BOOSTING ON THE RESPONSES WITH TWEEDIE LOSS FUNCTIONS

Michel Denuit, Julien Trufin, Thomas Verdebout

LIDAM Discussion Paper ISBA
2022 / 39

UCLouvain

LIDAM
ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Boosting on the responses with Tweedie loss functions

Michel Denuit
Institute of Statistics, Biostatistics and Actuarial Science
UCLouvain
Louvain-la-Neuve, Belgium

Julien Trufin
Department of Mathematics
Université Libre de Bruxelles (ULB)
Brussels, Belgium

Thomas Verdebout
Department of Mathematics
Université Libre de Bruxelles (ULB)
Brussels, Belgium

January 28, 2022

Abstract

Hainaut et al. (2022) showed that boosting can be conducted directly on the response under Tweedie loss function and log-link, by adapting the weights at each iteration step. In this short note, we complete the results obtained in Hainaut et al. (2022) by showing that among the usual link functions, the log-link function is actually the only one for which boosting can be regarded as an iteratively re-weighted procedure applied to the original data.

1 Introduction

Boosting emerged from the field of machine learning and became rapidly popular among insurance analysts. Broadly speaking, boosting is an iterative fitting procedure using weak, or base learners. In a regression context, weak learners have rigid parametric forms and cannot accurately adapt to the data under consideration. In each iteration, boosting fits a weak learner that improves the fit of the overall model such that the ensemble arrives at an accurate prediction. We refer the readers to Denuit et al. (2019b, 2020) for an extensive treatment of boosting in the context of insurance, with applications to tree-based methods and neural networks.

Boosting requires that the analyst first decides which metric should be optimized for estimating the score. In the boosting terminology, this metric is usually referred to as the "loss function". To ease numerical aspects, boosting is often applied on gradients of the loss function, that is, on the gradients of the deviance function in insurance applications. However, Hainaut et al. (2022) showed that there is often no need to boost gradients in insurance applications. As proved in their Proposition 3.1, boosting can easily be performed directly with responses under Tweedie loss function and log-link.

In this short note, we show that under Tweedie loss function, the log-link function is actually the only link function among the common ones for which boosting can be regarded as an iteratively re-weighted, or re-offseted procedure applied to the original data. Common link functions can be found in Table 4.3 in Denuit et al. (2019a).

2 Boosting with Tweedie loss functions

In actuarial pricing, the aim is to evaluate the pure premium as accurately as possible. The target is the conditional expectation $\mu(\boldsymbol{X}) = \mathrm{E}[Y|\boldsymbol{X}]$ of the response Y (claim number or claim amount for instance) given the available information summarized in a vector $\boldsymbol{X}$ of features $X_1, X_2, \ldots, X_p$. The function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is unknown to the actuary and is approximated by a working predictor $\boldsymbol{x} \mapsto \widehat{\mu}(\boldsymbol{x})$ entering premium calculation.

Lack of accuracy for $\widehat{\mu}(\boldsymbol{x})$ is defined by the generalization error

$$Err(\widehat{\mu}) = \mathrm{E}\left[L(Y, \widehat{\mu}(\boldsymbol{X}))\right],$$

where $L(.,.)$ is the loss function measuring the discrepancy between its two arguments and the expected value is over the joint distribution of $(Y, \boldsymbol{X})$. The goal is to find a function $\boldsymbol{x} \mapsto \widehat{\mu}(\boldsymbol{x})$ of the features minimizing the generalization error.

