Lecture 20: Boosting, Adaboost

CS 3780/5780

1 Boosting Theorem

The boosting theorem says that if weak learning hypothesis is satisfied by some weak learning algorithm, then Adaboost algorithm will ensemble the weak hypothesis and produce a classifier with 0 training error. Whats more, it also provides a bound on number of such weak learning hypothesis we would need to ensemble.

Theorem 1. *If Weak Learning Hypothesis holds with some margin $\gamma > 0$, then Adaboost will find an ensembled classifier with 0 training error on sample D within*

$$T \leq \frac{\log(n)}{2\gamma^2} \text{ iterations.}$$

Proof. Recall that ϵ_t is the w_t weighted error of the weak learner given by $\epsilon_t = \sum_{i=1}^n w_t[i]\mathbf{1}\{h_t(x_i) \neq y_i\}$. By weak learning hypothesis,

$$\epsilon_t = \frac{1}{2} - \gamma_t < \frac{1}{2} - \gamma \qquad \text{(better than random guess)}$$

Now we will analyze the training error.

$$\begin{aligned}
\text{err}_D(h_{\text{Boost}}) &= \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{h_{\text{Boost}}(x_i)y_i < 0\} \\
&\leq \frac{1}{n}\sum_{i=1}^n \exp\left(-h_{\text{Boost}}(x_i)y_i\right) \qquad \text{(Exp loss upper bounds classification loss)} \\
&= \frac{1}{n}\sum_{i=1}^n \exp\left(-\sum_{t=1}^T \alpha_t h_t(x_i)y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^n \prod_{t=1}^T \exp\left(-\alpha_t h_t(x_i)y_i\right)
\end{aligned}$$

Now let us derive a simplified form for the expression above. Note that the weights of Adaboost were given by

$$w_{t+1}[i] \propto w_t[i]\exp\left(-\alpha_t y_i h_t(x_i)\right)$$

and since w_{t+1} is a probability vector that sums to 1,

$$w_{t+1}[i] = \frac{w_t[i]\exp\left(-\alpha_t y_i h_t(x_i)\right)}{\sum_{j=1}^n w_t[j]\exp\left(-\alpha_t y_j h_t(x_j)\right)}$$

Let us denote the normalizing factor in the denominator as $Z_t = \sum_{j=1}^n w_t[j]\exp\left(-\alpha_t y_j h_t(x_j)\right)$. Note note that:

$$
\begin{aligned}
Z_T &= \sum_{j=1}^{n} w_T[j] \exp\left(-\alpha_T y_j h_T(x_j)\right) \\
&= \sum_{j=1}^{n} \frac{1}{Z_{T-1}} w_{T-1}[j] \exp\left(-\alpha_{T-1} y_j h_{T-1}(x_j)\right) \exp\left(-\alpha_T y_j h_T(x_j)\right) \\
&= \sum_{j=1}^{n} \frac{1}{Z_{T-2}} w_{T-2}[j] \exp\left(-\alpha_{T-2} y_j h_{T-2}(x_j)\right) \exp\left(-\alpha_{T-1} y_j h_{T-1}(x_j)\right) \exp\left(-\alpha_T y_j h_T(x_j)\right) \\
&\dots \\
&= \frac{1}{Z_1 \cdot Z_2 \cdot \dots \cdot Z_{T_1}} \sum_{j=1}^{n} w_1[j] \prod_{t=1}^{T} \exp\left(-\alpha_t y_j h_t(x_j)\right) \\
&= \frac{1}{n} \frac{1}{Z_1 \cdot Z_2 \cdot \dots \cdot Z_{T_1}} \sum_{j=1}^{n} \prod_{t=1}^{T} \exp\left(-\alpha_t y_j h_t(x_j)\right)
\end{aligned}
$$

Thus we can conclude that:

$$\prod_{t=1}^{T} Z_t = \frac{1}{n} \sum_{i=1}^{n} \prod_{t=1}^{T} \exp\left(-\alpha_t h_t(x_i) y_i\right)$$

Thus using this in the bound on training error we can conclude that:

$$\text{err}_D(h_{\text{Boost}}) \leq \prod_{t=1}^{T} Z_t \tag{1}$$

Now note that

$$
\begin{aligned}
z_t &= \sum_{j=1}^{n} w_t[j] \exp\left(-\alpha_t y_j h_t(x_j)\right) \\
&= \sum_{j=1}^{n} w_t[j] \exp\left(-\alpha_t\right) \mathbf{1}\{h_t(x_j) = y_j\} + \sum_{j=1}^{n} w_t[j] \exp\left(\alpha_t\right) \mathbf{1}\{h_t(x_j) \neq y_j\} \\
&= \exp\left(-\alpha_t\right) \left(\sum_{j=1}^{n} w_t[j] \mathbf{1}\{h_t(x_j) = y_j\} \right) + \exp\left(\alpha_t\right) \left(\sum_{j=1}^{n} w_t[j] \mathbf{1}\{h_t(x_j) \neq y_j\} \right) \\
&= \exp\left(-\alpha_t\right) (1 - \epsilon_t) + \exp\left(\alpha_t\right) \epsilon_t
\end{aligned}
$$

Plugging in $\alpha_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right) = \log\left(\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\right)$ we get:

$$
\begin{aligned}
z_t &= \sqrt{\frac{\epsilon_t}{1 - \epsilon_t}} (1 - \epsilon_t) + \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} \epsilon_t \\
&= 2\sqrt{\epsilon_t (1 - \epsilon_t)}
\end{aligned}
$$

Hence using this in Equation 1 we conclude that:

$$\begin{aligned}
\mathrm{err}_D(h_{\mathrm{Boost}}) &\le \prod_{t=1}^{T} Z_t \\
&= 2\prod_{t=1}^{T}\sqrt{\epsilon_t(1-\epsilon_t)} \\
&= 2\prod_{t=1}^{T}\sqrt{\left(\frac{1}{2}-\gamma_t\right)\left(\frac{1}{2}+\gamma_t\right)} \\
&= 2\prod_{t=1}^{T}\sqrt{\frac{1}{4}-\gamma_t^2} \\
&= \prod_{t=1}^{T}\sqrt{1-4\gamma_t^2} \\
&< \prod_{t=1}^{T}\sqrt{1-4\gamma^2} \\
&= (1-4\gamma^2)^{T/2} \\
&\le \exp\left(-4\gamma^2\right)^{T/2} \\
&= \exp\left(-2\gamma^2 T\right)
\end{aligned}$$

Thus the error decreases exponentially with number of iterations T. Now note that if $T = \frac{\log(n)}{2\gamma^2}$ then

$$\mathrm{err}_D(h_{\mathrm{Boost}}) < \frac{1}{n}$$

But we are dealing with zero-one loss and so if average loss over n points is smaller than $1/n$ then it can only be the case that $\mathrm{err}_D(h_{\mathrm{Boost}}) = 0$. Thus we have our theorem. □