Mehryar Mohri
Foundations of Machine Learning 2015
Courant Institute of Mathematical Sciences
Homework assignment 3
November 24, 2015
Due: December 07, 2015

A. Boosting-type Algorithm

1. Show that for all $u \in \mathbb{R}$ and integer $p > 1$, $1_{u \leq 0} \leq \Phi_p(-u)$ where $\Phi_p(u) = \max((1+u)^p, 0)$. Show that Φ_p is convex and differentiable.

2. Use Φ_p to derive a boosting-type algorithm using coordinate descent. You should give a full description of your algorithm, including the pseudocode, details for the choice of the step and direction, as well as a generalization bound.

B. L_2-Regularized Maxent

This problem studies L_2-regularized Maxent. We will use the notation introduced in class and will denote by J_S the dual objective function minimized given a sample S:

$$J_S(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \mathop{\mathrm{E}}_{x \sim S}\big[-\log \mathsf{p}_{\mathbf{w}}[x]\big],$$

where $\lambda > 0$ is a regularization parameter. We will assume that the feature vector is bounded: $\|\mathbf{\Phi}(x)\|_2 \leq r$ for all $x \in \mathcal{X}$, for some $r > 0$.

1. Use McDiarmid's inequality to prove that for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds:

$$\Big\| \mathop{\mathrm{E}}_{x \sim \mathcal{D}}[\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x \sim S}[\mathbf{\Phi}(x)] \Big\|_2 \leq \sqrt{\frac{2r^2}{m}}\left(1 + \sqrt{\log \frac{1}{\delta}}\right).$$

2. Let $\widehat{\mathbf{w}}$ be the L_2-regularized maxent solution for a sample S and $\mathbf{w}_{\mathcal{D}}$ the solution for an infinite sample:

$$\widehat{\mathbf{w}} = \mathop{\mathrm{argmin}}_{\mathbf{w} \in \mathbb{R}^N} J_S(\mathbf{w}) \quad \text{and} \quad \mathbf{w}_{\mathcal{D}} = \mathop{\mathrm{argmin}}_{\mathbf{w} \in \mathbb{R}^N} J_{\mathcal{D}}(\mathbf{w}).$$

where $J_{\mathcal{D}}(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \mathrm{E}_{x\sim\mathcal{D}}\big[-\log \mathsf{p}_{\mathbf{w}}[x]\big]$. Use the definition of $\widehat{\mathbf{w}}$ and $\mathbf{w}_{\mathcal{D}}$ as minimizers (use gradients) to prove that the following inequality holds:

$$\|\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}\|_2 \leq \frac{\Big\| \mathrm{E}_{x\sim S}[\mathbf{\Phi}(x)] - \mathrm{E}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)]\Big\|_2}{\lambda}.$$

3. For any $\mathbf{w}$ and any distribution $\mathcal{Q}$ define $\mathcal{L}_{\mathcal{Q}}(\mathbf{w})$ by $\mathcal{L}_{\mathcal{Q}}(\mathbf{w}) = \mathrm{E}_{x\sim\mathcal{Q}}[-\log \mathsf{p}_{\mathbf{w}}[x]]$. Show that

$$\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}}) \leq (\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}) \cdot \Big[\mathop{\mathrm{E}}_{x\sim S}[\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)]\Big] + \frac{\lambda}{2}\|\mathbf{w}_{\mathcal{D}}\|_2^2 - \frac{\lambda}{2}\|\widehat{\mathbf{w}}\|_2^2.$$

4. Use that to show that the following inequality holds for any $\mathbf{w}$:

$$\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) \leq \frac{1}{\lambda}\Big\| \mathop{\mathrm{E}}_{x\sim S}[\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)]\Big\|_2^2 + \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2.$$

5. Conclude by proving that for any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds:

$$\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) \leq \inf_{\mathbf{w}\in\mathbb{R}^N} \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{2r^2}{\lambda m}\left(1 + \sqrt{\log\frac{1}{\delta}}\right)^2.$$

C. Randomized Halving

In class, we showed that, in the realizable scenario (at least one expert is always correct), the number of mistakes made by Halving is upper bounded by $\log_2 N$. Here, we consider for the same realizable scenario a randomized version of Having defined as follows.

As for Halving, let H_t denote the set of remaining experts at the beginning of round t, with $H_1 = H$ the full set of N experts. At each round, let r_t be the fraction of experts in H_t predicting 1. Then, the prediction $\widehat{y}_t$ made by the algorithm is 1 with probability

$$p_t = \left[\frac{1}{2}\log_2\frac{1}{1-r_t}\right]1_{r_t\leq\frac{3}{4}} + 1_{r_t>\frac{3}{4}},$$

0 with probability $1-p_t$. The true label y_t is then received and H_{t+1} is derived from H_t by removing all experts who made a mistake.

1. Write the pseudocode of the algorithm.

2. Define the potential function $\Phi_t = \log_2 |H_t|$. Let $\mu_t = 1_{y_t \neq \widehat{y}_t}$, prove that for all $t \geq 1$, $\mathrm{E}[\mu_t] \leq \frac{\Phi_t - \Phi_{t+1}}{2}$.

3. Show that the expected number of mistakes made by randomized Halving is at most $\frac{1}{2} \log_2 N$.

4. (Bonus question) Prove that no randomized algorithm makes fewer than $\lfloor \frac{1}{2} \log_2 N \rfloor$ mistakes, in expectation.