AdaBoost as a forward stagewise additive model

two-class classification problem $y_i \in \{-1, 1\}$

x_i = p-dimensional vector of covariates

using as a loss-function the exponential loss: $L(y, f(x)) = \exp\{-y f(x)\}$

In a forward stagewise additive procedure, at step m → step length

$$(\beta_m, G_m) = \underset{\beta, G}{\operatorname{argmin}} \sum_{i=1}^{N} L\left(y_i, \underbrace{\hat{f}^{[m-1]}(x_i)}_{\text{previous estimate}} + \underbrace{\beta\, G(x_i)}_{\text{update}}\right)$$

→ weak classifier

in the case of exponential loss,

$$(\beta_m, G_m) = \underset{\beta, G}{\operatorname{argmin}} \sum_{i=1}^{N} \exp\left\{-y_i\left(\hat{f}^{[m-1]}(x_i) + \beta\, G(x_i)\right)\right\}$$

$$= \underset{\beta, G}{\operatorname{argmin}} \sum_{i=1}^{N} \underbrace{\exp\left\{-y_i \hat{f}^{[m-1]}(x_i)\right\}}_{w_i^{[m]} \text{ does not depend on } \beta \text{ or } G} \exp\left\{-y_i \beta\, G(x_i)\right\}$$

We proceed with the minimization in two steps

→ minimise in G given $\beta > 0$ (choice of the weak learner)

→ minimise in β (choice of step length)

if $y_i = G(x_i)$, then $y_i \cdot G(x_i) = 1$

if $y_i \neq G(x_i)$, then $y_i\, G(x_i) = -1$

$$\hat{G}_m = \underset{G}{\operatorname{argmin}} \sum_{i=1}^{N} w_i^{[m]} \exp\left\{-y_i \beta\, G(x_i)\right\}$$

$$= \underset{G}{\operatorname{argmin}} \left(\sum_{i: y_i = G(x_i)} w_i^{[m]} e^{-\beta} + \sum_{i: y_i \neq G(x_i)} w_i^{[m]} e^{\beta}\right)$$

$$= \underset{G}{\operatorname{argmin}} \left(\sum_{i: y_i = G(x_i)} w_i^{[m]} e^{-\beta} + \sum_{i: y_i \neq G(x_i)} w_i^{[m]} e^{-\beta} - \sum_{i: y_i \neq G(x_i)} w_i^{[m]} e^{-\beta} + \sum_{i: y_i \neq G(x_i)} w_i^{[m]} e^{\beta}\right)$$

$$= \underset{G}{\operatorname{argmin}} \left(\sum_{i=1}^{N} w_i^{[m]} e^{-\beta} + \left(e^{\beta} - e^{-\beta}\right) \sum_{i: y_i \neq G(x_i)} w_i^{[m]}\right)$$

$$= \underset{G}{\operatorname{argmin}} \left(\sum_{i=1}^{N} w_i^{[m]} e^{-\beta} + \left(e^{\beta} - e^{-\beta}\right) \sum_{i=1}^{N} w_i^{[m]}\, \mathbb{1}\left(y_i \neq G(x_i)\right)\right)$$

$$\Rightarrow \hat{G}_m = \underset{G}{\operatorname{argmin}} \left\{\sum_{i=1}^{N} w_i^{[m]}\, \mathbb{1}\left(y_i \neq G(x_i)\right)\right\}$$

that is the classifier used by AdaBoost (minimise the weighted misclassified observations)

About β

$$\hat{\beta}_m = \arg\min_{\beta} \sum_{i=1}^{N} w_i^{[m]} \exp\{-y_i \beta G(x_i)\}$$

$$y_i = G(x_i) \rightarrow \sum_{i: y_i = G(x_i)} w_i^{[m]} e^{-\beta}$$

$$y_i \neq G(x_i) \rightarrow \sum_{i: y_i \neq G(x_i)} w_i^{[m]} e^{\beta}$$

$$L(\beta) = \sum_{i: y_i = G(x_i)} w_i^{[m]} e^{-\beta} + \sum_{i: y_i \neq G(x_i)} w_i^{[m]} e^{\beta}$$

