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: April 4, 2023

Contents

1 Introduction

In the context of prediction, the goal is to estimate the conditional expectation $E(Y \mid X = x)$ for i.i.d. observation pairs $(Y_1, X_1), \ldots, (Y_n, X_n)$. Depending on the data structure, a variety of standard statistical models can be employed. For instance, if Y is binary, a parametric model like logistic regression might be suitable. The task of identifying true statistical model $\mathcal{P}$ for which $(Y, X) \sim P \in \mathcal{P}$, is challenging, and perhaps infeasible when we only have a limited amount data. It may therefore be motivating ot utilize non-parametric and data-driven regression methods, such as tree-based algorithms like XGBoost or random forests to estimate the conditional mean. However, the underlying assumptions of tree-ensemble methods regarding the data generating process are not explicit, and they may not have probabilistic interpretations. We can nevertheless incorporate these methods as a part of our repertoire, but it is important that we can compare and choose the best method that most effectively accompishes our goal, for example prediction.
The 'super learner' is the answer to how we can effectively select the 'best' learner (method or algorithm) among the learners that we have in our library of learners. The cross-validation selector, which evaluates learners on their cross-validated (empirical) risk and chooses the one with the lowest risk, is asymptotically equivalent to the oracle selector. The oracle selector finds the learner with the lowest true risk – the risk obtained by knowing the true distribution.
The discrete super learner is obtained by applying the cross-validation selector on our data. The asymptotic result shows that the cross-validation selector will select the same learner as the oracle selector when the number of observations goes to ∞. The discrete super learner is not a fixed learner from our library, but rather depends on the available data. It represents the learner chosen by the cross-validation selector, which can vary depending on the amount of data at hand.
We first present the general theory and our goal, which is to estimate the conditional expectation $E(Y \mid X = x)$ for Y, X being some outcome-covariate pair. More specifically, we focus on the case where we regress on a binary outcome $Y \in \{0, 1\}$. The conditional expectation of Y given X exactly becomes the conditional probability $P(Y = 1 \mid X = x)$.

2 Background

Our setup closely models what is described in [VDL06] and [LD03]. Let $O_1, \ldots, O_n$ 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 $O_1 = (Y_1, X_1), \ldots, O_n = (Y_n, X_n) \in \mathcal{O} = \mathbb{R} \times \mathcal{X}$ 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

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

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

Example 3 (Logistic regression model). Let $O_1 = (Y_1, X_n), \ldots, O_n = (Y_n, X_n) \in \mathcal{O} = \{0,1\} \times \mathcal{X}$ 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 $O_1, \ldots, O_2 \in \mathcal{O}$ 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 $T = h(O_1, \ldots, O_n)$. 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.

3 The discrete super learner, dSL

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 h)**.** An estimator algorithm is a measurable map $h : \mathcal{O}^n \to \Theta$ for $n \in \mathbb{N}$.

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

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

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

Note that the empirical measure is a random variable. Thus, we can write $h(P_n)$ as an alternative representation of the estimator $h(O_1, \ldots, O_n)$, 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 learner algorithms, $h_1, \ldots, h_n$. From these algorithms, we can define the set of learners $\{\hat{\theta}_q = h_q(P_n) | 1 \leq q \leq p\}$, where our goal is to find $\hat{\theta}_{\hat{q}}(P_n)$, which denotes the learner 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 the random binary vector $S = (S_1, \ldots, S_n) \in \{0, 1\}^n$ be 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.

