Stacking/Tips

機械学習の経済学への応用

川田恵介

# Stacking

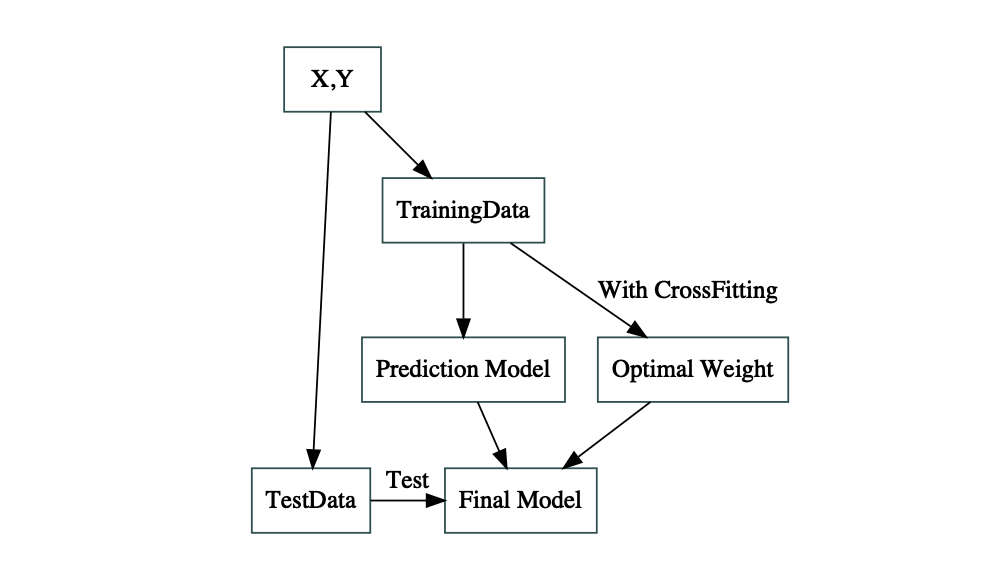
## Algorithm Selection

* 多くの有力なアルゴリズム: OLS, LASSO/Ridge/ElasticNet, (Pruned) Tree, RandomForest
  + 時間があれば、 Boosting, Bayesian Additive Regression Trees (BART)
  + NeuralNet, DeepLearningも選択肢
* どれを使う？
  + CrossValidation (BenchMark Test)
* Model Aggregation (Stacking) も有力な選択肢

## Stacking

* 複数のアルゴリズムから得られた予測結果の加重平均を最終予測モデル
  + RandomForest: **Tree** アルゴリズムから得られた予測結果の**単純**平均
* 最終予測モデル

## アルゴリズム



## Optimal Weight

* 例えば以下をOLS推定を行うことで、Weightを推定
* : アルゴリズム 、事例 を**含まない**データで推定された予測値
* 交差推定が利用可能

## Theoretical Guarantee

* Le and Clarke (2022) : 個々のアルゴリズムと比べて、 Stackingは**漸近的**に性能が良い
  + モデル選択ではなく、集計を推奨

## Practical Suggession

* 実践上では、微妙な改善に改善にとどまることが多いが、致命的な誤選択を避けれる
* 因果推論への応用でも強く推奨される (例えば Vansteelandt and Dukes (2022) をめぐる[Discussion](https://biblio.ugent.be/publication/8762862))

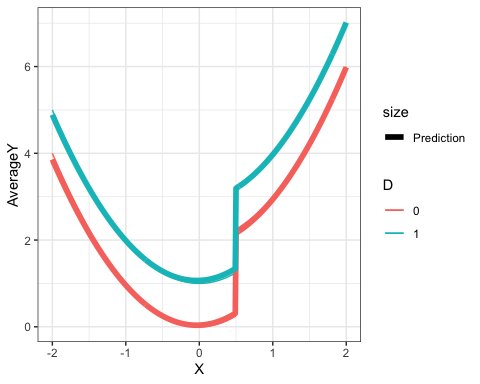
## 例

* SimulationWrong = X,D,Xの二乗項を用いる
* SimulationCollect = X,D,Xの二乗項、 ダミーを入れる
* OLSと OLSとRandomForestのStackingを比較

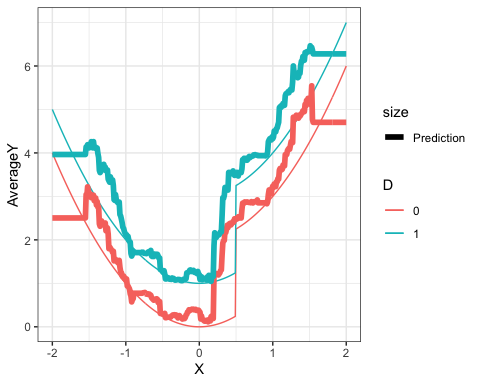
## BenchMark

| task\_id | learner\_id | regr.rsq |
| --- | --- | --- |
| SimulaitonWrong | regr.lm | 0.425 |
| SimulaitonWrong | regr.ranger | 0.420 |
| SimulaitonWrong | Stacking | 0.429 |
| SimulaitonCollect | regr.lm | 0.446 |
| SimulaitonCollect | regr.ranger | 0.408 |
| SimulaitonCollect | Stacking | 0.411 |

## Estimated OLS Model (5000)



## Estimated Stacking Model (5000)



## Tips

* 異なるベースモデルを採用したアルゴリズムを入れる(Tree系統と線形モデル系統)
* OLSや単純平均なども入れる
* Phillips et al. (2022) : より細かいSugestion
* Some Variant
  + [SuperLearner](https://github.com/tlverse/sl3) (Laan, Polley, and Hubbard 2007) : OLS with non-negative coeffiients を集計に用いる
  + mlr3pipelines : より柔軟
    - 予測値 + 元の変数も用いることが可能、 任意のアルゴリズムを集計に利用可能

