University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2006 *Paper* 206

Super Learning: An Application to Prediction of HIV-1 Drug Susceptibility

Sandra E. Sinisi* Maya L. Petersen† Mark J. van der Laan‡

*University of California, Berkeley, sinisi54@alum.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, mayaliv@berkeley.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/ucbbiostat/paper206

Copyright ©2006 by the authors.

Super Learning: An Application to Prediction of HIV-1 Drug Susceptibility

Sandra E. Sinisi, Maya L. Petersen, and Mark J. van der Laan

Abstract

Many statistical methods exist that can be used to learn a predictor based on observed data. Examples include decision trees, neural networks, support vector regression, least angle regression, Logic Regression, and the Deletion/Substitution/Addition algorithm. The optimal algorithm for prediction will vary depending on the underlying data-generating distribution. In this article, we introduce a "super learner," a prediction algorithm that applies any set of candidate learners and uses cross-validation to select among them. Theory shows that asymptotically the super learner performs essentially as well or better than any of the candidate learners. We briefly present the theory behind the super learner, before providing an example based on research aimed at predicting the in vitro phenotypic susceptibility of the HIV virus to antiretroviral drugs based on viral mutations. We apply the super learner to predict susceptibility to one protease inhibitor, nelfinavir, using a set of database-derived nonpolymorphic treatment-selected protease mutations.

1 Introduction

Numerous methods exist to learn from data the best predictor of a given outcome. A few examples include decision trees, neural networks, support vector regression, least angle regression, Logic Regression, and the Deletion/Substitution/Addition (D/S/A) algorithm. Such algorithms, or learners, may be characterized by the mechanism used to search the parameter space. For example, the D/S/A algorithm (Sinisi and van der Laan, 2004) uses polynomial basis functions, while Logic Regression (Ruczinski et al., 2003) constructs Boolean expressions of binary covariates. The relative performance of a given learner depends on how extensive each learner must search over subspaces (reflected in the variance) in order for the employed mechanism to achieve a comparable approximation of the truth (reflected in the bias). Thus, the relative performance of various learners will depend on the true data-generating distribution. In practice, it is generally impossible to know *a priori* which learner will perform best for a given prediction problem and data set.

The framework for unified loss-based estimation (van der Laan and Dudoit, 2003) suggests a solution to this problem in the form of a new estimator, which we call the "super learner." This estimator is itself a prediction algorithm, which applies a set of candidate learners to the observed data, and chooses the optimal learner for a given prediction problem based on cross-validated risk. Theoretical results show that such a super learner will perform asymptotically as well or better than any of the candidate learners (van der Laan and Dudoit, 2003; van der Laan et al., 2004). We present the super learner in the context of unified loss-based estimation in Section 2, and illustrate its performance in the context of a known data-generating distribution using a simulated example in Section 3.

In Section 4, we apply the super learner to research drawn from the treatment of Human Immunodeficiency Virus Type 1 (HIV-1). HIV frequently develops resistance to the antiretroviral drugs being used to treat it, resulting in loss of viral suppression and therapeutic failure. While over 15 licensed antiretroviral drugs exist, the majority fall into three classes: protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors (NRTIs), and non-nucleoside reverse transcriptase inhibitors (NNRTIs). There is a high-level of cross-resistance within drug classes; a virus that has developed resistance to one drug in a class may also be resistant to other drugs in the

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

same class. Thus, selecting a new "salvage" drug regimen for an individual who has developed resistance to his or her current regimen is not straightforward. Improved understanding of the genetic basis of resistance to specific antiretroviral drugs has the potential to guide selection of an effective salvage regimen.

In the data example presented in this paper, the goal is to relate mutations in HIV-1 protease and reverse transcriptase to changes in *in vitro* susceptibility to antiretroviral drugs. The outcome of interest is phenotypic drug susceptibility, and the predictors consist of mutations in the protease and reverse transcriptase enzymes. Rhee et al. (2006) applied five different learning methods to predict phenotypic drug susceptibility based on viral genotype (the presence or absence of mutations): (1) decision trees, (2) neural networks, (3) support vector regression, (4) least squares regression, and (5) least angle regression. We applied the super learner to the dataset used by Rhee et al. (2006), where we used Least Angle Regression, Least Squares Regression, the D/S/A algorithm, and Logic Regression as candidate learners.

