Double Machine Learning

Cholyeon Cho

February 2024

1 Introduction and General Explanation

Double Machine Learning is a general framework for using causal inference that achieves a root-n consistent estimator along with nice statistical descriptions like confidence intervals.

The key idea that is surrounding double machine learning is:

1. Machine learning can be viewed, in the statistical point of view, as a non-parametric or a semi-parametric estimator
2. Many studies are already done on non-parametric and semi-parametric estimators. These studies can be used in conjunction with machine learning to scale well on larger datasets where traditional statistics are less adequate.

2 Notations and Definitions

1. D: Treatment variable.
2. Z: Covariate matrix that is assumed to include all possible confounders.
3. Y: Outcome variable

The assumptions are that the outcome follows a linear model in respect to treatment. - Although there are versions of DML that work without the linearity assumption.

$$Y = D\theta_0 + g_0(Z) + U$$

Where

$$E[U|Z, D] = 0$$

Where the treatment is defined by the covariates and a residual.

$$D = m_0(Z) + V$$

where

$$E[V|Z] = 0$$

The equations $m_0(\cdot)$ and $g_0(\cdot)$ should be keeping track of effects of confoundedness.

3 Advantage of Using Machine Learning Methods

As mentioned above, machine learning methods can be viewed as non-parametric or semi-parametric estimators. This means that no strong assumptions have to be made to estimate the functional forms of $m_0(\cdot)$ and $g_0(\cdot)$.

4 The Naive Estimator

Why do we have to use machine learning two times instead of only once to estimate the average treatment effect then? The naive estimator is the approach of taking only one step of machine learning to estimate the parameter θ_0

The process of the Naive Estimator would be to:

1. Estimate $g_0(\cdot)$ with a machine learning algorithm

2. Estimate θ_0 with OLS

The problem with the Naive approach is that the estimator of the outcome variable is biased. That is, given the estimation θ_0 with OLS and the true parameter of interest θ,

$$E[\theta_0 - \theta] \neq 0$$

Intuitively, this naive approach means that we are trying to estimate $E[Y|Z]$. However, without the effects of the treatment taken into consideration, $E[U|Z] \neq 0$. Moreover, qua orthogonalization, regressing D on a function of Z would exclude any sort of effect that Z has on D, which makes V whatever that cannot be predicted by Z. The correlation graph means that this V is the aspect that is driving the treatment effect of Y.

5 Sources of Bias

There are largely two sources of bias when considering Double Machine Learning:

1. Regularization bias. This would be solved by orthogonalization.

2. Overfitting. This would be solved by cross-fitting

5.1 Regularization

The theorem that will be used here is the Frisch-Waugh-Lovell theorem. Given the linear model:

$$Y =_0 +_1 D +_2 Z + \mu$$

The following approaches yield the same result as regressing Y on D and Z.

1. Regress D on Z.

2. Regress Y on Z.

3. Regress the residual from step two on the residual from step one.

For the machine learning procedure, these steps would be as the following:

1. Predict D based on Z.

2. Predict Y based on Z.

3. Linear Regression of the residual by step two onto step one.

Note that the result of this procedure is exactly the equation that was assumed for the potential outcome.

The score function can be defined for DML with linear assumption.

$$\psi = (D - m_0(Z)) \times (Y - g_0(Z) - (D - m_0(Z))\theta)$$

This means the product of the residual of treatment multiplied by the residual of outcome that cannot be predicted by neither Z nor D.

The moment condition is that $\psi = 0$. This means that the residual of treatment on covariates is orthogonal to the residual of outcome on covariates and treatment.

The Neyman Orthogonality Condition implies the following equation:

$$\partial\eta E[\psi(W;\theta_0,\eta_0)][\eta - \eta_0] = 0$$

Setting this derivative to zero means that the score function would be robust to small perturbations on the parameter η.

5.2 Sample Splitting and Cross Fitting

Both the sample splitting method and the cross fitting method counteracts overfitting bias, cross-fitting being a more optimal method than sample splitting, while they both maintain similar methodologies.

5.2.1 Sample Splitting Approach

Applying the sample splitting method follows the following steps:

1. Randomly partition data into two subsets

2. Fit machine learning models for D and Y on first subset.

3. Obtain estimates for θ_0 by using the model from step two on the second dataset

The major downside of this approach is that it diminishes statistical power and efficiency by cutting the data into half.

5.2.2 Cross Splitting Approach

The downsides of sample splitting is overcame by the cross splitting approach.

1. Randomly partition data into two sets
2. Fit models for D and Y on the first dataset
3. Estimate $\theta_{0,1}$ from the second dataset using the model from step two
4. Fit machine learning models on dataset two
5. Estimate $\theta_{0,2}$ from dataset one using model from step four
6. Obtain θ_0 by taking the average of the two estimates of θ.

6 Algorithm Without Linear Assumption

It's possible to use double machine learning without linear assumption on the potential outcome.

$$Y = g_0(D, Z) + \zeta$$

$$D = m_0(Z) + \upsilon$$

Where

$$E[\zeta|D, Z] = 0$$

$$E[\upsilon|D] = 0$$

The score function here is:

$$\psi(W, \theta, \eta) = (g(1, Z) - g(0, Z)) + \frac{D(Y - g(1, Z))}{m(Z)} - \frac{(1 - D)(Y - g(0, Z))}{1 - m(Z)} - \theta$$

$$\eta(Z) = (g(0, Z), g(1, Z), m(Z))$$

$$\eta_0(Z) = (g_0(0, Z), g_0(1, Z), m_0(Z))'$$

This score function is Neyman Orthogonal.