Model-based boosting in R

Families in mboost

Nikolay Robinzonov

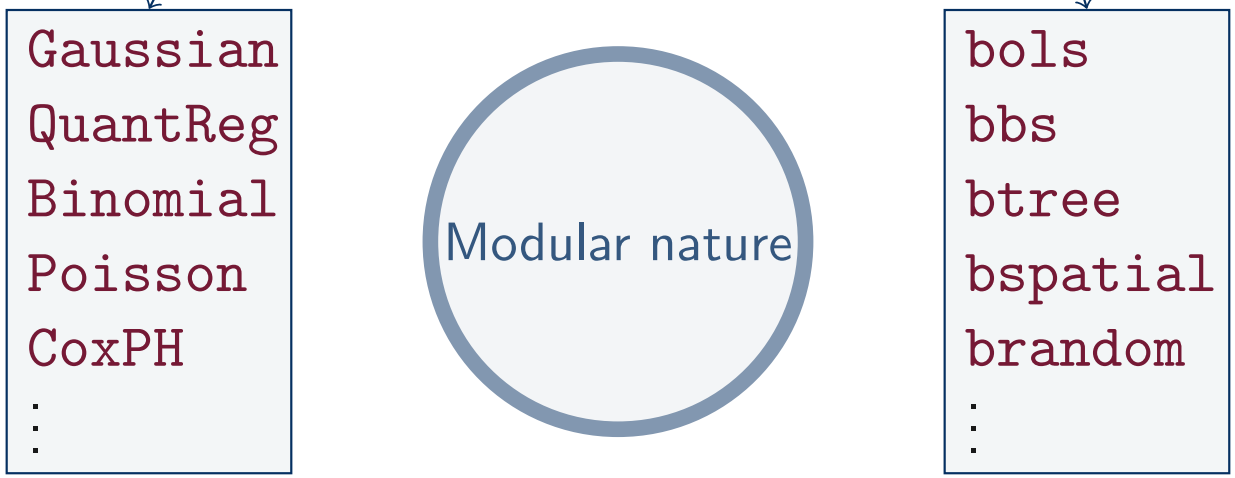
nikolay.robinzonov@stat.uni-muenchen.de
Institut für Statistik
Ludwig-Maximilians-Universität München

Statistical Computing 2011

Model-based boosting

$$\xi(Y|X = x) = \sum_{j=1}^{p} f_j(x)$$

- The right hand side is a sum of components taking a subset of the covariates into account through base-learners.
- The left hand side describes some characteristic of the conditional distribution of the response through the loss function (family).

