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.

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.

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

   (c) Write the full pseudocode of the algorithm.

   (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.

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 $\Phi\colon \mathbf{x} \mapsto \|(\mathbf{x})_+\|_\alpha^2 = \left[\sum_{i=1}^{N} (x_i)_+^\alpha\right]^{\frac{2}{\alpha}}$.

1. Show that Φ is twice differentiable over $\mathbb{R}^N - B$, where B is defined as follows:
$$B = \{\mathbf{u} \in \mathbb{R}^N \colon (\mathbf{u})_+ = 0\}.$$

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).

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 $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, $|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\|_p \|\mathbf{v}\|_q$).

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

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

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

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

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.