Boosting the First-Hitting-Time Regression Model

Vegard Stikbakke

May 14, 2018

Abstract

Empty.

Acknowledgements

Empty.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

sec:intro

In this thesis, we work with boosting for regression in the first hitting time model. First hitting time is a model in survival analysis which serves as an alternative to the proportional hazards model, typically known as Cox regression. Developments in FHT regression are relatively recent, and there has to our knowledge been no attempt at tackling it in the high-dimensional case, in which boosting is an appropriate choice of method.

Chapter 2

First hitting time regression models

2.1 Survival analysis and time-to-event models

sec:survival

Lifetimes and time-to-event data are of interest in many applications. Oncologists, doctors who study cancer, are interested in how quickly patients die after cancer has been discovered. Sociologists might be interested in the duration of marriages before divorce. We say that a lifetime T ends when an event occurs. In the previous examples, the events in question are death and divorce, say. We are usually interested in making inference about this lifetime, and in particular what factors it depends upon. In biomedical fields, this is known as survival analysis, while in engineering it is called reliability analysis. These are much studied fields. The main part of our thesis is applicable in both areas. In the former case, we may consider the time before a component of a system breaks and must be replaced. Let the lifetime or time-to-event T be a continuous non-negative random variable following a cumulative distribution function $F(t)$, such that

$$F(t) = \Pr(T < t)$$

is the probability of the event having happened before time t. We define the survival function $S(t)$ to be the converse, namely

$$S(t) = 1 - F(t).$$

Hence, $S(t)$ denotes the probability that the event has not yet happened at time t. If the cumulative distribution function is differentiable, we define the probability density function of T to be $f(t)$. Another much studied property of lifetime distributions is the hazard function $h(t)$. Somewhat informally, it denotes the probability of the event happening at some time t, assuming it has not happened yet. More formally, it is the limit of the conditional probability that the event will occur in a small interval $[t, t + \Delta t)$, conditional on the event not having happened yet,

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{F(t)}.$$

Censored data

In theory, a lifetime will always end. In the real world, however, we are constrained with finite time. Thus, when we observe lifetime data, it is not necessarily the case that the lifetime has ended yet. For example, some cancer patients survive, and die of old age. Similarly, some marriages do not end in divorce. In these cases, it is likely that the lifetime has not ended at the time we stop observing data. Here we cannot say anything about the lifetime itself. We can only say that this lifetime lasted at least until now. We call these observations censored observations. Specifically, these are right-censored observations, since they are censored at the right end of the time scale.

sec:surv-data

Data structures

Assume we observe $t_{(i)}$, $i = 1, \ldots, n$ independent and identically distributed (*iid*) observations from a random variable with density distribution f. For a single individual event i, we might observe the following. The time $t_{(i)}$ of the observation. Covariates $\mathbf{x}_{(i)}$ describing the individual. An indicator $\delta_{(i)}$ of whether the individual event has occurred ($\delta_{(i)} = 1$) or not ($\delta_{(i)} = 0$). The latter events, corresponding to $\delta_{(i)} = 0$, are censored. We are interested in setting up the likelihood, where we incorporate the covariates $\mathbf{x}_{(i)}$ into a parameter $\boldsymbol{\theta}$. If the event has occurred, the indicator $\delta_{(i)}$ is 1. We can then use the information about the lifetime distribution, such that the single individual i contributes

$$f(t_{(i)}|\mathbf{x}_{(i)}) \tag{2.1}$$

{eq:f}

to the likelihood. If the event has not (yet) occurred, the observation is censored, and $\delta_{(i)}$ is 0. Since we do not have the actual lifetime, we cannot use the lifetime distribution. We must use the survival distribution. As such, this observation contributes

$$S(t_{(i)}|\mathbf{x}_{(i)}) \tag{2.2}$$

{eq:S}

to the likelihood. Obviously, since an observation can only be either censored or not censored at the same time, $\delta_{(i)}$ is either 0 or 1. If the event has occurred, $\delta_{(i)}$ is 1, and then $1 - \delta_{(i)}$ is 0. Similarly, if the event has not occurred and the event is censored, then $\delta_{(i)}$ is 0, and then $1 - \delta_{(i)}$ is 1. This allows us to take the product of (2.1) and (2.2) where we take these to the power of $\delta_{(i)}$ and $1 - \delta_{(i)}$, respectively, so that a single observation makes a contribution of

$$f(t_{(i)}|\mathbf{x}_{(i)})^{\delta_{(i)}} S(t_{(i)}|\mathbf{x}_{(i)})^{1-\delta_{(i)}}$$

to the likelihood. Since we assumed the observations were independent, the likelihood of the observed sample as a whole is the product of the single likelihoods. The likelihood becomes

$$L(\boldsymbol{\theta}|\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}) = \prod_{i=1}^{n} f(t_{(i)}|\mathbf{x}_{(i)}, \boldsymbol{\theta})^{\delta_{(i)}} S(t_{(i)}|\mathbf{x}_{(i)}, \boldsymbol{\theta})^{1-\delta_{(i)}}. \tag{2.3}$$

