Boosting for classification: Loss Functions

Adrian Płoszczyca

Plan prezentacji

- Boosting - idea
- Algorytm AdaBoost
- Forward Stagewise Additive Modeling
- Właściwości wykładniczej funkcji straty
- Porównanie funkcji straty

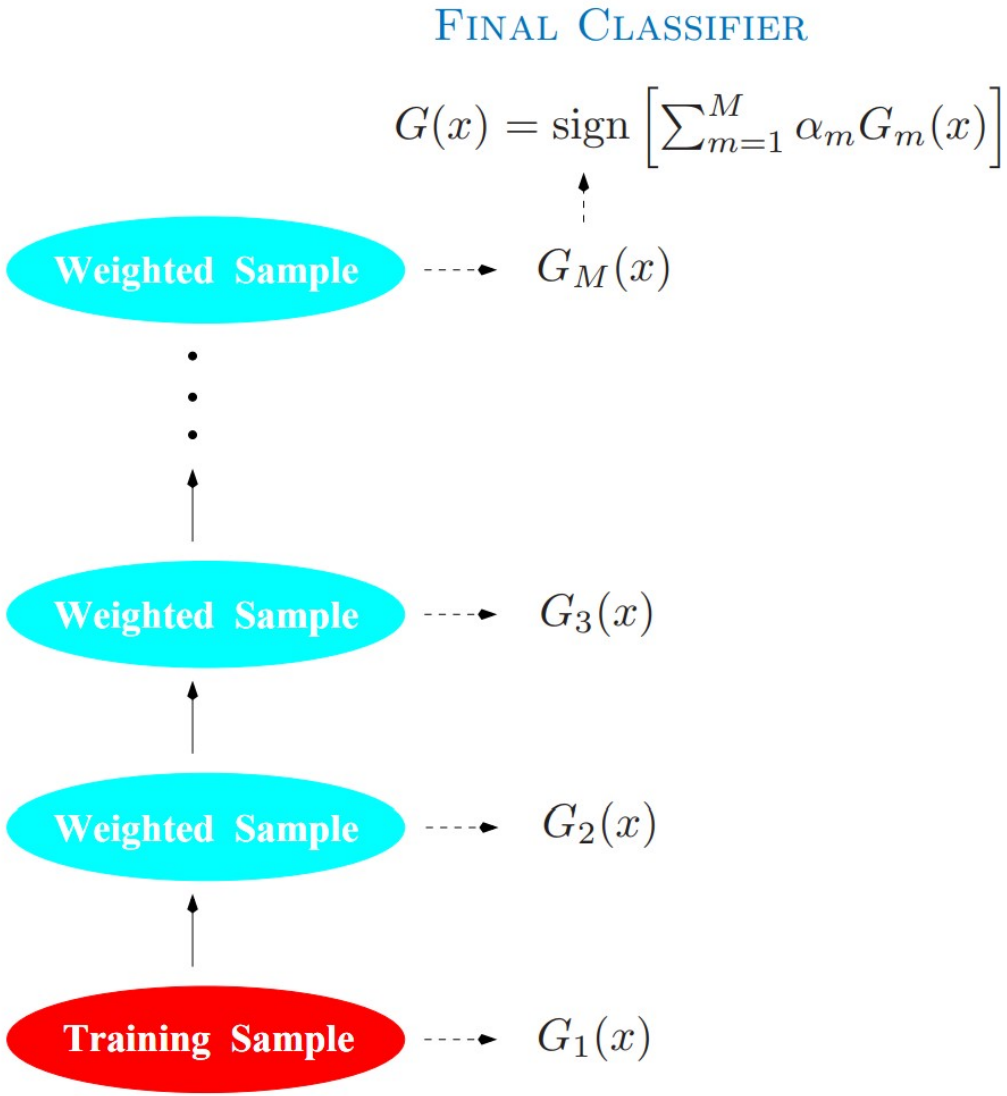
Boosting

Ideą boostingu jest połączenie wielu słabych klasyfikatorów w celu uzyskania klasyfikatora charakteryzującego się małym błędem klasyfikacji.

