1 Boosting recap

Recall that the boosting algorithm looks like this

**Algorithm 1** AdaBoost

1: for $t$ in 1, ..., $T$ (to create ensemble with $T$ classifiers) do
2:   if $t = 1$ then
3:     Initialize weights to $D_{1,i} = 1/N$
4:   else
5:     Set weights $D_{t,i} \propto D_{t-1,i} \exp(-\alpha_{t-1} y_i h_{t-1}(x_i))$
6:   end if
7: Train hypothesis $h_t$ by minimizing error $\mathcal{D}$ weighted by $D_t$
8: Evaluate weighted error $\epsilon_t = \sum_{i=1}^{N} D_{t,i} \delta(h_t(x_i) \neq y_i)$
9: Put a weight $\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)$ on $h_t$
10: end for
11: Final classifier is given by $H(x) = \text{sign} \left( \sum_{t=1}^{T} \alpha_t h_t(x) \right)$

There are two decisions for AdaBoost that we need to analyze: the choice of weight update for weights $D_t$ and the choice of classifier weight $\alpha_t$. The weight update is

$$D_{t+1,i} = \frac{D_{t,i} \exp(-\alpha_t y_i h_t(x_i))}{\sum_{i' = 1}^{N} D_{t,i'} \exp(-\alpha_t y_i' h_t(x_i'))} = \frac{1}{Z_t} D_{t,i} \exp(-\alpha_t y_i h_t(x_i)),$$

where we’ve defined $Z_t = \sum_{i' = 1}^{N} D_{t,i'} \exp(-\alpha_t y_i' h_t(x_i'))$. The choice of $\alpha_t$ is:

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)$$

2 Boosting: formal result

We can show that this choice of $\alpha_t$ actually minimizes the error of the final ensemble classifier on the training set. To start, note that we can bound the total number of errors on the dataset made by the final classifier $H(x)$, if we let
\[ f(x) = \sum_{t=1}^{T} \alpha_t h_t(x), \text{ such that } H(x) = \text{sign}(f(x)) : \]

\[ \frac{1}{N} \sum_{i=1}^{N} \delta(y^i \neq H(x^i)) \leq \frac{1}{N} \sum_{i=1}^{N} \exp(-y^i f(x^i)). \]

This interesting relationship follows from the fact that when \( y^i f(x^i) > 0 \) and the point is classified correctly, \( \exp(-y^i f(x^i)) > 0 \) and \( \delta(y^i \neq H(x^i)) = 0 \). So for all correctly classified points, the right hand side is larger than the left side. For all incorrectly classified points, we have \( \exp(-y^i f(x^i)) > 1 \), since the exponential of a positive number is greater than 1, while \( \delta(y^i \neq H(x^i)) = 1 \). So that means that for both the correct and incorrect points, the right hand side is bigger than the left, and the bound holds.

Now we can express the bound in terms of \( \alpha_t \) as follows. First, let’s substitute \( \sum_{t=1}^{T} \alpha_t h_t(x) \) for \( f(x) \) into the right-hand side:

\[ \frac{1}{N} \sum_{i=1}^{N} \exp(-y^i f(x^i)) = \frac{1}{N} \sum_{i=1}^{N} \exp(-y^i \sum_{t=1}^{T} \alpha_t h_t(x)) \]

\[ = \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \exp(-\alpha_t y^i h_t(x)) \]

Now recall that the weights at step \( T+1 \) would be given by

\[ D_{T+1,i} = \frac{1}{Z_T} D_{T,i} \exp(-\alpha_T y^i h_T(x^i)) \]

That means that we can rearrange the terms to get

\[ Z_T D_{T+1,i} = D_{T,i} \exp(-\alpha_T y^i h_T(x^i)) \]

We can substitute exactly the same thing for \( D_{T,i} \) to get

\[ Z_T D_{T+1,i} = \frac{1}{Z_{T-1}} D_{T-1,i} \exp(-\alpha_{T-1} y^i h_{T-1}(x^i)) \exp(-\alpha_T y^i h_T(x^i)) \]

and then again push \( Z_{T-1} \) to the left side to get

\[ Z_{T-1} Z_T D_{T+1,i} = D_{T-1,i} \exp(-\alpha_{T-1} y^i h_{T-1}(x^i)) \exp(-\alpha_T y^i h_T(x^i)) \]

We can keep doing this for all \( t \) to get

\[ \left[ \prod_{t=1}^{T} Z_t \right] D_{T+1,i} = \frac{1}{N} \prod_{t=1}^{T} \exp(-\alpha_t y^i h_t(x^i)) \]

Since this holds for all \( i \), we can sum both sides over \( i \) to get

\[ \left[ \prod_{t=1}^{T} Z_t \right] \sum_{i=1}^{N} D_{T+1,i} = \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \exp(-\alpha_t y^i h_t(x^i)) \]
Note however that $D_{T+1,i}$ is normalized, so the sum on the left side is just one, which gives us

$$ \prod_{t=1}^{T} Z_t = \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \exp(-\alpha_t y^t h_t(x^i)) $$

Substituting this into our bound above, we have

$$ \frac{1}{N} \sum_{i=1}^{N} \delta(y^i \neq H(x^i)) \leq \frac{1}{N} \sum_{i=1}^{N} \exp(-y^i f(x^i)) = \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \exp(-\alpha_t y^i h_t(x^i)) = \prod_{t=1}^{T} Z_t $$

This is interesting, because it shows that we can minimize our overall training error simply by minimizing the product $\prod_{t=1}^{T} Z_t$. At iteration $t$ of boosting, our choice of $\alpha_t$ only affects $Z_t$, so we simply need to minimize $Z_t$. I won’t go through this derivation in the lecture, but we can in fact show that if we set

$$ \alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon}{\epsilon} \right), $$

we can minimize $Z_t$. This requires taking the derivative of $Z_t$ and setting it to zero. To get some intuition for this, note that

$$ Z_t = \sum_{i=1}^{T} D_{t,i} \exp(-\alpha_t y^i h_t(x^i)) $$

and for some $i$, $y^i h_t(x^i)$ is positive, while for others, it’s negative. So we have to choose $\alpha_t$ so as to balance correct and incorrect classifications. Intuitively, if all samples are correct, then we simply set $\alpha_t$ to $\infty$ to minimize $Z_t$. Incidentally, the choice of training the classifier $h_t$ to minimize error on the weighted dataset can also be shown to minimize $Z_t$ and therefore the bound on the training error.