Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 3
Due: May 08, 2009

A. Perceptron algorithm

Let S be a labeled sample of N points in $\mathbb{R}^N$ with

$$x_i = (\underbrace{(-1)^i, \ldots, (-1)^i, (-1)^{i+1}}_{i \text{ first components}}, 0, \ldots, 0) \quad \text{and} \quad y_i = (-1)^{i+1}. \tag{1}$$

- [50 points] Show that the perceptron algorithm makes $\Omega(2^N)$ updates before finding a separating hyperplane, regardless of the order in which it receives the points.

  Let w be the weight vector. Since each update is of the form $w \leftarrow w + y_i x_i$ and since the components of the sample points are integers, the components of w are also integers.

  Let $n_1, \ldots, n_N \in \mathbb{Z}$ denote the components of w. w correctly classifies all points iff $y_i(w \cdot x_i) > 0$ for $i = 1, \ldots, m$, that is

$$\begin{cases} n_1 > 0 \\ n_1 - n_2 < 0 \\ -n_1 - n_2 + n_3 > 0 \\ \ldots \\ (-1)^N(n_1 + n_2 + \ldots + n_{N-1} - n_N) < 0 \end{cases} \Leftrightarrow \begin{cases} n_1 > 0 \\ n_2 > n_1 \\ n_3 > n_1 + n_2 \\ \ldots \\ n_N > n_1 + n_2 + \ldots + n_{N-1}. \end{cases}$$

  These last inequalities show that the data is linearly separable with $w = (1, 2, \ldots, 2^{N-1})$. They also imply that $n_1 \geq 1, n_2 \geq 2, n_3 \geq 4, \ldots, n_N \geq 2^{N-1}$. Since each update can at most increment n_N by 1, the number of updates is at least $2^{N-1} = \Omega(2^N)$.

B. Boosting

This problem considers an algorithm similar to AdaBoost but with a different objective function. Assume that the training data is given as m labeled examples $(x_1, y_1), \ldots, (x_m, y_m) \in X \times \{-1, +1\}$. Let $\Phi \colon \mathbb{R} \to \mathbb{R}$ be the function defined by

$$\Phi(u) = \begin{cases} (1+u)^2 & \text{if } u \geq -1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

[50 points]

- [20 points] Consider the objective function F defined by $F(\alpha) = \sum_{i=1}^{m} \Phi(-y_i f(x_i))$ where f is a linear combination of base classifiers: $f = \sum_{t=1}^{T} \alpha_t h_t$ as for AdaBoost. Show that F is convex and differentiable.

  To show that F is differentiable it suffices to show that Φ is differentiable since F is a sum of functions obtained by composition of Φ with a linear function of α. Φ is continuously differentiable on the interval $]-\infty, -1[$ with $\Phi'(u) = 0$, and is continuously differentiable on the interval $]-1, +\infty[$ with $\Phi'(u) = 2(1+u)$ and $\lim_{x \to^+ -1} \Phi'(u) = \lim_{x \to^- -1} \Phi'(u) = 0$, thus Φ is continuously differentiable over $\mathbb{R}$.

  Since its differential is always non-negative, it is an increasing function. Φ is twice differentiable over $]-1, +\infty[$ with $\Phi''(u) = 2$ and $\Phi''(u) = 0$ on $]-\infty, -1[$ and is continuous at -1, thus Φ is convex. F is thus convex as a sum of function obtained by composing an increasing and convex function with a linear function.

- [30 points] Derive a new boosting algorithm using the objective function F. Characterize the best base classifier h_u to select at each round of boosting if we use coordinate descent.

  - [20 points] Let I denote the set of indices i for which $\Phi'(-y_i f(x_i)) \neq 0$: $I = \{i \in [1, m] \colon \Phi'(-y_i f(x_i)) \neq 0\}$. The direction e_u taken by coordinate descent after $T-1$ rounds is the argmin_u of:

$$\begin{aligned}
\left.\frac{dF(\alpha + \beta e_u)}{d\beta}\right|_{\beta=0} &= -\sum_{i=1}^{m} y_i h_u(x_i) \Phi'(-y_i f(x_i)) \\
&\propto -\sum_{i=1}^{m} y_i h_u(x_i) \frac{\Phi'(-y_i f(x_i))}{\sum_{i \in I} \Phi'(-y_i f(x_i))} \\
&\propto -\sum_{i=1}^{m} y_i h_u(x_i) D_{T-1}(i) \\
&= -(1 - 2\epsilon_u),
\end{aligned}$$

    where $D_{T-1}(i) = \frac{1}{|I|} \frac{\Phi'(-y_i f(x_i))}{\sum_{i \in I} \Phi'(-y_i f(x_i))}$, and $\epsilon_u = \Pr_{D_{T-1}}[h_u(x_i) \neq y_i]$. Thus, the base classifier h_u selected at each round is the one with the minimal ϵ_u.

– [10 points] The step size β is given by:

$$\frac{dF(\alpha+\beta e_u)}{d\beta} = -\sum_{i=1}^{m} y_i h_u(x_i)\Phi'(-y_i f(x_i) - \beta y_i h_u(x_i)) = 0$$
$$\Leftrightarrow \quad \sum_{i\in I'}^{m} (1 - y_i f(x_i) - \beta y_i h_u(x_i)) = 0,$$

where $I' = \{i \in [1,m] \colon \Phi'(-y_i f(x_i) - \beta y_i h_u(x_i)) \neq 0\}$, that is $I' = \{i \in [1,m] \colon y_i f(x_i) + \beta y_i h_u(x_i) < 0\}$.