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 18, 2023

Contents

1 Introduction

Our setup closely models what is described in [VDL06] and [LD03]. Let $O_1, \dots, O_n$ be n-i.i.d. observations distributed according to $P \in \mathcal{P}$ on some measurable space $(\mathcal{O}, \mathcal{A})$ where $O_i \in \mathcal{O}$ for each i and $\mathcal{P}$ is our statistical model. For a parameter set Θ we define the corresponding loss function $L : \mathcal{O} \times \Theta \to [0, \infty)$ 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

$$R(\theta) = \int L(x, \theta) dP(x) = EL(O_1)$$

The parameter set Θ can be Euclidean, but for the focus of this thesis we will consider it as a collection of functions of the form $\theta : \mathcal{O} \to \mathbb{R}$.

Example 1 (Regression functions Θ). Let $O_1 = (Y_1, X_1), \dots, O_n = (Y_n, X_n) \in \mathcal{O} = \mathbb{R} \times \mathcal{X}$ be i.i.d. observations distributed according to some $P \in \mathcal{P}$ such that they satisfy the model

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

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

Example 2 (Parameteric family). Consider the initial setup from example 1. If Y_i is $\mathcal{B}(\mathbb{R}) - \mathcal{B}(\mathbb{R})$ measurable and X_i is $\mathcal{F} - \mathcal{B}(\mathbb{R})$ measurable for some sigma-algebra $\mathcal{F}$ on $\mathcal{X}$, then a **generalized regression model** could be considered as parametrized family of distributions, $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$, given that Θ is finite-dimensional.
We can parametrize the conditional probability distributions for Y_1 given $X_1 = x$ as $\mathcal{Q} = \{Q_{\theta(x)} \mid \theta \in \Theta\}$ such that $Q_{\theta(x)}$ is a valid probability distribution on $\mathcal{B}(\mathbb{R})$ for each $x \in X$ and $\theta \in \Theta$. For a given $P_\theta \in \mathcal{P}$ there will exist a $Q_\theta \in \mathcal{Q}$ such that

$$P_\theta(Y \in A \mid X = x) = Q_{\theta(x)}(A) \qquad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

If we assume that X_1 is distributed according to some H_0 on $\mathcal{X}$, then the distribution P_θ over our observations (the joint over Y and X) will be

$$P_{\theta,\eta}(X \in A, Y \in B) = \int_A Q_{\theta(x),\eta}(B) dH_0(x)$$

for every $A \in \mathcal{F}$ and $B \in \mathcal{B}(\mathbb{R})$.

Example 3 (Logistic regression model). Let $O_1 = (Y_1, X_n), \dots, O_n = (Y_n, X_n) \in \mathcal{O} = \{0,1\} \times \mathcal{X}$ be i.i.d. observations from some distribution $P_{\theta_0} \in \mathcal{P}$, where Y_i is binary and $\mathcal{X} \subseteq \mathbb{R}^k$. We would like to estimate the parameter function $\theta_0 \in \Theta$

$$\theta_0(x) = E(Y_1 \mid X_1 = x) = P_{\theta_0}(Y_1 = 1 \mid X_1 = x),$$

In logistic regression we assume that $\Theta = \{x \mapsto \text{expit}(\beta x) \mid \beta \in \mathbb{R}^k\}$, so $\theta_0(x) = \text{expit}(\beta_0 x)$, then the goal becomes to estimate the k-dimensional parameter β_0, in this case the $\mathbb{R}^k$ parameter β_0 completely determines θ_0, so Θ is also k-dimensional. The conditional distributions of Y_1 given $X_1 = x$ are Bernoulli distributions and can be parametrized as $\mathcal{Q} = \{\text{Ber}(\text{expit}(\beta x)) \mid \beta \in \mathbb{R}^k\}$. Now from example 2 we know that the statistical model, $\mathcal{P}$, can be parametrized through β, in particular we have

$$\begin{aligned} P_\beta(Y_1 = 1, X_1 \in A) &= \int_A Q_{\theta(x)}(\{1\}) dH_0(x) \\ &= \int_A \text{expit}(\beta x) dH_0(x) \end{aligned}$$