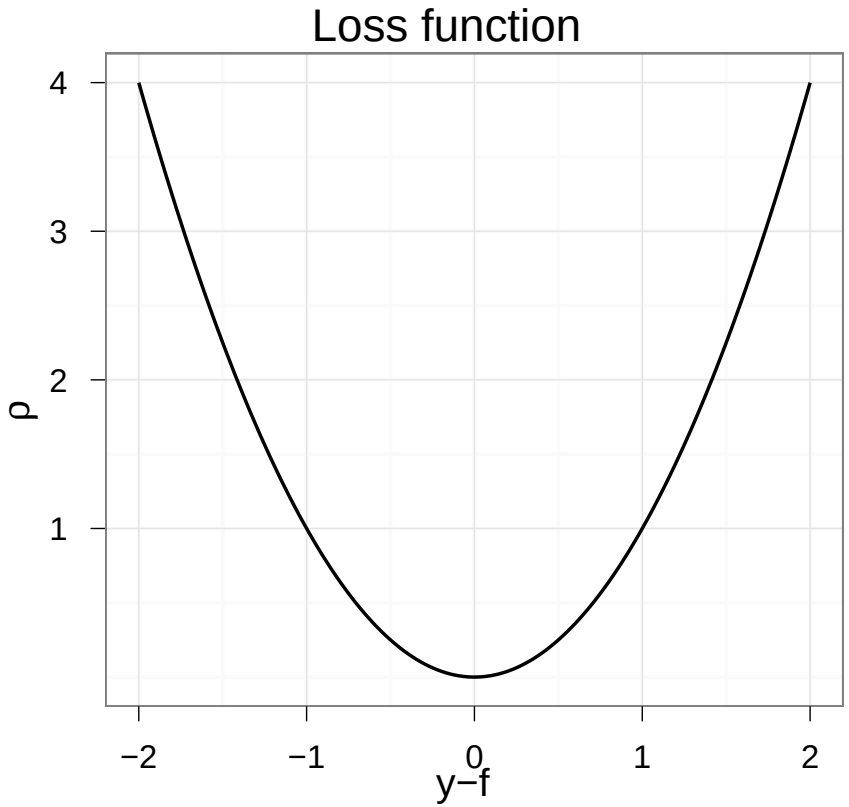


Families for Continuous Response

1. L_2-loss, `Gaussian()`

$$\rho(y, f) = \frac{1}{2}(y - f)^2$$

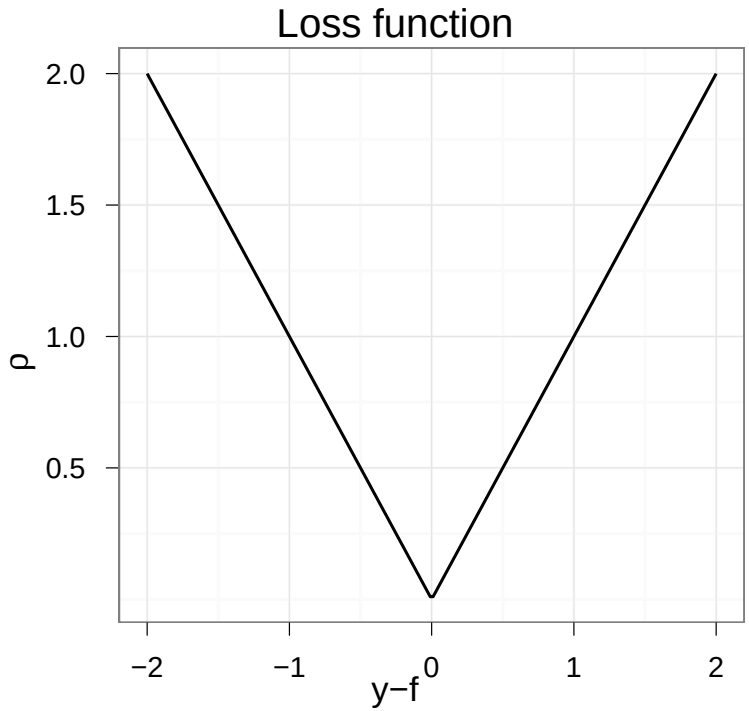
```
R> Gaussian()

 Squared Error (Regression)

Loss function: (y - f)^2

R> Gaussian()@ngradient

function (y, f, w = 1)
y - f
```



- "L_2-Boosting".
- Default `family` in `glmboost()` and `gamboost()`.
- Normally distributed response.

2. L_1-loss, `Laplace()`

$$\rho(y, f) = |y - f|$$



- A median regression approach. Resistance to long-tailed error distributions and outliers (robustness).

```
R> fit <- gamboost(100/mpg ~ bols(wt) + bols(hp), data = mtcars,
+                  family = Laplace())
```

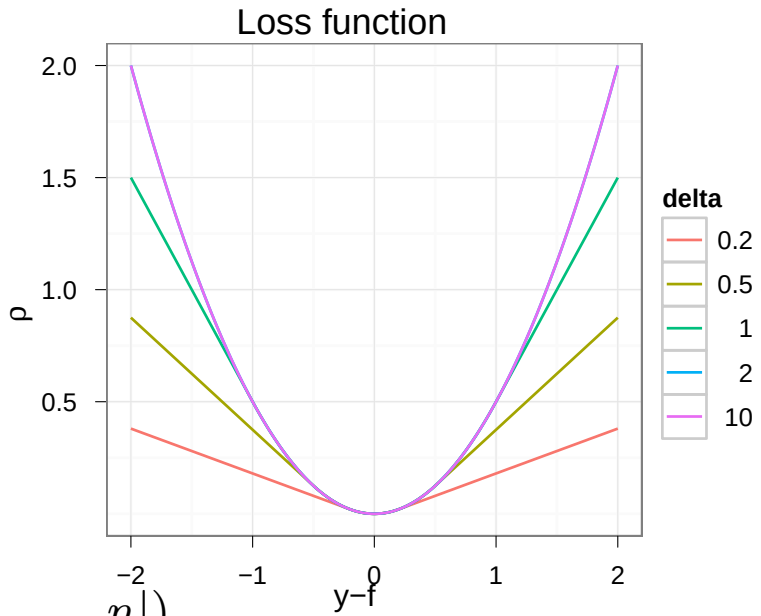
3. Huber-loss, `Huber()`

$$\rho(y, f; \delta) = \begin{cases} (y - f)^2/2 & \text{if } |y - f| \leq \delta, \\ \delta(|y - f| - \delta/2) & \text{if } |y - f| > \delta \end{cases}$$

```
R> fit <- gamboost(100/mpg ~ bols(wt) + bols(hp), data = mtcars,
+                  family = Huber(d = NULL))
```

