Boosting

[60 points]

Suppose we simplify AdaBoost by setting the parameter $\alpha_t$ to a fix value $\alpha_t = \alpha > 0$, independent of the boosting round $t$.

1. [20 points] Let $\gamma$ be such that $(\frac{1}{2} - \epsilon_t) \geq \gamma > 0$ where $\epsilon_t$ is defined as in class. Find the best value of $\alpha$ as a function of $\gamma$ by analyzing the empirical error.

As in class, we can show that

$$\hat{\text{error}}(H) \leq \prod_{t=1}^{T} Z_t,$$

and that

$$Z_t = (1 - \epsilon_t)e^{-\alpha} + \epsilon_t e^\alpha.$$

By definition of $\gamma$ and the fact that $e^\alpha - e^{-\alpha} > 0$ for all $\alpha > 0$,

$$Z_t = \epsilon_t(e^\alpha - e^{-\alpha}) + e^{-\alpha} \leq (1 - \gamma)(e^\alpha - e^{-\alpha}) + e^{-\alpha} = (\frac{1}{2} - \gamma)e^\alpha + (\frac{1}{2} + \gamma)e^{-\alpha} = u(\alpha).$$

$u(\alpha)$ is minimized for

$$\frac{1}{2} - \gamma)e^\alpha = (\frac{1}{2} + \gamma)e^{-\alpha},$$

that is for

$$\alpha = \frac{1}{2} \log \frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma}.$$
Tighter bounds on the product of the $Z_t$s can lead to better values for $\alpha$.

2. [20 points] For that value of $\alpha$, does the algorithm assign the same probability mass to correctly classified and misclassified examples at each round? Which set is assigned a higher probability mass?

As in the proof given in class, at round $t$, the probability mass assigned to correctly classified points is $p_+ = (1 - \epsilon_t) e^{-\alpha}$ and the probability mass assigned to the misclassified points is $p_- = \epsilon_t e^\alpha$. Thus,

$$\frac{p_-}{p_+} = \frac{\epsilon_t}{1 - \epsilon_t} \leq \frac{\frac{1}{2} - \gamma}{\frac{1}{2} + \gamma} = 1.$$  \hspace{1cm} (8)

This contrasts with AdaBoost’s property.

3. [20 points] Using the previous value of $\alpha$, give a bound on the empirical error of the algorithm that depends only on $\gamma$ and the number of rounds of boosting $T$.

$$Z_t \leq \left( \frac{1}{2} - \gamma \right) e^\alpha + \left( \frac{1}{2} + \gamma \right) e^{-\alpha}$$  \hspace{1cm} (9)

$$= \left( \frac{1}{2} - \gamma \right) \sqrt{\frac{1}{2} + \gamma} + \left( \frac{1}{2} + \gamma \right) \sqrt{\frac{1}{2} - \gamma}$$  \hspace{1cm} (10)

$$= 2 \sqrt{\left( \frac{1}{2} + \gamma \right) \left( \frac{1}{2} - \gamma \right)}. \hspace{1cm} (11)$$

Thus, the empirical error can be bounded as follows:

$$\text{error}(H) \leq \prod_{t=1}^T Z_t$$  \hspace{1cm} (12)

$$\leq \left[ 2 \sqrt{\left( \frac{1}{2} + \gamma \right) \left( \frac{1}{2} - \gamma \right)} \right]^T$$  \hspace{1cm} (13)

$$= (1 - 4\gamma^2)^{T/2}$$  \hspace{1cm} (14)

$$\leq e^{-2\gamma^2 T}. \hspace{1cm} (15)$$

4. [20 points] Using the previous bound, show that for $T > \frac{\log m}{2\gamma^2}$, the resulting hypothesis is consistent with the sample of size $m$. 

2
If \( \hat{\text{error}}(H) = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i f(x_i) \leq 0} \leq \frac{1}{m} \), then clearly \( \hat{\text{error}}(H) = 0 \).

Using the bound obtained in the previous question, if \( e^{-2\gamma^2 T} < \frac{1}{m} \), the empirical error is zero. This can be rewritten as

\[
T > \frac{\log m}{2\gamma^2}. \tag{16}
\]

5. [20 points] Let \( s \) be the VC dimension of the base learners used. Give a bound on the generalization error of the consistent hypothesis obtained after \( T = \left\lfloor \frac{\log m}{2\gamma^2} \right\rfloor + 1 \) rounds of boosting (\textit{hint}: you can use the fact that the VC dimension of the family of functions \( \{ \text{sgn}(\sum_{t=1}^{T} \alpha_t h_t) : \alpha_t \in \mathbb{R} \} \) is bounded by \( 2(s + 1)T \log_2(eT) \)). Suppose now that \( \gamma \) varies with \( m \). Based on the bound derived, what can you say if \( \gamma(m) = O(\sqrt{\frac{\log m}{m}}) \)?

Using the bound proved in class for the consistent case,

\[
\Pr[\text{error}_D(H) > \epsilon] \leq 2\Pi_C(2m)2^{-\frac{m}{d}} \leq 2\left(\frac{2e^m}{d}\right)^{d}2^{-\frac{m}{d}}. \tag{17}
\]

Setting the right-hand side to \( \delta \), with probability at least \( 1 - \delta \), the following bound holds for that consistent hypothesis:

\[
\text{error}_D(H) \leq \frac{2}{m} (d \log_2 \frac{2e^m}{d} + \log_2 \frac{2}{\delta} ), \tag{18}
\]

with \( d = 2(s + 1)T \log_2(eT) \) and \( T = \left\lfloor \frac{\log m}{2\gamma^2} \right\rfloor + 1 \).

The bound is vacuous for \( \gamma(m) = O(\sqrt{\frac{\log m}{m}}) \). This could suggest overfitting.