2 Methods

2.1 Loss-Based Estimation

The motivating methodology behind the concept of super learning comes from the loss-based estimation theory introduced in van der Laan and Dudoit (2003). We provide a brief description of this estimation road map before introducing the super learner.

van der Laan and Dudoit (2003) provide a general framework for parameter estimation problems. The data consist of realizations of random variables, $X_1, \ldots, X_n$, from an unknown *data generating distribution*, $F_{X,0}$. The goal is to use the data to estimate a *parameter* ψ_0 of the distribution $F_{X,0}$, where ψ_0 is defined as some function of $F_{X,0}$. That is, we wish to obtain an *estimator*, or function of the data, $\hat{\psi}$, that is *close* (in risk distance) to the parameter ψ_0. For example, in our HIV-1 data example, Y denotes a continuous measurement of drug susceptibility, and W is a d-dimensional vector of binary variables indicating the presence or absence of a mutation. X_i consists of the pair $X_i = (W_i, Y_i)$, measured on a sequence i. The parameter of interest ψ_0

Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

corresponds to the conditional expected value of drug susceptibility Y given the mutation profile W.

The general strategy for loss-based estimation is driven by the choice of a *loss function* and relies on *cross-validation* for estimator selection and performance assessment. The proposed estimation road map can be stated in terms of the following three main steps (van der Laan and Dudoit, 2003).

1. *Definition of the parameter of interest in terms of a loss function.* For the full data structure, define the parameter of interest as the minimizer of the expected loss, or *risk*, for a *loss function* chosen to represent the desired measure of performance (e.g., mean squared error in regression).

2. *Construction of candidate estimators based on a loss function.* Define a finite collection of candidate estimators for the parameter of interest.

3. *Cross-validation for estimator selection and performance assessment.* Use cross-validation to estimate risk based on the observed data loss function and to select an optimal estimator among the candidates in Step 2.

In the regression setting, our parameter of interest is $E(Y|W)$, which we denote $\psi(W)$. The loss function for our parameter of interest is the squared error loss function, $(Y - \psi(W))^2$. We will use various learning methods to construct candidate estimators needed for Step 2, and then use cross-validation as described in Step 3 to choose the optimal estimator among the candidates. We propose a super learner to perform Steps 2 and 3.

2.2 Candidate Learning Algorithms

Least Angle Regression (**LARS**) (Efron et al., 2004) is a model selection algorithm available in the `lars()` package of `R` (http://www.r-project.org). *Logic Regression* (Ruczinski et al., 2003) is an adaptive regression methodology that attempts to construct predictors as Boolean combinations of binary covariates available in the `LogicReg()` package of `R`. The *Deletion/Substitution/Addition (D/S/A) algorithm* (**DSA**) (Sinisi and van der Laan, 2004) for polynomial regression data-adaptively generates candidate predictors as

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

polynomial combinations of continuous and/or binary covariates. It is available as an `R` package at http://www.stat.berkeley.edu/users/laan/Software/. All of these methods have the option to carry out selection using v-fold cross-validation. The selected fine-tuning parameter(s) can include the ratio of the $L1$ norm of the coefficient vector in LARS; the number of logic trees and leaves in Logic Regression; and the number of terms and a complexity measure on each of the terms in DSA.

2.3 The Cross-Validation Selector

Cross-validation divides the available *learning* set into a *training* set and a *validation* set. Observations in the training set are used to construct (or *train*) the estimators, and observations in the validation set are used to assess the performance of (or *validate*) these estimators. The cross-validation selector selects the learner with the best performance on the validation sets. In v-fold cross-validation, the learning set is divided into v mutually exclusive and exhaustive sets of as nearly equal size as possible. Each set and its complement play the role of the validation and training sample, respectively, giving v splits of the learning sample into a training and corresponding validation sample. For each of the v splits, the estimator is applied to the training set, and its risk is estimated with the corresponding validation set. For each estimator/learner the v risks over the v validation sets are averaged resulting in the so-called *cross-validated risk*. The estimator with the minimal cross-validated risk is selected.

2.4 Super Learner

It is helpful to consider each learner as an algorithm applied to empirical distributions. Thus, if we index a particular learner with an index k, then this learner can be represented as a function $P_n \to \hat{\Psi}_k(P_n)$ from empirical probability distributions P_n to functions of the covariates. Consider a collection of $K(n)$ learners $\hat{\Psi}_k$, $k = 1, \ldots, K(n)$. The super learner is a new estimator defined as

$$\hat{\Psi}(P_n) \equiv \hat{\Psi}_{\hat{K}(P_n)}(P_n),$$

where $\hat{K}(P_n)$ denotes the cross-validation selector described above which simply selects the learner which performed best in terms of cross-validated

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

risk. Specifically,

$$\hat{K}(P_n) \equiv \arg\min_k E_{B_n} \sum_{i,B_n(i)=1} (Y_i - \hat{\Psi}_k(P^0_{n,B_n}))^2,$$

where $B_n \in \{0,1\}^n$ denotes a random binary vector whose realizations define a split of the learning sample into a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$. Here P^1_{n,B_n} and P^0_{n,B_n} are the empirical probability distributions of the validation and training sample, respectively.