The Tweedie family of distributions regroups the members of the Exponential Dispersion family having power variance functions $V(\mu) = \mu^\xi$ for some ξ. From e.g. Denuit et al. (2019a, Table 4.7), the Tweedie deviance loss function is given by

$$L(Y, \widehat{\mu}(\boldsymbol{X})) = \begin{cases} (Y - \widehat{\mu}(\boldsymbol{X}))^2 & \text{if } \xi = 0, \\ 2\left(Y \ln \frac{Y}{\widehat{\mu}(\boldsymbol{X})} - (Y - \widehat{\mu}(\boldsymbol{X}))\right) & \text{if } \xi = 1, \\ 2\left(-\ln \frac{Y}{\widehat{\mu}(\boldsymbol{X})} + \frac{Y}{\widehat{\mu}(\boldsymbol{X})} - 1\right) & \text{if } \xi = 2, \\ 2\left(\frac{\max\{Y,0\}^{2-\xi}}{(1-\xi)(2-\xi)} - \frac{Y\widehat{\mu}(\boldsymbol{X})^{1-\xi}}{1-\xi} + \frac{\widehat{\mu}(\boldsymbol{X})^{2-\xi}}{2-\xi}\right) & \text{otherwise } (\xi > 0). \end{cases} \tag{2.1}$$

For $\xi = 0$, we recover the L^2 loss function whereas $\xi = 1$ and 2 correspond to the Poisson and Gamma deviance functions, respectively.

Ensemble techniques assume structural models of the form

$$\mu(\boldsymbol{x}) = g^{-1}\left(\text{score}(\boldsymbol{x})\right) = g^{-1}\left(\sum_{m=1}^{M} T(\boldsymbol{x}; \mathbf{a}_m)\right), \tag{2.2}$$

where g is the link function (assumed to be monotone and differentiable) and $T(\boldsymbol{x}; \mathbf{a}_m)$, $m = 1, 2, \ldots, M$, are usually simple functions of the features $\boldsymbol{x}$, characterized by parameters

$\mathbf{a}_m$. In (2.2), the score is the function of features $\boldsymbol{x}$ mapped to $\mu(\boldsymbol{x})$ by the inverse of the link function g.

Let

$$\mathcal{D} = \{(\nu_1, y_1, \boldsymbol{x}_1), (\nu_2, y_2, \boldsymbol{x}_2), \ldots, (\nu_n, y_n, \boldsymbol{x}_n)\},$$

be the set of observations used to fit the model $\widehat{\mu}$, called training set, where ν_i denotes the weight of observation i. Estimating a score of the form

$$\text{score}(\boldsymbol{x}) = \sum_{m=1}^{M} T(\boldsymbol{x}; \mathbf{a}_m),$$

by minimizing the corresponding training sample estimate of the generalized error

$$\min_{\{\mathbf{a}_m\}_1^M} \sum_{i=1}^{n} \nu_i L\left(y_i, g^{-1}\left(\sum_{m=1}^{M} T(\boldsymbol{x}_i; \mathbf{a}_m)\right)\right) \tag{2.3}$$

is in general infeasible. It requires computationally intensive numerical optimization techniques. One way to overcome this problem is to approximate the solution to (2.3) by using a greedy forward stagewise approach, also called boosting.

Forward stagewise additive modeling consists in sequentially fitting a single function and adding it to the expansion of prior fitted terms. Each fitted term is not readjusted as new terms are added into the expansion, contrarily to a stepwise approach where previous terms are each time readjusted when a new one is added. Specifically, we start by computing

$$\widehat{\mathbf{a}}_1 = \underset{\mathbf{a}_1}{\text{argmin}} \sum_{i=1}^{n} \nu_i L\left(y_i, g^{-1}\left(\widehat{\text{score}}_0(\boldsymbol{x}_i) + T(\boldsymbol{x}_i; \mathbf{a}_1)\right)\right),$$

where $\widehat{\text{score}}_0(\boldsymbol{x})$ is an initial guess (for instance, just an intercept). Then, at each iteration $m \geq 2$, we solve the subproblem

$$\widehat{\mathbf{a}}_m = \underset{\mathbf{a}_m}{\text{argmin}} \sum_{i=1}^{n} \nu_i L\left(y_i, g^{-1}\left(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i) + T(\boldsymbol{x}_i; \mathbf{a}_m)\right)\right), \tag{2.4}$$

with

$$\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i) = \widehat{\text{score}}_{m-2}(\boldsymbol{x}_i) + T(\boldsymbol{x}_i; \widehat{\mathbf{a}}_{m-1}).$$

Boosting is thus an iterative method based on the idea that combining many simple functions should result in a powerful one. In a boosting context, the simple functions $T(\boldsymbol{x}; \mathbf{a}_m)$ are called weak learners or base learners.

