Adaboost.M1
exp-loss forward
stagewise modelling

[derivation "stolen/borrowed" from STKIN4300 UiO
by Riccardo De Bin]

$Y = \{-1, 1\}$

$X \in \mathbb{R}^p$

exponential loss

0) Loss-function $L(y, f(x)) = \exp(-y f(x))$

$$\sum_{i=1}^{N} L(y_i, f(x_i))$$

Forward stagewise modelling:

$$\underset{\beta, \gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L\Big(y_i, \underbrace{\hat{f}^{(m-1)}(x_i)}_{\text{current classifier - not to change}} + \overbrace{\beta b(x_i, \gamma)}^{\text{to be changed}}\Big)$$

1) What does this look like at step m?

$$(\beta_m, b_m) = \underset{\beta, f}{\operatorname{argmin}} \sum_{i=1}^{N} \exp\left(-y_i \sum_{k=1}^{m} \beta^{(k)} b^{(k)}(x_i)\right)$$

$$= \underset{\beta, b}{\operatorname{argmin}} \sum_{i=1}^{N} \exp\left\{-y_i \left(\sum_{k=1}^{m-1} \beta^{(k)} \hat{f}^{(k-1)}(x_i) + \beta b(x_i)\right)\right\}$$

$$= \underset{\beta, b}{\operatorname{argmin}} \sum_{i=1}^{N} w_i^{(m-1)} \cdot \exp(-y_i \beta b(x_i))\}$$

$$w_i^{(m-1)} = \exp\{-y_i \hat{f}^{(m-1)}(x_i)\}$$

known - from last iteration -

Observe: if $y_i = b(x_i)$ then $y_i b(x_i) = +1$
$\neq$ -1

Next: two step procedure

where a) minimize wrt b
b) minimize wrt β

2)
$$b_m = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{N} w_i^{(m-1)} \cdot \exp(-y_i \beta b(x_i))$$

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

add and subtract

$$\sum_{i: y_i \neq b(x_i)} w_i^{(m-1)} e^{-\beta}$$

$$= \underset{b}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i=1}^{N} w_i^{(m-1)} e^{-\beta}}_{\text{no } b \text{ here}} + (e^{\beta} - e^{-\beta}) \sum_{\underbrace{i: y_i = b(x_i)}_{I(y_i \neq b(x_i))}} w_i^{(m-1)} \right\}$$

$$\hat{b}_m = \underset{b}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} w_i^{(m-1)} I(y_i \neq b(x_i)) \right\}$$

3)

$$\beta^{(m)} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} w_i^{(m-1)} \exp\{-y_i \beta b(x_i)\}$$

$$y_i = b(x_i) \rightarrow \sum_{i:y_i=b(x_i)} w_i^{(m-1)} e^{-\beta}$$

$$y_i \neq b(x_i) \rightarrow \sum_{i:y_i\neq b(x_i)} w_i^{(m-1)} e^{\beta}$$

$$L = \sum_{i:y_i=b(x_i)} w_i^{(m-1)} e^{-\beta} + \sum_{i:y_i\neq b(x_i)} w_i e^{\beta}$$

$$\frac{\partial L}{\partial \beta} = -\sum_{i:y_i=b(x_i)} w_i^{(m-1)} e^{-\beta} + \sum_{i:y_i\neq b(x_i)} w_i^{(m-1)} e^{\beta} = 0$$

multiply e^{β}

$$-\sum_{i:y_i=b(x_i)} w_i^{(m-1)} + e^{2\beta} \sum_{i:y_i\neq b(x_i)} w_i^{(m-1)} = 0$$

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

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

scale each term with $\sum_{i=1}^{N} w_i^{(m-1)}$ to get

$$e^{2\beta} = \frac{1 - err_m}{err_m}$$

$$2\hat{\beta}_m = \ln\left(\frac{1-err_m}{err_m}\right)$$

where $err_m = \dfrac{\sum_{i: y_i \neq b(x_i)} w_i^{(m-1)}}{\sum_{i=1}^{N} w_i^{(m-1)}}$

In the Adaboost.M1: $\alpha_m = 2\beta_m$

$$\hat{\alpha}_m = \ln\left(\frac{1-err_m}{err_m}\right)$$

4) In 2) we found that $\hat{b}_m$ is minimizing the "weighted misclassification error" and in 3) we found

$\hat{\beta}_m = \frac{1}{2}\ln\left(\frac{1-err_m}{err_m}\right)$, or $\hat{\alpha}_m = \ln\left(\frac{1-err_m}{err_m}\right)$

This gives

Our classifier is updated as

$$\hat{f}^{(m)}(x_i) = \hat{f}^{(m-1)}(x_i) + \hat{\beta}_m \cdot \hat{b}_m(x_i)$$

and the weights (from 1))

$$w_i^{(m)} = w_i^{(m-1)} \exp\left(-y_i \hat{\beta}_m \hat{b}_m(x_i)\right)$$

But in the algorithm the weight update is written differently "$w_i \leftarrow w_i \cdot \exp(\alpha_m \cdot I(y_i \neq b_m(x_i))$"

Final "adjustment", look at exponent:

$$-y_i \cdot \hat{b}_m(x_i) = -I(y_i = \hat{b}_m(x_i)) + I(y_i \neq \hat{b}_m(x_i))$$

$$= -I(y_i = \hat{b}_m(x_i)) + I(y_i \neq \hat{b}_m(x_i)) - I(y_i \neq \hat{b}(x_i)) + I(y_i \neq \hat{b}_m(x_i))$$

$$= -1 + 2 \cdot I(y_i \neq \hat{b}_m(x_i))$$

Insert info:

$$w_i^{(m)} = w_i^{(m-1)} \cdot \exp\left(\hat{\beta}_m \left(-1 + 2I(y_i \neq \hat{b}_m(x_i))\right)\right)$$

$$= w_i^{(m-1)} \exp\Big(\underbrace{2\hat{\beta}_m}_{\hat{\alpha}_m} \cdot I(y_i \neq \hat{b}_m(x_i)) - \hat{\beta}_m \Big)$$

$$= w_i^{(m-1)} \exp\big(\hat{\alpha}_m \cdot I(y_i \neq \hat{b}_m(x_i)) \big) \cdot \exp\Big(-\frac{\hat{\alpha}_m}{2}\Big)$$

an extra factor = the same value is multiplied is for each weight w_i = a scaling