Meta Learners

CATE / HTE

$$\tau(x) = E[Y(1) - Y(0) \mid X = x]$$

where

$$\mu_0(x) = E[Y(0) \mid X = x] \qquad \mu_1(x) = E[Y(1) \mid X = x]$$

$$\Rightarrow \quad \tau(x) = \mu_1(x) - \mu_0(x)$$

Eval metric: Expected Mean Square Error (EMSE)

Meta learners

"Meta" → able to choose different estimators based on the type of problem. Domain knowledge plays a role here

- Use ML methods to estimate causal effect (CATE)
- General approach: find $\mu_1(x)$, $\mu_0(x)$ then compute CATE

S-learner (single learner)

- Treats the treatment variable as another covariate when training the model.

$$\mu(x, w) = E[Y^{obs} \mid X = x, W = w]$$

- Train single model that predicts $\mu(x,w)$ using training variable as a feature
- Estimate CATE:

$$\boxed{\hat{\tau}_S(x) = \hat{\mu}(x,1) - \hat{\mu}(x,0)}$$

Procedure

1. Set $W_i = 1$, predict
2. Set $W_i = 0$, predict
3. Compute CATE

→ Data efficient
→ Simple
→ Treatment plays no role; baseline can ignore if uninformative, which happens if CATE is zero in many places

→ Can "learn away" treatment variable
→ Can be biased towards zero if learner ignores treatment
→ May not make sense for some algos (k-NN) to have treatment indicator as a covariate

T-learner (Two-learner)

- Models treatment variable and control variable separately, forcing a split.

1. Split data into control, treatment subsets.
2. Train two models, one on each subset
3. Make individual predictions → calculate CATE

$$\boxed{\hat{\tau}_{CATE_i} = \hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)}$$

→ Ensures treatment is considered
→ Does well when trends differ a lot from control to treatment set
→ Handles complex treatment

→ Data inefficiency: both models trained on subsets of dataset, could harm fit
→ Higher variance from above, especially if one treatment arm has fewer obs.
→ Does not capture trends across treatment groups

X-learner

- Both S, T estimate response functions (factuals) separately and subtract them to compute CATE.
  → never estimate/use counterfactual at individual level.
- X-learner estimates CATE directly
- Allows it to utilise sparsity, smoothness while predicting.
- Works especially well for imbalanced datasets

Process (3 steps, 5 models):

1. Build models for $\hat{\mu}_1, \hat{\mu}_0$ as before
   $\hat{\mu}_0 \to E[Y(0) \mid X = x]$
   $\hat{\mu}_1 \to E[Y(1) \mid X = x]$
2. Build pseudo-treatment outcomes [model counterfactuals for individuals]
   - For $i \in$ control:
     $\hat{D}_i = \hat{\mu}_1(x_i) - y^{obs}$
   - For $i \in$ treated:
     $\hat{D}_i = y^{obs} - \hat{\mu}_0(x_i)$
   - $\hat{D}_i$ → estimate of ITE for a given individual. Build two ITE models to predict this: $\hat{\tau}_1, \hat{\tau}_0$
3. Take a weighted average of the models for your final result

$$\boxed{\hat{\tau}_X(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)}$$

Models: $\hat{\mu}_1, \hat{\mu}_0, \hat{\tau}_1, \hat{\tau}_0, \hat{e}$

$g \in [0,1]$ → weight function, usually propensity score

Xs-learner (simplified)

3 steps, 3 models

1. Models for $\hat{\mu}_1(x)$, $\hat{\mu}_0(x)$ again
2. Pseudo-treatment outcome:
   - If $i \in$ control:
     $\hat{D}_i = \hat{\mu}_1(x) - y^{obs}$
   - If $i \in$ treated:
     $\hat{D}_i = \hat{\mu}_0(x) - Y^{obs}$
   - Fit a single model η on $\hat{D}_i$

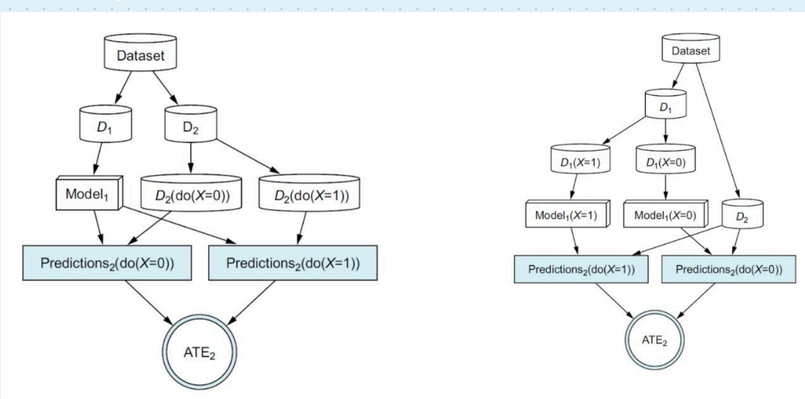
$$\boxed{\tau_{Xs}(x) = \eta(x)}$$

→ Adapts to structure
→ Works well with imbalanced groups
→ Captures heterogeneity thanks to pseudo-effects