By a general finite sample oracle inequality presented in (van der Laan and Dudoit, 2003; van der Laan et al., 2004), it follows that if $K(n)$ is polynomial in sample size, then either this estimator asymptotically outperforms any of the candidate learners, or, in the case one or more of the candidate learners converge at a parametric rate, then this super learner achieves the almost parametric rate of convergence $\log n/n$.

In our application of the super learner, we have a collection of four candidate learners: the LARS learner, DSA learner, Logic Regression learner, and a least squares fit. We are using the super learner to select the optimal candidate learner which performs best in terms of cross-validated risk. Figure 1 shows a depiction of the super learner estimation process.

3 Simulation

In this section, we apply the super learner to a simulated data set with a known data generating distribution. The following model was used to generate 500 observations ($i = 1, \ldots, 500$):

$$y_i = 2w_1w_{10} + 4w_2w_7 + 3w_4w_5 - 5w_6w_{10} + 3w_8w_9 + N(0,1), \qquad (1)$$

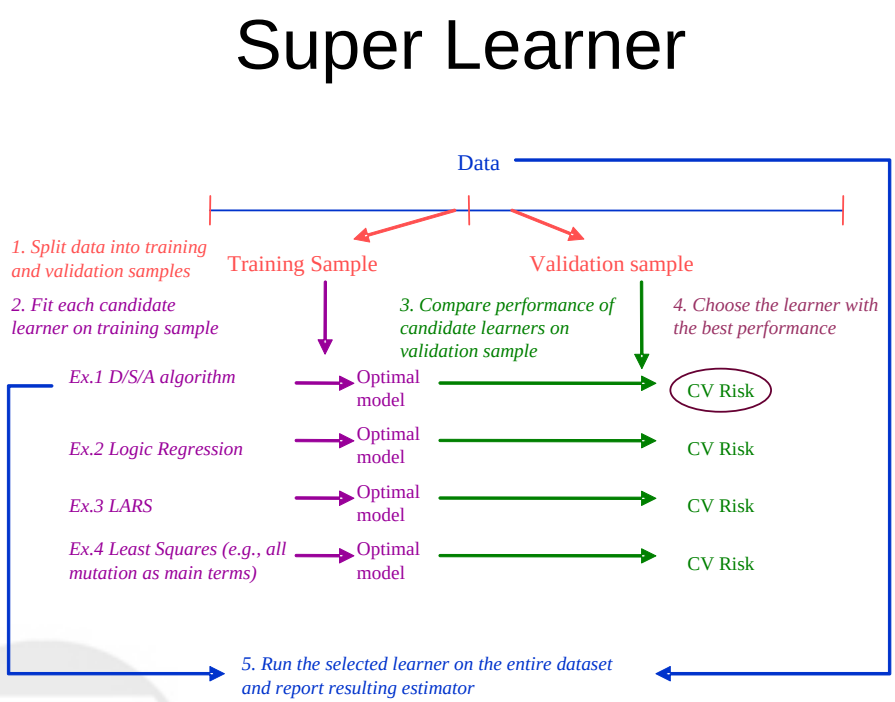
where $w_j \sim Bin(0.4)$, $j = 1, \ldots, 10$. The data for a given observation thus consists of a 10-dimensional vector of covariates W, and an outcome Y. These 500 observations constitute the learning set.

We applied the super learner with 10-fold cross-validation to the learning set to estimate $E(Y|W)$. This involved partitioning the learning set into 10 parts. Each part in turn served as the validation set, while the other 9/10ths of the data served as the training set. The super learner applied the following set of candidate learners to each of the 10 training sets: LARS,

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Figure 1: Schematic Diagram of the Super Learner



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

least squares, DSA, and Logic Regression. For least squares and LARS, two sets of input variables were used. One consisted of all main terms, and the other consisted of all main terms $w_1, \ldots, w_{10}$ and all two-way interactions $w_1, \ldots, w_{10}, w_1w_2, \ldots, w_9w_{10}$. Internal (10-fold) cross-validation was used to select the optimal fraction in LARS. Internal 10-fold cross-validation was also used to select the fine-tuning parameters for each candidate Logic Regression and DSA learner:

