Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 3
October 31, 2016
Due: A. November 11, 2016; B. November 22, 2016

A. Boosting

1. Implement AdaBoost with boosting stumps and apply the algorithm to the `spambase` data set of HW2 with the same training and test sets. Plot the average cross-validation error plus or minus one standard deviation as a function of the number of rounds of boosting T by selecting the value of this parameter out of $\{10, 10^2, \ldots, 10^k\}$ for a suitable value of k, as in HW2. Let T^* be the best value found for the parameter. Plot the error on the training and test set as a function of the number of rounds of boosting for $t \in [1, T^*]$. Compare your results with those obtained using SVMs in HW2.

   Solution:

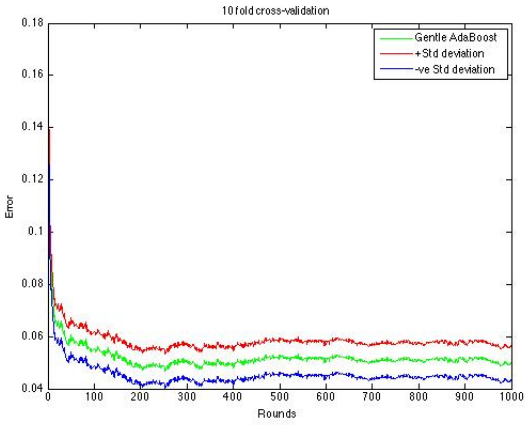

Figure 1: Cross validation error shown with ± 1 standard deviation (Figure courtesy of Chaitanya Rudra).

In figure 1 we see a typical plot of the average cross-validation performance shown with ± 1 standard deviation. Note that the error descreases exponen-

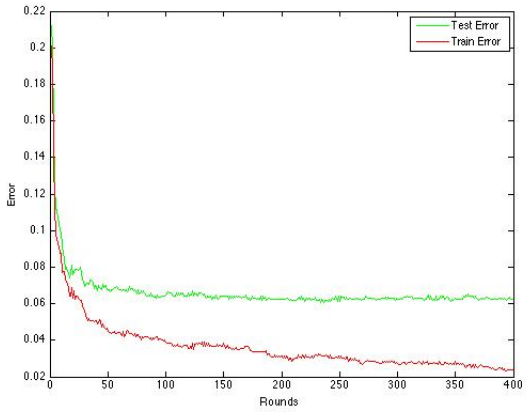

Figure 2: Test and training error (Figure courtesy of Chaitanya Rudra).

tially and eventually levels out after roughly 400 iterations. In figure 2 we see the training and test error after $T^* = 400$ rounds. Note that the test error eventually levels off, while the training error continues to decrease towards zero. Overall, the AdaBoost algorithm performs only slightly better than SVM algorithm increasing accuracy by about 1%.

2. Consider the following variant of the classification problem where, in addition to the positive and negative labels $+1$ and -1, points may be labeled with 0. This can correspond to cases where the true label of a point is unknown, a situation that often arises in practice, or more generally to the fact that the learning algorithm incurs no loss for predicting -1 or $+1$ for such a point. Let $\mathcal{X}$ be the input space and let $\mathcal{Y} = \{-1, 0, +1\}$. As in standard binary classification, the loss of $f \colon \mathcal{X} \to \mathbb{R}$ on a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is defined by $1_{yf(x)<0}$.

   Consider a sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ and a hypothesis set H of base functions taking values in $\{-1, 0, +1\}$. For a base hypothesis $h_t \in H$ and a distribution D_t over indices $i \in [1, m]$, define ϵ_t^s for $s \in \{-1, 0, +1\}$ by $\epsilon_t^s = \mathrm{E}_{i \sim D_t}[1_{y_i h_t(x_i) = s}]$.

   (a) Derive a boosting-style algorithm for this setting in terms of ϵ_t^ss, using the same objective function as that of AdaBoost. You should carefully justify the definition of the algorithm.

