MATH 547: HOMEWORK 2
DUE ON: FRIDAY, OCTOBER 4.

Please type the solutions in LaTeX, or write **very clearly** if you do it by hand. Lack of clarity in presentation and writing might result in a lower score.

Problem 1, 10 points: distance to a hyperplane:

Let $u \in \mathbb{R}^d$ be a unit vector, $c \in \mathbb{R}$ - a scaler, and

$$H_{u,c} = \{x \in \mathbb{R}^d : \langle u, x\rangle + c = 0\} \text{ - a hyperplane in } \mathbb{R}^d.$$

Show that for any $z \in \mathbb{R}^d$, the distance between z and $H_{u,c}$ equals

$$\text{dist}(z, H_{u,c}) = |\langle u, z\rangle + c|\,.$$

Here, $\text{dist}(z, H_{u,c})$ is the distance between z and its orthogonal projection on $H_{u,c}$.

Problem 2, 20 points: consistency of SVM:

Recall that the generalization error of a binary classifier of the form $g(x) = \text{sign}(f(x))$ can be expressed as

$$L(g) := P(Y \neq g(X)) = \mathbb{E}I\{Yf(X) < 0\}.$$

We showed in class that replacing the binary loss by the exponential loss $\ell(y) = e^{-y}$ and minimizing $\mathbb{E}\ell(Yf(X))$ over all real-valued functions f yields the function f_* such that $\text{sign}(f_*(x)) = \text{sign}(\eta(x))$, so that $\text{sign}(f_*)$ is a Bayes classifier. Show that the same conclusion holds if we use the *hinge loss*

$$\ell(y) = \max(1 - y, 0)$$

instead of the exponential loss. As we will see in class, this problem is related to Support Vector Machines (SVM).

Problem 3, 35 points: more on Boosting:

Recall that on every step t, *AdaBoost* generates certain weights $w_1^{(t+1)}, \ldots, w_n^{(t+1)}$ associated with the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The idea is that "hard to classify" points get bigger weights, thus becoming more "important" while easy to classify data get smaller weights. The goal of this problem is to make this intuition precise and to understand how exactly these weights change.

We showed in class that the weights are updated at each step according to the formula

$$w_j^{(t+1)} = \frac{w_j^{(t)} \exp(-\alpha_t Y_j g_t(X_j))}{Z_t}$$

where $Z_t = \sum_{j=1}^n w_j^{(t)} \exp(-\alpha_t Y_j g_t(X_j))$ is the "normalizing factor,"

$$\alpha_t = \frac{1}{2}\log\left(\frac{1 - e_{n,w^{(t)}}(g_t)}{e_{n,w^{(t)}}(g_t)}\right) \text{ and } e_{n,w^{(t)}}(g_t) = \sum_{j=1}^n w_j^{(t)} I\left\{Y_j \neq g_t(X_j)\right\}$$

is the **weighted empirical risk**.

(a) **(10 pts)** Read the class notes and prove that Z_t can be alternatively expressed as

$$Z_t = 2\sqrt{e_{n,w^{(t)}}(g_t))(1 - e_{n,w^{(t)}}(g_t))}.$$

This was already mentioned in class but you need to fill in the missing details.

(b) **(10 pts)** Simplify the expression for $w_j^{(t+1)}$ using this expression for Z_t (consider two cases separately: $Y_j = g_t(X_j)$ and $Y_j \neq g_t(X_j)$, and show that $w_j^{(t+1)} = w_j^{(t)} \frac{1}{2(1-e_{n,w^{(t)}}(g_t))}$ in the first case and $w_j^{(t+1)} = w_j^{(t)} \frac{1}{2e_{n,w^{(t)}}(g_t)}$ in the second case).

(c) **(5 pts)** Recall that, by the "weak learnability" assumption of AdaBoost, $e_{n,w^{(t)}}(g_t) < 1/2$. Use this fact together with the simplified expression for $w_j^{(t+1)}$ you obtained above to show that

(a) if g_t classifies X_j correctly, then $w_j^{(t+1)} < w_j^{(t)}$,

(b) if g_t classifies X_j incorrectly, then $w_j^{(t+1)} > w_j^{(t)}$.

(d) **(10pts)** Show that the distribution $(w_1^{(t+1)}, \ldots, w_n^{(t+1)})$ is the "hardest" for g_t, in a sense that

$$\sum_{j=1}^{n} w_j^{(t+1)} I\{Y_j \neq g_t(X_j)\} = \frac{1}{2}.$$

Problem 4, 20 points: expressive power of the convex combinations ("majority vote"):

Recall that AdaBoost looks for solutions in the class

$$\mathcal{F} = \left\{ \sum_{j=1}^{k} \alpha_j g_j,\ k \geq 1, \alpha_1, \ldots, \alpha_j \geq 0,\ g_1, \ldots, g_k \in \mathcal{G} \right\}$$

where $\mathcal{G}$ is some "base class" of binary classifiers. How complex, or expressive, can the class of binary classifiers $\{\text{sign}(f),\ f \in \mathcal{F}\}$ associated with $\mathcal{F}$ be? We look at one example in this exercise. Let $\mathcal{G}$ be a class of threshold classifiers, or "decision stumps,"

$$\mathcal{G} = \{g_{\theta,b}(x) = \text{sign}(x-\theta) \cdot b,\ \theta \in \mathbb{R},\ b \in \{-1,+1\}\}.$$

In other words, $g_{\theta,b}(x) = \begin{cases} b, & x \geq \theta, \\ -b, & x < \theta. \end{cases}$

(1) **(10 pts)** Let $-\infty = \theta_0 < \theta_1 < \ldots < \theta_r = \infty$ be a sequence of real numbers, and define

$$f_r(x) = \sum_{j=1}^{r} \alpha_j I\{x \in (\theta_{j-1}, \theta_j]\}$$

where $\alpha_j = (-1)^j$. Draw the graph of a generic function f_r for, say, $r = 4$ (you can pick θ's as you wish).

(2) **(10 pts)** Show that any function f_r can be realized as an element of the class $\mathcal{F}$ (after taking the sign). Specifically, let $h(x) = \text{sign}\left(\sum_{j=1}^{r} w_j \text{sign}(x-\theta_{j-1})\right)$ where $w_1 = -\frac{1}{2}$ and $w_j = (-1)^j$ for $j > 1$, and show that $h(x) = f_r(x)$. First, check it for the function that you drew in part (a), and then consider the general case.