- Logic Regression
  - trees $\in \{1, \ldots, 5\}$
  - leaves $\in \{1, \ldots, 20\}$
- DSA
  - terms $\in \{1, \ldots, 10\}$
  - order-of-interactions $\in \{1, 2\}$

Application of each candidate learner to the 10 training sets yielded a set of 10 estimators for each candidate learner; in the case of Logic Regression, LARS, and DSA these optimal estimators were indexed by fine-tuning parameters selected using internal cross-validation. The cross-validated risk for each of these candidate estimators was estimated by evaluating each estimator to the corresponding validation set. The resulting cross-validated risks for each estimator averaged across validation sets are displayed in Table 1.

Based on the table of cross-validated risks, DSA and Logic Regression were identified as the top learners. Table 2 shows a more detailed comparison of these two learners, illustrating variation in the selection of fine-tuning parameters within distinct partitions of the learning set and the associated cross-validated risks. 10-fold cross-validation within each of the 10 training sets consistently selected 5 trees for Logic Regression, and 2-way interactions for DSA. However, the number of leaves selected for Logic Regression and the number of terms selected by DSA varied across the 10 training sets. The winning learner between these two competitors varied across partitions of the learning set, with the lowest-cross validated risk achieved sometimes by DSA and sometimes by Logic Regression. On average, however, Logic Regression outperformed DSA (average cross-validated risk of 0.958 versus

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 1: *Simulated Example.* **Super Learner: Cross-Validated Risks of Candidate Learners**. "Least Squares/LARS (1)" refers to a least squares/LARS fit with main terms only, "Least Squares/LARS (2)" refers to a least squares/LARS fit with main terms and all 2-way interactions.

Method	Median	Mean	Std Error
Least Squares (1)	4.986	4.700	1.17
Least Squares (2)	0.960	1.056	0.22
LARS (1)	5.213	4.985	1.31
LARS (2)	0.952	1.062	0.23
Logic Regression	0.934	0.958	0.17
DSA	0.930	0.966	0.18

Table 2: *Simulated Example.* **Super Learner: Comparing Logic Regression and the D/S/A Algorithm**. Shows the fine-tuning parameters selected (number of leaves for Logic Regression, number of terms for DSA) and the associated cross-validated risks across the 10 partitions of the learning set into training and validation sets. Note: The additional fine-tuning parameters were selected consistently across the 10 training sets (cross-validation always selected 5 trees for Logic Regression and 2-way interactions for DSA).

	Logic Regression		D/S/A Algorithm	
Sample	Leaves	CV Risk	Terms	CV Risk
1	10	0.981	8	1.002
2	20	0.805	5	0.805
3	18	0.916	9	0.877
4	17	1.247	7	1.274
5	17	0.838	6	0.908
6	18	0.775	5	0.775
7	10	0.952	5	0.952
8	20	0.763	6	0.739
9	18	1.123	5	1.123
10	10	1.181	6	1.203
ave	15.8	0.958	6.2	0.966

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

0.966, respectively). Thus, the super learner selected Logic Regression as the optimal learner.

As the winning learner, Logic Regression was then applied to the entire learning sample. The final logic tree is displayed in Figure 2 and can be written as:

```
-3.09 * ((not X9) or (not X8)) +4.58 * ((not X10) or (not X6)) +4.17
* (((not X6) and X6) or (X7 and X2)) -3.09 * ((not X5) or (not X4))
+0.839 * X1
```

This fit has an R^2 of 0.874.

