Machine Learning for Economists
and Business Analysts

Debiased/Double Machine Learning

Anthony Strittmatter

Reference

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017):
"Double/Debiased/Neyman Machine Learning of Treatment Effects",
American Economic Review, 107 (5), pp. 261-265, download.

Overview

Inverse Probability Weighting

T-Learner

Double/Debiased Machine Learning

Potential Outcome Framework

Notation:

- ▶ D_i binary treatment dummy (e.g., assignment to training program)
- ▶ $Y_i(1)$ potential outcome under treatment (e.g., earnings under participation in training)
- ▶ $Y_i(0)$ potential outcome under non-treatment (e.g., earnings under non-participation in training)

Infeasible parameter:

- ▶ Individual causal effect: $\delta_i = Y_i(1) - Y_i(0)$

Feasible parameters:

- ▶ Average Treatment Effect (ATE): $\delta = E[Y_i(1) - Y_i(0)] = E[\delta_i]$
- ▶ Average Treatment Effect on the Treated (ATET): $\rho = E[\delta_i | D_i = 1]$

Identifying Assumptions for ATE

▶ **Stable Unit Treatment Value Assumption (SUTVA):**

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$$

▶ **No Support Problems:**

$$\epsilon < Pr(D_i = 1|X_i = x) = p(x) < 1 - \epsilon$$

for some small $\epsilon > 0$ and all x in the support of X_i

▶ **Conditional Independence Assumption (CIA):**

$$Y_i(1), Y_i(0) \perp\!\!\!\perp D_i|X_i = x$$

for all x in the support of X_i

Modified Outcome Method for ATE

Inverse Probability Weighting:

$$Y_{i,IPW}^* = \frac{D_i}{p(X_i)} Y_i - \frac{1 - D_i}{1 - p(X_i)} Y_i = \frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))} Y_i$$

with the propensity score $p(x) = Pr(D_i = 1 | X_i = x)$.

ATE: $\delta = E[Y_{i,IPW}^*]$ and $\hat{\delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{Y}_{i,IPW}^*$

Proof of Identification

$$\delta = E\left[Y_i(1)\right] - E\left[Y_i(0)\right] \overset{LIE}{=} \int E\left[Y_i(1)|X_i = x\right] - E\left[Y_i(0)|X_i = x\right] f_X(x)dx$$

$$\overset{CIA}{=} \int E\left[Y_i(1)|D_i = 1, X_i = x\right] - E\left[Y_i(0)|D_i = 0, X_i = x\right] f_X(x)dx$$

$$= \int E\left[Y_i|D_i = 1, X_i = x\right] - E\left[Y_i|D_i = 0, X_i = x\right] f_X(x)dx$$

$$= \int E\left[D_i Y_i|D_i = 1, X_i = x\right] - E\left[(1 - D_i)Y_i|D_i = 0, X_i = x\right] f_X(x)dx$$

$$\overset{LIE}{=} \int E\left[\left.\frac{D_i Y_i}{p(X_i)}\right| X_i = x\right] - E\left[\left.\frac{(1 - D_i)Y_i}{1 - p(X_i)}\right| X_i = x\right] f_X(x)dx$$

$$= \int E\left[\left.\frac{D_i Y_i}{p(X_i)} - \frac{(1 - D_i)Y_i}{1 - p(X_i)}\right| X_i = x\right] f_X(x)dx$$

$$= \int E\left[\left.\frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))} Y_i\right| X_i = x\right] f_X(x)dx \overset{LIE}{=} E\left[\frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))} Y_i\right]$$

Reference: Horvitz and Thompson (1952)

Modified Outcome Method with IPW

Advantages:

- ▶ Generic approach

- ▶ Sparsity assumptions can be avoided by appropriate choice of estimator for propensity score

- ▶ Heterogeneous treatment effects

Disadvantages:

- ▶ Potentially omitting "weak outcome confounders"

- ▶ Shows weak performance in simulations (Knaus, Lechner, and Strittmatter, 2018)