   Solution: Say a 'boosting-style algorithm' is just AdaBoost with a pos-

sibly different step size α_t. Recall these definitions from the description of AdaBoost: The final hypothesis is $f(x) = \sum_t \alpha_t h_t(x)$ and the normalization constant in round t is $Z_t = \sum_i D_t(i) \exp(-\alpha_t y_i h_t(x_i))$. We proved in class that

$$\frac{1}{m} \sum_i 1_{y_i f(x_i) < 0} \leq \frac{1}{m} \sum_i \exp(-y_i f(x_i)) = \prod_t Z_t$$

and that AdaBoost's step size can be derived by minimizing this objective in each round t. Taking that same approach, observe that

$$Z_t = \sum_i D_t(i) \exp(-\alpha_t y_i h_t(x_i)) = \epsilon_t^0 + \epsilon_t^- \exp(\alpha_t) + \epsilon_t^+ \exp(-\alpha_t).$$

Differentiating the right-hand side with respect to α_t and setting equal to zero shows that Z_t is minimized by letting $\alpha_t = \frac{1}{2} \log \left(\frac{\epsilon_t^+}{\epsilon_t^-} \right)$.

(b) What is the weak-learning assumption in this setting?

Solution: One possible assumption is $\frac{\epsilon_t^+ - \epsilon_t^-}{\sqrt{1-\epsilon_t^0}} \geq \gamma > 0$. Informally, this assumption says that the difference between the accuracy and error of each weak hypothesis is non-negligible relative to the fraction of examples on which the hypothesis makes any prediction at all. In part (d) we will prove that this assumption suffices to drive the training error to zero.

(c) Write the full pseudocode of the algorithm.

Solution:

1. Given: Training examples $((x_1, y_1), \ldots, (x_m, y_m))$.
2. Initialize D_1 to the uniform distribution on training examples.
3. for $t = 1, \ldots, T$:
   a. $h_t \leftarrow$ base classifier in H.
   b. $\alpha_t \leftarrow \frac{1}{2} \log \left(\frac{\epsilon_t^+}{\epsilon_t^-} \right)$.
   c. For each $i = 1, \ldots, m$: $D_{t+1}(i) \leftarrow \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$, where $Z_t \leftarrow \sum_i D_t(i) \exp(-\alpha_t y_i h_t(x_i))$ is the normalization constant.

i. $f \leftarrow \sum_{t=1}^T \alpha_t h_t$.

4. Return: $\mathrm{sign}(f)$.

(d) Give an upper bound on the training error of the algorithm as a function of the number of rounds of boosting and ϵ_t^ss.

Solution: Plug in the value of α_t from part (a) into $Z_t = \epsilon_t^0 + \epsilon_t^- \exp(\alpha_t) + \epsilon_t^+ \exp(-\alpha_t)$ to obtain $Z_t = \epsilon_t^0 + 2\sqrt{\epsilon_t^- \epsilon_t^+}$. Therefore

$$\frac{1}{m}\sum_i 1_{y_i f(x_i)<0} \leq \prod_t Z_t = \prod_t \left(\epsilon_t^0 + 2\sqrt{\epsilon_t^- \epsilon_t^+}\right).$$

Morever, if the weak learning assumption from part (b) is satisfied then

$$\begin{aligned}
\epsilon_t^0 + 2\sqrt{\epsilon_t^- \epsilon_t^+} &= \epsilon_t^0 + \sqrt{(1-\epsilon_t^0)^2 - (\epsilon_t^+ - \epsilon_t^-)^2} \\
&= \epsilon_t^0 + (1-\epsilon_t^0)\sqrt{1 - \frac{(\epsilon_t^+ - \epsilon_t^-)^2}{(1-\epsilon_t^0)^2}} \\
&\leq \sqrt{1 - \frac{(\epsilon_t^+ - \epsilon_t^-)^2}{1-\epsilon_t^0}} \\
&\leq \sqrt{1-\gamma^2}.
\end{aligned}$$

The first equality follows from $(\epsilon_t^+ + \epsilon_t^-)^2 - (\epsilon_t^+ - \epsilon_t^-)^2 = 4\epsilon_t^+\epsilon_t^-$ (just multiply and gather terms) and $\epsilon_t^+ + \epsilon_t^- = 1 - \epsilon_t^0$. The first inequality follows from the fact that square root is concave on $[0, \infty)$, and thus $\lambda\sqrt{x} + (1-\lambda)\sqrt{y} \leq \sqrt{\lambda x + (1-\lambda)y}$ for $\lambda \in [0,1]$. The last inequality follows from the weak learning assumption.