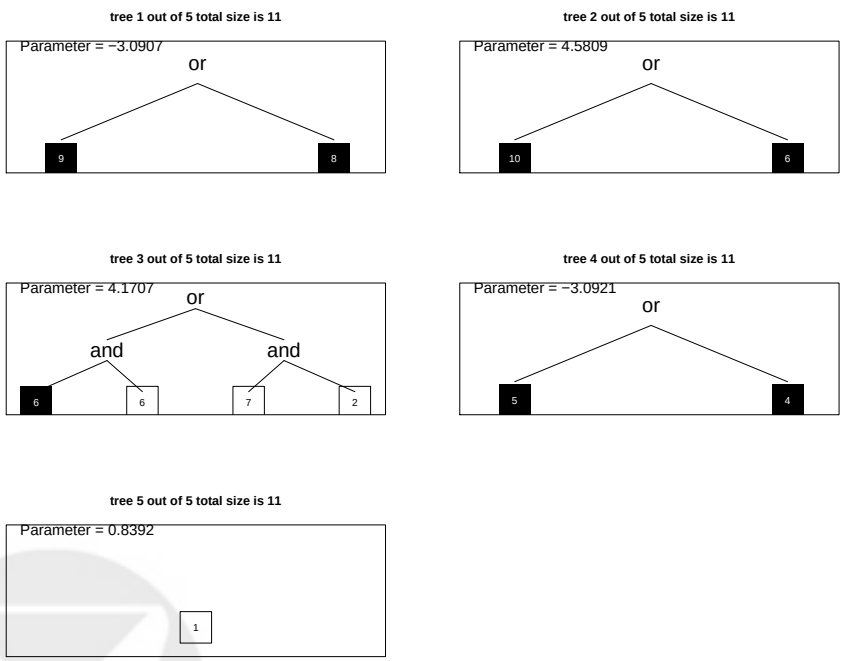


Figure 2: *Simulated Example.* Logic Regression Fit

Even though the super learner did not select DSA as the optimal learner, given the close competition between Logic Regression and DSA, we also

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

applied DSA to the learning sample. The final DSA fit had nine terms with eight two-way interactions, and an R^2 of 0.913:

$$\begin{aligned}\hat{y} &= 0.087 - 4.906w_6w_{10} + 4.211w_2w_7 + 3.205w_8w_9 + 3.107w_4w_5 \\ &\quad +1.984w_1w_{10} - 0.406w_7w_8 - 0.359w_6 + 0.406w_3w_6 - 0.325w_9w_{10}\end{aligned} \tag{2}$$

This can be compared to the true model which had 5 two-way interaction terms. All 5 of these interaction terms were included in the final 9 term DSA fit, with coefficients extremely comparable to those of the true model.

To assess the performance of the two estimators, we simulated 5000 observations from the true model to generate an independent test set. We evaluated the performance of the candidate estimators (see Figure 2 for the final Logic Regression fit and Equation 2 for the final DSA fit) on this set of 5000 observations. The Logic Regression fit yielded a mean squared prediction error (MSPE) of 1.37 with an R^2 of 0.84 while the DSA fit yielded a MSPE of 1.05 with an R^2 of 0.88.

4 Data Analysis

A description of the data used in our analysis is available in Rhee et al. (2006). The HIV-1 sequences were obtained from publicly available isolates in the Stanford HIV Reverse Transcriptase and Protease Sequence Database. We focus on predicting viral susceptibility to protease inhibitors (PIs) based on mutations in the protease region of the viral strand.

Mutations were defined as amino acid differences from the subtype B consensus wild type sequence at positions 1-99 in protease. We used a subset of these mutations, the non-polymorphic treatment-selected mutations (TSMs), as predictors. The TSMs were previously identified by (Rhee et al., 2005, 2006) as those significantly associated with antiretroviral therapy in persons infected with subtype B viruses. The association of these mutations with previous treatment is thought to result from selection due to their contributions to resistance, suggesting this is a promising set of candidate predictors. The 58 TSMs used, occurring at 34 positions in protease, are listed in Table 3. Mutations are referred to by position followed by amino acid substitution; for example, 90M refers to the occurrence of methionine at position 90.

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

The outcome of interest was standardized log fold change in drug susceptibility, defined as the ratio of IC_{50} of an isolate to a standard wildtype control isolate. IC_{50} is the concentration of a drug needed to inhibit viral replication by 50% where IC stands for *inhibitory concentration*. We applied our super learner to predicting susceptibility to a single PI, nelfinavir (NFV). A mutation profile and corresponding NFV susceptibility was available for 740 viral isolates; this constituted the learning sample.

4.1 Super Learner Results

We applied the super learner with 10-fold cross-validation to select the optimal learner given the following set of candidates: LARS, Logic Regression, DSA, and a least squares fit including all 58 mutations as main terms. We found no difference in risk when using DSA to search over 1-way or 2-way interactions. Similarly, Rhee et al. (2006) found that including all two-way interactions among the mutations as input variables did not improve the prediction accuracy. Therefore, DSA did not consider interactions and used 10-fold cross-validation to select between 1 and 50 main terms.

Table 4 shows the cross-validated risks of the candidate learners. The least squares fit yielded estimators with the lowest average risk, 0.187, although the average cross-validated risk for DSA was comparable at 0.188. The super learner thus selected the least squares fit as the optimal learner. Therefore, we fit a least squares model on all 740 observations. Tables 5 and 6 display the super learner estimator, in this case the least squares model of all main terms, fit on the entire learning sample.

