AdaBoost Training Error

David Helmbold

University of California, Santa Cruz
dph@soe.ucsc.edu

Spring ’11
AdaBoost

Takes sample $S = \{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}$ where $y_i \in \{+1, -1\}$. Initialize $D_1(i) = 1/\ell$ for each $i \in \{1, \ldots, \ell\}$

\begin{algorithm}
\begin{algorithmic}
\FOR{$t = 1$ \text{ to } $T$}
\STATE Train using $D_t$ over $S$ to get weak hypothesis $h_t$
\STATE $\varepsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$ \hfill $\triangleright$ training error of $h_t$ wrt $D_t$
\STATE $\alpha_t = \frac{1}{2} \ln \left( \frac{1-\varepsilon_t}{\varepsilon_t} \right)$ \hfill $\triangleright$ weight or vote of $h_t$
\FOR{all $i \in \{1, \ldots, \ell\}$}
\STATE $D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{\text{norm}_t}$
\ENDFOR
\ENDFOR
\STATE Output $H(x) = \text{sign} \left( \sum_t \alpha_t h_t(x) \right)$
\end{algorithmic}
\end{algorithm}
Facts about $\text{norm}_t$:

1. $\text{norm}_t = \sum_{i=1}^{\ell} D_t(i) \exp(-\alpha_t y_i h_t(x_i))$

2. $\ell \prod_{t=1}^{T} \text{norm}_t = \sum_{i=1}^{\ell} \exp(m_i) > \#\text{ensemble training errors}$

3. When $\alpha_t$ optimized, $\text{norm}_t = 2\sqrt{\epsilon_t(1 - \epsilon_t)} < 1$ (for $\epsilon_t < 1/2$)

4. Bound on training error decreases exponentially, $O(\ln \ell)$ rounds until ensemble perfect on training set (if $\epsilon_t$ bounded below 1/2)
Unroll $D_{T+1}(i)$:

$$D_{T+1}(i) = \frac{D_T(i) \exp(-\alpha_t y_i h_t(x_i))}{\text{norm}_T}$$

$$= \frac{D_{T-1}(i) \exp(-\alpha_{T-1} y_i h_{T-1}(x_i)) \exp(-\alpha_T y_i h_T(x_i))}{\text{norm}_{T-1}} \frac{1}{\text{norm}_T}$$

$$D_1(i) \prod_{t=1}^{T} \exp(-\alpha_t y_i h_t(x_i))$$

$$= \frac{D_1(i) \prod_{t=1}^{T} \exp(-\alpha_t y_i h_t(x_i))}{\prod_{t=1}^{T} \text{norm}_t}$$

$$= \frac{\exp\left(\sum_{t=1}^{T} \alpha_t y_i h_t(x_i)\right)}{\ell \prod_{t=1}^{T} \text{norm}_t}$$

margin $m_i$
Facts about $\text{norm}_t$:

1. $\text{norm}_t = \sum_{i=1}^{\ell} D_t(i) \exp(-\alpha_t y_i h_t(x_i))$

2. $\ell \prod_{t=1}^{T} \text{norm}_t = \sum_{i=1}^{\ell} \exp(m_i) > \#\text{ensemble training errors}$

3. When $\alpha_t$ optimized, $\text{norm}_t = 2 \sqrt{\epsilon_t (1 - \epsilon_t)} < 1$ (for $\epsilon_t < 1/2$)

4. Bound on training error decreases exponentially, $O(\ln \ell)$ rounds until ensemble perfect on training set (if $\epsilon_t$ bounded below 1/2)
Functional or un-normalized margin \( m_i = \sum_{t=1}^{T} \alpha_t y_t h_t(x_i) > 0 \).

- A training error on \( x_i \) only if \( m_i \leq 0 \)
- If \( m_i \leq 0 \) then \( \exp(-m_i) \geq 1 \)

\[
\sum_{i=1}^{\ell} \exp(-m_i) \text{ is an overestimate of } \# \text{ training errors!}
\]

\[
\sum_{i=1}^{\ell} \exp(-m_i) \text{ starts at } \ell, \text{ how fast does it drop with } t? 
\]

\[
\sum_{i=1}^{\ell} \exp(-m_i) = \ell \prod_{t} \text{norm}_t, \text{ want norm}_t \text{'s small.}
\]
Facts about \( \text{norm}_t \):

