Evaluation Exercise 2

Master in Data Science Upf - CI & ML - Part I

Delivery

- Team Work (1, 2 or 3 people)
- Sunday February 27th 2023
- Admissible formats: Rendered Rmarkdown or Jupyter Notebook

Objective

Calculating the ATE of the RHC dataset, implementing the Double ML technique with cross-fitting

https://hbiostat.org/data/repo/rhc

We will use the Right Heart Catheterization(RHC) dataset which can be used to study the impact of performing an RHC, an invasive test that provides many measurements such as blood pressure and cardiac output, into the chances of survival of the critically ill patients.

Variables

Treatment (D): swang1

Outcome (Y): death

Confounders (X): confounder.yml

Steps

1. Train ML model Y~X and calculate residuals using cross-fitting
2. Train ML model D~X and calculate residuals using cross-fitting
3. Run linear regression $Y_{residuals} \sim D_{residuals}$

Finally, check that the results are similar than the ones DoubleML would give (RHC data ATE estimation with DoubleML (Python).ipynb)

Training models

Train a model means:

- Execute cross-validation (train-test split) over a subset of hyperparameters, and choose the one with lower error. You can use functions that simplify the search of optimal parameters, such as caret in (R) or GridSearchCV in scikit learn
- You can use RandomForests or Boosting as the base model.

Using cross-fitting means:

- Split the data into two equally sized groups. Train a model on the first one, and predict on the second one. Then switch roles.