Hainaut et al. (2022) showed in their Proposition 3.1 that under Tweedie loss function and log-link, the subproblem (2.4) can be rewritten as

$$\widehat{\mathbf{a}}_m = \underset{\mathbf{a}_m}{\text{argmin}} \sum_{i=1}^{n} \nu_{i,m} L\left(\tilde{r}_{i,m}, \exp\left(T(\boldsymbol{x}_i; \mathbf{a}_m)\right)\right),$$

where $\nu_{i,m} = \nu_i \exp(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))^{2-\xi}$ and $\tilde{r}_{i,m} = \frac{y_i}{\exp(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))}$. In words, the mth iteration of the boosting procedure reduces to build a single weak learner on the working training set

$$\mathcal{D}^{(m)} = \{(\nu_{i,m}, \tilde{r}_{i,m}, \boldsymbol{x}_i), i = 1, \ldots, n\}$$

using the Tweedie deviance loss and the log-link function. The weights are each time updated together with the responses.

There is an easy probabilistic interpretation to this boosting algorithm. Recall that if Y obeys the Tweedie distribution with mean μ, power parameter ξ and weight ν then for any positive constant c, cY is Tweedie with mean $c\mu$, the same power parameter and modified weight $\nu(c)^{\xi-2}$. One says that the Tweedie distributions are closed under this type of scale transformation. See e.g. Denuit et al. (2019a) for a proof. Since the essence of boosting consists in treating $\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i)$ as a constant to estimate $\mathbf{a}_m$, this means that we can equivalently work with response $\tilde{r}_{i,m}$ obeying the Tweedie distribution with adapted weight $\nu_{i,m}$ to perform that estimation. This is a direct application of the result recalled earlier with $c = 1/\exp(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))$. This process can even be performed in an iterative way, by dividing the response $\tilde{r}_{i,m}$ with $\exp(T(\boldsymbol{x}_i;\mathbf{a}_m))$ and multiplying $\nu_{i,m}$ with $\exp((2-\xi)T(\boldsymbol{x}_i;\mathbf{a}_m))$ at each step.

Notice that in the Gaussian case ($\xi = 0$), the boosting procedure described here differs from the classical gradient boosting algorithm with L^2 loss, which uses identity link and raw residuals (current estimate subtracted from the response) whereas here, we work with ratios under log-link.

At each step of the boosting algorithm, the response is thus Tweedie distributed so that the loss function selected for the original responses y_i (and weights ν_i) is still the right choice at iteration m for new responses $\tilde{r}_{i,m}$ (and weights $\nu_{i,m}$). This is a direct consequence of the closure property of Tweedie distributed responses Y under scale transformation of the type cY. The next results shows that the subproblem (2.4) can be rewritten as

$$\widehat{\mathbf{a}}_m = \underset{\mathbf{a}_m}{\operatorname{argmin}} \sum_{i=1}^{n} \nu_i (h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))^{2-\xi} L\left(\frac{y_i}{h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))}, f\left(T(\boldsymbol{x}_i;\mathbf{a}_m)\right)\right) \tag{2.5}$$

for some positive real functions h and f if, and only if, the link function g is such that

$$g^{-1}(s+t) = h(s)f(t). \tag{2.6}$$

In other words, the following proposition, whose aim is to complete the results obtained in Hainaut et al. (2022), tells us that the only way to see the boosting procedure as successive fits of single weak learners on Tweedie distributed responses is to consider link functions satisfying condition (2.6). Note that the log-link function is such that the latter condition hold with $h = f = g^{-1}$.