{eq:surv-lik}

Proportional hazards

The most used method for doing regression on survival data is the Cox proportional hazards (PH) regression. It is based on an assumption that is often

called the PH property or the PH assumption, namely that

$$h(t|x) = h_0(t)g(\mathbf{x}), \tag{2.4}$$

where $h_0(t)$ is a baseline hazard function, which is common for all individuals. This means that at any two time points t_1 and t_2, the ratio between the hazard functions of any two $\mathbf{x}_1$ and $\mathbf{x}_2$ will be the same:

$$\frac{h(t_1|x_1)}{h(t_1|x_2)} = \frac{h(t_2|x_1)}{h(t_2|x_2)} \tag{2.5}$$

This is a strong assumption to make, and it will not always be the case in practice (Lee and Whitmore, 2010). Therefore there is a need for more flexible methods in survival analysis, where this assumption is not necessary.

2.2 The First Hitting Time (FHT) Model

sec:fht

Time-to-event data analysis in biomedical sciences is dominated by Cox regression, which is a semiparametric proportional hazards model, and which directly estimates the hazard rate (Stogiannis and Caroni, 2013). A class of parametric models which has got increasingly more attention recently is the first hitting time (FHT) model, originally developed by Whitmore in 1986 (Whitmore, 1986; Lee and Whitmore, 2006). The FHT model has been applied successfully to different kinds of data, especially in biomedial sciences. Examples include modelling lung cancer risk in railroad workers (Lee et al., 2004), ... In an FHT model, the health of an individual is modelled as a stochastic process, which, when it reaches some threshold (or, more generally, an absorbing state), triggers the event, at which point the lifetime ends. The time-to-event, or lifetime, becomes the time it takes for the process to reach this state. This is an attractive model because it models the process instead of the hazard rate (Aalen and Gjessing, 2001). It is also conceptually appealing, because it makes sense to imagine that there is some process governing when events happen, such that for two living individuals, they might have different distances, so to speak, away from death. We now descibe the key components in the FHT model. There is a parent stochastic process $\{Y(t)\}$, $Y \in \mathcal{Y}$, time t non-negative, $t \in \mathcal{T}$, with $t \geq 0$. The process has initial value $Y(0) = y_0$. There is a boundary set $\mathcal{B} \subset \mathcal{T}$, which is at times referred to as a boundary, barrier, or threshold, all of which are synonymous. The preferred term varies with which interpretation we want to use and what connotations we want to evoke. The choice of process is flexible. It might have continuous or discrete sample paths. We define the first hitting time S to be the first time t that the process Y reaches the absorbing state $B \in \mathcal{B}$,

hmm ... is this good enough?

$$S = \inf\{t \colon Y(t) \in \mathcal{B}\}. \tag{2.6}$$

Note that by definition $y_0 \notin \mathcal{B}$, since the event has not yet happened at time $t = 0$. Note also that it is quite possible that the process does not reach an absorbing state, and so that $P(S < \infty) < 1$. The FHT model does not require the PH assumption, and is hence more flexible. In fact, the PH model may be obtained by constructing the first hitting time model in a specific way (Lee and Whitmore, 2010). Different choices for the process Y lead to different kinds of distributions for the first hitting time. We now look at some common choices for the process.

sec:wiener

Wiener process

The Wiener process, also known as the standard Brownian motion process, is a process which is continuous in time and space. It is so far the most commonly used process in FHT literature. The Wiener process is a fairly simple process: It has three parameters, the drift μ, the variance σ^2, and the initial value $Y(0) = y_0$. It has independent increments, such that $Y(t_2) - Y(t_1)$ and $Y(t_4) - Y(t_3)$ are independent for any disjoint intervals (t_1, t_2) and (t_3, t_4). Each increment is normally distributed and with both the mean and the standard deviation proportional to the length of the interval, i.e., for any interval (t_1, t_2),

$$Y(t_2) - Y(t_1) \sim N(\mu(t_2 - t_1), \sigma^2(t_2 - t_1)). \tag{2.7}$$

If we consider Y to be a Wiener process in the FHT model, we will typically choose the boundary $\mathcal{B}$ to be $\mathcal{B} = (-\infty, 0]$, i.e., the event is triggered when $Y \leq 0$. Accordingly, we then also assume that the initial level y_0 is positive. This is a very conceptually appealing model, because it assumes that individuals might have different initial levels, and that also the drift might be different between individuals. It is also attractive because it has closed-form probability and cumulative density functions, and its likelihood is computationally simple. There are no restrictions on the movements of the process, meaning, it is non-monotonic. If we do want a monotonic restriction on the movement of the process, we may use a gamma process.