If H_0 has density f w.r.t. Lebesgue measure, we can write

$$P_\beta(Y_1 = 1, X_1 \in A) = \int_A \text{expit}(\beta x) f(x) dm(x)$$

We will now turn our attention to statistical estimators. Statistical literature commonly write that an estimator is stochastic variable taking values in our parameter space $\hat{\theta} \in \Theta$. An estimator is achieved by considering i.i.d. observations $O_1, \dots, O_2 \in \mathcal{O}$ distributed according to some measure P from some statistical model $\mathcal{P}$. We leave the model unspecified as it can be both parametric or nonparametric. Now let $h : \mathcal{O}^n \to \Theta$ be a measurable map, an estimator created from h is the random variable $T = h(O_1, \dots, O_n)$. For $\Theta \subseteq \mathbb{R}^k$ the canonical σ-algebra on Θ is the Borel algebra, but when the parameter set is a set of functions, the σ-algebra can only be chosen after careful consideration of constraints on Θ.

In the following section we introduce the terminology "estimator algorithm" which corresponds to the measurable map h from our finite sample observation space to our parameter space.

Definition 1 (Estimator algorithm $\boldsymbol{\theta}$)**.** An estimator algorithm is a measurable map $\boldsymbol{\theta} : \mathcal{O}^n \to \Theta$ for $n \in \mathbb{N}$.

Definition 2 (Statistical Estimator $\hat{\theta}$)**.** Let $O_1, \dots, O_n \in \mathcal{O}$ be i.i.d. observations distributed according to some $P \in \mathcal{P}$ for a statistical model $\mathcal{P}$ on $\mathcal{O}$. Let $\boldsymbol{\theta} : \mathcal{O}^n \to \Theta$ be an estimator algorithm. An estimator is the random variable $\hat{\theta} = \boldsymbol{\theta}(O_1, \dots, O_n) \in \Theta$.

There is a one-to-one correspondence between the tuples of i.i.d. observations $(O_1, \dots, O_n) \in \mathcal{O}^n$ and the empirical measures over n observations on $(\mathcal{O}, \mathcal{A})$ defined as

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(O_i) \qquad \text{for } A \in \mathcal{A}.$$

Note that the empirical measure is a random variable. Thus, we can write $\boldsymbol{\theta}(P_n)$ as an alternative representation of the estimator $\boldsymbol{\theta}(O_1, \dots, O_n)$, by adjusting the notation without introducing ambiguity.

Example 4 (Prediction algorithm). Consider the setup from example 3, where we have i.i.d. observations $O_1 = (Y_1, X_1), \ldots, O_n = (Y_n, X_n)$ such that $Y_i \in \{0,1\}$ and $\mathcal{X} \in \mathbb{R}^k$ and our goal is to estimate the probability $\theta(x) = P_\theta(Y_1 = 1 \mid X_1 = x)$...

We would now like to consider the scenario where we have a library (set) of estimator algorithms, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$. From these algorithms, we can define the set of estimators $\{\hat{\theta}_q = \boldsymbol{\theta}_q(P_n) | 1 \leq q \leq p\}$, where our goal is to 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 $S = (S_1, \ldots, S_n) \in \{0,1\}^n$ independent of $X_1, \ldots, X_n$ 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

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

Where P is the true distribution for our data X.

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

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

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

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

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 $\hat{q} : \mathcal{X}^n \to \{1, \dots, p\}$ which finds the estimator that minimizes the cross-validation risk given our data $X \in \mathcal{X}^n$.

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

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

$$R(\theta_0) = \int L(x, \theta_0) dP(x),$$

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

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

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*

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

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.