1. \( \text{norm}_t = \sum_{i=1}^{\ell} D_t(i) \exp(-\alpha_t y_i h_t(x_i)) \)

2. \( \ell \prod_{t=1}^{T} \text{norm}_t = \sum_{i=1}^{\ell} \exp(m_i) > \#\text{ensemble training errors} \)

3. when \( \alpha_t \) optimized, \( \text{norm}_t = 2\sqrt{\epsilon_t(1-\epsilon_t)} < 1 \) (for \( \epsilon_t < 1/2 \))

4. Bound on training error decreases exponentially, \( O(\ln \ell) \) rounds until ensemble perfect on training set (if \( \epsilon_t \) bounded below 1/2)
\[ \text{norm}_t = \sum_{i=1}^{\ell} D_t(i) \exp(-\alpha_t y_i h_t(x_i)) \]

\[ = \sum_{i: h_t(x_i) = y_i} D_t(i) \exp(-\alpha_t) + \sum_{i: h_t(x_i) \neq y_i} D_t(i) \exp(\alpha_t) \]

\[ = \exp(-\alpha_t) (1 - \epsilon_t) + \exp(\alpha_t) \epsilon_t \]

Convex, minimized (wrt \( \alpha_t \)) when derivative = 0:
Setting derivative of \( \exp(-\alpha_t) (1 - \epsilon_t) + \exp(\alpha_t) \epsilon_t \) to 0:

\[
0 = -\exp(-\alpha_t) (1 - \epsilon_t) + \exp(\alpha_t) \epsilon_t
\]

\[
\exp(-2\alpha_t) = \frac{\epsilon_t}{1 - \epsilon_t}
\]

\[
\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)
\]
Substitute $\alpha_t = \frac{1}{2} \ln \left( \frac{1-\epsilon_t}{\epsilon_t} \right)$ gives $\exp(\alpha_t) = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$ and

$$\text{norm}_t = \exp(-\alpha_t) (1 - \epsilon_t) + \exp(\alpha_t) \epsilon_t$$

$$= \sqrt{\frac{\epsilon_t}{1 - \epsilon_t}} \cdot (1 - \epsilon_t) + \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} \cdot \epsilon_t$$

$$= 2 \sqrt{\epsilon_t (1 - \epsilon_t)}$$

$\epsilon_t < 1/2$ means $\text{norm}_t < 1$
Facts about $\text{norm}_t$:

1. $\text{norm}_t = \sum_{i=1}^{\ell} D_t(i) \exp(-\alpha_t y_i h_t(x_i))$

2. $\ell \prod_{t=1}^{T} \text{norm}_t = \sum_{i=1}^{\ell} \exp(m_i) > \#\text{ensemble training errors}$

3. When $\alpha_t$ optimized, $\text{norm}_t = 2 \sqrt{\epsilon_t (1 - \epsilon_t)} < 1$ (for $\epsilon_t < 1/2$)

4. Bound on training error decreases exponentially, $O(\ln \ell)$ rounds until ensemble perfect on training set (if $\epsilon_t$ bounded below 1/2)
Assume all $\epsilon_t \leq c < 1/2$, then all norm$_t \leq b = 2\sqrt{c(1-c)} < 1$

then $\ell \prod_{t=1}^{T} \text{norm}_t \leq \ell b^t$, decreases exponentially in $t$

as does the $\sum_{i=1}^{\ell} \exp(-m_i)$ overestimate of training error

in $O(\ln \ell)$ iterations, ensemble perfect on training set
(but $\frac{1}{\ln(1/b)}$ constant hidden in big-O)
Facts about \( \text{norm}_t \): 

1. \( \text{norm}_t = \sum_{i=1}^{\ell} D_t(i) \exp(-\alpha_t y_i h_t(x_i)) \)

2. \( \ell \prod_{t=1}^{T} \text{norm}_t = \sum_{i=1}^{\ell} \exp(m_i) > \#\text{ensemble training errors} \)

3. when \( \alpha_t \) optimized, \( \text{norm}_t = 2 \sqrt{\epsilon_t(1 - \epsilon_t)} < 1 \) (for \( \epsilon_t < 1/2 \))

4. Bound on training error decreases exponentially, \( O(\ln \ell) \) rounds until ensemble perfect on training set (if \( \epsilon_t \) bounded below \( 1/2 \))