Proof of bound of perceptron algorithm by Simon (ms9144)

Let

$\vec{x}^{(i)} = feature\ vector\ of\ i^{th} data$

$y^{(i)} = label\ of\ i^{th} data\ \in \{-1, 1\}$

$\vec{w}_t = weight\ vector\ on\ the\ t^{th} iteration\ or\ mistake\ (\vec{w}_1 = \vec{0}, \vec{w}_{t+1} = \vec{w}_t + y^{(i)}\vec{x}^{(i)})$

$\vec{w}^* = weight\ vector\ of\ unit\ length\ that\ linearly\ separates\ all\ data$

$\gamma = geometric\ margin\ (\exists \vec{w}^*\ s.t. |\vec{w}^*| = 1, \vec{y}^{(i)}\left(\vec{x}^{(i)} \cdot \vec{w}^*\right) \geq \gamma\ \forall i)$

$R = upper\ bound\ of\ |\vec{x}^{(i)}|\ \forall \vec{x}^{(i)}$

We will prove two lemmas first.

| $(lemma\ 1)\ \|\vec{w}_{t+1}\|^2 \leq tR^2$ | |
|---|---|
| | $LHS = \|\vec{w}_{t+1}\|^2$ |
| from the definition of $\vec{w}_t$ | $= \left(\vec{w}_t + y^{(i)}\vec{x}^{(i)}\right) \cdot \left(\vec{w}_t + y^{(i)}\vec{x}^{(i)}\right)$ |
| Distributive property | $= \vec{w}_t \cdot \vec{w}_t + 2y^{(i)}\vec{x}^{(i)} \cdot \vec{w}_t + \left(y^{(i)}\right)^2 \vec{x}^{(i)} \cdot \vec{x}^{(i)}$ |
| $y^{(i)}(\vec{x}^{(i)} \cdot \vec{w}_t) \leq 0$ since it's a mistake | $\leq \vec{w}_t \cdot \vec{w}_t + \left(y^{(i)}\right)^2 \vec{x}^{(i)} \cdot \vec{x}^{(i)}$ |
| from definition of $y^{(i)}$ and $R$ | $\leq \vec{w}_t \cdot \vec{w}_t + R^2$ |
| | $= \|\vec{w}_t\|^2 + R^2$ |
| By induction | $\leq tR^2 \quad \cdots (*)$ |

The last step is proved by induction:

| $Induction\ Hypothesis$: $\|\vec{w}_t\|^2 + R^2 \leq tR^2$ or $\|\vec{w}_t\|^2 \leq (t-1)R^2$ | |
|---|---|
| $(Base\ case)\ t = 1$ | |
| from the definition of $\vec{w}_t$ | $\|\vec{w}_1\|^2 + R^2 = R^2$ |
| | $\leq tR^2 = R^2$ |
| $(General\ Case)$ | |
| | $LHS = \|\vec{w}_{t+1}\|^2$ |
| from $(*)$ | $= \|\vec{w}_t\|^2 + R^2$ |
| from the induction hypothesis | $\leq (t-1)R^2 + R^2$ |
| | $= tR^2$ |

| $(lemma\ 2)\ \vec{w}_{t+1} \cdot \vec{w}^* \geq t\gamma$ | |
|---|---|
| | $LHS = \vec{w}_{t+1} \cdot \vec{w}^*$ |
| from definition of $\vec{w}_t$ | $= \left(\vec{w}_t + y^{(i)}\vec{x}^{(i)}\right) \cdot \vec{w}^*$ |
| distributive property | $= \vec{w}_t \cdot \vec{w}^* + y^{(i)}\vec{x}^{(i)} \cdot \vec{w}^*$ |
| from the definition of $\gamma$ | $\geq \vec{w}_t \cdot \vec{w}^* + \gamma$ |
| by induction | $\geq t\gamma$ |

The last step follows from another induction proof:

| | |
|---|---|
| $Induction\ hypothesis: \vec{w}_t \cdot \vec{w}^* + \gamma \geq\ t\gamma$ | |
| $(Base\ case)\ t = 1$ | |
| from the definition of $\vec{w}_t$ | $\vec{w}_1 \cdot \vec{w}^* + \gamma = \gamma$<br>$\geq t\gamma = 1\gamma$ |
| $(General\ Case)$ | |
| from the definition of $\vec{w}_t$<br>distributive property<br>The first two terms is the LHS of the induction hypothesis and the last term is bounded by $\gamma$. | $\vec{w}_{t+1} \cdot \vec{w}^* + \gamma = \left(\vec{w}_t + y^{(i)}\vec{x}^{(i)}\right) \cdot \vec{w}^* + \gamma$<br>$= \vec{w}_t \cdot \vec{w}^* + \gamma + y^{(i)}\vec{x}^{(i)} \cdot \vec{w}^*$<br>$= t\gamma + \gamma$<br><br>$=(t+1)\gamma$<br>$\geq t\gamma$ |

$(lemma\ 1)\ |\vec{w}_{t+1}|^2 \leq tR^2$

$(lemma\ 2)\ \vec{w}_{t+1} \cdot \vec{w}^* \geq t\gamma$

Now that we proved the two lemmas above, we can use them to arrive at the conclusion:

| | |
|---|---|
| From lemma 1 | $\sqrt{t}R \geq |\vec{w}_{t+1}|$ |
| $|\vec{w}^*| = 1$ by definition; $cos\theta \leq 1$ | $\geq |\vec{w}_{t+1}||\vec{w}^*|cos\theta$ |
| | $= \vec{w}_{t+1} \cdot \vec{w}^*$ |
| From lemma 2 | $\geq t\gamma$ |

It follows from $\sqrt{t}R \geq t\gamma$ that $t \leq R^2/\gamma^2$and we can conclude that the number of iterations $t$ is bounded by $R^2/\gamma^2$.

**Reference:**

1. Professor Ross' Slides on Perceptron Algorithm
2. Andrew Ng's CS229 lecture note