Learning From Data
Problem 1.3
Proof

Oleksii Symon

January 16, 2018

Task

Problem 1.3 Prove that the PLA eventually converges to a linear separator for separable data. The following steps will guide you through the proof. Let $\mathbf{w}^*$ be an optimal set of weights (one which separates the data). The essential idea in this proof is to show that the PLA weights $\mathbf{w}(t)$ get "more aligned" with $\mathbf{w}^*$ with every iteration. For simplicity, assume that $\mathbf{w}(0) = \mathbf{0}$.

(a) Let $\rho = \min_{1 \le n \le N} y_n(\mathbf{w}^{*\mathsf{T}}\mathbf{x}_n)$. Show that $\rho > 0$.

(b) Show that $\mathbf{w}^{\mathsf{T}}(t)\mathbf{w}^* \ge \mathbf{w}^{\mathsf{T}}(t-1)\mathbf{w}^* + \rho$, and conclude that $\mathbf{w}^{\mathsf{T}}(t)\mathbf{w}^* \ge t\rho$. *[Hint: Use induction.]*

(c) Show that $\|\mathbf{w}(t)\|^2 \le \|\mathbf{w}(t-1)\|^2 + \|\mathbf{x}(t-1)\|^2$.

[Hint: $y(t-1) \cdot (\mathbf{w}^{\mathsf{T}}(t-1)\mathbf{x}(t-1)) \le 0$ *because* $\mathbf{x}(t-1)$ *was misclas sified by* $\mathbf{w}(t-1)$*.]*

(d) Show by induction that $\|\mathbf{w}(t)\|^2 \le tR^2$, where $R = \max_{1 \le n \le N} \|\mathbf{x}_n\|$.

(continued on next page)

(e) Using (b) and (d), show that

$$\frac{\mathbf{w}^{\intercal}(t)}{\|\mathbf{w}(t)\|}\mathbf{w}^* \geq \sqrt{t}\cdot\frac{\rho}{R},$$

and hence prove that

$$t \leq \frac{R^2\|\mathbf{w}^*\|^2}{\rho^2}.$$

$\left[\textit{Hint: } \frac{\mathbf{w}^{\intercal}(t)\mathbf{w}^*}{\|\mathbf{w}(t)\|\,\|\mathbf{w}^*\|} \leq 1. \textit{ Why?}\right]$

In practice, PLA converges more quickly than the bound $\frac{R^2\|\mathbf{w}^*\|^2}{\rho^2}$ suggests. Nevertheless, because we do not know ρ in advance, we can't determine the number of iterations to convergence, which does pose a problem if the data is non-separable.

Solution

0.1 (a)

$$\rho = \min_{1\leq n\leq N} y_n(w^{*\intercal}x_n)$$

As $y_n = sign(w^{*\intercal}x_n)$ we get 2 cases:

$y_n = -1 \rightarrow w^{*\intercal}x_n = -1 \rightarrow -1 * -1 = 1$

$y_n = 1 \rightarrow w^{*\intercal}x_n = 1 \rightarrow 1 * 1 = 1$

0.2 (b)

$w^{\intercal}(t)w^* \geq w^{\intercal}(t-1)w^* + \rho$

Let's expand left and right sides of equation

$(w(t-1) + y(t-1)x(t-1))^{\intercal}w^* \geq w^{\intercal}(t-1)w^* + \rho$

As $y(t-1)x(t-1)w^* > 0$ we get

$w^{\intercal}(t-1)w^* + y(t-1)x^{\intercal}(t-1)w^* \geq w^{\intercal}(t-1)w^* + \rho$

Subtract $w^{\intercal}(t-1)w^*$ from both parts of equation

$y(t-1)x^{\intercal}(t-1)w^* \geq \rho$

This inequality is correct because ρ is a minimum in the entire data set

Now we will show that $w^{\intercal}(t)w^* \geq t\rho$. We will prove this using induction.

