UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

A Bachelor of Science thesis

Super Learners

and their oracle properties

Jinyang Liu

Supervised by Prof. Thomas Gerds
Co-supervised by Prof. Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen, Denmark

Submitted: March 2, 2023

Contents

1 Introduction

Our setup closely models what is described in [VDL06] and [LD03]. Let $X_1, ..., X_n$ be n-i.i.d. observations distributed according to $P \in \mathcal{P}$ on some measurable space $(\mathcal{X}^n, \mathcal{A})$ where $X_i \in \mathcal{X}$ for each i and $\mathcal{P}$ is our statistical model. For a parameter set Θ we define the corresponding loss function $L : \mathcal{X} \times \Theta \to [0, \infty)$ as a measurable map such that our goal is to find an estimator $\hat{\theta}$ that minimizes the true risk function $R : \Theta \to \mathbb{R}$ given as

$$R(\theta) = \int L(x, \theta) \mathrm{d}P(x) = EL(X_1)$$

The parameter set Θ can be Euclidean, but it can also be considered as a collection of functions of the form $\theta : \mathcal{X}^n \to \mathbb{R}$.

Example 1 (Regression functions Θ). Let $(Y_1, X_1), ..., (Y_n, X_n) \in \mathbb{R} \times \mathcal{X}^n$ be i.id. observations such that satisfy the model

$$Y_1 = \theta_0(X_1) + \varepsilon,$$

for an unobservable stochastic error term ε. The goal is to estimate the **regression function** $\theta_0 \in \Theta$ where $\Theta = \{\theta \mid \theta : \mathcal{X}^n \to \mathbb{R}\}$, is the set of regression functions each having $\mathcal{X}^n$ as their domain. [VDL06]

Example 2 (Parameteric family). Consider the setup from example 1. A generalized regression model could be considered as parametrized family of distributions $\mathcal{Q} = \{Q_{\theta,\eta} \mid \theta \in \Theta, \eta \in \mathbb{R}^m\}$ such that

$$Y_1 \mid X_1 = x \sim Q_{\theta(x),\eta},$$

where $\eta \in \mathbb{R}^m$ is some nuisance parameter that could relate to ε. The conditional distributions $\mathcal{Q}$ are said to be parametrized by θ and η [Bic+93]

Definition 1 (Estimator of θ_0)**.** An estimator for $\theta_0 \in \Theta$ is a measurable map $\hat{\theta} : \mathcal{X}^n \to \Theta$.

Definition 2 (Prediction model)**.**

We would like to consider a set estimators $\{\hat{\theta}_q(P_n) | 1 \leq q \leq p\}$, where we find $\hat{\theta}_{\hat{q}}(P_n)$, which denotes the estimator that minimizes R and $\hat{q}$ may depend on the observations.
In order to find $\hat{q}$ we have to proceed via cross validation. In cross validation, we randomly split our data into a training set and a test set. Let $S = (S_1, ..., S_n) \in \{0, 1\}^n$ independent of $X_1, ..., X_n$ such that $S_i = 0$ indicates that X_i should be in the training set and $S_i = 1$ indicates that X_i belongs to the test set. We can define the empirical distributions over these two

subsets, $P^0_{n,S}$ and $P^1_{n,S}$ as

$$P^0_{n,S} = \frac{1}{n_0} \sum_{i:S_i=0} \delta_{X_i}$$
$$P^1_{n,S} = \frac{1}{1-n_0} \sum_{i:S_i=1} \delta_{X_i}$$

Where n_0 would be the number of S_i's that are marked 0.

Definition 3 (True risk of q'th estimator averaged over splits)**.** Given the data $X \in \mathcal{X}^n$ and a set of estimators $\{\hat{\theta}_q \mid 1 \leq q \leq p\}, p \in \mathbb{N}$. The risks of these estimator averaged over the splits specified by some S is given as a function of q

$$q \mapsto E_S \int L(x, \hat{\theta}_q(P^0_{n,S}))\mathrm{d}P(x) = E_S R(\hat{\theta}_q(P^0_{n,S}))$$

Where P is the true distribution for our data X.