Due to the similarity in cross-validated risk of least squares and DSA, we also applied DSA to the learning sample. Cross-validation selected a final DSA estimator with 40 main terms. This can be contrasted to the super learner estimator, corresponding to the least squares estimator with 58 main terms. The similarity in cross-validated risk between these two estimators suggests that prediction is only marginally improved by including the other 18 mutations in the prediction model.

We investigated the size of the mutation set needed to achieve a predictor with comparable cross-validated risk by examining the cross-validated risks for the best main term model of each size (as fit by DSA). The resulting plot, shown in Figure 3, illustrates that the decline in cross-validated risk flattens

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 3: Nonpolymorphic treatment-selected protease mutations

W	pr Mutation	W	pr Mutation
1	10F	30	55R
2	10R	31	58E
3	11I	32	66F
4	20I	33	67F
5	20T	34	71I
6	20V	35	73A
7	23I	36	73C
8	24I	37	73S
9	30N	38	73T
10	32I	39	74A
11	33F	40	74P
12	34Q	41	74S
13	35G	42	76V
14	43T	43	79A
15	46I	44	82A
16	46L	45	82F
17	46V	46	82S
18	47V	47	82T
19	48M	48	84A
20	48V	49	84C
21	50L	50	84V
22	50V	51	85V
23	53L	52	88D
24	54A	53	88S
25	54L	54	88T
26	54M	55	89V
27	54S	56	90M
28	54T	57	92R
29	54V	58	95F

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

Table 4: Super Learner: Cross-Validated Risks

Method	Median	Mean	Std Error
Least Squares	0.175	0.187	0.039
LARS	0.192	0.205	0.052
Logic Regr	0.219	0.256	0.115
DSA	0.174	0.188	0.041

sharply as the regression models reach approximately 20 main terms. This suggests that while the truly optimal predictor may use all 58 treatment-selected mutations as main terms, the majority of the predictive information can be captured by much smaller models of around 20 main terms.

We examined the best model of each size selected by DSA in order to investigate which mutations provide the most predictive information. The best models of each size selected by DSA happened to be nested. For example, the best model of size 1 contained the mutation 90M, the best model of size 2 contained the mutations 90M and 30N, the best model of size 3 contained the mutations 90M, 30N, and 54V, etc. The best models of size 1-20 are summarized in Table 7.

The p-values for the coefficients from the least squares fit (Tables 5,6) and the list of the best DSA models of each size (Table 7) provide alternative rankings of the importance of each candidate mutation for resistance to NFV. The two approaches produce quite comparable insight into the set of mutations key to predicting susceptibility to NFV; 18 of the 20 mutations selected in the DSA models of size 1-20 were among the top 20 mutations as ranked by p-value in the least squares model. Both rankings successfully identified the majority of mutations known to contribute significantly to NFV resistance, including 90M, 30N, 88S, and 84A.

5 Discussion

We have presented a super learner that uses cross-validation to select an optimal learner among a set of candidate learners. Theoretical results show that the super learner asymptotically will outperform any of the candidate estimators it employs as long as the number of candidate learners is

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 5: Least Squares Model (Significance codes: 0 = ***, 0.001 = **, 0.01 = *, 0.05 = .)

Term	$\hat{\beta}$	SE	t-stat	p-value
(Intercept)	-1.00710	0.02264	-44.488	< 2e-16 ***
10F	0.17563	0.05588	3.143	0.001743 **
10R	0.04947	0.20014	0.247	0.804860
11I	-0.13140	0.14178	-0.927	0.354371
20I	0.48555	0.06452	7.525	1.67e-13 ***
20T	0.18773	0.09784	1.919	0.055438 .
20V	0.06381	0.22119	0.288	0.773056
23I	0.11457	0.11941	0.959	0.337687
24I	0.41711	0.08131	5.130	3.78e-07 ***
30N	1.25778	0.08336	15.088	< 2e-16 ***
32I	0.26527	0.10747	2.468	0.013818 *
33F	-0.14549	0.07144	-2.036	0.042099 *
34Q	0.10543	0.26001	0.405	0.685243
35G	-0.27326	0.27505	-0.993	0.320820
43T	0.23510	0.09581	2.454	0.014384 *
46I	0.29382	0.04603	6.383	3.21e-10 ***
46L	0.10619	0.06182	1.718	0.086304 .
46V	-0.01298	0.23730	-0.055	0.956410
47V	-0.08614	0.13087	-0.658	0.510622
48M	0.53284	0.26823	1.986	0.047379 *
48V	0.22471	0.09295	2.418	0.015886 *
50L	-0.84575	0.11336	-7.461	2.63e-13 ***
50V	0.13139	0.11513	1.141	0.254199
53L	0.04208	0.08835	0.476	0.633978
54A	1.64700	0.39027	4.220	2.77e-05 ***
54L	0.21353	0.10413	2.051	0.040689 *
54M	0.44536	0.13024	3.419	0.000665 ***
54S	1.46285	0.32910	4.445	1.03e-05 ***
54T	1.75388	0.22332	7.854	1.57e-14 ***
54V	0.56756	0.05273	10.764	< 2e-16 ***