1. $w^\intercal(0)w^* \geq 0\rho$
2. $w^\intercal(1)w^* \geq 1\rho$
   $(w(0) + y(0)x(0))^\intercal w^* \geq \rho$
   $y(0)x^\intercal(0)w^* \geq \rho$
3. $w^\intercal(t)w^* \geq t\rho$
   $(w(t-1) + y(t-1)x(t-1))^\intercal w^* \geq t\rho$
   $(w(t-2) + y(t-2)x(t-2) + y(t-1)x(t-1))^\intercal w^* \geq t\rho$
   $(y(0)x^\intercal(0) + \ldots + y(t-2)x^\intercal(t-2) + y(t-1)x^\intercal(t-1))w^* \geq t\rho$

0.3 (c)

To prove part (c) we should firstly prove the Cauchy-Schwarz inequality.

$|\vec{x} \cdot \vec{y}\,| \leq \|\vec{x}\,\|\,\|\vec{y}\,\|$ and $|\vec{x} \cdot \vec{y}\,| = \|\vec{x}\,\|\,\|\vec{y}\,\| \leftrightarrow \vec{x} = c\vec{y}$

Let's define a function $p(t) = \|t\vec{y} - \vec{x}\,\|^2 \geq 0$

$\|t\vec{y} - \vec{x}\,\|^2 = (t\vec{y} - \vec{x}\,) \cdot (t\vec{y} - \vec{x}\,)$

$= \vec{y} \cdot \vec{y}\,t^2 - 2\vec{x} \cdot \vec{y}\,t + \vec{x} \cdot \vec{x} \geq 0$

Let's define $\vec{y} \cdot \vec{y} = a$ and $2\vec{x} \cdot \vec{y} = b$,and $\vec{x} \cdot \vec{x} = c$

$at^2 - bt + c \geq 0$

$p\left(\dfrac{b}{2a}\right) = \dfrac{ab^2}{4a^2} - \dfrac{b^2}{2a} + c \geq 0$

$-\dfrac{b^2}{4a} + c \geq 0$

$4ac \geq b^2$

Substitute back defined values

$\cancel{4}(\vec{y} \cdot \vec{y}\,)(\vec{x} \cdot \vec{x}\,) \geq \cancel{4}(\vec{x} \cdot \vec{y}\,)^2$

$\sqrt{\|y\|^2\,\|x\|^2} \geq \sqrt{(\vec{x} \cdot \vec{y}\,)^2}$

$\|\vec{x}\,\|\,\|\vec{y}\,\| \geq |\vec{x} \cdot \vec{y}\,|$

Now we will use Cauchy-Schwarz inequality to prove triangle inequality.

$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$
and $\|\vec{x} + \vec{y}\| = \|\vec{x}\| + \|\vec{y}\| \leftrightarrow \vec{x} = 0$ and $\vec{y} = 0$
$\|\vec{x} + \vec{y}\|^2 = (\vec{x} + \vec{y}) \cdot (\vec{x} + \vec{y}) =$
$\vec{x} \cdot \vec{x} + 2\vec{x} \cdot \vec{y} + \vec{y} \cdot \vec{y} =$
$\|\vec{x}\|^2 + 2\vec{x} \cdot \vec{y} + \|\vec{y}\|^2$
Now let's take a look at $\vec{x} \cdot \vec{y}$
It can be observed that $\vec{x} \cdot \vec{y} \leq |\vec{x} \cdot \vec{y}|$
For example when $\vec{x}$ has all negative values and $\vec{y}$ has all positive ones
Also from Cauchy-Schwarz's inequality we know that $|\vec{x} \cdot \vec{y}| \leq \|x\| \|y\|$
So we get $\vec{x} \cdot \vec{y} \leq |\vec{x} \cdot \vec{y}| \leq \|x\| \|y\|$
$\|x\|^2 + 2\vec{x} \cdot \vec{y} + \|y\|^2 \leq \|x\|^2 + 2\|\vec{x}\| \|\vec{y}\| + \|y\|^2$
$\|\vec{x} + \vec{y}\|^2 \leq (\|x\| + \|y\|)^2$
$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$

Now let's start proving part (c). Firstly we will open the left part of inequality.

