Convergence Proof for the Perceptron Algorithm

Michael Collins

Figure 1 shows the perceptron learning algorithm, as described in lecture. In this note we give a convergence proof for the algorithm (also covered in lecture).

The convergence theorem is as follows:

Theorem 1 *Assume that there exists some parameter vector* $\underline{\theta}^*$ *such that* $||\underline{\theta}^*|| = 1$*, and some* $\gamma > 0$ *such that for all* $t = 1 \ldots n$*,*

$$y_t(\underline{x}_t \cdot \underline{\theta}^*) \geq \gamma$$

Assume in addition that for all $t = 1 \ldots n$*,* $||\underline{x}_t|| \leq R$*.*

Then the perceptron algorithm makes at most

$$\frac{R^2}{\gamma^2}$$

errors. (The definition of an error is as follows: an error occurs whenever we have $y' \neq y_t$ *for some* (j, t) *pair in the algorithm.)*

Note that for any vector $\underline{x}$, we use $||\underline{x}||$ to refer to the Euclidean norm of $\underline{x}$, i.e., $||\underline{x}|| = \sqrt{\sum_i x_i^2}$.

Proof: First, define $\underline{\theta}^k$ to be the parameter vector when the algorithm makes its k'th error. Note that we have

$$\underline{\theta}^1 = \underline{0}$$

Next, assuming the k'th error is made on example t, we have

$$\begin{aligned} \underline{\theta}^{k+1} \cdot \underline{\theta}^* &= (\underline{\theta}^k + y_t \underline{x}_t) \cdot \underline{\theta}^* & (1) \\ &= \underline{\theta}^k \cdot \underline{\theta}^* + y_t \underline{x}_t \cdot \underline{\theta}^* & (2) \\ &\geq \underline{\theta}^k \cdot \underline{\theta}^* + \gamma & (3) \end{aligned}$$

Eq. 1 follows by the definition of the perceptron updates. Eq. 3 follows because by the assumptions of the theorem, we have

$$y_t \underline{x}_t \cdot \underline{\theta}^* \geq \gamma$$

Definition: $\text{sign}(z) = 1$ if $z \geq 0$, -1 otherwise.

Inputs: number of iterations, T; training examples $(\underline{x}_t, y_t)$ for $t \in \{1 \ldots n\}$ where $\underline{x} \in \mathbb{R}^d$ is an input, and $y_t \in \{-1, +1\}$ is a label.

Initialization: $\underline{\theta} = \underline{0}$ (i.e., all parameters are set to 0)

Algorithm:

- For $j = 1 \ldots T$
  - For $t = 1 \ldots n$
    1. $y' = \text{sign}(\underline{x}_t \cdot \underline{\theta})$
    2. If $y' \neq y_t$ Then $\underline{\theta} = \underline{\theta} + y_t \underline{x}_t$, Else leave $\underline{\theta}$ unchanged

Output: parameters $\underline{\theta}$

Figure 1: The perceptron learning algorithm.

It follows by induction on k (recall that $||\underline{\theta}^1|| = 0$), that

$$\underline{\theta}^{k+1} \cdot \underline{\theta}^* \geq k\gamma$$

In addition, because $||\underline{\theta}^{k+1}|| \times ||\underline{\theta}^*|| \geq \underline{\theta}^{k+1} \cdot \underline{\theta}^*$, and $||\underline{\theta}^*|| = 1$, we have

$$||\underline{\theta}^{k+1}|| \geq k\gamma \tag{4}$$

In the second part of the proof, we will derive an upper bound on $||\underline{\theta}^{k+1}||$. We have

$$\begin{aligned} ||\underline{\theta}^{k+1}||^2 &= ||\underline{\theta}^k + y_t \underline{x}_t||^2 && (5) \\ &= ||\underline{\theta}^k||^2 + y_t^2 ||\underline{x}_t||^2 + 2 y_t \underline{x}_t \cdot \underline{\theta}^k && (6) \\ &\leq ||\underline{\theta}^k||^2 + R^2 && (7) \end{aligned}$$

The equality in Eq. 5 follows by the definition of the perceptron updates. Eq. 7 follows because we have: 1) $y_t^2 ||\underline{x}_t||^2 = ||\underline{x}_t||^2 \leq R^2$ by the assumptions of the theorem, and because $y_t^2 = 1$; 2) $y_t \underline{x}_t \cdot \underline{\theta}^k \leq 0$ because we know that the parameter vector $\underline{\theta}^k$ gave an error on the t^{th} example.

It follows by induction on k (recall that $||\underline{\theta}^1||^2 = 0$), that

$$||\underline{\theta}^{k+1}||^2 \leq kR^2 \tag{8}$$

Combining the bounds in Eqs. 4 and 8 gives

$$k^2\gamma^2 \leq ||\underline{\theta}^{k+1}||^2 \leq kR^2$$

from which it follows that

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

□