Machine Learning Fundamentals Homework 2

Predrag Pilipovic (predrag.pilipovic@grenoble-inp.org)

November 2, 2019

An analysis of the Margin-based Perceptorn Algorithm

Given a training set $S = \{(\mathbf{x}_i, y_i) \mid i \in \{1, 2, ..., m\}\}$ that is linearly separable with a hyperplane having a normal vector $\mathbf{w}^*$ and a maximum margin

$$\rho = \min_i \frac{y_i \langle \mathbf{x}_i, \mathbf{w}^* \rangle}{||\mathbf{w}^*||},$$

Novikoff's theorem states that the Perceptron algorithm run cyclically over S is guaranteed to converge after at most $\frac{R^2}{\rho^2}$ updates, where R is the radius of the hypersphere containing the training points. However, this theorem does not guarantee that the hyperplane solution of the Perceptron algorithm achieves a margin close to ρ. Suppose we modify the Perceptron algorithm to ensure that the margin of the hyperplane solution is at least $\frac{\rho}{2}$. In particular, consider the algorithm described in Algorithm 1. Let Υ denote the set of times at which the algorithm makes an update and let $U = |\Upsilon|$ be the total number of updates.

Question 1.

Explain why we have

$$U\rho \leqslant \frac{\left\langle \mathbf{w}^*, \sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)} \right\rangle}{||\mathbf{w}^*||}.$$

Applying the Cauchy-Schwarz inequality we get (to not be proven)

$$U\rho \leqslant \left|\left|\sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)}\right|\right|.$$

If we suppose that $\mathbf{w}^{(0)} = \mathbf{0}$, deduce then

$$U\rho \leqslant \left|\left|\mathbf{w}^{(T+1)}\right|\right|.$$

Conclude that if $\left|\left|\mathbf{w}^{(T+1)}\right|\right| < \frac{4R^2}{\rho}$ then $U \leqslant \frac{4R^2}{\rho^2}$. (For the remainder of this problem, we will assume that $\left|\left|\mathbf{w}^{(T+1)}\right|\right| \geqslant \frac{4R^2}{\rho}$.)

For the first part of the question we used bilinearity of scalar product, and property of the margin. So, we have

$$\frac{\left\langle \mathbf{w}^*, \sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)} \right\rangle}{||\mathbf{w}^*||} = \sum_{t\in\Upsilon} \frac{y^{(t)}\left\langle \mathbf{w}^*, \mathbf{x}^{(t)} \right\rangle}{||\mathbf{w}^*||} \geqslant \sum_{t\in\Upsilon} \rho = |\Upsilon|\rho = U\rho.$$

Algorithm 1 Margin-based Perceptron

Training set $S = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, ..., m\}$
Initialize the weights $\mathbf{w}^{(0)} \leftarrow \mathbf{0}$
Maximum number of iterations $T > 1$
$t \leftarrow 0$
Learning rate $\eta \leftarrow 1$
repeat
    Choose randomly an example $(\mathbf{x}^{(t)}, y^{(t)}) \in S$
    if $\frac{y^{(t)}\langle \mathbf{w}^{(t)}, \mathbf{x}^{(t)}\rangle}{||\mathbf{w}^{(t)}||} < \frac{\rho}{2}$ **then**
        $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + y^{(t)} \cdot \mathbf{x}^{(t)}$
    else
        $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)}$
    end if
    $t \leftarrow t + 1$
until $t > T$
Return $\mathbf{w}^{(T+1)}$

Knowing that Cauchy-Schwarz inequality states $\langle u, v\rangle \leqslant ||u|| \cdot ||v||$, for all vectors u, v, we have

$$U\rho \leqslant \frac{\left\langle \mathbf{w}^*, \sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)} \right\rangle}{||\mathbf{w}^*||} \leqslant \frac{||\mathbf{w}^*|| \cdot \left|\left|\sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)}\right|\right|}{||\mathbf{w}^*||} \leqslant \left|\left|\sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)}\right|\right|.$$

Having in mind that we need an update rule for indexes $t \in \Upsilon$, sometimes there will be an update and sometimes there will not. So going from $\mathbf{w}^{(T)}$ to $\mathbf{w}^{(0)}$ there will be $|\Upsilon|$ changes. After applying inductive steps we have

$$\begin{aligned}
\mathbf{w}^{(T+1)} &= \mathbf{w}^{(T)} + y^{(T)}\mathbf{x}^{(T)} \\
&= \mathbf{w}^{(T-1)} + y^{(T-1)}\mathbf{x}^{(T-1)} + y^{(T)}\mathbf{x}^{(T)} \\
&\vdots \\
&= \mathbf{w}^{(0)} + \sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)} = \sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)}.
\end{aligned}$$