Example 5 (Random splits). For $n = 9$ observations one could for example define the distribution of the random vector S as

$$P(S = (0, 0, 0, 0, 0, 0, 1, 1, 1)) = \frac{1}{3}$$
$$P(S = (0, 0, 0, 1, 1, 1, 0, 0, 0)) = \frac{1}{3}$$
$$P(S = (1, 1, 1, 0, 0, 0, 0, 0, 0)) = \frac{1}{3},$$

i.e. 3-fold cross-validation.
In general for n observations we have 2^n ways of choosing which observations should be in the training set and in the validation set. It might not be desirable to define the discrete probabilities for S over $\{0, 1\}^n$ simply as $\frac{1}{2^n}$ for each possible combination of training/validation data, since that would also include the combination where $n_1 = 0$. To ensure that we always have $n_1 > 0$, then let n_1 be given, then we see that there are $\binom{n}{n_1}$ ways of choosing both the validation and training set. We can therefore define the distribution of S as

$$P(S = s) = \binom{n}{n_1}^{-1} \qquad \text{for each } s \in \{0, 1\}^n \text{ where } \sum_i s_i = n_1$$

3.1 Risks and selectors

Definition 3 (True risk of q'th learner averaged over splits)**.** Given the data $O_1, \ldots, O_n \in \mathcal{O}$ and a set of learners $\{\theta_q(P^0_{n,S}) \mid 1 \leq q \leq k\}, k \in \mathbb{N}$ applied to our training data $P^0_{n,S}$. The risks of these learners averaged over some split-variable S is given as a function of q

$$q \mapsto E_S \int L(o, \theta_q(P^0_{n,S}))dP(o) = E_S R(\theta_q(P^0_{n,S}))$$

Where P is the true distribution for our data X.

Definition 4 (Oracle selector)**.** The oracle selector is a function $\tilde{q} : \mathcal{O}^n \to \{1, \ldots, k\}$ which finds the learner that minimizes the true risk given our data $O_1, \ldots, O_n \in O$.

$$\tilde{q}(O_1, \ldots, O_n) = \arg\min_{1 \leq q \leq k} E_S R(\theta_q(P^0_{n,S}))$$

Where $P^0_{n,s}$ is the empirical distribution over the training set of $O_1, \ldots, O_n$ as specified by some split-variable S.

In similar manner to the above the defintions, we can define the cross-validation risk and the cross-validation selector for our learners

Definition 5 (Cross-validation risk of q'th learner averaged over splits)**.** Given the data $O_1, \ldots, O_n \in \mathcal{O}$ and a set of learners $\{\theta_q(P^n_{n,S}) \mid 1 \leq q \leq k\}, k \in \mathbb{N}$. The cross-validation risks of these learners averaged over some split-variable S is given as a function of q

$$q \mapsto E_S \int L(o, \theta_q(P^0_{n,S}))dP^1_{n,S}(o) = E_S \hat{R}(\theta_q(P^0_{n,S}))$$

Where $P^1_{n,S}$ is the empircal distribution over the validation of $O_1, \ldots, O_n$. We write $\hat{R}$ for the empirical risk over the validation set.

Definition 6 (Cross-validation selector)**.** The cross-validation selector is a function $\hat{q} : \mathcal{O}^n \to \{1, \ldots, k\}$ which finds the learner that minimizes the cross-validation risk given our data $O_1, \ldots, O_n \in \mathcal{O}$.

$$\hat{q}(O_1, \ldots, O_n) = \arg\min_{1 \leq q \leq k} E_S \hat{R}(\theta_q(P^0_{n,S}))$$

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 $O_1, \ldots, O_n$ 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

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

and clearly it is the case that $R(\theta_0) \leq R(\theta)$ for any learner θ of θ_0. Given a set of learners we define the centered conditional risk as the difference

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

3.2 Oracle inequalities

We introduce the notation Pf for the integral $\int f dP$ of an integrable function f with respect to P. Additionally, if P_n represents the empirical measure of $O_1, \ldots, O_n$, we denote the empirical process indexed over an appropriate class of functions $\mathcal{F}$ as $G_n f = \sqrt{n}(P_n f - Pf)$. Furthermore, we extend this notation to $G^i_{n,S} f = \sqrt{n}(P^i_{n,S} - Pf)$ for the empirical processes that correspond to applying the empirical measure over either the training sample or validation sample.

