Mehryar Mohri
Introduction to Machine Learning
Courant Institute of Mathematical Sciences
Midterm exam
October 5th, 2011.

A. Perceptron algorithm

In class, we saw that when the training sample S is linearly separable with a maximum margin $\rho > 0$, then the Perceptron algorithm run cyclically over S is guaranteed to converge after at most R^2/ρ^2 updates, where R is the radius of the sphere containing the sample points.

This does not guarantee however that the hyperplane solution of the Perceptron achieves a margin close to ρ. Suppose we modify the Perceptron algorithm to ensure that the margin of the hyperplane solution is at least $\rho/2$ by updating the weight vector not only when the prediction is incorrect but also when the margin $\frac{y_t \mathbf{w}_t \cdot \mathbf{x}_t}{\|\mathbf{w}_t\|}$ on point $\mathbf{x}_t$ is less than $\rho/2$. Figure 1 gives the pseudocode of the resulting algorithm, MPerceptron.

The objective of this problem is to show that the algorithm MPerceptron converges after at most $16R^2/\rho^2$. Let I denote the set of times $t \in [1, T]$ at which the algorithm makes an update and let $M = |I|$ be the total number of updates made.

1. Using an analysis similar to the one given in class for the Perceptron algorithm, show that $M\rho \leq \|\mathbf{w}_{T+1}\|$. Conclude that if $\|\mathbf{w}_{T+1}\| < \frac{4R^2}{\rho}$, then $M < 4R^2/\rho^2$. In what follows, we will assume that $\|\mathbf{w}_{T+1}\| \geq \frac{4R^2}{\rho}$.

2. Show that for any $t \in I$ (including $t = 0$), the following holds:

$$\|\mathbf{w}_{t+1}\|^2 \leq (\|\mathbf{w}_t\| + \rho/2)^2 + R^2.$$

3. Infer from that that for any $t \in I$, we have

$$\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \rho/2 + \frac{R^2}{\|\mathbf{w}_t\| + \|\mathbf{w}_{t+1}\| + \rho/2}.$$

4. Using the previous question, show that for any $t \in I$ such that either $\|\mathbf{w}_t\| \geq \frac{4R^2}{\rho}$ or $\|\mathbf{w}_{t+1}\| \geq \frac{4R^2}{\rho}$, we have

$$\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \frac{3}{4}\rho.$$

MPerceptron()

$\mathbf{w}_1 \leftarrow \mathbf{0}$
for $t \leftarrow 1$ **to** T **do**
    Receive$(\mathbf{x}_t)$
    Receive(y_t)
    if $\big((\mathbf{w}_t = 0)$ **or** $(\frac{y_t \mathbf{w}_t \cdot \mathbf{x}_t}{\|\mathbf{w}_t\|} < \frac{\rho}{2})\big)$ **then**
        $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t \mathbf{x}_t$
    else $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t$
return $\mathbf{w}_{T+1}$

Figure 1: MPerceptron algorithm.

5. Show that $\|\mathbf{w}_1\| \leq R \leq 4R^2/\rho$. Since by assumption we have $\|\mathbf{w}_{T+1}\| \geq \frac{4R^2}{\rho}$, conclude that there must exist a largest time $t_0 \in I$ such that $\|\mathbf{w}_{t_0}\| \leq \frac{4R^2}{\rho}$ and $\|\mathbf{w}_{t_0+1}\| \geq \frac{4R^2}{\rho}$.

6. Show that $\|\mathbf{w}_{T+1}\| \leq \|\mathbf{w}_{t_0}\| + \frac{3}{4}M\rho$. Conclude that $M \leq 16R^2/\rho^2$.

B. Nearest-neighbor algorithm

Consider a learning task where the input space $\mathcal{X}$ is one-dimensional: $\mathcal{X} = \mathbb{R}$. There are $n > 1$ classes, $\mathcal{Y} = \{y_1, \ldots, y_n\}$, all equally probable: $\Pr[y_i] = 1/n$ for all $i \in [1, n]$. Let r be a positive real number with $r < \frac{n-1}{n}$. Let I_0 be the interval

$$I_0 = [0, \eta[\,,$$

where $\eta = \frac{nr}{n-1}$ and, for any $i \in [1, n]$, let I_i be the interval of length $1 - \eta$ defined by

$$I_i = [2i - 1 - 2(i-1)\eta, 2i - (2i-1)\eta[\,.$$

The conditional probability for each class y_i, $i \in [1, n]$, is defined by the following:

$$\begin{aligned}
&\Pr[x \in I_0 \mid y_i] = \eta \\
&\Pr[x \in I_i \mid y_i] = 1 - \eta \\
&\Pr\big[x \notin (I_0 \cup I_i) \,\big|\, y_i\big] = 0.
\end{aligned}$$

1. Show that the Bayes error R^* is equal to r.

2. Suppose we have a training sample S containing at least one point falling in each of the intervals I_i, $i \in [1, n]$. What is the error rate of the nearest-neighbor algorithm trained on S? Justify your answer.