Applying the previous calculation with inequality we know from before, we have

$$\left|\left|\mathbf{w}^{(T+1)}\right|\right| = \left|\left|\sum_{t\in\Upsilon} y^{(t)}\mathbf{x}^{(t)}\right|\right| \geqslant U\rho.$$

If we assume that

$$\left|\left|\mathbf{w}^{(T+1)}\right|\right| < \frac{4R^2}{\rho},$$

then from the previous two lines we can conclude

$$U < \frac{4R^2}{\rho^2}.$$

Question 2.

Show that for any $t \in \Upsilon$ (including $t = 0$), the following holds

$$\left|\left|\mathbf{w}^{(t+1)}\right|\right|^2 \leqslant \left(\left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}\right)^2 + R^2.$$

Again, we are working just with points for which we need an update rule, so we know that

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y^{(t)}\mathbf{x}^{(t)}.$$

That also means we have

$$\frac{y^{(t)}\left\langle \mathbf{w}^{(t)}, \mathbf{x}^{(t)} \right\rangle}{\left|\left|\mathbf{w}^{(t)}\right|\right|} < \frac{\rho}{2}.$$

Having that in mind, we can bound $\left|\left|\mathbf{w}^{(t+1)}\right|\right|^2$.

$$\begin{aligned}
\left|\left|\mathbf{w}^{(t+1)}\right|\right|^2 &= \left|\left|\mathbf{w}^{(t)} + y^{(t)}\mathbf{x}^{(t)}\right|\right|^2 \\
&= \left|\left|\mathbf{w}^{(t)}\right|\right|^2 + \underbrace{2y^{(t)}\left\langle \mathbf{w}^{(t)}, \mathbf{x}^{(t)} \right\rangle}_{<\rho\cdot\left|\left|\mathbf{w}^{(t)}\right|\right|} + \underbrace{\left|\left|\mathbf{x}^{(t)}\right|\right|^2}_{\leqslant R^2} \\
&< \left|\left|\mathbf{w}^{(t)}\right|\right|^2 + \rho\cdot\left|\left|\mathbf{w}^{(t)}\right|\right| + R^2 \\
&< \left|\left|\mathbf{w}^{(t)}\right|\right|^2 + \rho\cdot\left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho^2}{4} + R^2 \\
&= \left(\left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}\right)^2 + R^2.
\end{aligned}$$

Note that for $t = 0$, we have $\mathbf{w}^{(0)} = \mathbf{0}$ and we cannot normalize it, but for convention we can say that the normalization of zero vector is again zero vector. So we have

$$y^{(0)}\left\langle \frac{\mathbf{w}^{(0)}}{||\mathbf{w}^{(0)}||}, \mathbf{x}^{(0)} \right\rangle = 0.$$

This yields an upgrade rule $\mathbf{w}^{(1)} = y^{(1)}\mathbf{x}^{(1)}$. So, we have

$$\left|\left|\mathbf{w}^{(1)}\right|\right|^2 = \left|\left|y^{(1)}\mathbf{x}^{(1)}\right|\right|^2 = \left|\left|\mathbf{x}^{(1)}\right|\right|^2 \leqslant R^2 < \frac{\rho^2}{4} + R^2 = \left(\left|\left|\mathbf{w}^{(0)}\right|\right| + \frac{\rho}{2}\right)^2 + R^2.$$

Question 3.

From Question 2, infer that for any $t \in \Upsilon$ we have

$$\left|\left|\mathbf{w}^{(t+1)}\right|\right| \leqslant \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2} + \frac{R^2}{\left|\left|\mathbf{w}^{(t+1)}\right|\right| + \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}}.$$

Applying the difference of squares formula on inequality from Question 2, we have

$$
\begin{aligned}
&\left|\left|\mathbf{w}^{(t+1)}\right|\right|^2 \leqslant \left(\left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}\right)^2 + R^2 \\
\Leftrightarrow &\left|\left|\mathbf{w}^{(t+1)}\right|\right|^2 - \left(\left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}\right)^2 \leqslant R^2 \\
\Leftrightarrow &\left(\left|\left|\mathbf{w}^{(t+1)}\right|\right| - \left|\left|\mathbf{w}^{(t)}\right|\right| - \frac{\rho}{2}\right)\left(\left|\left|\mathbf{w}^{(t+1)}\right|\right| + \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}\right) \leqslant R^2 \\
\Leftrightarrow &\left|\left|\mathbf{w}^{(t+1)}\right|\right| - \left|\left|\mathbf{w}^{(t)}\right|\right| - \frac{\rho}{2} \leqslant \frac{R^2}{\left|\left|\mathbf{w}^{(t+1)}\right|\right| + \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}} \\
\Leftrightarrow &\left|\left|\mathbf{w}^{(t+1)}\right|\right| \leqslant \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2} + \frac{R^2}{\left|\left|\mathbf{w}^{(t+1)}\right|\right| + \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}}.
\end{aligned}
$$

Question 4.