Therefore we have $\frac{1}{m}\sum_i 1_{y_i f(x_i)<0} \leq \left(\sqrt{1-\gamma^2}\right)^T \leq \exp\left(-\frac{\gamma^2 T}{2}\right)$, where we used $1 + x \leq \exp(x)$.

B. On-line learning

The objective of this problem is to show how another regret minimization algorithm can be defined and studied. Let L be a loss function convex in its first argument and taking values in $[0, M]$.

We will adopt the notation used in the lectures and assume $N > e^2$. Additionally, for any expert $i \in [1, N]$, we denote by $r_{t,i}$ the instantaneous regret of that expert at time $t \in [1, T]$, $r_{t,i} = L(\widehat{y}_t, y_t) - L(y_{t,i}, y_t)$, and by $R_{t,i}$ his cumulative regret up to time t: $R_{t,i} = \sum_{s=1}^t r_{t,i}$. For convenience, we also define $R_{0,i} = 0$ for all $i \in [1, N]$. For any $x \in \mathbb{R}$, $(x)_+$ denotes $\max(x, 0)$, that is the positive part of x, and for $\mathbf{x} = (x_1, \ldots, x_N)^\top \in \mathbb{R}^N$, $(\mathbf{x})_+ = ((x_1)_+, \ldots, (x_N)_+)^\top$.

Let $\alpha > 2$ and consider the algorithm that predicts at round $t \in [1, T]$ according to $\widehat{y}_t = \frac{\sum_{i=1}^n w_{t,i} y_{t,i}}{\sum_{i=1}^n w_{t,i}}$, with the weight $w_{t,i}$ defined based on the αth power of the regret up to time $(t-1)$: $w_{t,i} = (R_{t-1,i})_+^{\alpha-1}$. The potential function we use to analyze the algorithm is based on the function Φ defined over $\mathbb{R}^N$ by $\Phi\colon \mathbf{x} \mapsto \|(\mathbf{x})_+\|_\alpha^2 = \big[\sum_{i=1}^N (x_i)_+^\alpha\big]^{\frac{2}{\alpha}}$.

1. Show that Φ is twice differentiable over $\mathbb{R}^N - B$, where B is defined as follows:
$$B = \{\mathbf{u} \in \mathbb{R}^N \colon (\mathbf{u})_+ = 0\}.$$
Solution: For $\mathbf{x} \notin B$, we can write $\Phi(x) = \phi_2(\sum_{i=1}^N \phi_1(x_i))$, where $x_1, \ldots, x_N$ denote the components of $\mathbf{x}$ and $\phi_1 \colon \mathbb{R} \to \mathbb{R}$ and $\phi_2 \colon (0, \infty_+) \to \mathbb{R}$ are the functions defined by
$$\forall u \in \mathbb{R}, \phi_1(u) = (u)_+^\alpha \tag{1}$$
$$\forall u \in (0, \infty_+), \phi_2(u) = u^{\frac{2}{\alpha}}. \tag{2}$$
It is not hard to see that both ϕ_1 and ϕ_2 are twice (continously) differentiable, which shows, by composition, the same property for Φ over $\mathbb{R}^N - B$. □

2. For any $t \in [1, T]$, let $\mathbf{r}_t$ denote the vector of instantaneous regrets, $\mathbf{r}_t = (r_{t,1}, \ldots, r_{t,N})^\top$, and similarly $\mathbf{R}_t = (R_{t,1}, \ldots, R_{t,N})^\top$. We define the potential function as $\Phi(\mathbf{R}_t) = \|(\mathbf{R}_t)_+\|_\alpha^2$. Compute $\nabla\Phi(\mathbf{R}_{t-1})$ for $\mathbf{R}_{t-1} \notin B$ and show that $\nabla\Phi(\mathbf{R}_{t-1}) \cdot \mathbf{r}_t \leq 0$ (*hint*: use the convexity of the loss with respect to the first argument).