→ More moving parts
→ Computationally intensive
→ Less stable with small datasets since pseudo-effects become noisy when imputation is weak

Crossfitting

S-learner



T-learner

Propensity Scores

Allows you to get causal insights from observational data

- Similar propensity score for two individuals ⟹ Distribution of meaningful covariates is the same

Generation

Logistic regression → most popular $\qquad e(x_i) = P(W_i = 1 \mid X_i = x)$

Matching

- Pick treated-untreated pairs with closest propensity score
- Ensure no untreated subject mapped to more than one treated → matching without replacement

Stratification

- Group data into bins based on propensity
- Once done, we can compare treated and untreated within given stratum directly

Inverse Propensity Treatment Weighting (IPTW)

$\frac{1}{\text{propensity}}$ $\qquad$ weight treatment by $\frac{1}{\text{propensity}}$

Observational data has bias → Need to be able to assign treatment to account for this; only then "random" → Use propensity to do this

Essentially give every instance a weight based on the "surprisingness" of their treatment

$$e(x_i) = P(W_i \mid X_i = x) \;\{\; P(\text{getting treatment given } x)$$

lower the value of e, lower the likelihood of treatment

⟹ If they ARE treated, need to weighted MORE because this is a SIGNIFICANT result.

Procedure

1. Assign weights:

$$\boxed{w_{i,t} = \frac{1}{e_i}} \qquad \boxed{w_{i,u} = \frac{1}{1 - e_i}}$$

} why? read the whole above explanation again and figure it out, it's obvious

2. Aggregate outcome for treated and untreated using weights

$$\boxed{\hat{Y}_t = \frac{\sum_i w_{i,t} Y_{i,t}}{N}} \qquad \boxed{\hat{Y}_u = \frac{\sum_i w_{i,u} Y_{i,u}}{N}}$$

3. $\boxed{ATE = \hat{Y}_t - \hat{Y}_u}$

Limitations

Only accounts for measured confounders unlike RCT

When IPTW

- Confounding exists w/o unmeasured confounders
- No positivity violations

} general

- You do not need individual level predictions, only average effect
- You understand the treatment assignment process well

} over other models

ATT: Average Treatment Effect on Treated

$$E[Y_i(1) - Y_i(0) \mid W_i = 1] \quad \text{or} \quad E[Y_i(1) - Y_i(0) \mid X_i = 1]$$

ATU: Average Treatment Effect on Untreated / Control

$$E[Y_i(1) - Y_i(0) \mid W_i = 0] \quad \text{or} \quad E[Y_i(1) - Y_i(0) \mid X_i = 0]$$

Difference b/w ATE, ATT

- Generally, ATT > ATE
- Covarying factors that make ATE, ATT different

Selection bias

$$\boxed{ATE = (\pi_t \cdot ATT) + (\pi_u \cdot ATU)}$$

where π denotes proportion

The difference between ATE and ATT can be written as:

$$\boxed{ATE = ATT + \text{Selection bias}}$$

$$E[Y \mid T=1] - E[Y \mid T=0] = E[Y_1 \mid T=1] - E[Y_0 \mid T=0]$$

$$= E[Y_1 \mid T=1] - E[Y_0 \mid T=1] + E[Y_0 \mid T=1] - E[Y_0 \mid T=0]$$

$$\boxed{ATE = \underbrace{E[Y_1 - Y_0 \mid T=1]}_{ATT} + \underbrace{E[Y_0 \mid T=1] - E[Y_0 \mid T=0]}_{Bias}}$$

$$\boxed{\text{Ideally } E[Y_0 \mid T=1] = E[Y_0 \mid T=0]}$$

- The estimation can be improved and is found to be somewhat close to the previous estimation if we subgroup + consider the confounder

$$\text{Effect}_{\text{Stratified}} = (\pi_{\text{Treated}} \times \hat{Y}_{\text{Treated}}) - (\pi_{\text{Untreated}} \times \hat{Y}_{\text{Untreated}})$$

$$\text{Effect}_{\text{Old}} = \frac{80 + 75 + 85}{3} - \frac{60}{1} = 20$$

$$\text{Effect}_{\text{Young}} = \frac{75}{1} - \frac{80 + 100 + 80}{3} = -11.667$$

$$\text{ATE} = (\tfrac{4}{8} \times 20) + (\tfrac{4}{8} \times -11.667) = 4.1667$$

	Confounder	Treatment	Realized
	Age	Treated	Outcome
ID	Z_i	X_i	Y_i
1	Old	1	80
2	Old	1	75
3	Old	1	85
4	Old	0	60
5	Young	1	75
6	Young	0	80
7	Young	0	100
8	Young	0	80