Boosting a Generalized Poisson Hurdle Model

Vera Hofer

University of Graz

Paris, 23/08/2010

Ensemble Techniques

- Aim at improving the predictive performance of fitting techniques by
  - by constructing multiple function predictions from the data by means of a “weak” base procedure
  - and then using a convex combination of them for final aggregated prediction
- Random forest, boosting and bagging most famous ensemble techniques
- Originally designed for classification
- Gradient descent approximation in function space (Breiman, 1998, 1999) is an easy tool to use boosting in regression

Usual Regression

- Let $Y \in \mathbb{R}$ be a random variable and $\mathbf{x} \in \mathbb{R}^p$ a vector of predictor values
- Let f be a regression function such that $\hat{Y} = f(\mathbf{x})$.
- Let $L(Y, f(\mathbf{x}))$ be the loss function that measures goodness of fit. For example $L(Y, f(\mathbf{x})) = (Y - F(\mathbf{x}))^2$, known as L^2-loss.
- The regression function f is found from minimizing the the expected loss

$$f(\mathbf{x}) = \arg\min_F \mathbb{E}_{Y|\mathbf{x}}(L(Y, F(\mathbf{x})) \,|\, \mathbf{x} = \mathbf{x}))$$

Boosting

- Boosting attemts to find a regression function f of the form

  $$f(\mathbf{x}) = \sum_{i=0}^{m} f_m(\mathbf{x})$$

  by minimizing expected loss using gradient descent techniques, i.e. following the steepest descent with respect to f of the loss function in a forward stagewise manner.
- f_m are simple functions of $\mathbf{x}$ ("base learners").
- Choice of the loss function and the type of base learners yield a variety of different boosted regression models.

Gradient Descent

- Start with initial function $f_0(\mathbf{x})$.
- In step $m \geq 1$, the current argument f_{m-1} is changed into the direction of the negative gradient of expected loss

$$\begin{aligned} U_m(\mathbf{x}) &= -\frac{\partial}{\partial f} \mathbb{E}_{Y|\mathbf{x}}(L(Y, F(\mathbf{x})) \,|\, \mathbf{x} = \mathbf{x})) \,|_{f=f_{m-1}(\mathbf{x})} = \\ &= \mathbb{E}_{Y|\mathbf{x}}(-\nabla L(y, f)) \,|_{f=f_{m-1}(\mathbf{x})} \end{aligned}$$

such that $f_m = f_{m-1} + \nu\, U_m$, where ∇L is the gradient of the loss function with respect to f, and ν is the shrinkage parameter.

Sample Version of Gradient Descent

- f_0 is traditionally chosen as $f_0 = \arg\min_c \sum_{i=1}^{N} L(y_i, c)$.
- The conditional mean of the negative gradient is found from regression:
    - The negative gradient of the loss function, $V_i = -\nabla L(y_i, f_{m-1}(\mathbf{x}_i))$, is evaluated at the given sample.
    - This "pseudo-response" is fitted to the predictors $\mathbf{x}_i$ by the "base learner" u_m to get the direction $\hat{U}_m(\mathbf{x}) = u_m(\mathbf{x})$.
    - The regression function then becomes $f_m = f_{m-1} + \nu \, u_m$.
    - The process is iterated until $m = M$.

Tuning Parameters

- M can be determined by cross validation.
- ν is of minor importance unless it is not too large. Typically, $\nu = 0.1$. Smaller values of ν favor better test error but need a larger number of iterations.
- As "base learner" simple models such as regression tree or componentwise linear least squares (CLLS) are used. CLLS are very fast in calculation, wheras tree can cope with nonlinear structures.

Count Data Regression

- Common models: Poisson, negative binomial
- Alternative model: The generalised Poisson distribution (Consul and Jain (1970); Consul (1979))
- To address overdispersion caused by an excess of zeros, zero-inflated models were introduced (Johnson and Kotz, 1969; Mullahy, 1986; Lambert, 1992).
  - Derived from mixing a count distribution and a point mass at zero.
  - Problem: different sources of zeros impede interpretation
- Alternative model: hurdle models consist of a hurdle component to account for zeros, and a zero-truntcated count component to account for non-zeros. The zero-truncated component follows any zero-truncated count distribution.

