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 $\|\mathbf{w}_{T+1}\| \geq \frac{4R^2}{\rho}$.

   Solution: By assumption, there exists $\mathbf{v} \in \mathbb{R}^N$ such that for all $t \in [1, T]$, $\rho \leq \frac{y_t(\mathbf{v}\cdot\mathbf{x}_t)}{\|\mathbf{v}\|}$, where ρ is the maximum margin achievable on S. Summing up these inequalities gives

$$\begin{aligned} M\rho \leq \frac{\mathbf{v} \cdot \sum_{t \in I} y_t \mathbf{x}_t}{\|\mathbf{v}\|} &\leq \Big\| \sum_{t \in I} y_t \mathbf{x}_t \Big\| && \text{(Cauchy-Schwarz inequality)} \\ &= \Big\| \sum_{t \in I} (\mathbf{w}_{t+1} - \mathbf{w}_t) \Big\| && \text{(definition of updates)} \\ &= \|\mathbf{w}_{T+1}\| && \text{(telescoping sum, } \mathbf{w}_0 = 0). \end{aligned}$$

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

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

Solution: For any $t \in I$, by definition of the update, $\mathbf{w}_{t+1} = \mathbf{w}_t + y_t \mathbf{x}_t$, thus

$$\begin{aligned} \|\mathbf{w}_{t+1}\|^2 &= \|\mathbf{w}_t\|^2 + \|\mathbf{x}_t\|^2 + 2y_t \mathbf{w}_t \cdot \mathbf{x}_t \\ &\leq \|\mathbf{w}_t\|^2 + \|\mathbf{x}_t\|^2 + \|\mathbf{w}_t\|\rho \qquad \text{(def. of update condition)} \\ &\leq \|\mathbf{w}_t\|^2 + R^2 + \|\mathbf{w}_t\|\rho + \rho^2/4 \\ &= (\|\mathbf{w}_t\| + \rho/2)^2 + R^2. \end{aligned}$$

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

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

Solution: In view of the previous result, $\|\mathbf{w}_{t+1}\|^2 - (\|\mathbf{w}_t\| + \rho/2)^2 = R^2$, that is

$$\begin{aligned} &(\|\mathbf{w}_{t+1}\| - \|\mathbf{w}_t\| + \rho/2)(\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2) \leq R^2 \\ \implies &(\|\mathbf{w}_{t+1}\| - \|\mathbf{w}_t\| + \rho/2) \leq \frac{R^2}{\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2} \\ \implies &\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \rho/2 + \frac{R^2}{\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2}. \end{aligned}$$

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

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

Solution: If $\|\mathbf{w}_t\| \geq \frac{4R^2}{\rho}$ or $\|\mathbf{w}_{t+1}\| \geq \frac{4R^2}{\rho}$, then $\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2 \geq \frac{4R^2}{\rho}$, thus

$$\frac{R^2}{\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2} \leq \frac{R^2}{4R^2/\rho} = \frac{\rho}{4}.$$

In view of this, the inequality of the previous question implies

$$\begin{aligned} &\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \rho/2 + \frac{R^2}{\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2} \\ \implies &\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \rho/2 + \frac{\rho}{4} = \|\mathbf{w}_t\| + \frac{3}{4}\rho. \end{aligned}$$

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 $\|\mathbf{w}_{t_0+1}\| \geq \frac{4R^2}{\rho}$.

   Solution: Since $\mathbf{w}_1 = y_1\mathbf{x}_1$, $\|\mathbf{w}_1\| = \|\mathbf{x}_1\| \leq R$. The margin ρ is at most twice the radius R, thus, $\rho \leq 2R$ and $2R/\rho \geq 1$. This implies that $\|\mathbf{w}_1\| \leq R \leq 2R^2/\rho$. Since $\|\mathbf{w}_1\| \leq 2R^2/\rho$ and $\|\mathbf{w}_{T+1}\| \geq \frac{4R^2}{\rho}$, there must exist at least one update time $t \in I$ at which $\|\mathbf{w}_t\| \leq \frac{4R^2}{\rho}$ and $\|\mathbf{w}_{t+1}\| \geq \frac{4R^2}{\rho}$. The set of such times t is non empty and thus admits a largest element t_0.

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

   Solution: By definition of t_0, for any $t \geq t_0$, $\|\mathbf{w}_{t+1}\| \geq \frac{4R^2}{\rho}$. Thus, by the inequality of question 4, the following holds for any $t \geq t_0$,

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

   This implies that

$$\begin{aligned}
\|\mathbf{w}_{T+1}\| &\leq \|\mathbf{w}_{t_0}\| + \Big|[t_0, T+1[\cap I\Big| \frac{3}{4}\rho \\
&\leq \|\mathbf{w}_{t_0}\| + M\frac{3}{4}\rho \\
&\leq \frac{4R^2}{\rho} + M\frac{3}{4}\rho.
\end{aligned}$$

   By the first question $M\rho \leq \|\mathbf{w}_{T+1}\|$, therefore

$$M\rho \leq \frac{4R^2}{\rho} + M\frac{3}{4}\rho \iff M\rho/4 \leq 4R^2/\rho \iff M \leq 16R^2/\rho^2.$$

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

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

```
MPerceptron()
w_1 ← 0
for t ← 1 to T do
    Receive(x_t)
    Receive(y_t)
    if ((w_t = 0) or (y_t w_t · x_t / ||w_t|| < ρ/2)) then
        w_{t+1} ← w_t + y_t x_t
    else  w_{t+1} ← w_t
return w_{T+1}
```

Figure 1: MPerceptron algorithm.

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.

   Solution: A Bayes classifier h^* can be defined by $h^*(x) = y_i$ for $x \in I_i$, $i \geq 1$, which guarantees zero error on these intervals. For I_0, since all classes are equiprobable, we can just choose one class: $h^*(x) = y_1$ for $x \in I_0$. Its error is then $(n-1)/n \cdot nr/(n-1) = r$ over this interval. Thus, the overall error of h^* is $R^* = r$.

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.

   Solution: First observe that for $x \in I_i$, $i \in [0, n]$, no point $x' \in I_j$, $j \neq i$ is closer to x than a point x'' in I_i. Thus, the nearest neighbor rule labels I_i with the label of the points falling in I_i for $i > 0$. Since at least one point falls in each of these intervals, the nearest neighbor algorithm labels them all correctly. For $i = 0$, regardless of the labeling, since all classes are equiprobable, its error is r. Thus, the nearest-neighbor algorithm's overall error rate is $R^* = r$, which is optimal.