Lemma 7 (Lemma 2.1 in [VDL06]). *For $\delta > 0$ it holds that*

$$\begin{aligned} E_S \int L(o, \theta_{\hat{q}}(P^0_{n,S}))dP(o) \leq (1+2\delta)E_S \int L(o, \theta_{\tilde{q}}(P^0_{n,S}))dP(o) \\ + \frac{1}{\sqrt{n_1}} E_S \max_{1\leq q\leq k} \int L(o, \theta_q(P^0_{n,S}))d((1+\delta)G^1_{n,S} - \delta\sqrt{n_1}P)(o) \\ - \frac{1}{\sqrt{n_1}} E_S \max_{1\leq q\leq k} \int L(o, \theta_q(P^0_{n,S}))d((1+\delta)G^1_{n,S} + \delta\sqrt{n_1}P)(o) \end{aligned}$$

Proof. See appendix □

Theorem 8 (Theorem 2.3 in [VDL06]). *For $\theta \in \Theta$ let $(M(\theta), v(\theta))$ be a Bernstein pair for the function $o \mapsto L(o, \theta)$ and assume that $R(\theta) = \int L(o, \theta)dP(o) \geq 0$ for every $\theta \in \Theta$. Then for $\delta > 0$ and $1 \leq p \leq 2$ it holds that*

$$\begin{aligned} ER(\theta_{\hat{q}}(P^0_{n,S})) \leq & (1+2\delta)ER(\theta_{\tilde{q}}(P^0_{n,S}))+ \\ & (1+\delta)E\left(\frac{16}{n_1^{1/p}}\log(1+k)\sup_{\theta\in\Theta}\left[\frac{M(\theta)}{n_1^{1-1/p}} + \left(\frac{v(\theta)}{R(\theta)^{2-p}}\right)^{1/p}\left(\frac{1+\delta}{\delta}\right)^{2/p-1}\right]\right), \end{aligned}$$

where k is the number of learners in our library $\{\theta_q(P^0_{n,S}) \mid 1 \leq p \leq k\}$.

Example 6 (Binary regression). Consider the case where we have i.id. observations $O_1 = (Y_1, X_1), \ldots, O_n = (Y_n, X_n)$ such that $Y_i \in \{0, 1\}$ and $X \in \mathbb{R}^d$ distributed according some $P \in \mathcal{P}$. We would like to estimate the conditional expectation $\theta_0(x) = E(Y \mid X = x) = P(Y = 1 \mid X = x)$. Let $\Theta = \{\theta \mid \theta : \mathcal{X} \to [0, 1]\}$ and choose the quadratic loss function $L((Y, X), \theta) = (Y - \theta(X))^2$.
We observe that the quadratic loss is bounded by 1 for all choices of $\theta \in \Theta$ and $O \in \mathcal{O}$, it follows that for every $\theta \in \Theta$ that $M(\theta) = 1$ and $v(\theta) = 1$ is a valid Bernstein pair for the function $o \mapsto L(o, \theta)$. It is clear that $R(\theta) = \int L(o, \theta)dP(o) \geq 0$, if we plug these numbers in theorem 8, then we have

$$\begin{aligned} ER(\theta_{\hat{q}}(P^0_{n,S})) &\leq (1+2\delta)ER(\theta_{\tilde{q}}(P^0_{n,S}))+ \\ &\quad (1+\delta)E\left(\frac{16}{n_1}\log(1+k)\sup_{\theta\in\Theta}\left[1 + \frac{1}{R(\theta)}\frac{1+\delta}{\delta}\right]\right) \\ &= (1+2\delta)ER(\theta_{\tilde{q}}(P^0_{n,S}))+ \\ &\quad (1+\delta)E\left(\frac{16}{n_1}\log(1+k)\sup_{\theta\in\Theta}\left[1 + \frac{1}{R(\theta)}\frac{1+\delta}{\delta}\right]\right) \end{aligned}$$

The following result is due to [LD03]:

Theorem 9 (Asymptotic equality)**.** *The cross validation selector $\hat{q}$ performs asymptotically as well as the oracle selector $\tilde{q}$ in the sense that*

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

4 The ensemble super learner, eSL

5 Simulation results

6 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.