- $Y \in \{-1, 1\}$ - zmienne objaśniane
- X - zmienne objaśniające
- $G(X)$ - klasyfikator
- błąd klasyfikacji

$$\overline{\text{err}} = \frac{1}{N} \sum_{i=1}^{N} I(y_i \neq G(x_i))$$

AdaBoost - schemat



AdaBoost.M1. - algorytm

1. $w_i = 1/N,\ i = 1, 2, \ldots, N.$

2. Dla $m = 1, \ldots, M$:

   a) Przypisujemy klasy danym treningowym na podstawie klasyfikatora $G_m(x)$

   b) Obliczamy

$$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$

   c) $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m).$

   d) $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G(x_i)), i = 1, \ldots, N].$

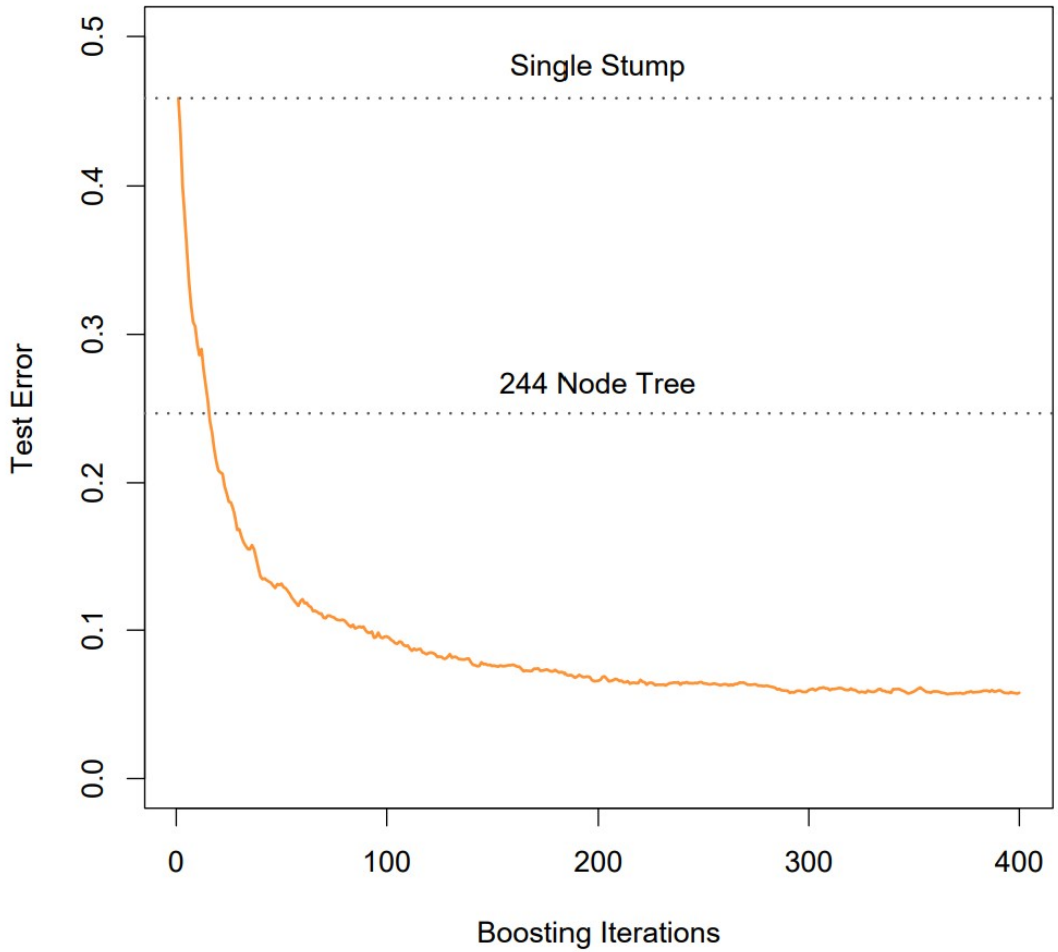
3. $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right].$