$$\frac{\partial L(\beta)}{\partial \beta} = -\sum_{i: y_i = G(x_i)} w_i^{[m]} e^{-\beta} + \sum_{i: y_i \neq G(x_i)} w_i^{[m]} e^{\beta} = 0$$

multiply both sides by e^{β}

$$-\sum_{i: y_i = G(x_i)} w_i^{[m]} + e^{2\beta} \sum_{i: y_i \neq G(x_i)} w_i^{[m]} = 0$$

$$e^{2\beta} \sum_{i: y_i \neq G(x_i)} w_i^{[m]} = \sum_{i=1}^{N} w_i^{[m]} - \sum_{i: y_i \neq G(x_i)} w_i^{[m]}$$

$$e^{2\beta} = \frac{\sum_{i=1}^{N} w_i^{[m]} \Big/ \sum_{i=1}^{N} w_i^{[m]} - \sum_{i: y_i \neq G(x_i)} w_i^{[m]} \Big/ \sum_{i=1}^{N} w_i^{[m]}}{\sum_{i: y_i \neq G(x_i)} w_i^{[m]} \Big/ \sum_{i=1}^{N} w_i^{[m]}}$$

$$= \frac{1 - err^{[m]}}{err^{[m]}} \quad \Rightarrow \hat{\beta}_m = \frac{1}{2} \log \frac{1 - err^{[m]}}{err^{[m]}}$$

where $$err^{[m]} = \frac{\sum_{i: y_i \neq G(x_i)} w_i^{[m]}}{\sum_{i=1}^{N} w_i^{[m]}} = \frac{\sum_{i=1}^{N} w_i^{[m]} \mathbb{1}(y_i \neq G(x_i))}{\sum_{i=1}^{N} w_i^{[m]}}$$

When our classifier is updated

$$\hat{f}^{[m]}(x_i) = \hat{f}^{[m-1]}(x_i) + \hat{\beta}_m \hat{G}_m(x_i)$$

we need to update the weights

$$\begin{aligned} w_i^{[m+1]} &= \exp\left\{-y_i \hat{f}^{[m]}(x_i)\right\} \\ &= \exp\left\{-y_i \left(\hat{f}^{[m-1]}(x_i) + \hat{\beta}_m \hat{G}_m(x_i)\right)\right\} \\ &= \exp\left\{-y_i \hat{f}^{[m-1]}(x_i)\right\} \exp\left\{-y_i \hat{\beta}_m \hat{G}_m(x_i)\right\} \\ &= w_i^{[m]} \exp\left\{-y_i \hat{\beta}_m \hat{G}_m(x_i)\right\} \end{aligned}$$

$$y_i G(x_i) = \begin{cases} 1 & \text{if } y_i = \hat{G}_m(x_i) \\ -1 & \text{if } y_i \neq \hat{G}_m(x_i) \end{cases}$$

$$\begin{aligned} &= \mathbb{1}(y_i = G(x_i)) - \mathbb{1}(y_i \neq G(x_i)) \\ &= \underbrace{\mathbb{1}(y_i = G(x_i)) + \mathbb{1}(y_i \neq G(x_i))}_{1} - \mathbb{1}(y_i \neq G(x_i)) - \mathbb{1}(y_i \neq G(x_i)) \\ &= 1 - 2\mathbb{1}(y_i \neq G(x_i)) \end{aligned}$$

$$\begin{aligned} &= w_i^{[m]} \exp\left\{\hat{\beta}_m \left(2\,\mathbb{1}(y_i \neq G(x_i)) - 1\right)\right\} \\ &= w_i^{[m]} \exp\left\{2\hat{\beta}_m \mathbb{1}(y_i \neq G(x_i))\right\} \underbrace{\exp\left\{\hat{\beta}_m\right\}}_{\text{const.}} \\ &\propto w_i^{[m]} \exp\left\{2\hat{\beta}_m \mathbb{1}(y_i \neq G(x_i))\right\} \end{aligned}$$

→ disappear with the normalization