# まとめ

* 母平均 を頑張って近似:
* 一般に、特に有限サンプルに対して、 “one size fits all” アルゴリズムは存在しない
* 予測モデルの”構造”を真剣に論じるべきかどうかは、慎重に検討
  + 現状、あまりおすすめしない

## PreProcess

* 人間が背景知識を用いて、手助けできる
* 線形モデルにおいては、母平均のkinkを上手く近似できない
  + 背景知識からkinkが生じるポイントが予想できる場合も

## PreProcess

* 機械学習を用いるとしても、データクリーニングは重要
* 経済データではしばしば大量のカテゴリーがある変数が存在 (例: 所在地、テキスト、出身大学+学部)
  + ダミー変数として処理すると、ほとんど0(情報量が少ない)の大量の変数が出来上がる
  + 学習性能が悪い
* より少ない変数で近似したい(低次元のベクトル空間に埋め込む)
  + カテゴリー変数 (Johannemann et al., n.d.)
  + テキスト (Gentzkow, Kelly, and Taddy 2019)

## Cross Validation

* サンプルサイズが巨大な場合、交差検証ではなく単純な分割 (検証データと訓練データ)が現実的
* サンプルサイズが小さい時は？
  + 非常に細かく分割する (leave one out)
  + 「ある程度の規模で分割し評価する」を繰り返す (repeated cross validation)
  + 後者をsuggest

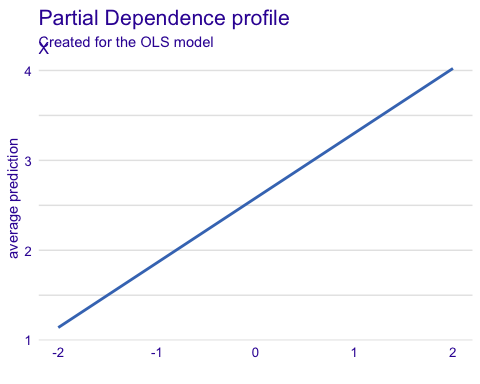
## 推定モデル

* 通常の母集団についての統計的推論(信頼区間推定)は、極めて困難
  + 格差や不平等、因果効果研究に用いることは極めて危険
* 推定された予測モデルの性質 ( 母集団の性質) を可視化する手法は、多く議論されている
  + Model Interpretation
  + なぜそのような予測になるのか、“モデルを説明する”
  + “Fair” Algorithmに関連して盛り上がる

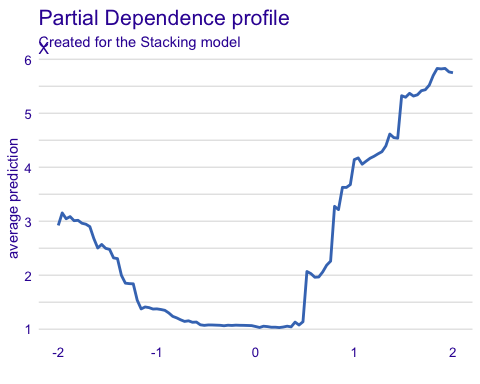
## 例

* [DALEXパッケージ](https://dalex.drwhy.ai/) を利用
* X,Dを用いる
* Partial Dependent Plot:
  + モデル全体の特徴を説明する
* Shapley values:
  + 個別事例に対する予測結果を説明する

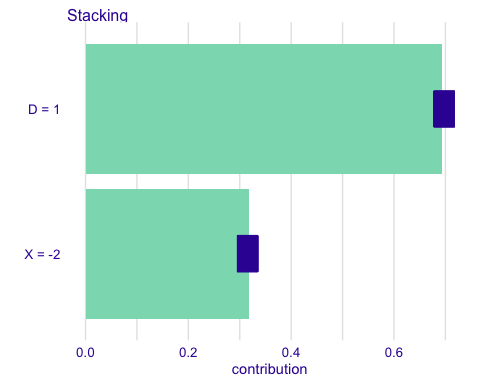
## Partial Dependent Plot



## Partial Dependent Plot



## Shapley Values



## Reference

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. “Text as Data.” *Journal of Economic Literature* 57 (3): 535–74. <https://doi.org/10.1257/jel.20181020>.

Johannemann, Jonathan, Vitor Hadad, Susan Athey, and Stefan Wager. n.d. “Sufficient Representations for Categorical Variables.” <https://doi.org/10.48550/arXiv.1908.09874>.

Laan, Mark J. van der, Eric C. Polley, and Alan E. Hubbard. 2007. “Super Learner.” *Statistical Applications in Genetics and Molecular Biology* 6 (1). <https://doi.org/10.2202/1544-6115.1309>.

Le, Tri M., and Bertrand S. Clarke. 2022. “Model Averaging Is Asymptotically Better Than Model Selection for Prediction.” *Journal of Machine Learning Research* 23 (33): 1–53. <http://jmlr.org/papers/v23/20-874.html>.

Phillips, Rachael V., Mark J. van der Laan, Hana Lee, and Susan Gruber. 2022. “Practical Considerations for Specifying a Super Learner.” <https://doi.org/10.48550/arXiv.2204.06139>.

Vansteelandt, Stijn, and Oliver Dukes. 2022. “Assumption-Lean Inference for Generalised Linear Model Parameters.” *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. <https://doi.org/10.48550/arXiv.2006.08402>.