Other processes

The gamma process is suitable for modelling a process which we would require to be monotonic, typically a physical degradation, i.e. where the damage cannot mend itself, unlike a patient's health. The first hitting time that arises from the gamma process is an inverse gamma first hitting time distribution (Lee and Whitmore, 2006). Other choices of processes include Markov chain state models, the Bernoulli process, and the Ornstein-Uhlenbeck process. For a complete review, see Lee and Whitmore (2006). Due to its simplicity and flexibility, we will in thesis focus on the Wiener process as the choice of process in the FHT model. However, large parts of our results can easily be extended for other processes. For brevity, we will in the sequel say "the FHT" when we in fact mean the FHT with the Wiener process.

2.3 First hitting time regression based on underlying Wiener process

It can be shown that the first hitting time of the Wiener process follows an inverse Gaussian distribution (Chhikara, 1988),

$$f(t|y_0, \mu, \sigma^2) = \frac{y_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left[-\frac{(y_0 + \mu t)^2}{2\sigma^2 t}\right]. \tag{2.8}$$

{eq:fht-ig}

TODO!

See Appendix for the mathematical derivation. If the drift μ is positive, then it is not certain that the process will reach 0, so in this case it is an improper pdf. In most cases, y is not measured directly. If that is the case, then the scale of y is arbitrary. Thus, we may fix one parameter in the distribution. As

per convention we choose to set the variance to unity, i.e., $\sigma^2 = 1$ (Lee and Whitmore, 2006; Caroni, 2017). While μ and y_0 have simple interpretations in terms of the underlying process, they do not in terms of the lifetime distribution. The mean lifetime is $\frac{y_0}{|\mu|}$, and its variance is $\frac{y_0}{|\mu|^3}$ (Caroni, 2017). This

The cumulative distribution function of the FHT is (Xiao et al., 2015)

$$F(t|\mu,\sigma^2,y_0) = \Phi\left[-\frac{(\mu t + y_0)}{\sqrt{\sigma^2 t}}\right] + \exp\left(-\frac{2y_0\mu}{\sigma^2}\right)\Phi\left[\frac{\mu t - y_0}{\sqrt{\sigma^2 t}}\right], \tag{2.9}$$

{eq:cumulative}

where $\Phi(x)$ is the cumulative distribution function of the standard normal, i.e.,

$$\Phi(x) = \int_{-\infty}^{x} \exp(-y^2/2)/\sqrt{2\pi}\ \mathrm{d}y. \tag{2.10}$$

The survival function $S(t)$ is

$$\begin{aligned} S(t) &= 1 - \Phi\left[-\frac{(\mu t + y_0)}{\sqrt{\sigma^2 t}}\right] - \exp\left(-\frac{2y_0\mu}{\sigma^2}\right)\Phi\left[\frac{\mu t - y_0}{\sqrt{\sigma^2 t}}\right] \\ &= \Phi\left[\frac{\mu t + y_0}{\sqrt{\sigma^2 t}}\right] - \exp\left(-\frac{2y_0\mu}{\sigma^2}\right)\Phi\left[\frac{\mu t - y_0}{\sqrt{\sigma^2 t}}\right]. \end{aligned} \tag{2.11}$$

{eq:survival}

In the last step we use the fact that $1 - \Phi(-x) = \Phi(x)$, since the standard normal distribution is symmetric around 0.

Regression

We may introduce effects from covariates by allowing μ and y_0 to depend on covariates $\mathbf{x}$ and $\mathbf{z}$. A simple and much used model is to use the identity and the logarithm link function for the drift μ and the initial level y_0, respectively.

$$\begin{aligned} \mu &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \\ \ln y_0 &= \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z} \end{aligned} \tag{2.12}$$

{eq:coeffs}

$\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression coefficients. Note that we may let $\mathbf{x}$ and $\mathbf{z}$ share none, some, or all elements.

2.4 Likelihood

sec:lik

In formila (2.3), we reported the likelihood of lifetime regression models in its most general formulation. For an inverse Gaussian FHT this then becomes (inserting (2.8) and (2.11) into (2.3))

$$\begin{aligned} L(\boldsymbol{\theta}) = &\left(\frac{y_0}{\sqrt{2\pi\sigma^2 t^3}}\exp\left[-\frac{(y_0+\mu t)^2}{2\sigma^2 t}\right]\right)^{\delta_i} \\ &\times\left[\Phi\left(\frac{\mu_t + y_0}{\sqrt{\sigma^2 t}}\right) - \exp\left(-\frac{2y_0\mu}{\sigma^2}\right)\Phi\left(\frac{\mu t - y_0}{\sqrt{\sigma^2 t}}\right)\right]^{1-\delta_i} \end{aligned} \tag{2.13}$$

{eq:fht-lik}

We can now substitute the covariates in (2.12) into this.

This can be expanded a lot more!

Optimization

not really!!