$$\|w(t)\|^2 = \|w(t-1) + y(t-1)x(t-1)\|^2$$

We know that $\|\vec{x} + \vec{y}\|^2 = \|\vec{x}\|^2 + 2\vec{x} \cdot \vec{y} + \|\vec{y}\|^2$. So we will plug our variables into this equation.

$$\|w(t-1) + y(t-1)x(t-1)\|^2 = \|w(t-1)\|^2 + 2w(t-1)y(t-1)x(t-1) + \|y(t-1)x(t-1)\|^2$$

We know that $w^{\intercal}(t-1)x(t-1)$ was classified incorrectly. So $w(t-1)y(t-1)x(y-1) \leq 0$. Let $w(t-1)y(t-1)x(t-1) = C$ - some constant.

$$\|w(t-1)\|^2 - 2C + \|y(t-1)x(t-1)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2$$

$$\|y(t-1)x(t-1)\|^2 = \|\pm x(t-1)\|^2 = \|x(t-1)\|^2$$

0.4 (d)

Let's show that $\|w(t)\|^2 \leq tR^2$ where $R = \max_{1 \leq n \leq N} \|x_n\|$

1. $\|w(0)\|^2 \leq 0R^2$ as $w(0) = 0$
2. $\|w(1)\|^2 \leq R^2$
   $\|w(0) + y(0)x(0)\|^2 \leq R^2$
   $\|y(0)x(0)\|^2 = \|x(0)\|^2 \leq R^2$
   This is also true as R is the biggest x_n in all set.

3. $\|w(t)\|^2 \leq tR^2$
   $\|w(t-1) + y(t-1)x(t-1)\|^2 \leq tR^2$
   Let's open left part of inequality according to triangle inequality as in (c).
   $\|w(t-1)\|^2 + 2w(t-1)y(t-1)x(t-1) + \|y(t-1) + x(t-1)\| \leq tR^2$
   At the end we will have
   $\|w(0)\|^2 + 2w(0)y(0)x(0) + \|y(0)x(0)\|^2 \ldots + 2w(t-1)y(t-1)x(t-1) + \|y(t-1)x(t-1)\|^2 \leq tR^2$
   As $w(n)y(n)x(n) \leq 0$ because of misclassification the above will be true.

0.5 (e)

From second part of (b) we know that $w^\intercal(t)w^* \geq t\rho$.
From (d) we know that $\|w(t)\|^2 \leq tR^2$ and hence $\|w(t)\| \leq \sqrt{t}R$.
So we can combine this results and find t - number of iterations needed for PLA to converge.

$$\frac{w^\intercal(t)w^*}{\|w(t)\|} \geq \frac{t\rho}{\sqrt{t}R}$$
$$\frac{w^\intercal(t)w^*}{\|w(t)\|} \geq \frac{\sqrt{t}\rho}{R}$$
$$\sqrt{t} \leq \frac{w^\intercal(t)w^*R}{\|w(t)\|\,\rho}$$
$$t \leq \frac{w^{2\intercal}(t)w^{*2}R^2}{\|w(t)\|^2\,\rho^2}$$
$$\frac{t}{\|w^*\|^2} \leq \frac{w^{2\intercal}(t)w^{*2}R^2}{\|w(t)\|^2\,\|w^*\|^2\,\rho^2}$$

Let's take a look at $\dfrac{w^{2\intercal}(t)w^{*2}}{\|w(t)\|^2\,\|w^*\|^2}$

$$\|w(t)\|^2\,\|w^*\|^2 = \left(w^2(t)_1 + \ldots + w^2(t)_n\right) * \left(w^{*2}(t)_1 + \ldots + w^{*2}(t)_n\right)$$
$$w^{2\intercal}(t)w^{*2} = w^{2\intercal}(t)_1 * w_1^{*2} + \ldots + w^{2\intercal}(t)_n * w_n^{*2}$$

So it should be clear that $\dfrac{w^{2\intercal}(t)w^{*2}}{\|w(t)\|^2\,\|w^*\|^2} \leq 0$ and we can get rid of it

$$t \leq \frac{\|w^*\|^2\,R^2}{\rho^2}$$