Generalized Poisson Distribution of Y

- Probability density function, $p(y \mid \mu, \phi)$, with mean μ, and dispersion parameter ϕ

$$p(y \mid \mu, \phi) = \frac{\mu\, W^{y-1}}{y!}\, \phi^{-y}\, e^{-\frac{W}{\phi}}$$

  where $W = \mu + (\phi - 1)\, y$ and $\mu > 0$.
- Assume $\phi > 1$. Otherwise ϕ must be restricted to guarantee that $p(y \mid \mu, \phi) \geq 0$.
- $\phi > 1$ indicates overdispersion, whereas $\phi < 1$ indicates underdispersion.
- For $\phi = 1$ the GP reduces to the Poisson distribution
- Mean and variance of the GP are:

$$\mathbb{E}(Z) = \mu \qquad \mathrm{Var}(Y) = \phi^2\, \mu$$

Generalized Poisson Hurdle Distribution (1)

- Two-component model: a hurdle component to model zeros versus nonzeros, and a zero-tructated count component to account for the nonzeros.
- The hurdle at zero is assumed to be a Bernoulli variable $B(\omega, 1)$ where $\omega = P(Y_0 = 0)$.
- The zero-truncated component $Y_T \sim GP_T(\mu, \phi, p)$ with probability density function

$$p_T(y \,|\, \mu, \phi) = \frac{p(y \,|\, \mu, \phi)}{p(0 \,|\, \mu, \phi)} = \frac{p(y \,|\, \mu, \phi)}{1 - e^{-\mu/\phi}}\,.$$

where $p(y \,|\, \mu, \phi)$ is the GP probability density function

Generalized Poisson Hurdle Distribution (2)

- Probability density function of a generalised Poisson hurdle distribution (GPH):

$$p_H(y \mid \mu, \phi, \omega) = 1_{(y==0)} \cdot \omega + 1_{(y>0)} \cdot (1-\omega) \frac{p(y \mid \mu, \phi)}{1 - e^{-\mu/\phi}} ,$$

- Mean and variance of GPH are

$$\begin{aligned} \mathbb{E}(Z) &= \frac{(1-\omega)\,\mu}{1 - e^{-\mu/\phi}} \\ \mathrm{Var}(Z) &= \frac{\phi^2\,\mu\,(1-\omega)}{1 - e^{-\mu/\phi}} + \frac{\mu^2\,(1-\omega)(\omega - e^{-\mu/\phi})}{(1 - e^{-\mu/\phi})^2} . \end{aligned}$$

Regression Model

- $Y_i \overset{iid}{\sim} GPH(\mu_i, \phi_i, \omega_i)$.
- $\log(\mu_i) = g(\mathbf{x}_i)$
- $\log(\phi_i - 1) = h(\mathbf{x}_i)$
- $\log\left(\frac{\omega_i}{1-\omega_i}\right) = l(\mathbf{x}_i)$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is a vector of predictor values.

Loss Function

The loglikelihood function serves as a loss function for determining the predictors g, h, and l:

$$
\begin{aligned}
&L(Y, g, h, l) = \\
&= -1_{(Y=0)} \left(-\log\left(1 + e^{-l}\right)\right) - 1_{(Y>0)} \left(-\log(1 + e^{l}) + g + \right. \\
&+(Y-1)\log(e^{g} + e^{h}\,Y) - \log(Y!) - Y\log(1 + e^{h}) \\
&\left. -\frac{e^{g} + e^{h}\,Y}{1 + e^{h}} - \log\left(1 - \exp\left(-\frac{e^{g}}{1 + e^{h}}\right)\right)\right)
\end{aligned}
$$

Boosting Generalized Poisson Hurdle Model (1)

- Common boosting methods are based on a loss function that involves only one ensemble. Thus, they can only be applied when a regression function is fit only for one parameter.
- The GPH model requires estimating a regression function on all three parameters.
- When using ensemble techniques, three ensembles must be fit simultaneously.
- The loss function of the GPH model depends on three inter-related regression functions, g, h, and l. Thus, the gradient of the GPH boost is a three components vector.