Until now, mainly numerical maximum likelihood methods have been used to find optimal parameters, via direct maximization of the likelihood. Finding a closed-form solution for the maximum likelihood is not possible. It is only feasible to do numerical optimization of the maximum likelihood in the low-dimensional case, since it will optimize the entire parameter space at once. Therefore it is necessary to develop methods which can deal with high-dimensional cases. That is what we intend to do in the main part of the thesis. For our purposes, we need to differentiate the logarithm with respect to the parameters μ and y_0. Since the logarithm is monotone, it preserves optimality, and hence we can take the logarithm of (2.13), and we get

$$\begin{aligned} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_{(i)} \Bigg(& \ln y_0 - \frac{1}{2} \ln\Big(2\pi\sigma^2 t_{(i)}^3\Big) - \frac{\big(y_0 + \mu t_{(i)}\big)^2}{2\sigma^2 t_{(i)}} \Bigg) \\ & + (1 - \delta_{(i)}) \ln\left(\Phi\left(\frac{\mu t_{(i)} + y_0}{\sqrt{\sigma^2 t_{(i)}}} \right) - \exp\left(-\frac{2 y_0 \mu}{\sigma^2} \right) \Phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}} \right) \right) \end{aligned} \tag{2.14}$$

{eq:loglik}

See the appendix A for the full derivation.

Chapter 3

Statistical boosting

3.1 Statistical learning theory

sec:learning-theory

Assume we have a joint distribution $(\mathbf{X}, Y)$, $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$, and $Y = F(\mathbf{X}) + \varepsilon$, $F(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x})$ We wish to estimate the underlying $F(\mathbf{X})$. For an estimate $\hat{F}(\cdot)$, we measure the loss, or the difference, with a loss function

$$\mathrm{L}(Y, \hat{F}(\mathbf{X})).$$

A common loss function for regression is the squared loss, also known as the L_2 norm,

$$\mathrm{L}(Y, \hat{F}(\mathbf{X})) = (Y - \hat{F}(\mathbf{X}))^2.$$

For a $\hat{F}(X)$, we wish to estimate the expected loss, also known as the generalization or test error,

$$\mathrm{Err}_\tau = \mathrm{E}[\mathrm{L}(Y, \hat{F}(\mathbf{X}))|\tau],$$

where (X, Y) is drawn randomly from their joint distribution and the training set τ is held fixed. It is infeasible to do effectively in practice and hence we must instead estimate the expected prediction error, because?

$$\mathrm{Err} = \mathrm{E}[\mathrm{Err}_\tau] = \mathrm{E}_\tau \left([\mathrm{L}(Y, \hat{F}(\mathbf{X}))|\tau] \right), \tag{3.1}$$

{eq:err}

i.e., average over many different test sets. In practice, we observe a sample $(\mathbf{x}_i, y_i)_{i=1}^N$. For this sample, we can calculate the training error,

$$\overline{\mathrm{err}}(F) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{L}(y_i, \hat{F}(\mathbf{x}_i)), \tag{3.2}$$

{eq:empirical-risk}

also known as the empirical risk. To estimate err (3.1), one can do two things. First, if the observed sample is large enough, one can choose a portion of this, say 20%, to be used as a hold-out test set. We then train/fit/estimate based on the other 80%, and estimate Err by

$$\hat{\mathrm{Err}} = \frac{1}{M} \sum_{i=1}^{M} L(y_i, \hat{F}(\mathbf{x}_i)),$$

where (x_i, y_i) here are from the test set.

3.2 The history of boosting

history

Boosting is one of the most promising methodological approaches for data analysis developed in the last two decades (Mayr et al., 2014). Boosting originated in the fields of computational learning theory and machine learning. Kearns and Valiant (1989), working on computational learning theory, posed a question: Could any weak learner be transformed to become a strong learner? A weak learner, sometimes also simple or base learner, means one which has a low signal-to-noise ratio, and which in general performs poorly. For classification purposes it is easy to give a good example: A weak learner is one which performs only slightly better than random uniform chance. Freund and Schapire (1996) then invented the AdaBoost algorithm for constructing a binary classifier. It was evidence that the answer to the original question was positive. The AdaBoost algorithm performs iterative reweighting of original observations. For each iteration, it gives more weight to misclassified observations, and then trains a new weak learner based on all weighted observations. It then adds the new weak learner to the final classifier. The resulting AdaBoost classifier is then a linear combination of these weak classifiers, i.e., a weighted majority vote. In its original formulation, the AdaBoost classifier does not have interpretable coefficients, and as such it is a so-called black-box algorithm. In statistics, however, we are interested in models which are interpretable.

"something to connect the sentences"???? how! new section on AdaBoost? also rephrase weak learner part!

3.3 Statistical boosting

sec:sboost

rephrase!