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

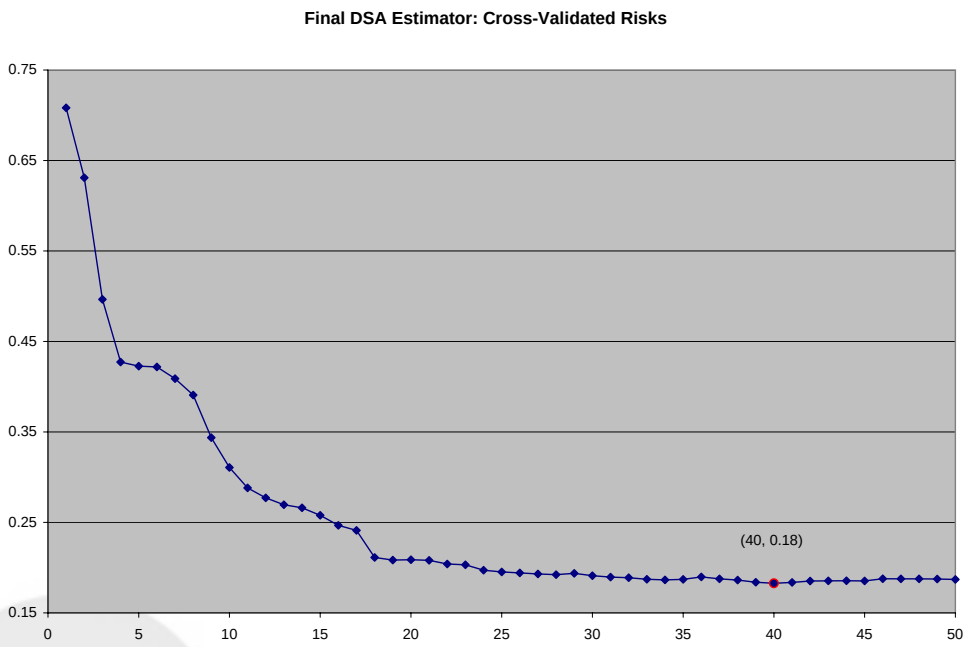
Table 6: Least Squares Model (cont'd)

Term	$\hat{\beta}$	SE	t-stat	p-value
55R	0.21188	0.10159	2.086	0.037384 *
58E	0.21815	0.08184	2.665	0.007870 **
66F	0.24775	0.17450	1.420	0.156138
67F	0.66268	0.24346	2.722	0.006656 **
71I	0.07484	0.11650	0.642	0.520804
73A	0.03265	0.21218	0.154	0.877737
73C	0.14335	0.15814	0.906	0.365016
73S	0.44710	0.06210	7.199	1.60e-12 ***
73T	0.46391	0.10172	4.560	6.05e-06 ***
74A	0.05345	0.39476	0.135	0.892345
74P	0.53279	0.15491	3.439	0.000618 ***
74S	0.45321	0.09666	4.689	3.32e-06 ***
76V	-0.09230	0.08718	-1.059	0.290075
79A	0.73175	0.41289	1.772	0.076799 .
82A	0.30910	0.05866	5.269	1.84e-07 ***
82F	0.61187	0.11130	5.497	5.45e-08 ***
82S	0.42036	0.29461	1.427	0.154085
82T	0.25881	0.07793	3.321	0.000945 ***
84A	2.17172	0.15347	14.151	< 2e-16 ***
84C	1.76901	0.14486	12.212	< 2e-16 ***
84V	0.31758	0.04599	6.906	1.15e-11 ***
85V	-0.21926	0.09819	-2.233	0.025868 *
88D	0.42180	0.08864	4.758	2.38e-06 ***
88S	1.09265	0.07317	14.933	< 2e-16 ***
88T	0.55475	0.40114	1.383	0.167139
89V	0.05987	0.15417	0.388	0.697893
90M	0.64667	0.04185	15.453	< 2e-16 ***
92R	0.04901	0.43799	0.112	0.910932
95F	0.31722	0.20472	1.550	0.121722

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Figure 3: DSA Estimator applied to learning sample, sizes $\in \{1, \ldots, 50\}$.



