Scalable Machine Learning for Epidemiological Studies

Eric Polley

polley.eric@mayo.edu

June 20, 2018

MAYO CLINIC

Why Machine Learning?

- Risk or prognosis predictors
- Identification of eligible participants
- Automated follow-up on cohort study participants

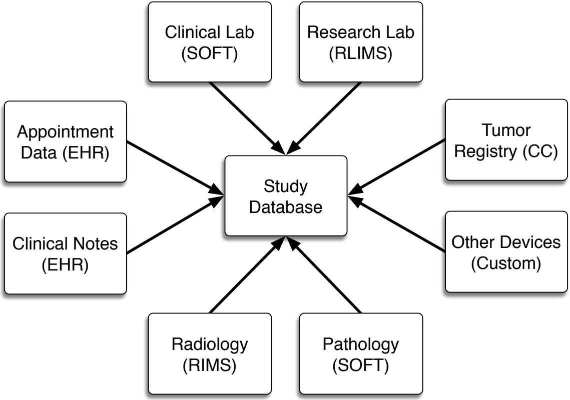
MAYO CLINIC

Components of Machine Learning

- Data Collection
- Data preprocessing (feature engineering)
- Algorithm selection
- Performance evaluation

Data Collection

- Data source/format used to train a predictor must match data for future cases
- Be aware of data velocity



MAYO CLINIC

Data Preprocessing

- Plan for missing data
- In some cases, a procedure order can be more predictive than the results
- Be wary of multidimensional outliers
- Utilize algorithms to convert clinical notes or images (radiology or digital pathology) to tabular data

MAYO CLINIC

Algorithm Menu

Ridge regression	CNN	ranger	MARS	adaboost
GAM	earth	BART	k−nearest neighbors	bartMachine
Leekasso	Neural Networks	FREE	Gradient Boosting	Random Forests
Support Vector Machines	Deep Neural Network	bayesglm	xgBoost	Relaxo
bagging	gbm	rpart	Elastic Net	Lasso

Figure 1: Machine Learning Bingo

Hyperparameter Selection

Not only many algorithms, but most have tuning parameters:

- Number of trees
- Regularization penalty scalar
- Degree of interactions
- etc.

Hyperparameter Optimization

- Using metadata to refine possible values for optimal hyperparameters
- `autoML` ideas like TPOT[1] or H2O.ai autoML[2]

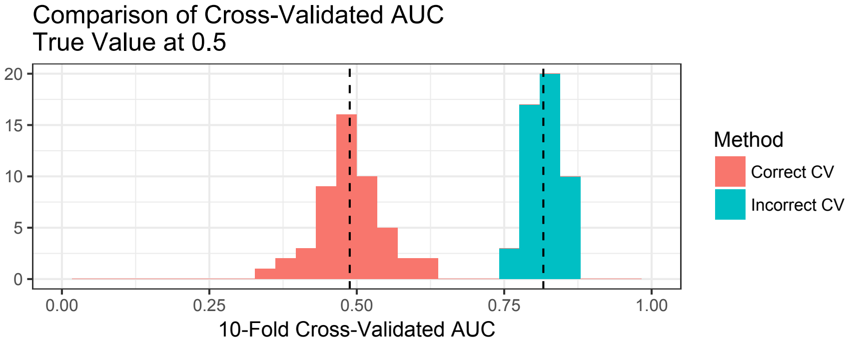
[1] `https://epistasislab.github.io/tpot/`
[2] `https://www.h2o.ai`

MAYO CLINIC

Performance Evaluation

- With large number of variables, could filter prior to estimating predictor
- Filtering steps need to be considered part of algorithm
- Example: Select variables with univariate association with outcome, then perform Random Forests predictor on subset of variables
- When using Cross-Validation to estimate performance, selection step must be nested (repeated) within data splits

Performance Evaluation



Ensembles

- How to train all these algorithms on a dataset & estimate performance
- SuperLearner framework[3]

$$f_{SL}(X) = \alpha_1 f_1(X) + \alpha_2 f_2(X) + \ldots + \alpha_p f_p(X) \tag{1}$$

- Implementations available in R (`SuperLearner`, `caretEnsemble`, and `sl3`) and Java (`H2O autoML`)

[3] van der Laan, Polley, Hubbard (2017)

Ensembles

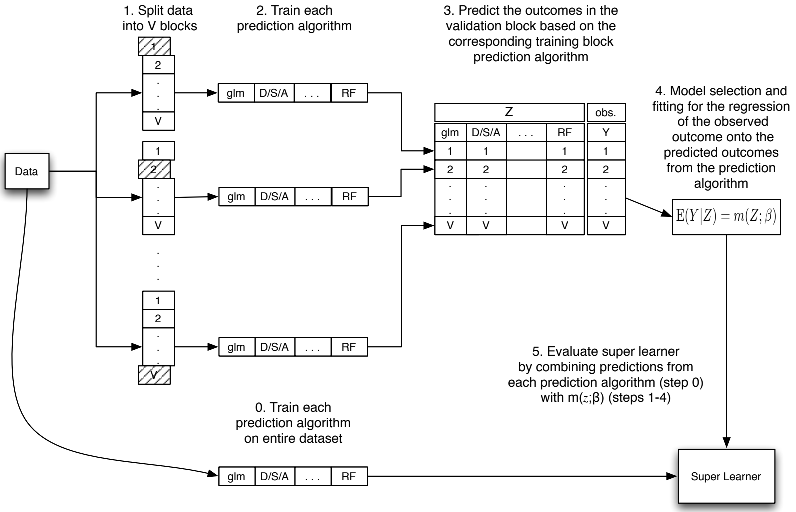


Figure 2: SuperLearner diagram

ML in the clinic

Two pressure points

- Estimating/Training the predictor
- Running the predictor on new participants

MAYO CLINIC

Thanks!

Email: Polley.Eric@mayo.edu

Slides and Code: https://github.com/ecpolley/SER2018

MAYO CLINIC