- A compromise between L_1 and L_2 loss.
- δ defines the outliers which are subject to absolute error loss.
- Friedman (2001) proposed a strategy for adaptively changing δ at each boosting step (default in `mboost`):

$$\delta^{[m]} = \text{median}(|y_i - f^{[m-1]}(x_i), i = 1, \ldots, n|)$$

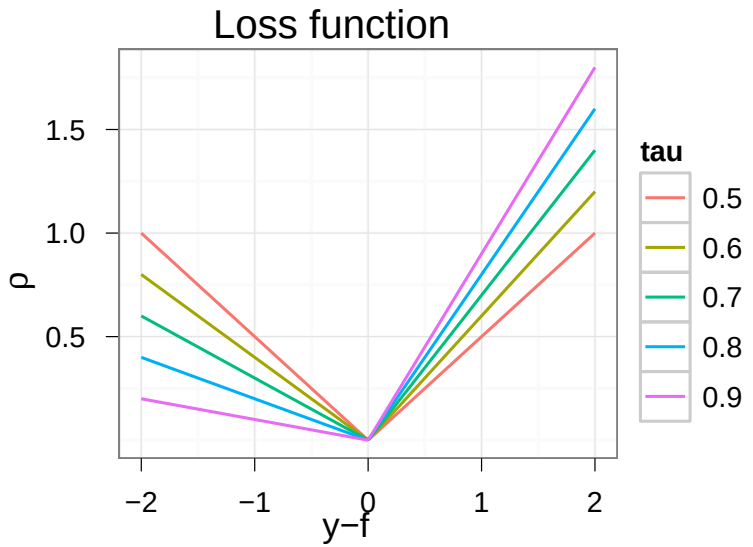


4. 'Check function'-loss, `QuantReg()`

$$\rho(y, f; \tau) = \begin{cases} (y - f) \cdot \tau & \text{if } (y - f) \geq 0, \\ (y - f) \cdot (\tau - 1) & \text{if } (y - f) < 0. \end{cases}$$

```
R> fit <- gamboost(100/mpg ~ bols(wt) + bols(hp), data = mtcars,
+                  family = QuantReg(tau = 0.5))
```

- One needs to specify the quantile τ.
- Note that `Laplace()` is essentially the same as `QuantReg(tau = 0.5)` with `nu = 0.2`.
- For further details see next talk and Fenske et al. (2011).



Families for Binary Response

Binary response

The `Binomial()` family uses the negative binomial log-likelihood as a loss function.

$$\hat{f} = \sum_{i}^{p} \hat{f}_i(x_i) \qquad \text{half of the log-odds ratio}$$