AdaBoost.M1. - przykład

$$X_j \sim \mathcal{N}(0,1), \quad j = 1, \ldots, 10$$

$$Y = \begin{cases} 1\,, & \text{dla } \sum_{j=1}^{10} X_j^2 > \chi^2_{10}(0.5), \\ -1, & \text{w p.p.} \end{cases}$$

AdaBoost.M1. - przykład



Boosting Fits an Additive Model

Boosting opiera się na funkcji

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m).$$

Dążymy do minimalizowania funkcji straty L

$$\min_{\{\beta_m, \gamma_m\}_1^M} \sum_{i=1}^{N} L\left(y_i, \sum_{m=1}^{M} \beta_m b(x_i; \gamma_m)\right).$$

Alternatywnie

$$\min_{\beta, \gamma} \sum_{i=1}^{N} L\left(y_i, \beta b(x_i; \gamma)\right).$$

Forward Stagewise Additive Modeling

1. $f_0(x) = 0$.
2. Dla $m = 1, \ldots, M$:

   a) Obliczamy

$$(\beta_m, \gamma_m) = \arg\min_{\beta,\gamma} \sum_{i=1}^{N} L\left(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)\right)$$

   b) $f_m(x) = f_{m-1}(x_i) + \beta_m b(x_i; \gamma_m)$.

Exponential Loss and AdaBoost

Rozważymy wykładniczą funkcję straty postaci

$$L(y, f(x)) = \exp(-yf(x)).$$

Chcemy zatem policzyć

$$(\beta_m, G_m) = \arg\min_{\beta, G} \sum_{i=1}^{N} \exp\left(-y_i(f_{m-1}(x_i) + \beta G(x_i))\right).$$

Niech

$$w_i^{(m)} = \exp(-y_i f_{m-1}(x_i)).$$

Mamy wtedy

$$(\beta_m, G_m) = \arg\min_{\beta, G} \sum_{i=1}^{N} w_i^{(m)} \exp\left(-\beta y_i G(x_i)\right).$$

Exponential Loss and AdaBoost

Rozwiązaniem

$$(\beta_m, G_m) = \arg\min_{\beta, G} \sum_{i=1}^{N} w_i^{(m)} \exp\left(-\beta y_i G(x_i)\right)$$

jest

$$G_m = \arg\min_{G} \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i))$$

$$\beta_m = \frac{1}{2} \log \frac{1 - \text{err}_m}{\text{err}_m},$$

gdzie

$$\text{err}_m = \frac{\sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i^{(m)}} \qquad (= \text{AdaBoost 2b}).$$

Exponential Loss and AdaBoost

Mamy wagi postaci

$$w_i^{(m+1)} = w_i^{(m)} \cdot e^{-\beta_m y_i G_m(x_i)}.$$

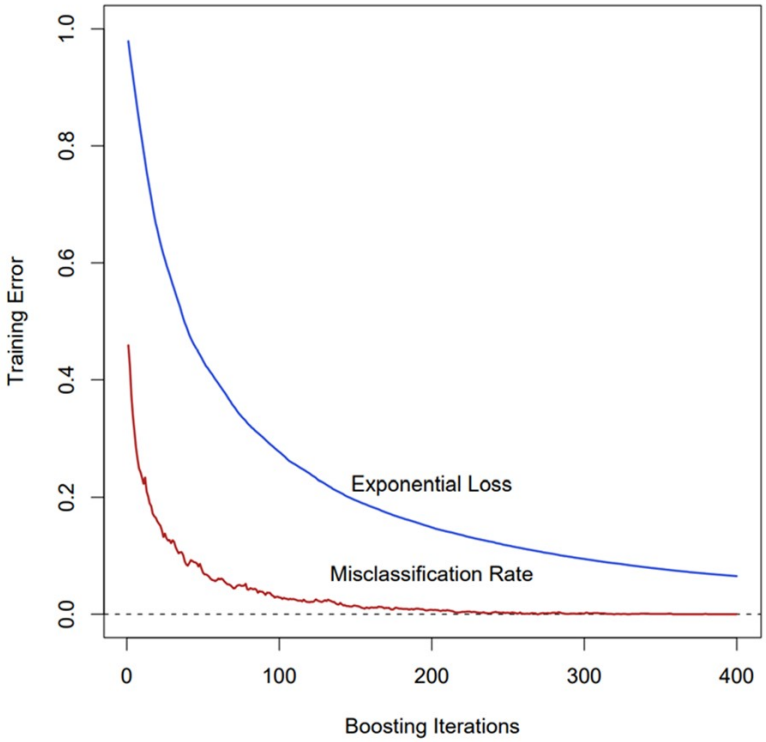
Podstawiając $-y_i G_m(x_i) = 2 \cdot I(y_i \neq G_m(x_i)) - 1$ otrzymujemy wzór na wagi postaci