In statistics, we are not just interested in prediction accuracy. We also want to estimate the relation between observed predictor variables and the expectation of the response,

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = F(\mathbf{x}). \tag{3.3}$$

{eq:exp-f}

In addition to using boosting for classification, like in the original AdaBoost, we would also like to use it in more general settings, and we therefore extend our discussion to a more general regression scheme where the outcome variable Y can be continuous. We are interested in interpreting how the different covariates of $\mathbf{X}$ affect $\mathrm{E}(\cdot)$. A choice for $F(\mathbf{X})$ which is amenable to such interpretation is the generalized additive model (GAM),

$$F(\mathbf{x}) = \alpha + \sum_{j=1}^{p} f_j(x_j), \tag{3.4}$$

{eq:gam}

where $\alpha \in \mathbb{R}$ is an offset and x_j is the j-th component of $\mathbf{x}$. $F(\mathbf{x})$ is a sum of component-wise functions f_j, and as such a GAM contains interpretable additive predictors. Friedman et al. (2000) showed that AdaBoost fits a GAM with a forward stagewise algorithm, for a particular exponential loss function. This provided a way of viewing boosting through a statistical lens. Friedman (2001) further investigated the topic, providing exceptional insight into boosting. However, we must first discuss how we find an approximate solution for $F(\mathbf{X})$ in (3.3).

3.4 Finding a solution

We wish to find $F(\mathbf{X})$ in (3.3), so we are interested in solving

$$\underset{F}{\operatorname{argmin}} \operatorname{E}\left[\operatorname{L}(Y, F(\mathbf{X}))\right],$$

where L is an appropriate loss function. But as discussed in chapter 3.1, we have in practice observed a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^N$, drawn from the joint distribution of $(\mathbf{X}, Y)$. Therefore we substitute the expected loss with the expected risk (3.2), and so we want a solution to

$$\underset{F}{\operatorname{argmin}} \overline{\operatorname{err}}(F). \tag{3.5}$$

{eq:argmin-risk}

Finding F is not easy. We will now discuss a general optimization algorithm used in boosting.

3.5 Gradient descent

Gradient descent is an optimization algorithm for minimizing a differentiable multivariate function F. The motivation behind the gradient descent algorithm is that in a small interval around a point $\mathbf{x}$, F is decreasing in the direction of the negative gradient at $\mathbf{x}$. Therefore, by moving slightly in that direction, F will decrease. Indeed, with a sufficiently small step length, gradient descent will always converge, albeit to a local minimum. More formally, the algorithm is

1. Start with an initial guess $\mathbf{x}_0$, e.g. $\mathbf{x}_0 = \mathbf{0}$. Let $m = 1$.
2. Calculate the direction to step in, $\mathbf{g}_{m-1} = -\nabla F(\mathbf{x}_{m-1})$.
3. Let $\mathbf{h}_m = \nu \mathbf{g}_{m-1}$, where ν is the best step length, found by
$$\nu = \underset{\nu}{\operatorname{argmin}}\, \mathbf{x}_{m-1} + \nu \mathbf{g}_{m-1}$$
4. Let $\mathbf{x}_m = \mathbf{x}_{m-1} + \mathbf{h}_{m-1}$
5. Increase m, and go to step 2. Repeat until $m = M$.
6. The resulting final guess is $\mathbf{x}_M = \mathbf{x}_0 + \sum_{m=1}^M \mathbf{h}_m(\mathbf{x}_m)$

Back to our goal of finding an F to minimize (3.5). Often we choose a parameterized model $F(\mathbf{X}; \boldsymbol{\beta})$. Finding the optimal $\boldsymbol{\beta}$ analytically might be infeasible. The gradient descent algorithm can then be used. In this case, we fix $\mathbf{X}$ and let $F(\mathbf{X})$ in the algorithm be $F(\boldsymbol{\beta}; \mathbf{X})$. Thus we use gradient descent to find an optimal $\boldsymbol{\beta}$. We would then say we are doing gradient descent in parameter space. We are now ready to reveal Friedman's useful insight.

3.6 Gradient boosting: Functional gradient descent

There is another possible way to use gradient descent, and that is the important insight by Friedman in 2001 (Friedman, 2001). He showed that boosting can be viewed as an optimization procedure in functional space. Briefly, we first

describe a naive way of doing this. Consider the function value at each $\mathbf{x}$ directly as a parameter, and use gradient descent directly on these parameters. However, this does not generalize to unobserved values $\mathbf{X}$, and we are after all interested in the population minimizer of (3.3). We can instead assume a parameterized form for F, e.g.,

needs work!

$$F(\mathbf{X}; \{\boldsymbol{\beta}\}_{m=1}^{M}) = \sum_{m=1}^{M} \nu H(\mathbf{X}; \boldsymbol{\beta}_m), \tag{3.6}$$