Solution: For any $i \in [1, N]$,
$$\frac{\partial\Phi(\mathbf{R}_{t-1})}{\partial R_{t-1,i}} = \frac{2}{\alpha}\alpha(R_{t-1,i})_+^{\alpha-1}\big[\sum_{i=1}^N (R_{t-1,i})_+^\alpha\big]^{\frac{2-\alpha}{\alpha}} = 2w_{t,i}\|(\mathbf{R}_{t-1})_+\|_\alpha^{2-\alpha}. \tag{3}$$
Thus, we can write
$$\begin{aligned}
\operatorname{sign}(\nabla\Phi(\mathbf{R}_t) \cdot \mathbf{r}_t) &= \operatorname{sign}\big(\sum_{i=1}^N w_{t,i} r_{t,i}\big) \\
&= \operatorname{sign}\big(\sum_{i=1}^N w_{t,i}(L(\widehat{y}_t, y_t) - L(y_{t,i}, y_t)\big) \\
&= \operatorname{sign}\Big(\sum_{i=1}^N \frac{w_{t,i}}{\sum_{i=1}^N w_{t,i}}(L(\widehat{y}_t, y_t) - L(y_{t,i}, y_t)\Big).
\end{aligned}$$

Now, by the convexity of the first argument of L, the following holds:

$$\sum_{i=1}^{N} \frac{w_{t,i}}{\sum_{i=1}^{N} w_{t,i}} (L(\widehat{y}_t, y_t) - L(y_{t,i}, y_t)) \\ = L(\widehat{y}_t, y_t) - \sum_{i=1}^{N} \frac{w_{t,i}}{\sum_{i=1}^{N} w_{t,i}} L(y_{t,i}, y_t) \leq 0,$$

since $\widehat{y}_t = \frac{\sum_{i=1}^{n} w_{t,i} y_{t,i}}{\sum_{i=1}^{n} w_{t,i}}$. □

3. (Bonus question) Prove the inequality $\mathbf{r}^\top [\nabla^2 \Phi(\mathbf{u})] \mathbf{r} \leq 2(\alpha - 1) \|\mathbf{r}\|_\alpha^2$ valid for all $\mathbf{r} \in \mathbb{R}^N$ and $\mathbf{u} \in \mathbb{R}^N - B$ (*hint*: write the Hessian $\nabla^2 \Phi(\mathbf{u})$ as a sum of a diagonal matrix and a positive semi-definite matrix multiplied by $(2 - \alpha)$. Also, use Hölder's inequality generalizing Cauchy-Schwarz: for any $p > 1$ and $q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, $|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\|_p \|\mathbf{v}\|_q$).

Solution: As already seen, for any $i \in [1, N]$, we have

$$\frac{\partial \Phi(\mathbf{u})}{\partial u_i} = 2(u_i)_+^{\alpha - 1} \|(\mathbf{u})_+\|_\alpha^{2-\alpha}. \tag{4}$$

For $j \neq i$, we obtain:

$$\frac{\partial^2 \Phi(\mathbf{u})}{\partial u_j \partial u_i} = 2(2 - \alpha)(u_j)_+^{\alpha - 1} (u_i)_+^{\alpha - 1} \|(\mathbf{u})_+\|_\alpha^{2(1-\alpha)}. \tag{5}$$