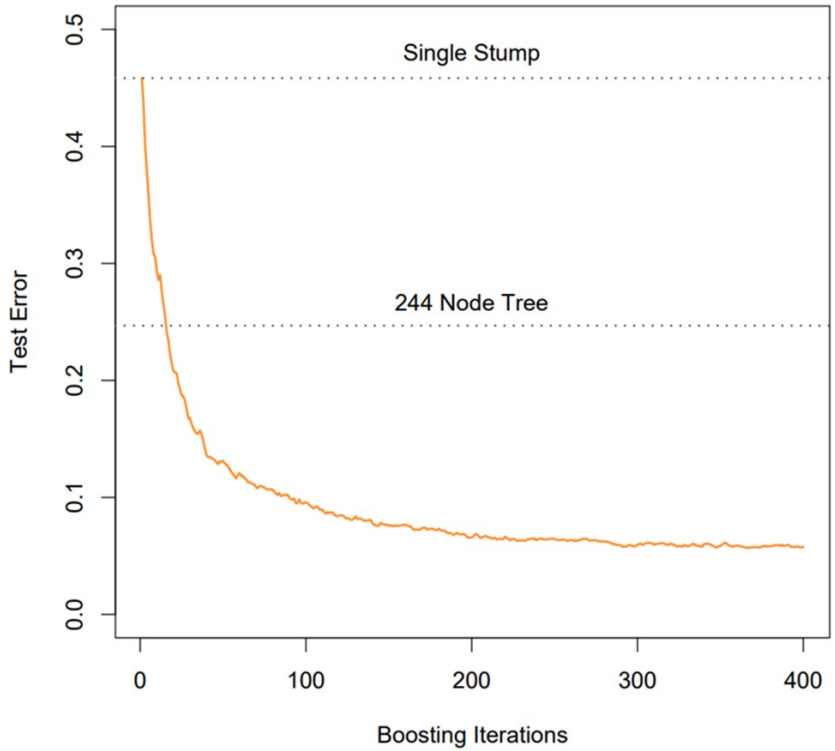
$$w_i^{(m+1)} = w_i^{(m)} \cdot e^{\alpha_m I(y_i \neq G_m(x_i))} \cdot e^{-\beta_m},$$

gdzie $\alpha_m = 2\beta_m$, a ponieważ $\beta_m = \frac{1}{2} \log \frac{1-\text{err}_m}{\text{err}_m}$, mamy $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$(= AdaBoost 2c).

Oprócz tego możemy zapisać $w_i^{(m+1)} = w_i^{(m)} \cdot e^{\alpha_m I(y_i \neq G_m(x_i))}$(= AdaBoost 2d).

