Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 4 – Solution
April 27, 2007

This is a shorter assignment to leave you more time to work on your project.

Boosting

[60 points]

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

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

   As in class, we can show that
   $$\widehat{error}(H) \leq \prod_{t=1}^{T} Z_t, \tag{1}$$
   and that
   $$Z_t = (1 - \epsilon_t)e^{-\alpha} + \epsilon_t e^{\alpha}. \tag{2}$$
   By definition of γ and the fact that $e^{\alpha} - e^{-\alpha} > 0$ for all $\alpha > 0$,
   $$\begin{aligned} Z_t &= \epsilon_t(e^{\alpha} - e^{-\alpha}) + e^{-\alpha} && (3)\\ &\leq (1 - \gamma)(e^{\alpha} - e^{-\alpha}) + e^{-\alpha} && (4)\\ &= (\frac{1}{2} - \gamma)e^{\alpha} + (\frac{1}{2} + \gamma)e^{-\alpha} = u(\alpha). && (5)\end{aligned}$$
   $u(\alpha)$ is minimized for
   $$(\frac{1}{2} - \gamma)e^{\alpha} = (\frac{1}{2} + \gamma)e^{-\alpha}, \tag{6}$$
   that is for
   $$\alpha = \frac{1}{2} \log \frac{\frac{1}{2} + \gamma}{\frac{1}{2} - \gamma}. \tag{7}$$

Tighter bounds on the product of the Z_ts can lead to better values for α.

2. [20 points] For that value of α, 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,

$$\frac{p_-}{p_+} = \frac{\epsilon_t}{1-\epsilon_t} \frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma} \leq \frac{\frac{1}{2}-\gamma}{\frac{1}{2}+\gamma} \frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma} = 1. \tag{8}$$

   This contrasts with AdaBoost's property.

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

$$\begin{aligned} Z_t &\leq (\frac{1}{2}-\gamma)e^{\alpha} + (\frac{1}{2}+\gamma)e^{-\alpha} && (9) \\ &= (\frac{1}{2}-\gamma)\sqrt{\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}} + (\frac{1}{2}+\gamma)\sqrt{\frac{\frac{1}{2}-\gamma}{\frac{1}{2}+\gamma}} && (10) \\ &= 2\sqrt{(\frac{1}{2}+\gamma)(\frac{1}{2}-\gamma)}. && (11) \end{aligned}$$

   Thus, the empirical error can be bounded as follows:

$$\begin{aligned} \widehat{error}(H) &\leq \prod_{t=1}^{T} Z_t && (12) \\ &\leq [2\sqrt{(\frac{1}{2}+\gamma)(\frac{1}{2}-\gamma)}]^T && (13) \\ &= (1-4\gamma^2)^{T/2} && (14) \\ &\leq e^{-2\gamma^2 T}. && (15) \end{aligned}$$

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.

If $\widehat{error}(H) = \frac{1}{m}\sum_{i=1}^{m} 1_{y_i f(x_i) \leq 0} \leq \frac{1}{m}$, then clearly $\widehat{error}(H) = 0$. 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 (*hint*: you can use the fact that the VC dimension of the family of functions $\left\{ \operatorname{sgn}(\sum_{t=1}^{T} \alpha_t h_t) : \alpha_t \in \mathbb{R} \right\}$ is bounded by $2(s+1)T\log_2(eT)$). Suppose now that γ 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,

$$\Pr[error_D(H) > \epsilon] \leq 2\Pi_C(2m)2^{-\frac{m\epsilon}{2}} \leq 2(\frac{2em}{d})^d 2^{-\frac{m\epsilon}{2}}. \tag{17}$$

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

$$error_D(H) \leq \frac{2}{m}\big(d\log_2\frac{2em}{d} + \log_2\frac{2}{\delta}\big), \tag{18}$$

   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.