{eq:gradboost}

where $H(\mathbf{X}; \boldsymbol{\beta})$ is also a function on the GAM form (3.4), but typically simpler, i.e., a base learner as discussed previously. We would like to minimize a data based estimate of the loss, i.e. the empirical risk, and so would choose $\{\boldsymbol{\beta}_m\}$ as the minimizers of

$$\underset{\{\boldsymbol{\beta}_m\}_{m=1}^{M}}{\operatorname{argmin}} \overline{\operatorname{err}}(H(\mathbf{x}; \{\boldsymbol{\beta}_m\})).$$

However, estimating these simultaneously may be infeasible. We can then use a greedy stagewise approach, where we at each step m choose the $\boldsymbol{\beta}_m$ which gives the best improvement, while not changing any of the previous $\{\boldsymbol{\beta}\}_{k=1}^{m-1}$. Hence at each step m the current solution is

$$F_m = F_{m-1} + \nu H(\mathbf{x}; \boldsymbol{\beta}_m),$$

where the parameters $\boldsymbol{\beta}_m$ are those in H minimizing the empirical risk when added to the fixed part F_{m-1}:

$$\boldsymbol{\beta}_m = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \overline{\operatorname{err}}(H(\mathbf{x}; \boldsymbol{\beta}_k) + H(\mathbf{x}; \boldsymbol{\beta})).$$

The final model is then the sum of these terms, like in (3.6). To find $\boldsymbol{\beta}_m$ in each step here, we use gradient descent. We have outlined a generic functional gradient descent algorithm. It can be stated more formally as follows.

1. Initialize $F_0(\mathbf{x})$, e.g., by setting it to zero for all components. Select a base learner H.
2. Compute the negative gradient vector,
$$U_i = -\frac{\partial \mathrm{L}(y_i, F_{m-1}(\cdot))}{\partial F_{m-1}(\cdot)}, \quad i = 1, \ldots, N.$$
3. Estimate $\hat{H}_m$ by fitting $(\mathbf{X}_i, U_i)$ using the base learner H (like in the previous algorithm):
$$\boldsymbol{\beta}_m = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathrm{L}(u_i, H(\mathbf{x}_i; \boldsymbol{\beta}))$$
4. Update $F_m(\cdot) = F_{m-1}(\cdot) + \nu H(\cdot; \boldsymbol{\beta}_m)$.
5. Repeat steps from 2 until $m = M$.

Note that while we call this functional gradient descent (FGD), this is exactly the gradient boosting algorithm.

3.7 L2Boost

Based on Friedman (2001)'s results, Bühlmann and Yu (2003) developed the L2Boost algorithm. It is a special case of the generic functional gradient descent (FGD) algorithm, where we choose the squared error loss to be the loss function,

$$\mathrm{L}(y, F(\mathbf{x})) = \frac{1}{2}\left(y - F(\mathbf{x})\right)^2.$$

The negative gradient vector of the loss then becomes the residual vector,

$$\frac{\partial \mathrm{L}(y, F(\mathbf{x}))}{\partial x_i} = (y - F(x_i)), \quad i = 1, \ldots, n,$$

and hence the boosting steps become repeated refitting of residuals (Friedman, 2001; Bühlmann and Yu, 2003). With $\nu = 1$ and $M = 2$, this had been proposed already by (Tukey, 1977), who called it "twicing".

1. Initialize $F_0(\mathbf{x})$, e.g., by setting it to zero for all components. Select a base learner H, such as ordinary least squares, stumps, etc.

2. Compute the residuals

$$U_i = (y_i, F_{m-1}(x_i)), \quad i = 1, \ldots, n$$

3. Estimate $\hat{H}_m$ by fitting $(\mathbf{X}_i, U_i)_{i=1}^N$ using the base learner H:

$$\boldsymbol{\beta}_m = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathrm{L}(u_i, H(\mathbf{x}_i; \boldsymbol{\beta}))$$

   Note that $\hat{H}(\cdot; \boldsymbol{\beta}_m)$ is then an estimate of the negative gradient vector.

4. Update $F_m(\cdot) = F_{m-1}(\cdot) + \nu H(\cdot; \boldsymbol{\beta}_m)$.

5. Repeat steps from 2 until $m = M$..

They also prove some important theoretical results for L2Boost.

more on L2Boost!!

3.8 Component-wise gradient boosting

In high-dimensional settings, it might often be infeasible, if not impossible, to use a base learner H which incorporates all p dimensions. Indeed, using least squares base learners, it is impossible, since the matrix which must be inverted is singular when $p > N$. Component-wise gradient boosting is a technique/algorithm which does work in these settings. First developed by Bühlmann and Yu (2003), and it has further been refined and explored, see e.g. Bühlmann (2006). The idea of the algorithm is to select a set of base learners, the most important property of which being that they are univariate: Each base learner is only a function of one component x_j of the data $\mathbf{X}$, i.e.,

$$h_j(\mathbf{x}) = h_j(x_j).$$

In each iteration, we fit the learners separately, and choose only the one which gives the best improvement to be added in the final model. The resulting model $F_m(\cdot)$ is then a sum of componentwise effects,

$$F_m(\mathbf{X}) = \sum_{j=1}^{p} f_j(x_j),$$

where

$$f_j(x_j) = \sum_{m=1}^{M} \mathbb{1}_{mj} h_j(x_j; \boldsymbol{\beta}_m),$$

better notation??!

where $\mathbb{1}_{mj}$ is an indicator function which is 1 if component j was selected at iteration m and 0 if not. Hence this model is a GAM (3.4). Crucially, if we stop sufficiently early, we will typically perform variable selection. It is likely that some base learners have never been added to the final model, and as such those components in $\mathbf{X}$ are not added. We now give a presentation of the algorithm.

more about variable selection!!

