Math 345 - Notes
PLA Convergence Theorem
September 12, 2018

PLA Convergence Theorem

Theorem 1. *Given a linearly separable data set $\mathcal{D}$, the PLA algorithm on $\mathcal{D}$ ends after finitely many iterations.*

Proof. Since $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ is linearly separable, there exists a weight vector, call it $\mathbf{w}^*$ so that

$$y_j = \text{sign}(\mathbf{w}^* \cdot \mathbf{x}_j) \qquad \text{for all } (\mathbf{x}_j, y_j) \in \mathcal{D}$$

It it important to note that, since y_j and $\mathbf{w}^* \cdot \mathbf{x}_j$ have the same sign,

$$y_j\, \mathbf{w}^* \cdot \mathbf{x}_j > 0 \qquad \text{for all } (\mathbf{x}_j, y_j) \in \mathcal{D} \tag{1}$$

Without loss of generality, we start the PLA with $\mathbf{w}(0) = \mathbf{0}$. Suppose we are at the update

$$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t) \tag{2}$$

Since $\mathbf{w}(t)$ comes from an iterative process, we can trace our steps back to time 0 as follows:

$$\begin{aligned}
\mathbf{w}(t+1) &= \mathbf{w}(t) + y(t)\mathbf{x}(t) \\
&= \mathbf{w}(t-1) + y(t-1)\mathbf{x}(t-1) + y(t)\mathbf{x}(t) \\
&= \mathbf{w}(t-2) + y(t-2)\mathbf{x}(t-2) + y(t-1)\mathbf{x}(t-1) + y(t)\mathbf{x}(t) \\
&= \dots \\
&= \mathbf{w}(0) + \sum_{j=0}^{t} y(j)\mathbf{x}(j) \\
&= \sum_{j=0}^{t} y(j)\mathbf{x}(j)
\end{aligned}$$

Taking the dot product of this with the ideal weight vector $\mathbf{w}^*$,

$$\mathbf{w}^* \cdot \mathbf{w}(t+1) = \mathbf{w}^* \cdot \sum_{j=0}^{t} y(j)\mathbf{x}(j) = \sum_{j=0}^{t} y(j)\mathbf{w}^* \cdot \mathbf{x}(j) \geq (t+1)m,$$

where $m = \min\limits_{1 \leq j \leq N} y_j \mathbf{w}^* \cdot \mathbf{x}_j$. We note that m is finite and positive, by (1) and the fact that there are $N < \infty$ elements in the data set. Using Cauchy-Schwarz,

$$m(t+1) \leq |\mathbf{w}^* \cdot \mathbf{w}(t+1)| \leq ||\mathbf{w}^*||\, ||\mathbf{w}(t+1)|| \quad \Rightarrow \quad ||\mathbf{w}(t+1)|| \geq \frac{m(t+1)}{||\mathbf{w}^*||} \tag{3}$$

Since $\mathbf{w}^* \neq \mathbf{0}$, we can safely divide by $||\mathbf{w}^*||$. This gives a lower bound on $||\mathbf{w}(t+1)||$.

For the upper bound, we will use the update rule (2). Also recall, from linear algebra, that the dot product of a vector with itself is the magnitude squared:

$$\begin{aligned}
||\mathbf{w}(t+1)||^2 &= \mathbf{w}(t+1)\cdot\mathbf{w}(t+1) \\
&= [\mathbf{w}(t)+y(t)\mathbf{x}(t)]\cdot[\mathbf{w}(t)+y(t)\mathbf{x}(t)] \\
&= ||\mathbf{w}(t)||^2 + 2y(t)\mathbf{w}(t)\cdot\mathbf{x}(t) + y(t)^2||\mathbf{x}(t)||^2 \\
&\leq ||\mathbf{w}(t)||^2 + 2y(t)\mathbf{w}(t)\cdot\mathbf{x}(t) + ||\mathbf{x}(t)||^2 && \text{because } |y(t)| = 1 \\
&\leq ||\mathbf{w}(t)||^2 + ||\mathbf{x}(t)||^2 && \text{since } (\mathbf{x}(t), y(t)) \text{ is mislableled} \\
&\leq ||\mathbf{w}(t-1)||^2 + ||\mathbf{x}(t-1)||^2 + ||\mathbf{x}(t)||^2 && \text{iterate} \\
&\leq \dots \\
&\leq ||\mathbf{w}(0)||^2 + ||\mathbf{x}(0)||^2 + \cdots + ||\mathbf{x}(t-1)||^2 + ||\mathbf{x}(t)||^2 \\
&\leq \sum_{j=0}^{t} ||\mathbf{x}(j)||^2
\end{aligned}$$

Let $M = \max_{1\leq j\leq N} ||\mathbf{x}_j||^2$, which is finite and positive. Then

$$||\mathbf{w}(t+1)||^2 \leq M(t+1) \tag{4}$$

Now we combine (3) and (4) to obtain the inequality

$$\frac{m^2(t+1)^2}{||\mathbf{w}^*||^2} \leq ||\mathbf{w}(t+1)||^2 \leq M(t+1),$$

which holds true only when

$$\frac{m^2(t+1)^2}{||\mathbf{w}^*||^2} \leq M(t+1) \quad \Rightarrow \quad (t+1) \leq \frac{M||\mathbf{w}^*||^2}{m^2} < \infty.$$

Thus, the number of iterations is bounded by the (unknown) constant $\dfrac{M||\mathbf{w}^*||^2}{m^2}$ which depends only on the data set $\mathcal{D}$. $\square$