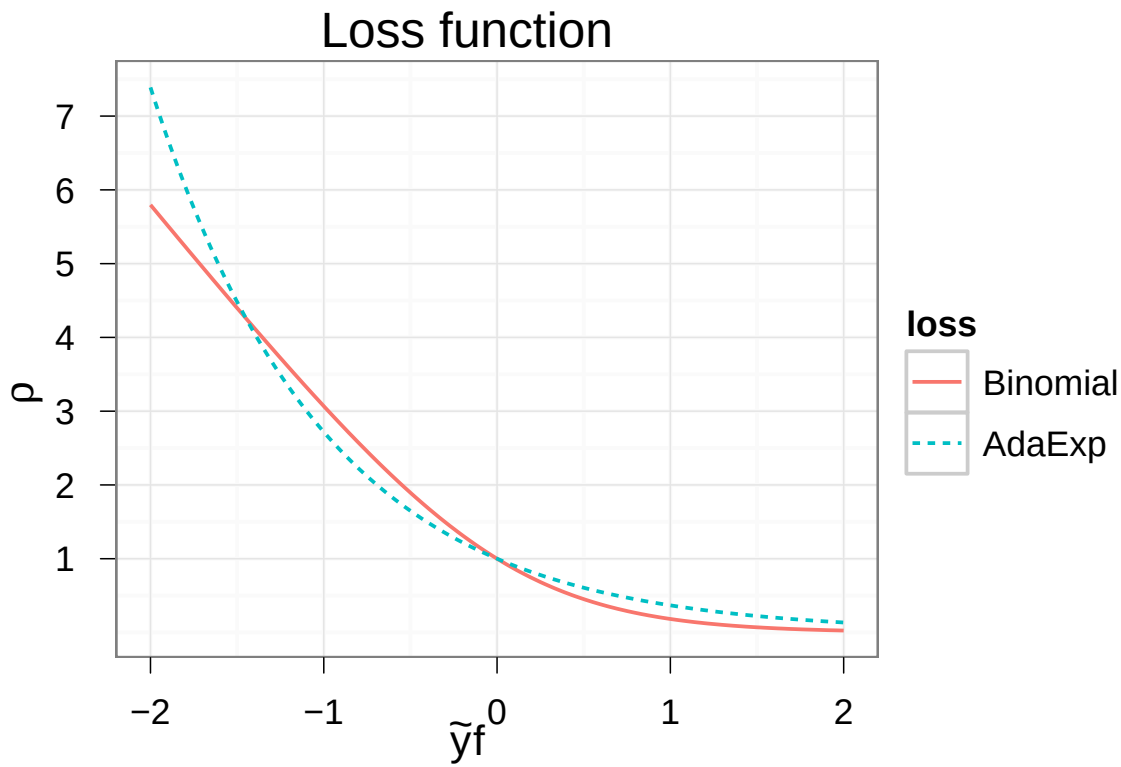
$$\pi(\hat{f}) = \mathbb{P}(Y = 1|x) = \text{logit}^{-1}(2\hat{f}) = \frac{e^{2\hat{f}}}{1 + e^{2\hat{f}}}$$

$$\begin{aligned} \rho(y, f) &= -[y \log(\pi(\hat{f})) + (1 - y) \log(1 - \pi(\hat{f}))] \\ &= \log(1 + \exp(-2\,\tilde{y}\hat{f})) \end{aligned}$$

- The binary response $Y \in \{0, 1\}$ is internally re-coded to $\tilde{Y} = 2\,Y - 1 \in \{-1, 1\}$.
- $\hat{f}$ equals the half of the logits (see next talks).
- You can choose between `Binomial(link = "logit")` and `Binomial(link = "probit")`.

Binary response

- `AdaExp()`
  - $\rho(y, f) = \exp(-\tilde{y}f)$
  - This essentially leads to the AdaBoost algorithm by Freund and Schapire (1996).
  - Similar to `Binomial()`.



Families for Count & Censored Responses

Count response

- `Poisson()` implements a family for fitting count data. The loss function is the negative Poisson log-likelihood. The natural link function $\log(\mu) = \eta$ is assumed as a link function.
- `NBinomial()` leads to regression models with a negative binomial conditional distribution of the response. Suitable for overdispersed count data (Schmid et al., 2010).

Censored response

- `CoxPH()` is suitable for survival models. This family implements the negative partial log-likelihood for Cox models resulting in a proportional hazards model.
- `Weibull()` is an implementation of the negative log-likelihood function of accelerated failure time models with Weibull outcomes (see Schmid and Hothorn, 2008, for further details).
- `Loglog()` and `Lognormal()` are similar to Weibull but designed for log-logistic outcome and log-normal outcome, respectively.

Take home notes

- A flexible framework addressing a rich set of scientific questions.
- Large number of possible combinations of families and base-learners.

$$\Rightarrow \underline{\text{New insights into understanding our data.}}$$

References

Nora Fenske, Thomas Kneib, and Torsten Hothorn. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Journal of the American Statistical Association, to appear, 2011.

Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, 148:156, 1996.

J.H. Friedman. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5):1189–1232, 2001.

Matthias Schmid and Torsten Hothorn. Flexible boosting of accelerated failure time models. BMC Bioinformatics, 9(269), 2008.

Matthias Schmid, Sergej Potapov, Annette Pfahlberg, and Torsten Hothorn. Estimation and regularization techniques for regression models with multidimensional prediction functions. Statistics and Computing, 20:139–150, 2010.