Using the inequality from Question 3, show that for any $t \in \Upsilon$ such that either

$$\left|\left|\mathbf{w}^{(t)}\right|\right| \geqslant \frac{4R^2}{\rho} \textbf{ or } \left|\left|\mathbf{w}^{(t+1)}\right|\right| \geqslant \frac{4R^2}{\rho},$$

we have

$$\left|\left|\mathbf{w}^{(t+1)}\right|\right| \leqslant \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{3}{4}\rho.$$

Without loss of generality let's assume that $\left|\left|\mathbf{w}^{(t)}\right|\right| \geqslant \frac{4R^2}{\rho}$. Using inequality from Question 3, we have

$$
\begin{aligned}
\left|\left|\mathbf{w}^{(t+1)}\right|\right| &\leqslant \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2} + \frac{R^2}{\left|\left|\mathbf{w}^{(t+1)}\right|\right| + \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2}} \\
&\leqslant \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2} + \frac{R^2}{\left|\left|\mathbf{w}^{(t)}\right|\right|} \\
&\leqslant \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2} + \frac{R^2}{\frac{4R^2}{\rho}} \\
&= \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{\rho}{2} + \frac{\rho}{4} = \left|\left|\mathbf{w}^{(t)}\right|\right| + \frac{3}{4}\rho.
\end{aligned}
$$

Question 5.

Show that $\left|\left|\mathbf{w}^{(1)}\right|\right| \leqslant R \leqslant \frac{4R^2}{\rho}$. Since by assumption we have $\left|\left|\mathbf{w}^{(T+1)}\right|\right| \geqslant \frac{4R^2}{\rho}$, conclude that there must exist a largest time $t_0 \in \Upsilon$, such that $\left|\left|\mathbf{w}^{(t_0)}\right|\right| \leqslant \frac{4R^2}{\rho}$ and $\left|\left|\mathbf{w}^{(t_0+1)}\right|\right| \geqslant \frac{4R^2}{\rho}$.

We have already seen that $\left|\left|\mathbf{w}^{(1)}\right|\right| \leqslant R$ in the end of the Question 2. So, we need to prove that $R \leqslant \dfrac{4R^2}{\rho}$. Knowing that every term in last inequality is greater than zero, it is equivalent to $\dfrac{\rho}{R} \leqslant 4$. We will use properties of R and ρ, so let us fix $\left(\mathbf{x}^{(t)}, y^{(t)}\right) \in S$. We know that

$$
\begin{aligned}
R &\geqslant \left|\left|\mathbf{x}^{(t)}\right|\right| \\
\rho &\leqslant \frac{y^{(t)}\left\langle \mathbf{x}^{(t)}, \mathbf{w}^*\right\rangle}{||\mathbf{w}^*||}.
\end{aligned}
$$

We also know that $y^{(t)} \leqslant 1$. Having in mind the previous and Cauchy-Schwarz inequality, we have

$$
\frac{\rho}{R} \leqslant \frac{y^{(t)}\left\langle \mathbf{x}^{(t)}, \mathbf{w}^*\right\rangle}{\left|\left|\mathbf{x}^{(t)}\right|\right| \cdot ||\mathbf{w}^*||} \leqslant \frac{\left|\left|\mathbf{x}^{(t)}\right|\right| \cdot ||\mathbf{w}^*||}{\left|\left|\mathbf{x}^{(t)}\right|\right| \cdot ||\mathbf{w}^*||} = 1 < 4.
$$

For the next part, we should use next inequalities

$$
\begin{aligned}
\left|\left|\mathbf{w}^{(1)}\right|\right| &\leqslant \frac{4R^2}{\rho}, \\
\left|\left|\mathbf{w}^{(T+1)}\right|\right| &\geqslant \frac{4R^2}{\rho}.
\end{aligned}
$$

