Proof of the Perceptron Convergence Theorem

Theorem 1 (Perceptron Convergence Theorem)**.** *If the training data is linearly separable, then the perceptron algorithm converges in a finite number of updates (mistakes). Specifically, the number of updates k satisfies:*

$$k \leq \frac{R^2}{\gamma^2},$$

where:

$R = \max_i \|\mathbf{x}_i\|$ *is the maximum input norm,*

$\gamma = \min_i t_i(\mathbf{w}^{*T}\mathbf{x}_i) > 0$ *is the minimum functional margin,*

$\mathbf{w}^*$ *is a separating weight vector with* $\|\mathbf{w}^*\| = 1$.

Proof

Assume the data is linearly separable. Then there exists a weight vector $\mathbf{w}^*$ such that for all i,

$$t_i(\mathbf{w}^{*T}\mathbf{x}_i) > 0.$$

Without loss of generality, assume $\|\mathbf{w}^*\| = 1$. Define:

$$R = \max_i \|\mathbf{x}_i\|, \quad \gamma = \min_i t_i(\mathbf{w}^{*T}\mathbf{x}_i) > 0.$$

Let $\mathbf{w}_k$ denote the weight vector after the k-th update (i.e., after the k-th misclassification). We initialize $\mathbf{w}_0 = \mathbf{0}$.

Step 1: Growth of Alignment with $\mathbf{w}^*$

Each time a mistake is made on example $(\mathbf{x}, t)$, the update is:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + t\mathbf{x}.$$

Taking the dot product with $\mathbf{w}^*$:

$$\mathbf{w}_{k+1}^T\mathbf{w}^* = \mathbf{w}_k^T\mathbf{w}^* + t\mathbf{x}^T\mathbf{w}^*.$$

Since the data is correctly classified by $\mathbf{w}^*$, we have $t(\mathbf{w}^{*T}\mathbf{x}) \geq \gamma$. Thus:

$$\mathbf{w}_{k+1}^T\mathbf{w}^* \geq \mathbf{w}_k^T\mathbf{w}^* + \gamma.$$

Unrolling this recursion from $\mathbf{w}_0 = \mathbf{0}$:

$$\mathbf{w}_k^T\mathbf{w}^* \geq k\gamma. \tag{1}$$

Step 2: Bounding the Norm of $\mathbf{w}_k$

Now consider the squared norm:

$$\|\mathbf{w}_{k+1}\|^2 = \|\mathbf{w}_k + t\mathbf{x}\|^2 = \|\mathbf{w}_k\|^2 + 2t\mathbf{w}_k^T\mathbf{x} + \|\mathbf{x}\|^2.$$

Since the example was misclassified, $t\mathbf{w}_k^T\mathbf{x} < 0$, so:

$$\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + \|\mathbf{x}\|^2 \leq \|\mathbf{w}_k\|^2 + R^2.$$

Unrolling from $\mathbf{w}_0 = \mathbf{0}$:

$$\|\mathbf{w}_k\|^2 \leq kR^2. \tag{2}$$

Step 3: Apply Cauchy-Schwarz Inequality

By the Cauchy-Schwarz inequality:

$$\mathbf{w}_k^T \mathbf{w}^* \leq \|\mathbf{w}_k\| \cdot \|\mathbf{w}^*\| = \|\mathbf{w}_k\|.$$

From (1), $k\gamma \leq \mathbf{w}_k^T \mathbf{w}^* \leq \|\mathbf{w}_k\|$, so:

$$k\gamma \leq \|\mathbf{w}_k\|.$$

Squaring both sides:

$$k^2\gamma^2 \leq \|\mathbf{w}_k\|^2.$$

Using (2), $\|\mathbf{w}_k\|^2 \leq kR^2$, we get:

$$k^2\gamma^2 \leq kR^2 \quad \Rightarrow \quad k\gamma^2 \leq R^2 \quad \Rightarrow \quad k \leq \frac{R^2}{\gamma^2}.$$

Conclusion

The number of updates (mistakes) k is bounded by $\frac{R^2}{\gamma^2}$, which is finite. Therefore, the perceptron algorithm converges in a finite number of steps when the data is linearly separable. After at most $\left\lfloor \frac{R^2}{\gamma^2} \right\rfloor$ mistakes, the weight vector $\mathbf{w}_k$ will correctly classify all training examples.