Evaluation Exercise 2

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

Delivery

- Team Work (1, 2 or 3 people)
- Sunday February 26nd 2023
- Admissible formats: Rendered Rmarkdown or Jupyter Notebook

Objective

Calculating the ATE using the T-learner on the 401 data set
https://docs.doubleml.org/stable/examples/py_double_ml_pension.html

Using:

- T-learner

- 2-fold cross fitting

- bootstrap for obtaining confidence intervals

Steps

Steps

1) Split the dataset into D1 and D2. Make sure there is similar proportion of treated (e401=1) in both groups.
2) Start with D1. Split D1 into two:
   a) D1_1 with e401 = 1
   b) D1_0 with e401 = 0
3) Train two models, one for D1_1, and another for D1_0, f1 and f0, respectively
4) Make predictions on the complementary dataset D2, and calculate the ATE_2 as the average of f1(D2) - f0(D2)
5) Repeat the process switching roles between D1 and D2, and calculate the ATE = (ATE_1 + ATE_2)/2

Steps

6) Include the previous process inside a ‘for’ loop.

7) Pick a random sample (with replacement) of the original dataset D, which we will call D_boostrap, and calculate the ATE with the process described in the previous slide of the D_boostrap

8) Get the results for all the resamplings and calculate the mean and confidence intervals (quantiles 2.5%, 97.5%) of the ATEs.

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 ase caret in (R) or GridSearchCV in scikit learn
- You can use RandomForests or Boosting as the base model.