Definition 4 (Oracle selector)**.** The oracle selector is a function $\tilde{q} : \mathcal{X}^n \to \{1, ..., p\}$ which finds the estimator that minimizes the true risk given our data $X \in \mathcal{X}^n$.

$$\tilde{q}(X) = \underset{1\leq q\leq p}{\arg\min}\, E_S R(\hat{\theta}_q(P^0_{n,S}))$$

Where $P^0_{n,s}$ is the empirical distribution over the training set of X as specified by some split-variable S.

In light of the above definitions, we will define the cross-validation risk and the cross-validation selector for our estimators

Definition 5 (Cross-validation risk of i'th estimator averaged over splits)**.** Given the data $X \in \mathcal{X}^n$ and a set of estimators $\{\hat{\theta}_q \mid 1 \leq q \leq p\}, p \in \mathbb{N}$. The cross-validation risks of these estimator averaged over the splits specified by some S is given as a function of q

$$q \mapsto E_S \int L(x, \hat{\theta}_q(P^0_{n,S}))\mathrm{d}P^1_{n,s}(x) = E_S \hat{R}(\hat{\theta}_q(P^0_{n,S}))$$

Where $P^1_{n,S}$ is the empircal distribution over the validation set for our data X. We write $\hat{R}$ for empirical risk over the validation set.

Definition 6 (Cross-validation selector)**.** The cross-validation selector is a function $\hat{q} : \mathcal{X}^n \to \{1, ..., p\}$ which finds the estimator that minimizes the cross-validation risk given our data $X \in \mathcal{X}^n$.

$$\hat{q}(X) = \underset{1\leq q\leq p}{\arg\min}\, E_S \hat{R}(\hat{\theta}_q(P^0_{n,S}))$$

Where $\hat{R}$ is the empirical risk over the validation set and $P^0_{n,s}$ is the empirical distribution over the training set of X as specified by some split-variable S.

We are interested in the risk difference between the cross-validation selector and and the oracle selector, we remark that the optimal risk is attained at the true value θ_0

$$R(\theta_0) = \int L(x, \theta_0) \mathrm{d}P(x),$$

and clearly it is the case that $R(\theta_0) \leq R(\hat{\theta})$ for any estimator $\hat{\theta}$ of θ_0. Given a set of estimators we define the centered conditional risk as the difference

$$\begin{aligned}
\Delta_S(\hat{\theta}_{\hat{q}}, \theta_0) &= R(\hat{\theta}_{\hat{q}}(P^0_{n,S})) - R(\theta_0) \\
&= E_S \int L(x, \hat{\theta}_{\hat{q}}(P^0_{n,S})) - L(x, \theta_0) \mathrm{d}P(x)
\end{aligned}$$

The following result is due to [LD03]:

Theorem 7 (Asymptotic equality)**.** *The cross validation selector $\hat{q}$ performs asymptotically as well as the oracle selector $\tilde{q}$ in the sense that*

$$\frac{\Delta_S(\hat{\theta}_{\hat{q}}, \theta_0)}{\Delta_S(\hat{\theta}_{\tilde{q}}, \theta_0)} \to 1 \qquad \text{in probability for } n \to \infty$$

2 The discrete super learner, dSL

2.1 Finite sample properties

3 The ensemble super learner, eSL

4 Simulation results

5 Discussion

Pellentesque tincidunt sodales risus, vulputate iaculis odio dictum vitae. Ut ligula tortor, porta a consequat ac, commodo non risus. Nullam sagittis luctus pretium. Integer vel nibh at justo convallis imperdiet sit amet ut lorem. Sed in gravida turpis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Sed in massa vitae ligula pellentesque feugiat vitae in risus. Cras iaculis tempus mi, sit amet viverra nulla viverra pellentesque.