Boosting Generalized Poisson Hurdle Model (2)

At any step $m > 0$ the pseudo-responses, (V_i^g, V_i^h, V_i^l), of the three ensembles, are obtained as the negative gradient of the loss function evaluated at the current values $(g_{m-1}, h_{m-1}, l_{m-1})$ of g, h and l

$$(V_i^g, V_i^h, V_i^w) = \left(-\frac{\partial L}{\partial g}, -\frac{\partial L}{\partial h}, -\frac{\partial L}{\partial w}\right)\Bigg|_{(y_i, g_{m-1}, h_{m-1}, w_{m-1})}$$

where

$$-\frac{\partial L}{\partial g} = 1_{(y>0)}\left(1 + \frac{(y-1)e^g}{e^g + y\, e^h} - \frac{e^g}{1+e^h} - \frac{\exp\left(-\frac{e^g}{1+e^h}\right)\frac{e^g}{1+e^h}}{1-\exp\left(-\frac{e^g}{1+e^h}\right)}\right)$$

Boosting Generalized Poisson Hurdle Model (3)

$$
\begin{aligned}
-\frac{\partial L}{\partial h} &= 1_{(y>0)}\left(\frac{y(y-1)e^h}{e^g+ye^h}-\frac{ye^h}{1+e^h}-\frac{e^h(y-e^g)}{(1+e^h)^2}+\right.\\
&\qquad\left.+\frac{\exp\left(-\frac{e^g}{1+e^h}\right)\frac{e^{g+h}}{(1+e^h)^2}}{1-\exp\left(-\frac{e^g}{1+e^h}\right)}\right)\\
-\frac{\partial L}{\partial l} &= 1_{(y=0)}\left(\frac{1}{1+e^l}\right)-1_{(y>0)}\left(\frac{1}{1+e^{-l}}\right)
\end{aligned}
$$

Multivariate Componentwise Least Squares (1)

- The three pseudo-responses are estimated by multivariate componentwise least squares (MCLLS).
- The methods assumes that all three ensemble have the same predictors.
- In each boosting step only one predictor variable is selected in the sense of Wilks' lambda.
  - Let $\mathbf{X}^{(j)}$ be the j-column of the design matrix, and let $\mathbf{V}$ be the matrix with ith row (V_i^g, V_i^h, V_i^l).
  - The "base learner" has the form $u_m(\mathbf{x}) = \boldsymbol{\beta}^{(s)}\, x^{(s)}$, where

$$\boldsymbol{\beta}^{(j)} = \left(\beta_g^{(s)}, \beta_h^{(s)}, \beta_l^{(s)}\right) = ||\mathbf{X}^{(j)}||^{-2}(\mathbf{X}^{(j)})^t\,\mathbf{V}$$

Multivariate Componentwise Least Squares (2)

$$s = \arg\min_{1\leq j\leq p} \frac{\det(\mathbf{V}^t\mathbf{V} - (\beta^{(j)})^t\,(\mathbf{X}^{(j)})^t\mathbf{V})}{\det(\mathbf{V}^t\mathbf{V} - n\overline{\mathbf{V}}^t\,\overline{\mathbf{V}})}$$

where $\overline{\mathbf{V}}$ is the mean gradient, and n stands for the sample size. This yields the coefficient $\beta_g^{(s)}$ for the μ-ensemble g, $\beta_h^{(s)}$ for the ϕ ensemble h, and $\beta_l^{(l)}$ for the ω ensemble l. Then the ensembles are updated as

$$\begin{aligned} g_m(\mathbf{x}) &= g_{m-1}(\mathbf{x}) + \nu\beta_g^{(s)}\,x^{(s_m)}\,, \\ h_m(\mathbf{x}) &= h_{m-1}(\mathbf{x}) + \nu\beta_h^{(s)}\,x^{(s_m)}\,, \\ w_m(\mathbf{x}) &= w_{m-1}(\mathbf{x}) + \nu\beta_l^{(s)}\,x^{(s_m)}\,. \end{aligned}$$

Initial Values (1)

- After M iterations the parameters take the form

$$\hat{\mu}_i = e^{g_m(\mathbf{x}_i)} \qquad \hat{\phi}_i = 1 + e^{h_m(\mathbf{x}_i)} \qquad \hat{\omega}_i = \frac{e^{l_m(\mathbf{x}_i)}}{1 + e^{l_m(\mathbf{x}_i)}}$$

