments: Results

Table: CATE MSE: Mean and (SE)

N	$K=3$	$K=6$	$K=10$	$K=15$	$K=20$
Pool Method					
200	2.107	1.588	1.391	1.298	1.025
	(0.419)	(0.379)	(0.378)	(0.330)	(0.245)
400	1.807	1.222	0.935	1.109	1.140
	(0.345)	(0.179)	(0.120)	(0.159)	(0.196)
800	1.535	1.461	0.958	1.445	1.152
	(0.203)	(0.234)	(0.133)	(0.334)	(0.331)
1200	1.385	1.438	1.222	1.047	0.924
	(0.233)	(0.414)	(0.223)	(0.158)	(0.143)
Proposed Method					
200	4.502	1.932	1.864	2.144	1.552
	(2.078)	(0.484)	(0.473)	(0.542)	(0.473)
400	1.851	0.598	0.573	0.490	0.206
	(0.712)	(0.278)	(0.194)	(0.211)	(0.116)
800	1.618	0.538	0.142	0.077	0.073
	(0.536)	(0.296)	(0.094)	(0.049)	(0.051)
1200	0.113	0.059	0.258	0.102	0.003
	(0.076)	(0.042)	(0.162)	(0.070)	(0.002)

Takeaways

- We relax the i.i.d. assumption from CATE estimation methods.
- We propose a method to select CATE models that are robust under a family of distributional changes in the data.
- The proposed method shows positive results on a limited number of simulation scenarios.