Optimal Prediction with SuperLearner

{ Nima Hejazi & Alan Hubbard } Division of Biostatistics, UC Berkeley

School of Public Health
University of California, Berkeley

Overview

1. This analysis seeks to generate optimal predictions for 5 genomic covariates across 67 sample observations, nearly evenly distributed across 5 *"knockout"* genotypes.
2. **Uninformative censoring:** the competition rules indicate that missingness was artificially introduced; this greatly simplifies the missing data problem to be solved via prediction methods.
3. Rather than use a single machine learning algorithm to estimate the missing values in the data set, a weighted combination of a library of learning algorithms is used to generate provably optimal predictions.
4. While the theory underlying the Super-Learner methodology is quite rich, at its core, the method simply uses cross-validation to rank learning algorithms within a provided library according to a meta-learner, building a weighted combination of learning algorithms for prediction.

Introduction

- The goal of this prediction challenge is to infer the withheld values of a single genomic covariate for a subset of individuals from 5 randomly selected "knock-out" conditions in the full data set provided.
- In order to predict the missing values in a *provably optimal* manner, this analysis relies on the SuperLearner algorithm, to generate asymptotically optimal prediction.
- The problem of overfitting with the individual (and ensemble) learners is avoided by employing V-fold cross-validation (where $V = 10$ in the results presented).
- The 5 genomic covariates that we provide predicted values for all have continuous measurements, thus, we use the squared error (L2) loss function with SuperLearner.

Methodology

The **SuperLearner** method works as follows:

- Start by defining a base library of L learners: $\Psi^1, \ldots, \Psi^L$ to be used within SuperLearner.
- Specify a meta-learning method (Φ), used to evaluate the base learners.
- Use V-fold cross validation in each estimation step ($V = 10$ in our case) to protect against overfitting and evaluate learners.
- Each base learner is used to generate fitted values for the training fold, generating a new matrix of subset-specific fits.
- Then, the meta-learner is used to find the optimal combination of these fits.

In the analysis for this competition, we have used:

- The full data set, iteratively predicting values for the 5 genomic covariates of interest
- In each run of SuperLearner, indicator variables are used to impute the missing values remaining in the training set.

Results I

To assess the performance of SuperLearner, we apply the mean squared error, discounting observations with missing data in each of the 5 genomic covariates of interest.

That is, for each covariate j, observations with missing values in y_j, and corresponding fitted values in $\hat{y_j}$, are removed, then MSE is applied:

$$MSE_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (\hat{y_j} - y_j)^2$$

Covariate	MSE
RBC Distribution Width	0.00004096
Neutrophil Differential Count	0.00846919
Lymphcyte Differential Count	0.06623846
Monocyte Cell Count	0.00005010
Basophil Differential Count	0.00007293

Table 1: "Competition" Covariates with MSE

Results II

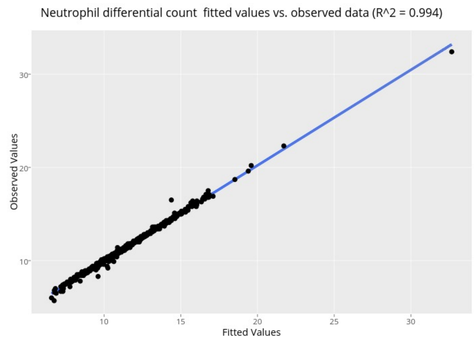

Figure 1: fitted vs true values for neutrophils

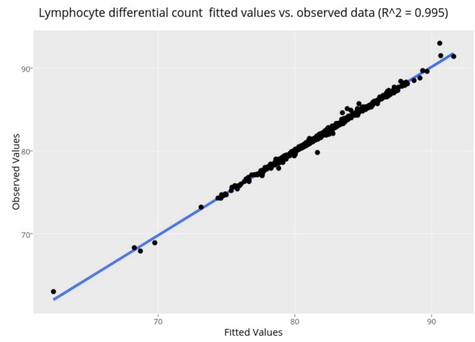

Figure 2: fitted vs true values for lymphocytes

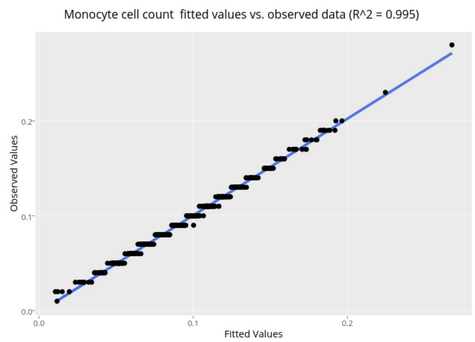

Figure 3: fitted vs true values for monocytes

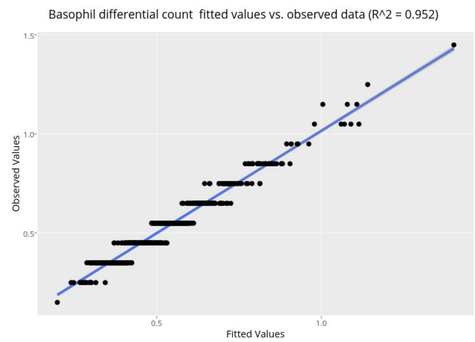

Figure 4: fitted vs true values for basophils

Conclusion

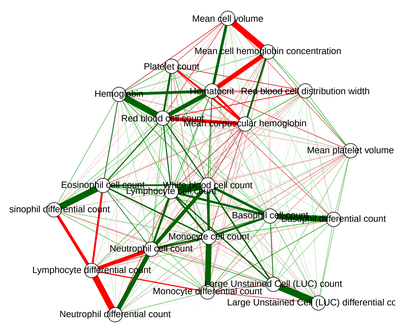

Figure 5: Graph of Correlation Structure

- In order to visualize the relationship between the genomic covariates in the observed data set, a graph is generated from the correlation matrix.
- We hold that a predictive analysis ***does not target causal parameters***. Thus, we refrain from providing a causal graph in our work.
- SuperLearner provides asymptotically optimal prediction, and our results display MSE values that substantiate this claim.

References

[1] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[2] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Contact Information

Web http://nimahejazi.org & hubbard.berkeley.edu

Email nhejazi@berkeley.edu & hubbard@berkeley.edu