This means that there must exists at least one t such that $\left|\left|\mathbf{w}^{(t)}\right|\right| \leqslant \frac{4R^2}{\rho}$ and $\left|\left|\mathbf{w}^{(t+1)}\right|\right| \geqslant \frac{4R^2}{\rho}$. We can call this t a change point. Let us prove that there exist at least one change point t. Suppose the opposite, for all $t \in \Upsilon$, it must be either

$$
\left|\left|\mathbf{w}^{(t)}\right|\right| > \frac{4R^2}{\rho} \text{ or } \left|\left|\mathbf{w}^{(t+1)}\right|\right| < \frac{4R^2}{\rho}.
$$

We know that for $t = 1$ is $\left|\left|\mathbf{w}^{(1)}\right|\right| \leqslant \frac{4R^2}{\rho}$, so it must be $\left|\left|\mathbf{w}^{(2)}\right|\right| < \frac{4R^2}{\rho}$. Now, when $t = 2$ it is $\left|\left|\mathbf{w}^{(2)}\right|\right| < \frac{4R^2}{\rho}$, so again we conclude that it must be $\left|\left|\mathbf{w}^{(3)}\right|\right| < \frac{4R^2}{\rho}$. After applying the same rule until $t = T$, we have $\left|\left|\mathbf{w}^{(T)}\right|\right| < \frac{4R^2}{\rho}$, so it must be $\left|\left|\mathbf{w}^{(T+1)}\right|\right| < \frac{4R^2}{\rho}$, which is a contradiction. So, of all change points $t \in \Upsilon$ we can choose the greatest one, and call it t_0.

Question 6.

Show that

$$
\left|\left|\mathbf{w}^{(T+1)}\right|\right| \leqslant \left|\left|\mathbf{w}^{(t_0)}\right|\right| + \frac{3}{4}U\rho.
$$

Conclude that $U \leqslant 16\dfrac{R^2}{\rho^2}$**.**

From the previous question we know that there exist the greatest change point $t_0 \in \Upsilon$ such that

$$
\left|\left|\mathbf{w}^{(t_0)}\right|\right| \leqslant \frac{4R^2}{\rho} \text{ and } \left|\left|\mathbf{w}^{(t_0+1)}\right|\right| \geqslant \frac{4R^2}{\rho}.
$$

This means that for all $t > t_0$ it must be

$$\left|\left|\mathbf{w}^{(t)}\right|\right| \geqslant \frac{4R^2}{\rho}.$$

Then, we can apply result from the Question 3 on every $\mathbf{w}^{(t)}$, for $t > t_0$, so we have

$$\begin{aligned}
\left|\left|\mathbf{w}^{(T+1)}\right|\right| &= \left|\left|\mathbf{w}^{(T)}\right|\right| + \frac{3}{4}\rho \\
&\leqslant \left|\left|\mathbf{w}^{(T-1)}\right|\right| + 2 \cdot \frac{3}{4}\rho \\
&\vdots \\
&\leqslant \left|\left|\mathbf{w}^{(t_0)}\right|\right| + \widetilde{U}\frac{3}{4}\rho,
\end{aligned}$$

where $\widetilde{U}$ denotes a number of updates from time t_0 until $T+1$. Note here that we cannot be sure if there was any update before t_0 because t_0 is the greatest change point, which means that there could be other change points as well. When a change point occurs, it implies that update happened. Anyway, it must be $\widetilde{U} \leqslant U$, because U denotes a number of updates from $t = 0$. So, we have

$$\left|\left|\mathbf{w}^{(T+1)}\right|\right| \leqslant \left|\left|\mathbf{w}^{(t_0)}\right|\right| + \frac{3}{4}\widetilde{U}\rho \leqslant \left|\left|\mathbf{w}^{(t_0)}\right|\right| + \frac{3}{4}U\rho.$$

Finally, in the Question 1 we saw that

$$U\rho \leqslant \left|\left|\mathbf{w}^{(T+1)}\right|\right|.$$

After applying it with the previous line and the fact that

$$\left|\left|\mathbf{w}^{(t_0)}\right|\right| \leqslant \frac{4R^2}{\rho},$$

we have

$$U\rho \leqslant \left|\left|\mathbf{w}^{(T+1)}\right|\right| \leqslant \left|\left|\mathbf{w}^{(t_0)}\right|\right| + \frac{3}{4}U\rho \leqslant \frac{4R^2}{\rho} + \frac{3}{4}U\rho.$$

This means that

$$U\rho \leqslant \frac{4R^2}{\rho} + \frac{3}{4}U\rho,$$

i.e. $U \leqslant 16\dfrac{R^2}{\rho^2}$.