We also have

$$\frac{\partial^2 \Phi(\mathbf{u})}{\partial u_i^2} = 2(\alpha - 1)(u_i)_+^{\alpha - 2} \|(\mathbf{u})_+\|_\alpha^{2-\alpha} + 2(2 - \alpha)(u_i)_+^{2(\alpha-1)} \|(\mathbf{u})_+\|_\alpha^{2(1-\alpha)}. \tag{6}$$

Consider the diagonal matrix $\mathbf{D} = \operatorname{diag}((u_1)_+^{\alpha-2}, \ldots, (u_N)_+^{\alpha-2})$ and the matrix $\mathbf{M}$ defined by $\mathbf{M}_{ij} = (u_j)_+^{\alpha-1} (u_i)_+^{\alpha-1}$. In view of the previous expressions, we can write

$$\nabla^2 \Phi(\mathbf{u}) = 2(\alpha - 1) \|(\mathbf{u})_+\|_\alpha^{2-\alpha} \mathbf{D} + 2(2 - \alpha) \|(\mathbf{u})_+\|_\alpha^{2(1-\alpha)} \mathbf{M}. \tag{7}$$

$\mathbf{M}$ is a positive semi-definite matrix since $\mathbf{M} = \mathbf{v}\mathbf{v}^\top$ with $\mathbf{v} = (u_1^{\alpha-1}, \ldots, u_N^{\alpha-1})^\top$. Thus, for any $\mathbf{r}$, $\mathbf{r}^\top \mathbf{M} \mathbf{r} \geq 0$. Since $2 - \alpha < 0$, we can write

$$\mathbf{r}^\top \nabla^2 \Phi(\mathbf{u}) \mathbf{r} \leq 2(\alpha - 1) \|(\mathbf{u})_+\|_\alpha^{2-\alpha} \mathbf{r}^\top \mathbf{D} \mathbf{r} \\ = 2(\alpha - 1) \|(\mathbf{u})_+\|_\alpha^{2-\alpha} \sum_{i=1}^{N} (u_i)_+^{\alpha-2} r_{t,i}^2.$$

By Hölder's inequality, the following holds:

$$\sum_{i=1}^{N}(u_i)_+^{\alpha-2}\, r_{t,i}^2 \le \Big[\sum_{i=1}^{N}((u_i)_+^{\alpha-2})^{\frac{\alpha}{\alpha-2}}\Big]^{\frac{\alpha-2}{\alpha}} \Big[\sum_{i=1}^{N}(r_i^2)^{\frac{\alpha}{2}}\Big]^{\frac{2}{\alpha}} = \|(\mathbf{u})_+\|_\alpha^{\alpha-2}\|\mathbf{r}\|_\alpha^2.$$

This implies that $\mathbf{r}^\top\nabla^2\Phi(\mathbf{u})\mathbf{r} \le 2(\alpha-1)\|\mathbf{r}\|_\alpha^2$. □

4. Using the answers to the two previous questions and Taylor's formula, show that for all $t \ge 1$, $\Phi(\mathbf{R}_t)-\Phi(\mathbf{R}_{t-1}) \le (\alpha-1)\|\mathbf{r}_t\|_\alpha^2$, if $\gamma\mathbf{R}_{t-1}+(1-\gamma)\mathbf{R}_t \notin B$ for all $\gamma \in [0,1]$.

   Solution: By assumption, the segment $[\mathbf{R}_{t-1}, \mathbf{R}_t]$ does not meet B and thus Φ is twice differentiable over the interior $(\mathbf{R}_{t-1}, \mathbf{R}_t)$. Thus, by Taylor's formula, there exists $\mathbf{u} \in (\mathbf{R}_{t-1}, \mathbf{R}_t)$ such that

$$\begin{aligned}\Phi(\mathbf{R}_t) - \Phi(\mathbf{R}_{t-1}) = \nabla(\mathbf{R}_{t-1})\cdot\mathbf{r}_t + \frac{1}{2}\mathbf{r}_t^T\nabla^2(\mathbf{u})\mathbf{r}_t &\le \frac{1}{2}\mathbf{r}_t^T\nabla^2(\mathbf{u})\mathbf{r}_t\\ &\le (\alpha-1)\|\mathbf{r}_t\|_\alpha^2,\end{aligned}$$

   where we used the inequalities obtained in the two previous questions. □

5. Suppose there exists $\gamma \in [0,1]$ such that $(1-\gamma)\mathbf{R}_{t-1} + \gamma\mathbf{R}_t \in B$. Show that $\Phi(\mathbf{R}_t) \le (\alpha-1)\|\mathbf{r}_t\|_\alpha^2$.