1. Start with an initial guess, e.g. $F_0 = \mathbf{0}$.
   Specify a set of base learners $h_1(\cdot), \ldots, h_p(\cdot)$.
2. Compute the negative gradient vector $\mathbf{U}$.
3. Fit $(\mathbf{X}_i, U_i)_{i=1}^N$ separately to every base learner to get $\hat{h}_1(x_1), \ldots, \hat{h}_p(x_p)$.
4. Select the component k which best fits the negative gradient vector.

$$k = \underset{j \in [1,p]}{\operatorname{argmin}} \sum_{i=1}^{N} (u_i - \hat{h}_j(\mathbf{x}_i))^2$$

5. Update $F_m(\cdot) = F_{m-1}(\cdot) + \nu h_k(x_k)$

In fact, Buhlmann believes that it is mainly in the case of high-dimensional predictors that boosting has a substantial advantage over classical approaches (Bühlmann, 2006).

3.9 The importance of stopping early

The number of iterations in the boosting procedure, M, is a tuning parameter. It acts as a regularizer.

Appendices

Appendix A

Appendix 1: Differentiating the IG FHT

appendix

First we have the likelihood,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{y_0}{\sqrt{2\pi\sigma^2 t_{(i)}^3}} \exp\left[-\frac{(y_0 + \mu t_{(i)})^2}{2\sigma^2 t_{(i)}} \right] \right)^{\delta_{(i)}} \times \left[1 - \Phi\left(-\frac{y_0 + \mu t_{(i)}}{\sqrt{\sigma^2 t_{(i)}}} \right) - \exp\left(-\frac{2y_0\mu}{\sigma^2} \right) \Phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}} \right) \right]^{1-\delta_{(i)}}, \tag{A.1}$$

{eq:fht-loglik}

with respect to parameters μ, and y_0. First, note that for any cumulative distribution function F that is symmetric around 0, and for $x \in \mathbb{R}$,

$$F(x) = 1 - (1 - F(x)) = 1 - F(-x), \tag{A.2}$$

and so in particular,

$$\Phi(x) = 1 - (1 - \Phi(x)) = 1 - \Phi(-x), \tag{A.3}$$

and thus we can rewrite (A.1) as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{y_0}{\sqrt{2\pi\sigma^2 t_{(i)}^3}} \exp\left[-\frac{(y_0 + \mu t_{(i)})^2}{2\sigma^2 t_{(i)}} \right] \right)^{\delta_{(i)}} \times \left[\Phi\left(\frac{y_0 + \mu t_{(i)}}{\sqrt{\sigma^2 t_{(i)}}} \right) - \exp\left(-\frac{2y_0\mu}{\sigma^2} \right) \Phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}} \right) \right]^{1-\delta_{(i)}}. \tag{A.4}$$

{eq:fht-loglik-proper}

It is easier to work with the log likelihood, so we take the log of (A.4) and get

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_{(i)} \left(\ln y_0 - \frac{1}{2} \ln\left(2\pi\sigma^2 t_{(i)}^3 \right) - \frac{\left(y_0 + \mu t_{(i)} \right)^2}{2\sigma^2 t_{(i)}} \right) + (1 - \delta_{(i)}) \ln\left(\Phi\left(\frac{\mu t_{(i)} + y_0}{\sqrt{\sigma^2 t_{(i)}}} \right) - \exp\left(-\frac{2y_0\mu}{\sigma^2} \right) \Phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}} \right) \right) \tag{A.5}$$

To make things easier, let us introduce some intermediate functions here. Let

$$f_i(\boldsymbol{\theta}) = \ln y_0 - \frac{1}{2}\ln\Big(2\pi\sigma^2 t_{(i)}^3\Big) - \frac{\big(y_0 + \mu t_{(i)}\big)^2}{2\sigma^2 t_{(i)}} \tag{A.6}$$

and

$$g_i(\boldsymbol{\theta}) = \Phi\left(\frac{\mu t_{(i)} + y_0}{\sqrt{\sigma^2 t_{(i)}}}\right) - \exp\Big(-\frac{2y_0\mu}{\sigma^2}\Big)\Phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}}\right). \tag{A.7}$$