- ▶ Not $\sqrt{N}$-consistent in high-dimensional setting

Overview

Conditional Mean Differences

Identification:

$$
\begin{aligned}
\delta &= E[Y_i(1)] - E[_i Y(0)] \\
&\overset{LIE}{=} \int E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] f_X(x) dx \\
&\overset{CIA}{=} \int E[Y_i(1)|D_i = 1, X_i = x] - E[Y_i(0)|D_i = 0, X_i = x] f_X(x) dx \\
&= \int \underbrace{E[Y_i|D_i = 1, X_i = x]}_{=\mu_1(x)} - \underbrace{E[Y_i|D_i = 0, X_i = x]}_{=\mu_0(x)} f_X(x) dx
\end{aligned}
$$

Estimator:

$$
\hat{\delta} = \frac{1}{N} \sum_{i=1}^{N} (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))
$$

with $\hat{\mu}_1(x) = \hat{E}[Y_i|D_i = 1, X_i = x]$ and $\hat{\mu}_0(x) = \hat{E}[Y_i(0)|D_i = 0, X_i = x]$ being the estimated conditional expectations of the potential outcomes.

Overview

Inverse Probability Weighting

T-Learner

Double/Debiased Machine Learning

Double/Debiased Machine Learning (DML)

Efficient Score:

$$Y_{i,DML}^* = \mu_1(X_i) - \mu_0(X_i) + \frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))} Y_i - \frac{D_i}{p(X_i)} \mu_1(X_i) + \frac{1 - D_i}{1 - p(X_i)} \mu_0(X_i)$$

$$= \mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - \mu_0(X_i))}{1 - p(X_i)}$$

ATE: $\delta = E[Y_{i,DML}^*]$ and $\hat{\delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{Y}_{i,DML}^*$

We can use standard ML methods to estimate $\hat{\mu}_1(x)$, $\hat{\mu}_0(x)$, and $\hat{p}(x)$.

Reference: Chernozhukov et al., 2017

Proof of Identification

$$
\begin{aligned}
\delta =& E\left[\mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1-D_i)(Y_i - \mu_0(X_i))}{1 - p(X_i)}\right] \\
=& E\left[\frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))}Y_i + \frac{(p(X_i) - D_i)\mu_1(X_i)}{p(X_i)} - \frac{(D_i - p(X_i))\mu_0(X_i)}{1 - p(X_i)}\right] \\
=& \int E\left[\frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))}Y_i + \frac{(p(X_i) - D_i)\mu_1(X_i)}{p(X_i)}\right. \\
& \left.-\frac{(D_i - p(X_i))\mu_0(X_i)}{1 - p(X_i)}\,\right|\, X_i = x\right] f_X(x)dx \\
=& \int \left(E\left[\frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))}Y_i\,\right|\, X_i = x\right] + \frac{E[p(X_i) - D_i | X_i = x]}{p(x)}\mu_1(x)\right. \\
& \left.-\frac{E[D_i - p(X_i) | X_i = x]}{1 - p(x)}\mu_0(x)\right) f_X(x)dx \\
=& \int E\left[\frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))}Y_i\,\right|\, X_i = x\right] f_X(x)dx = E\left[Y_i(1) - Y_i(0)\right]
\end{aligned}
$$

Reference: Robins and Rotnitzki (1995)

DML Cross-Fitting Algorithm

1. Partition the data randomly in samples S^A and S^B

2. Estimate the nuisance parameters $\hat{\mu}_1^A(x)$, $\hat{\mu}_0^A(x)$, and $\hat{p}^A(x)$ in S^A; and $\hat{\mu}_1^B(x)$, $\hat{\mu}_0^B(x)$, and $\hat{p}^B(x)$ in S^B with ML

3. Calculate the efficient scores in samples S^A and S^B, respectively:

$$\hat{Y}_{i,DML}^{A*} = \hat{\mu}_1^B(X_i^A) - \hat{\mu}_0^B(X_i^A) + \frac{D_i^A(Y_i^A - \hat{\mu}_1^B(X_i^A))}{\hat{p}^B(X_i^A)} - \frac{(1 - D_i^A)(Y_i^A - \hat{\mu}_0^B(X_i^A))}{1 - \hat{p}^B(X_i^A)}$$

$$\hat{Y}_{i,DML}^{B*} = \hat{\mu}_1^A(X_i^B) - \hat{\mu}_0^A(X_i^B) + \frac{D_i^B(Y_i^B - \hat{\mu}_1^A(X_i^B))}{\hat{p}^A(X_i^B)} - \frac{(1 - D_i^B)(Y_i^B - \hat{\mu}_0^A(X_i^B))}{1 - \hat{p}^A(X_i^B)}$$

