HTE

Daniel Saggau

2/28/2021

Motivation

- ML: Minimize out of sample loss.
- Economics: Don't care about prediction but causality and inference
- Important: Finding HTE
  - Policy Implication
  - Reduce ex-post data-mining (remove rigidity of pre-analysis plans)
- Sometimes the pre-analysis is too easy, not let evaluate from plan.
- With non-experimental data, there we have the risk of ex-post data mining.

Machine Learning Solutions

- Data Driven search for heterogeneity with valid standard errors
- Main difference, in economics we are concerned with policy implications and inference and valid standard errors is valuable allows us to judge to what extent we are able to infer from our results
- Tree based models are really important (causal trees and forests)
- HTE and double/orthogonal machine learning (useful with many dimensions)
- Deep instrumental variables

why many methods?

- Solutions
- Finding heterogeneity is not a standard problem and dependent on characteristics of data and DGP
- How effect impact covariates
- Another goal would be choosing optimal policy
- random trees better for non-linear effects
- low vs. high dimensional data
- If our main goal is estimating heterogenous effects as a function of covariates, tree based and forest based methods will be a good first choice

Outline

- Set-up Problem
- Algorithms: Causal tree and honest sample-splitting
- Algorithms: Causal Forest
- Overview of other methods

Set-up Problem

- Randomly assigned treatment W
- Can also define individual outcome
- Unobserved individual treatment effect
- Conditional Average treatment effect conditional on a certain individual

- Is difference outcome and treatment
- Theoretically we can apply our ML tools to predict the CATE
- Problem: Cannot just go and apply ML because we cannot evaluate the impurity/GOF
- True individual effect is unobserved
- Unsupervised learning tool for evaluation
- How to obtain **valid** standard errors

Benchmark models: Nearest meighbor matching model

- k-NN
- Can find k people who have exactly the same characteristics as this person
- Simply compare the difference between the treatment and control group
- Don't try to find single but group similar characteristics
- Methods allows for transparent asymptotic but only works with number of dimensions is small
- Curve of dimensionality
- Causal trees and forest try to determine the relevant dimensions for treatment heterogeneity

Algorithms:

Assume we have a 50/50 randomization. Idea of transformation will be clear in the second step. Recursive partitioning for heterogeneous causal effects. Transformed outcome trees:

- Try to find mapping of inputs.

- Generalize treatment with probability p.
- When we transform our outcome, we have the expected treatment effect.
- Simple and intuitive.
- By this transformation we try to solve one of the first challenges mentioned.
- We can take any machine learning approach, and try to find a mapping of our covariates to predict heterogeneous treatment effects.
- Splitting covariate space, it is common to apply cost-complexity pruning
- Tries to find optimal split of our covariate space, s.t. we have minimum criteria (MSE + stability via penalization when going too deep and overfitting)
- For each leaf of the tree we will know the Yi* we will eval prediction of tau_i*.

Disadvantage:

- When leaf is poor estimator of tau_i
- Instead of estimating our individual treatment effect
- We can sample average treatment effect within each leaf
- Outperformed by the causal tree algorithm
- Instead of prediction, we will have the treatment effect estimate

Algorithms: Causal tree and honest sample-splitting

- Main Idea is instead of choosing splits, the algorithm chooses split that max. treatment effect in each child node

- Not have objective best prediction, but have the most heterogeneous treatment in each child node

- Causal random forest usually better than causal tree

- Honest estimation for both constructing partition and cross validation change in its anticipation

- Able to have reasonable good standard error due to honest estimation

- negative MSE criterion

- Remember: Don't remember true label

- Proposed unbiased estimator: Squared predicted treatment effect
- Means that we try to max. variance/heterogeneity across leafs
- max on both sides of the split
- Optimal number is not growing to the end of tree, but instead cv and perform cost-complexity pruning

Honest vs. Adaptive

- Adaptive vs. honest algorithm: We need to adjust our criteria
- Adaptive uses model selection via and fit via same training data creating a bias
- Spurious correlation between covariates and outcomes affect the selected model
- In the end sample mean is more extreme than entire population due to training data
- Not always constant inflow of the data
- Argue better to use honest algorithm proceeding to steps
- One part of the data is used for the tree and another part which is arguable independent is to estimate treatment effects within leaves of the tree
- Eliminate bias but increases the variance of the estimators
- Lasso to select model and then do inference on same training data will run into same error
- Common use Lasso for selection but not OLS
- Focus less on bias but more in variance focus because honest makes bias irrelevant

Better inference because more weight on variance of estimators. Trade off personal prediction and variance. Has many applications in economics but also in CS.

Summary:

- Unbiased estimator for standard error
- CT are unstable prediction rules
- CV and prunning inferior compared to shallow tree

Algorithms: Causal Forest Algorithm

- Random forest: model averaging across many causal CART
- Biggest contribution of the paper has sound theory
- Show that are pointwise consistent
- Bagging/**subsampling** the training set

Summary

- Honest
- Subsampling (not bootstrapping, individual unbiased trees)
- Using 50 % selection and 50% for prediction
- This has produced unbiase but variable estimates
- Therefore, subsampling (50%) allows for honest individual trees
- There is no strong theory behind the split of the data
- Continuous covariates Xi
- Lipschitz response (cond. mean function)

Notes

- Uncover strong treatment heterogeneity

Overview of other methods