CSCI 5525: Machine Learning (Fall'15)

Homework 4, Due 12/04/15

1. **(35 points)** This problem considers developing an ADMM (Alternating Direction Method of Multipliers) for sparse 2-class logistic regression. Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ be the dataset under consideration, where $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{0, 1\}$. The objective function to be minimized is given by:

$$f(\mathbf{w}) = h(\mathbf{w}) + \lambda g(\mathbf{w}) = -\sum_{i=1}^{N} \left\{ y_i \mathbf{w}^T \mathbf{x}_i - \log(1 + \exp(\mathbf{w}^T \mathbf{x}_i)) \right\} + \lambda \|\mathbf{w}\|_2^2 \ ,$$

where $\lambda > 0$ is a positive constant, $\|\mathbf{w}\|_2^2 = \sum_{j=1}^{d} w_j^2$, $h(w)$ refers to the first term and $g(w)$ refers to the second term.

   (a) (20 points) Assume the dataset of size N has been grouped into m mini-batches of n points each, i.e., $N = mn$. Clearly outline the key steps of the ADMM algorithm for solving the problem. Note that you will have to introduce a separate w_h for each mini-batch $h = 1, \ldots, m$, and provide key steps for updating each primal and dual parameters.

   (b) (10 points) Which steps can be executed in parallel? Which steps require access to more than one mini-batch? Clearly explain your answers.

   (c) (5 points) Is the resulting algorithm a double-loop algorithm, i.e., an outer loop executing the ADMM updates, and an inner iterative loop solving one/more of the key steps of the ADMM algorithm? Clearly explain your answer.

2. **(35 points)** This problem considers two different approaches, respectively based on VC dimensions and Rademacher complexity, for measuring complexity of a function class $\mathcal{F} = \{f\}$ where each $f : \mathbb{R}^d \mapsto \{-1, +1\}$.

   (a) (20 points) Define the VC dimension $VC(\mathcal{F})$ of a function class $\mathcal{F}$. Show that the VC dimension of hyper-plane classifiers in $\mathbb{R}^d$, $\mathcal{F} = \{f : f(x) = \text{sign}(\mathbf{w}^T x + w_0), \mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}\}$ is $VC(\mathcal{F}) = d + 1$.

   (b) (15 points) Define the Rademacher complexity $R_n(\mathcal{F})$ of a function class $\mathcal{F}$ for n-samples. Show that for any two sets of independent samples x, x' of size n

$$E\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \rho_i (f(x_i') - f(x_i))\right] \leq 2R_n(\mathcal{F}) \ ,$$

   where the expectation is over $\mathbf{x} = \{x_1, \ldots, x_n\}, x' = \{x_1', \ldots, x_n'\}$, and the independent Rademacher variables $\{\rho_1, \ldots, \rho_n\}$.

3. **(30 points)** Consider an online learning scenario with n experts where the learner adaptively maintains a distribution $\mathbf{w}_t$ for $t = 1, \ldots, T$, over the experts. At each time-step t, the learner

receives a loss vector $\ell_t \in [0,1]^n$ from the environment, and incurs expected loss $\mathbf{w}_t^T \ell_t$. The probability distribution over experts is then updated as $w_{t+1}(i) = w_t(i)\beta^{\ell_t(i)}/Z_{t+1}$, where Z_{t+1} is the normalization constant and $\beta \in (0,1)$. Let L_i be the cumulative loss incurred by expert i (i.e. $L_i = \sum_{t=1}^T \ell_t(i)$) and let L_{adapt} be the cumulative expected loss incurred by the online learning algorithm where $L_{adapt} = \sum_{t=1}^T \mathbf{w}_t^T \ell_t$.

(a) (20 points) Assuming $\mathbf{w}_1$ is the uniform distribution so that $w_1(i) = 1/n$, show that

$$L_{adapt} \leq \frac{\log(1/\beta)}{1-\beta} \min_i L_i + \frac{1}{1-\beta} \log n \ .$$

For the analysis, you may use the following facts: for $\beta \in (0,1), \ell_t(i) \in [0,1]$, we have $\beta^{\ell_t(i)} \leq 1 - (1-\beta)\ell_t(i)$; and $(1+x) \leq \exp(x)$ for all x.

(b) (10 points) Assuming $\min_i L_i = L_{min}$, what value of β minimizes the overall regret, i.e., $L_{adapt} - L_{\min}$? Clearly explain your answer.