Thus we see that the partial derivatives are

$$\frac{\partial}{\partial y_0} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_{(i)} \frac{\partial}{\partial y_0} f_i(\boldsymbol{\theta}) + (1 - \delta_{(i)}) \frac{\frac{\partial}{\partial y_0} g_i(\boldsymbol{\theta})}{g_i(\boldsymbol{\theta})} \tag{A.8}$$

and

$$\frac{\partial}{\partial \mu} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_{(i)} \frac{\partial}{\partial \mu} f_i(\boldsymbol{\theta}) + (1 - \delta_{(i)}) \frac{\frac{\partial}{\partial \mu} g_i(\boldsymbol{\theta})}{g_i(\boldsymbol{\theta})}. \tag{A.9}$$

$$\begin{aligned}\frac{\partial}{\partial y_0} g_i(\boldsymbol{\theta}) = {} & \frac{1}{\sqrt{\sigma^2 t_{(i)}}} \phi\left(\frac{\mu t_{(i)} + y_0}{\sqrt{\sigma^2 t_{(i)}}}\right) + \frac{2\mu}{\sigma^2} \exp\Big(-\frac{2y_0\mu}{\sigma^2}\Big)\Phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}}\right) \\ & + \frac{1}{\sqrt{\sigma^2 t_{(i)}}} \exp\Big(-\frac{2y_0\mu}{\sigma^2}\Big)\phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}}\right)\end{aligned} \tag{A.10}$$

$$\frac{\partial}{\partial y_0} f_i(\boldsymbol{\theta}) = \frac{1}{y_0} - \frac{y_0 + \mu t_{(i)}}{\sigma^2 t_{(i)}} \tag{A.11}$$

$$\frac{\partial}{\partial \mu} f_i(\boldsymbol{\theta}) = -\frac{y_0 + \mu t_{(i)}}{\sigma^2} \tag{A.12}$$

$$\begin{aligned}\frac{\partial}{\partial \mu} g_i(\boldsymbol{\theta}) = {} & \frac{t_{(i)}}{\sqrt{\sigma^2 t_{(i)}}} \phi\left(\frac{\mu t_{(i)} + y_0}{\sqrt{\sigma^2 t_{(i)}}}\right) + \frac{2y_0}{\sigma^2} \exp\Big(-\frac{2y_0\mu}{\sigma^2}\Big)\Phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}}\right) \\ & - \frac{t_{(i)}}{\sqrt{\sigma^2 t_{(i)}}} \exp\Big(-\frac{2y_0\mu}{\sigma^2}\Big)\phi\left(\frac{\mu t_{(i)} - y_0}{\sqrt{\sigma^2 t_{(i)}}}\right)\end{aligned} \tag{A.13}$$

Bibliography

aalengjessing2001 Aalen, O. O. and Gjessing, H. K. (2001). Understanding the shape of the hazard rate: a process point of view (with comments and a rejoinder by the authors). *Statist. Sci.*, 16(1):1–22.

buhlmann2006 Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.*, 34(2):559–583.

buhlmann-yu Bühlmann, P. and Yu, B. (2003). Boosting with the l2 loss. *Journal of the American Statistical Association*, 98(462):324–339.

caroni2017 Caroni, C. (2017). *First Hitting Time Regression Models*. John Wiley & Sons, Inc.

chhikara1988 Chhikara, R. (1988). *The Inverse Gaussian Distribution: Theory: Methodology, and Applications*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis.

adaboost Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, pages 148–156, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

friedman2000 Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.*, 28(2):337–407.

friedman2001 Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232.

kearnsvaliant Kearns, M. and Valiant, L. G. (1989). Crytographic limitations on learning boolean formulae and finite automata. In *Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing*, STOC '89, pages 433–444, New York, NY, USA. ACM.

leewhitmore2006 Lee, M.-L. T. and Whitmore, G. A. (2006). Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statist. Sci.*, 21(4):501–513.

lee2010 Lee, M.-L. T. and Whitmore, G. A. (2010). Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Analysis*, 16(2):196–214.

leewhitmore2004 Lee, M. T., Whitmore, G. A., Laden, F., Hart, J. E., and Garshick, E. (2004). Assessing lung cancer risk in railroad workers using a first hitting time regression model. *Environmetrics*, 15(5):501–512.

mayr14a Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014). The evolution of boosting algorithms. from machine learning to statistical modelling. *Methods of Information in Medicine*, 53(6):419–427.

stogiannis-2013 Stogiannis, D. and Caroni, C. (2013). Issues in fitting inverse gaussian first hitting time regression models for lifetime data. *Communications in Statistics - Simulation and Computation*, 42(9):1948–1960.

tukey Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company.

whitmore1986 Whitmore, G. A. (1986). First-passage-time models for duration data: Regression structures and competing risks. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 35(2):207–219.

threg Xiao, T., Whitmore, G., He, X., and Lee, M.-L. (2015). The r package threg to implement threshold regression models. *Journal of Statistical Software, Articles*, 66(8):1–16.