Exponential Loss and AdaBoost





Exponential Loss Function

Argumentem minimalizującym $E_{Y|x}(e^{-Yf(x)})$ jest

$$f^*(x) = \frac{1}{2}\log\frac{\Pr(Y=1|x)}{\Pr(Y=-1|x)}.$$

Równoważnie możemy zapisać

$$\Pr(Y=1|x) = \frac{1}{1+e^{-2f^*(x)}}.$$

Binomial negative log-likelihood

Niech

$$p(x) = \mathsf{Pr}(Y = 1|x) = \frac{e^{f(x)}}{e^{-f(x)} + e^{f(x)}} = \frac{1}{1 + e^{-2f(x)}}$$

oraz

$$Y' = (Y + 1)/2 \in \{0, 1\}.$$

Wtedy mamy

$$l(Y, p(x)) = Y' \log p(x) + (1 - Y') \log(1 - p(x))$$

i równoważnie

$$-l(Y, f(x)) = \log\left(1 + e^{-2Yf(x)}\right).$$

Zatem

$\arg\min_{f(x)} E_{Y|x}(e^{-Yf(x)})$ jest taki sam jak $\arg\min_{f(x)} E_{Y|x}(-l(Y, f(x)))$.

Robust Loss Functions for Classification

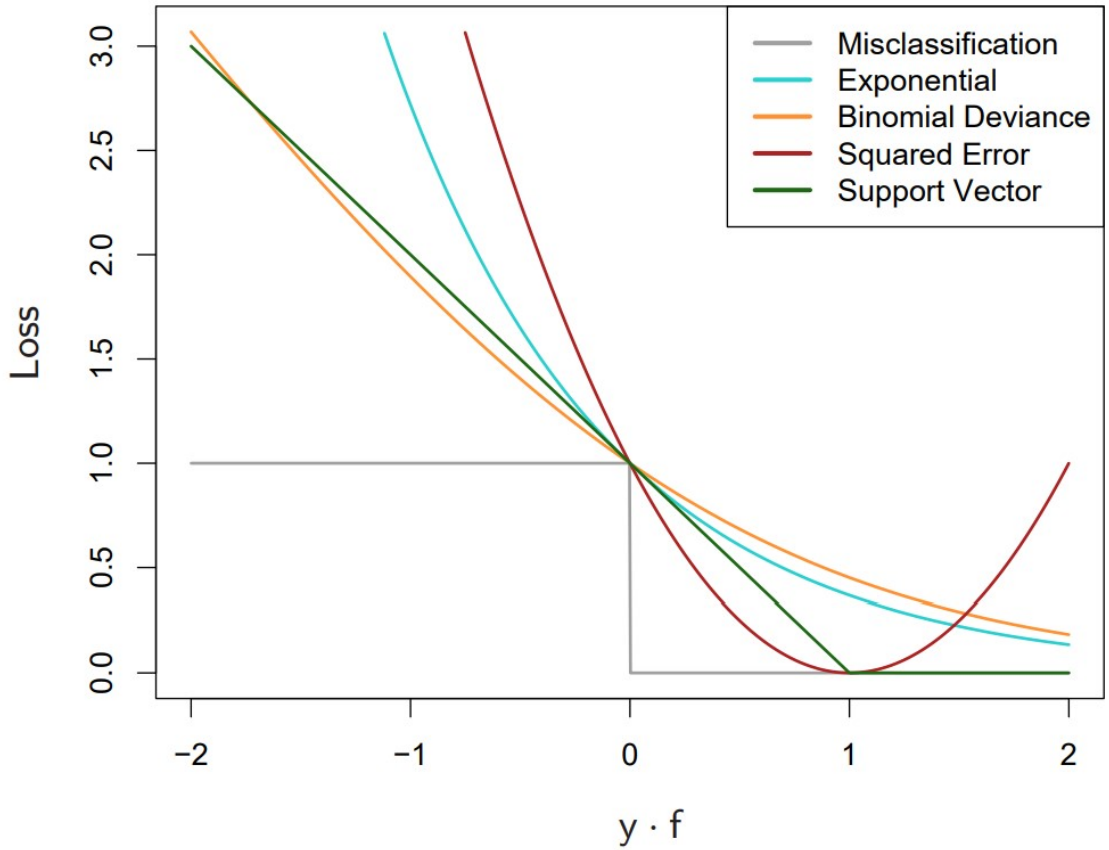
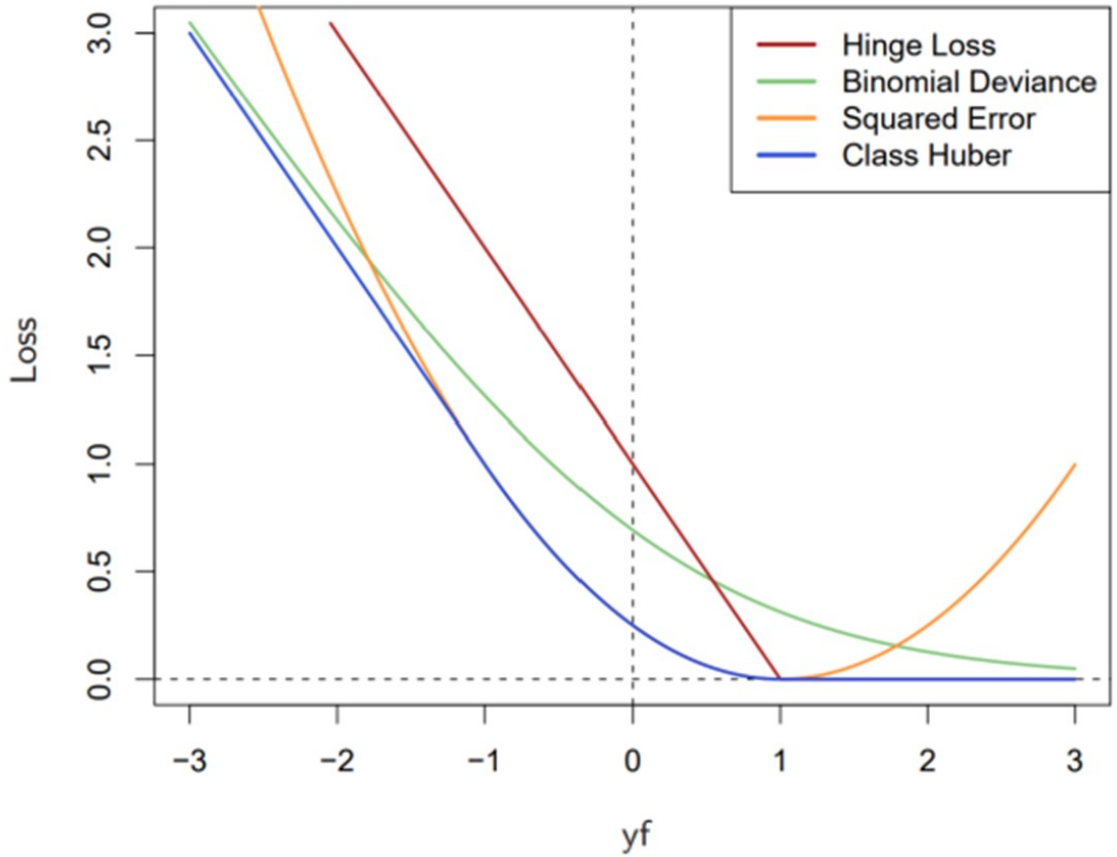


FIGURE 10.4. *Loss functions for two-class classification. The response is* $y = \pm 1$*; the prediction is* f*, with class prediction* $\text{sign}(f)$*. The losses are misclassification:* $I(\text{sign}(f) \neq y)$*; exponential:* $\exp(-yf)$*; binomial deviance:* $\log(1 + \exp(-2yf))$*; squared error:* $(y - f)^2$*; and support vector:* $(1 - yf)_+$ *(see Section 12.3). Each function has been scaled so that it passes through the point* $(0, 1)$*.*

Robust Loss Functions for Classification



$$L(y, f(x)) = \begin{cases} -4yf(x), & \text{dla } \; yf(x) < -1 \\ [1 - yf(x)]_+^2, & \text{w p.p.} \end{cases},$$

KONIEC

Dziękuję za uwagę