Proposition 2.1. *We have that*

$$L\left(y_i, g^{-1}(\widehat{score}_m(\boldsymbol{x}_i))\right) = (h(\widehat{score}_{m-1}(\boldsymbol{x}_i))^{2-\xi} L\left(\frac{y_i}{h(\widehat{score}_{m-1}(\boldsymbol{x}_i))}, f(T(\boldsymbol{x}_i;\widehat{\mathbf{a}}_m))\right)$$

for some positive real functions h and f if, and only if, the link function g is such that (2.6) *is fulfilled.*

Proof. Obviously, the Tweedie loss function $L(y,s)$ is $2-\xi$ homogeneous in the sense that for any $b > 0$,

$$L(by, bs)) = b^{2-\xi} L(y,s). \tag{2.7}$$

Therefore, if

$$L\left(y_i, g^{-1}(\widehat{\text{score}}_m(\boldsymbol{x}_i))\right) = (h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))^{2-\xi} L\left(\frac{y_i}{h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))}, f(T(\boldsymbol{x}_i; \widehat{\mathbf{a}}_m))\right),$$

we have that

$$L\left(y_i, g^{-1}(\widehat{\text{score}}_m(\boldsymbol{x}_i))\right) = L(y_i, h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i)) f(T(\boldsymbol{x}_i; \widehat{\mathbf{a}}_m)))$$

so that $g^{-1}(\widehat{\text{score}}_m(\boldsymbol{x}_i)) = h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i)) f(T(\boldsymbol{x}_i; \widehat{\mathbf{a}}_m))$, which is the desired result. Conversely, if the inverse link function g^{-1} is such that $g^{-1}(\widehat{\text{score}}_m(\boldsymbol{x}_i))$ factorizes into

$$g^{-1}(\widehat{\text{score}}_m(\boldsymbol{x}_i)) = h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i)) f(T(\boldsymbol{x}_i; \widehat{\mathbf{a}}_m)),$$

the homogeneity property in (2.7) leads to

$$\begin{aligned}
L\left(y_i, g^{-1}(\widehat{\text{score}}_m(\boldsymbol{x}_i))\right) &= L\left(y_i, h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i)) f(T(\boldsymbol{x}_i; \widehat{\mathbf{a}}_m))\right) \\
&= L\left(h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i)) \frac{y_i}{h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))}, h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i)) f(T(\boldsymbol{x}_i; \widehat{\mathbf{a}}_m))\right) \\
&= h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))^{2-\xi} L\left(\frac{y_i}{h(\widehat{\text{score}}_{m-1}(\boldsymbol{x}_i))}, f(\mathbf{T}(\mathbf{x}_i, \mathbf{a}_m))\right),
\end{aligned}$$

which concludes the proof. □

The only non-trivial inverse link functions satisfying condition (2.6) is necessarily of the form $g^{-1}(x) = \exp(ax)$ with $a \in \mathbb{R}$; see for instance Aczél (2014). As a result, the boosting procedure (2.5) with a Tweedie loss function can only be obtained with a log-link function.

References

- Aczél, J. (2014). On Applications and Theory of Functional Equations. Academic Press.

- Denuit, M., Hainaut, D., Trufin, J. (2019a). Effective Statistical Learning Methods for Actuaries I: GLM and Extensions. Springer Actuarial Lecture Notes Series.

- Denuit, M., Hainaut, D., Trufin, J. (2019b). Effective Statistical Learning Methods for Actuaries III: Neural Networks and Extensions. Springer Actuarial Lecture Notes Series.

- Denuit, M., Hainaut, D., Trufin, J. (2020). Effective Statistical Learning Methods for Actuaries II: Tree-based Methods and Extensions. Springer Actuarial Lecture Notes Series.

- Hainaut, D., Trufin, J., Denuit, M. (2022). Response versus gradient boosting trees, GLMs and neural networks under Tweedie loss and log-link. Accepted for publication in Scandinavian Actuarial Journal.