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

Table 7: DSA Estimator: *Best* Model of Sizes 1 to 20. (i.e., Best model of size 1: L90M, Best Model of Size 2: L90M and 30N, etc.)

Mutation	Order	Mutation	Order
90M	1	20I	11
30N	2	50L	12
54V	3	73S	13
46I	4	24I	14
84C	5	54S	15
84A	6	74S	16
88S	7	82F	17
54T	8	10F	18
84V	9	54M	19
82A	10	88D	20

polynomial in sample size (or, if one of the candidate estimators it employs achieves a parametric rate of convergence, the super learner will converge at an almost parametric rate). These results suggest that the investigator pays a very small price for considering multiple alternative learners. Currently, most researchers employ one, or at most a handful, of learning algorithms to answer prediction questions. A better approach would be to apply as many candidate learners as are feasible given time and computing limitations, and choose among them using the super learner.

Of course, in practical applications using finite samples, there is no guarantee that the super learner will always select the optimal learner for a given data application. Our simulation results illustrate this point well. Logic Regression was selected by the super learner as optimal, with a slightly lower average cross-validated risk than the D/S/A algorithm. However, when the performance of the two estimators was evaluated on an independent test set, DSA slightly outperformed the Logic Regression estimator. These results suggest that when employing the super learner, it is worthwhile to evaluate not only the final estimator provided by the winning learner but also competetive estimators. The results of the data example reinforce the utility of considering alternative learners with risks comparable to that of the optimal learner to provide additional insight into the data.

In our data example, both the super learner estimator, corresponding

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

to the least squares fit of all main terms, and the best models selected by the D/S/A algorithm provided a reasonable ranking of the importance of specific mutations for susceptibility to NFV. Such use of estimators aimed at prediction to provide a ranking of variable importance is common practice. We would like to point out, however, that if a ranking of variable importance is the goal of the analysis, then this implies a distinct parameter of interest from the optimal predictors focused on in this paper. Ideally, efforts to learn variable importance from the data will focus directly on the quantity of interest; such an approach is introduced in van der Laan (2006) and applied in Birkner and van der Laan (2005).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206

References

M. D. Birkner and M. J. van der Laan. Application of a Variable Importance Measure Method to HIV-1 Sequence Data. Technical Report 196, Division of Biostatistics, University of California, Berkeley, Nov. 2005. URL `http://www.bepress.com/ucbbiostat/paper196/`.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32(2), 2004.

S. Rhee, W. J. Fessel, A. R. Zolopa, L. Hurley, T. Liu, J. Taylor, D. P. Nguyen, S. Slome, D. Klein, M. Horberg, J. Flamm, S. Follansbee, J. M. Schapiro, and R. W. Shafer. HIV-1 Protease and Reverse-Transcriptase Mutations: Correlations with Antiretroviral Therapy in Subtype B Isolates and Implications for Drug-Resistance Surveillance. *The Journal of Infectious Disease*, 192:456–465, 2005.

S. Rhee, J. Taylor, G. Wadhera, J. Ravela, A. Ben-Hur, D. Brutlag, and R. W. Shafer. Genotypic Predictors of Human Immunodeficiency Virus Type 1 Drug Resistance. Technical report, Stanford University, 2006. (In preparation).

I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic Regression. *Journal of Computational and Graphical Statistics*, 12(3): 475–511, 2003. URL `http://www.biostat.jhsph.edu/~iruczins/publications/publications.html/`.

S. E. Sinisi and M. J. van der Laan. Deletion/Substitution/Addition Algorithm in Learning with Applications in Genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art18/`. Article 18.

M. J. van der Laan. Statistical Inference for Variable Importance. *The International Journal of Biostatistics*, 2, 2006. URL `http://www.bepress.com/ijb/vol2/iss1/2/`.

M. J. van der Laan and S. Dudoit. Unified Cross-Validation Methodology for Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. Technical Report 130, Division of Biostatistics, University of Califor-

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

nia, Berkeley, Nov. 2003. URL http://www.bepress.com/ucbbiostat/paper130/.

M. J. van der Laan, S. Dudoit, and A. W. van der Vaart. The Cross-Validated Adaptive Epsilon-Net Estimator. Technical Report 142, Division of Biostatistics, University of California, Berkeley, February 2004. URL http://www.bepress.com/ucbbiostat/paper142/.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/ucbbiostat/paper206