- Initial values g_0, h_0 and w_0 are obtained from a nonlinear system of equations:
  - Mean and variance of a zero-truncated GP are,

$$\mathbb{E}(Y_T) = \mu_T = \frac{\mu}{1 - e^{-\frac{\mu}{\phi}}} \qquad \mathrm{Var}(Y_T) = \sigma_T^2 = \frac{\mu\left(\mu + \phi^2\right)}{1 - e^{-\frac{\mu}{\phi}}} .$$

  - Using moment estimators

$$\hat{\mu}_T = \frac{1}{n_T} \sum_{y_i > 0} y_i \qquad \hat{\sigma}_T^2 = \frac{1}{n_T - 1} \sum_{y_i > 0} (y_i - \hat{\mu}_T)^2$$

where n_T is the number of nonzero observations. Let n_0 be the number of zeros and $n = n_0 + n_T$ the total sample size.

Initial Values (2)

- Estimations for the parameters μ and ϕ are then obtained from the nonlinear systems of equations with respect to $\hat{\mu}$ and $\hat{\phi}$:

$$\hat{\mu}_T = \frac{\hat{\mu}}{1 - e^{-\frac{\hat{\mu}}{\hat{\phi}}}} \qquad \hat{\sigma}_T^2 = \frac{\hat{\mu}\left(\hat{\phi}\left(1 - e^{-\frac{\hat{\mu}}{\hat{\phi}}}\right) - \hat{\mu}\, e^{-\frac{\hat{\mu}}{\hat{\phi}}}\right)}{\left(1 - e^{-\frac{\hat{\mu}}{\hat{\phi}}}\right)^2}$$

- Furthermore,

$$\hat{\omega}_0 = \frac{n_0}{n}$$

- Finally, $g_0(\mathbf{x}) = \log(\hat{\mu})$, $h_0(\mathbf{x}) = \log(\hat{\phi} - 1)$, and $l(\mathbf{x}) = \log(\hat{\omega}) - \log(1 - \hat{\omega})$.

Empirical Analysis

Two real datasets:

- Data from the US National Medical Expenditure Servey 1987/88 which was used by Deb and Trivedi (1997) to invesigate the number of physician/non-physician office and hospital outpatient visits of individuals aged 66 and over, who are covered by a particular public insurance program.
- Data from the German Socioeconomic Panel which was used in Riphahn et al (2003) to study the number of doctor visits in the last three months and the number of hospital visits in the last year.

Comparison

Compared models

- GP hurdle boost (GPH)
- Poisson hurdle (P)
- negative binomial (nB)
- negative binomial hurdle (nBH)

Characteristics:

- Loglikelihood (LogLik) and loglikelihood per sample (Avg LogLik) for training (train) and testing (test)
- Standard deviation of the loglikelihood per sample unit (Std Avg LogLik) for training and testing
- Root mean squared error of the number of zeros (RMSE zeros) is given for training and testing
- Vuong's test

Results (1)

US National Medical Expenditure Servey (M=9308)

	GPH	P	nB	nBH
LogLik train	-9776	-12897	-9735	-9668
LogLik test	-2452	-3250	-2437	-2423
Avg LogLik train	-2.7736	-3.6590	-2.7619	-2.7428
Avg LogLik test	-2.7823	-3.6883	-2.7657	-2.7502
Std Avg LogLik train	0.0027	0.0431	0.0056	0.0049
Std Avg LogLik test	0.0121	0.1776	0.0225	0.0200
RMSE zeros train	44.5515	0.0000	60.3376	0.0000
RMSE zeros test	12.5356	6.2778	16.0971	6.2778
Vuong test value model verus nBH	-1.3178	-13.3882*	-6.0281*	

Results (2)

German Socioeconomic Panel ($M = 1433$)

	GPH	P	nB	nBH
LogLik train	-46172.13	-60303.21	-46321.25	-45854.36
LogLik test	-11551.54	-15125.44	-11591.15	-11478.57
Avg LogLik train	-2.1121	-2.7585	-2.1189	-2.0976
Avg LogLik test	-2.1137	-2.7676	-2.1209	-2.1003
Std Avg LogLik train	0.0028	0.0201	0.0033	0.0031
Std Avg LogLik test	0.0111	0.0810	0.0132	0.0127
RMSE zeros train	420.1186	0.0000	316.1610	0.0000
RMSE zeros test	105.9726	9.9499	79.8624	9.9499
Vuong test value model verus nBH	-8.0397*	-25.8170*	-16.8593*	