   Solution: For that γ, by definition, we have $(\mathbf{R}_{t-1} + \gamma\mathbf{r}_t)_+ = \mathbf{0}$. Observe that for any two scalars c and d, $(c+d)_+ \le c_+ + d_+$ and therefore that

$$(\mathbf{R}_{t-1}+\mathbf{r}_t)_+ \le (\mathbf{R}_{t-1}+\gamma\mathbf{r}_t)_+ + (1-\gamma)(\mathbf{r}_t)_+ = (1-\gamma)(\mathbf{r}_t)_+ \le (\mathbf{r}_t)_+. \quad (8)$$

   $\mathbf{u} \mapsto \|\mathbf{u}\|_\alpha^2$ is a non-decreasing function of each component u_i, thus $\Phi(\mathbf{R}_t) = \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t) \le \|(\mathbf{r}_t)_+\|_\alpha^2$. Since $\alpha > 2$, this implies $\Phi(\mathbf{R}_t) \le (\alpha-1)\|(\mathbf{r}_t)_+\|_\alpha^2$. □

6. Using the two previous questions, derive an upper bound on $\Phi(\mathbf{R}_T)$ expressed in terms of T, N, and M.

   Solution: In view of the two previous questions, the inequality $\Phi(\mathbf{R}_t) - \Phi(\mathbf{R}_{t-1}) \le (\alpha-1)\|\mathbf{r}_t\|_\alpha^2$ holds in all cases. Summing up these inequalities for all $t \in [1,T]$, we obtain

$$\Phi(\mathbf{R}_T) = \Phi(\mathbf{R}_T) - \Phi(\mathbf{R}_0) = \sum_{t=1}^{T}\Phi(\mathbf{R}_t) - \Phi(\mathbf{R}_{t-1}) \le (\alpha-1)\sum_{t=1}^{T}\|\mathbf{r}_t\|_\alpha^2.$$

   Since $|r_{t,i}| \le M$ for all t and i, this yields $\Phi(\mathbf{R}_T) \le (\alpha-1)N^{\frac{2}{\alpha}}M^2T$. □

7. Show that $\Phi(\mathbf{R}_T)$ admits as a lower bound the square of the regret R_T of the algorithm.

   Solution: By definition of the regret,

   $$R_T = \max_{i \in [1,N]} R_{T,i} \leq \|\mathbf{R}_T\|_\alpha = \sqrt{\Phi(\mathbf{R}_T)}.$$

   Note that this implies that $\Phi(\mathbf{R}_T) \geq (R_T)_+^2$, which is not exactly the statement given in the question. However, it suffices for proving upper bounds on the regret, which is the motivation behind our construction of Φ.

8. Using the two previous questions give an upper bound on the regret R_T. For what value of α is the bound the most favorable? Give a simple expression of the upper bound on the regret for a suitable approximation of that optimal value.

   Solution: In view of the inequalities of the two previous questions, we have

   $$R_T \leq \sqrt{(\alpha - 1) N^{\frac{2}{\alpha}} M^2 T}.$$

   For $N > e^2$, the function $\alpha \mapsto (\alpha - 1) N^{\frac{2}{\alpha}}$ reaches its minimum over $(2, +\infty)$ at $\left(\log N + \sqrt{\log^2 N - \log N}\right)$, which is approximately $2 \log N$. Plugging in that value yields the bound

   $$R_T \leq M \sqrt{(2 \log N - 1) e T}.$$