4. Calculate ATE,

$$\hat{\delta} = \frac{1}{2}(\underbrace{\hat{E}[\hat{Y}_{i,DML}^{A*}|S^A]}_{=\hat{\delta}_A} + \underbrace{\hat{E}[\hat{Y}_{i,DML}^{B*}|S^B]}_{=\hat{\delta}_B}),$$

Asymptotic Results for ATE

- Main Regularity Condition: Convergence rate of nuisance parameters is at least $\sqrt[4]{N}$.

- ATE can be estimated $\sqrt{N}$-consistently

$$\sqrt{N}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \sigma^2)$$

with $\sigma^2 = Var(Y^*_{i,DML})$ and $Var(\hat{\delta}) = \sigma^2/N$

- Split sample estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{2}\left(\hat{\sigma}^2_A + (\hat{\delta}_A - \hat{\delta})^2\right) + \frac{1}{2}\left(\hat{\sigma}^2_B + (\hat{\delta}_B - \hat{\delta})^2\right)$$

for $\hat{\delta} = 1/2(\hat{\delta}_A + \hat{\delta}_B)$

Advantages of DML

Advantages compared to IPW and Conditional Mean Differences:

- ► Treatment and outcome equations are modelled explicitly

- ► Double robustness property

- ► $\sqrt{N}$-consistent and asymptotically normal even under high-dimensional confounding

Efficient Score for ATET

$$Y_{i,ATET}^* = \frac{D_i(Y_i - \mu_0(X_i))}{p} - \frac{p(X_i)(1 - D_i)(Y_i - \mu_0(X_i))}{p(1 - p(X_i))}$$

with $p = Pr(D_i = 1)$.

ATET: $\rho = E[Y_{i,ATET}^*]$ and $\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \hat{Y}_{i,ATET}^*$

Estimator of Asymptotic Variance:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{Y}_{i,ATET}^* - \hat{\rho} \right)^2$$

and $\hat{Var}(\hat{\rho}) = \hat{\sigma}^2/N$

References: Chernozhukov et al., 2017, Farrell, 2015

Proof of Identification for ATET

$$
\begin{aligned}
\rho =& E\left[\frac{D_i(Y_i - \mu_0(X_i))}{p} - \frac{p(X_i)(1-D_i)(Y_i - \mu_0(X_i))}{p(1-p(X_i))}\right] \\
=& \int E\left[\frac{D_i Y_i}{p} - \frac{p(X_i)(1-D_i)Y_i}{p(1-p(X_i))} - \frac{(D_i - p(X_i))\mu_0(X_i)}{p(1-p(X_i))}\,\bigg|\, X_i = x\right] f_X(x)dx \\
=& \int \left(\frac{E[D_i Y_i | X_i = x]}{p} - \frac{p(x)E[(1-D_i)Y_i | X_i = x]}{p(1-p(x))}\right. \\
& \left. - \frac{E[D_i - p(X_i)|X_i = x]}{p(1-p(x))}\mu_0(x)\right) f_X(x)dx \\
=& \int \left(\frac{E[D_i Y_i | X_i = x]}{p} - \frac{p(x)E[(1-D_i)Y_i | X_i = x]}{p(1-p(x))}\right) f_X(x)dx \\
=& \int \frac{p(x)}{p}\left(E[D_i Y_i | D_i = 1, X_i = x] - E[(1-D_i)Y_i | D_i = 0, X_i = x]\right) f_X(x)dx \\
=& \int \left(E[Y_i(1)|D_i = 1, X_i = x] - E[Y_i(0)|D_i = 0, X_i = x]\right) f_{X|D=1}(x)dx \\
=& \int \left(E[Y_i(1)|D_i = 1, X_i = x] - E[Y_i(0)|D_i = 1, X_i = x]\right) f_{X|D=1}(x)dx \\
=& E[Y_i(1) - Y_i(0)|D_i = 1]
\end{aligned}
$$