Note for Perceptron Learning Algorithm (Slide 2, Page 15)

Dongyuan Wu

August 21, 2020

1 What to prove

Start from $\mathbf{w}_0 = \mathbf{0}$, after T mistake corrections,

$$\frac{\mathbf{w}_f^T \mathbf{w}_T}{\left\|\mathbf{w}_f^T\right\| \left\|\mathbf{w}_T\right\|} \geq \sqrt{T} \cdot constant$$

2 What we have known

- In each round, we correct the mistake so that $\mathbf{w}_{t+1} = \mathbf{w}_t + y_{n(t)} \mathbf{x}_{n(t)}$.
- Exists perfect $\mathbf{w}_f$ such that $y_n = \operatorname{sign}(\mathbf{w}_f^T \mathbf{x}_n)$. In this way, every $\mathbf{x}_n$ correctly away from line:

  $$y_{n(t)} \mathbf{w}_f^T \mathbf{x}_{n(t)} \geq \min_n y_n \mathbf{w}_f^T \mathbf{x}_n > 0$$

  So we have

  $$\begin{aligned} \mathbf{w}_f^T \mathbf{w}_{t+1} &= \mathbf{w}_f^T (\mathbf{w}_t + y_{n(t)} \mathbf{x}_{n(t)}) \\ &= \mathbf{w}_f^T \mathbf{w}_t + y_{n(t)} \mathbf{w}_f^T \mathbf{x}_{n(t)} \\ &\geq \mathbf{w}_f^T \mathbf{w}_t + \min_n y_n \mathbf{w}_f^T \mathbf{x}_n \\ &> \mathbf{w}_f^T \mathbf{w}_t \end{aligned}$$

- $\mathbf{w}_t$ changed only when mistake:

  $$\operatorname{sign}(\mathbf{w}_t^T \mathbf{x}_{n(t)}) \neq y_{n(t)} \Leftrightarrow y_{n(t)} \mathbf{w}_t^T \mathbf{x}_{n(t)} \leq 0$$

  So we have

  $$\begin{aligned} \left\|\mathbf{w}_{t+1}\right\|^2 &= \left\|\mathbf{w}_t + y_{n(t)} \mathbf{x}_{n(t)}\right\|^2 \\ &= \left\|\mathbf{w}_t\right\|^2 + 2 y_{n(t)} \mathbf{w}_t^T \mathbf{x}_{n(t)} + \left\|y_{n(t)} \mathbf{x}_{n(t)}\right\|^2 \\ &\leq \left\|\mathbf{w}_t\right\|^2 + \left\|y_{n(t)} \mathbf{x}_{n(t)}\right\|^2 \\ &\leq \left\|\mathbf{w}_t\right\|^2 + \max_n \left\|y_n \mathbf{x}_n\right\|^2 \\ &= \left\|\mathbf{w}_t\right\|^2 + \max_n \left\|\mathbf{x}_n\right\|^2 \quad (\because y_n = \pm 1) \end{aligned}$$

3 Proof

We have

$$\begin{aligned}
\mathbf{w}_f^T\mathbf{w}_T &= \mathbf{w}_f^T(\mathbf{w}_{T-1} + y_{n(T-1)}\mathbf{x}_{n(T-1)}) \\
&\geq \mathbf{w}_f^T\mathbf{w}_{T-1} + \min_n y_n\mathbf{w}_f^T\mathbf{x}_n \\
&= \mathbf{w}_f^T(\mathbf{w}_{T-2} + y_{n(T-2)}\mathbf{x}_{n(T-2)}) + \min_n y_n\mathbf{w}_f^T\mathbf{x}_n \\
&\geq \mathbf{w}_f^T\mathbf{w}_{T-2} + 2\min_n y_n\mathbf{w}_f^T\mathbf{x}_n \\
&= \cdots \quad \text{(repeat totally } T \text{ times)} \\
&\geq \mathbf{w}_f^T\mathbf{w}_0 + T\min_n y_n\mathbf{w}_f^T\mathbf{x}_n \\
&\geq T\min_n y_n\mathbf{w}_f^T\mathbf{x}_n \quad (\because \mathbf{w}_0 = \mathbf{0})
\end{aligned}$$

and

$$\begin{aligned}
\|\mathbf{w}_T\|^2 &= \left\|\mathbf{w}_{T-1} + y_{n(T-1)}\mathbf{x}_{n(T-1)}\right\|^2 \\
&= \|\mathbf{w}_{T-1}\|^2 + 2y_{n(T-1)}\mathbf{w}_{T-1}^T\mathbf{x}_{n(T-1)} + \left\|y_{n(T-1)}\mathbf{x}_{n(T-1)}\right\|^2 \\
&\leq \|\mathbf{w}_{T-1}\|^2 + \left\|y_{n(T-1)}\mathbf{x}_{n(T-1)}\right\|^2 \\
&\leq \|\mathbf{w}_{T-1}\|^2 + \max_n \|y_n\mathbf{x}_n\|^2 \\
&= \left\|\mathbf{w}_{T-2} + y_{n(T-2)}\mathbf{x}_{n(T-2)}\right\|^2 + \max_n \|y_n\mathbf{x}_n\|^2 \\
&\leq \|\mathbf{w}_{T-2}\|^2 + 2\max_n \|y_n\mathbf{x}_n\|^2 \\
&= \cdots \quad \text{(repeat totally } T \text{ times)} \\
&\leq \|\mathbf{w}_0\|^2 + T\max_n \|y_n\mathbf{x}_n\|^2 \\
&= T\max_n \|\mathbf{x}_n\|^2 \quad (\because \mathbf{w}_0 = \mathbf{0} \text{ and } y_n = \pm 1)
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{\mathbf{w}_f^T\mathbf{w}_T}{\left\|\mathbf{w}_f^T\right\|\|\mathbf{w}_T\|} &\geq \frac{T\min_n y_n\mathbf{w}_f^T\mathbf{x}_n}{\left\|\mathbf{w}_f^T\right\|\sqrt{T\max_n \|\mathbf{x}_n\|^2}} \\
&= \sqrt{T}\frac{\min_n y_n\mathbf{w}_f^T\mathbf{x}_n}{\left\|\mathbf{w}_f^T\right\|\max_n \|\mathbf{x}_n\|} \\
&= \sqrt{T}\cdot \mathit{constant}
\end{aligned}$$

Because $\frac{\mathbf{w}_f^T\mathbf{w}_T}{\left\|\mathbf{w}_f^T\right\|\|\mathbf{w}_T\|} \leq 1$, the equation might be:

$$\sqrt{T}\frac{\min_n y_n\mathbf{w}_f^T\mathbf{x}_n}{\left\|\mathbf{w}_f^T\right\|\max_n \|\mathbf{x}_n\|} \leq 1 \Leftrightarrow T \leq \frac{\left\|\mathbf{w}_f^T\right\|^2\max_n \|\mathbf{x}_n\|^2}{(\min_n y_n\mathbf{w}_f^T\mathbf{x}_n)^2} \Leftrightarrow T \leq \frac{R^2}{\rho^2}$$

where $R^2 = \max_n \|\mathbf{x}_n\|^2$ and $\rho = \min_n y_n \frac{\mathbf{w}_f^T}{\|\mathbf{w}_f^T\|}\mathbf{x}_n$.

Since T mistake corrections increase the inner product by $\sqrt{T} \cdot constant$, the maximum number of corrected mistakes is $1/constant^2$.