Perceptron Algorithm

Narayana Santhanam

January 17, 2020

We adopt the convention that

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ -1 & x \le 0 \end{cases}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)}$ be the training examples in $\mathbb{R}^d$ (namely, each example is a vector with d real coordinates. Each $\mathbf{z}^{(i)}$ is normalized to length 1, namely $||\mathbf{z}^{(i)}|| = 1$ for all i. We are given the labels for the training examples, let them be $y_1, \ldots, y_T$.

The examples are linearly separable, namely there is a vector $\mathbf{w}^* \in \mathbb{R}^d$ such that for all $1 \le i \le T$,

$$\text{sign}(\mathbf{w}^* \cdot \mathbf{z}^{(i)}) = y_i.$$

For convenience, we set $||\mathbf{w}^*|| = 1$.

The perceptron training algorithm

Input: Training examples $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)}$ and labels $y_1, \ldots, y_T$
0. Initialize $\mathbf{w}^{(1)} = 0$
1. for $i = 1, \ldots, T$:
2. Estimate $\hat{y}_i = \text{sign}(\mathbf{w}^{(i)} \cdot \mathbf{z}^{(i)})$
3. if $\hat{y}_i = y_i$:
4. $\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)}$ (no error)
5. else:
6. $\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} + y_i \mathbf{z}^{(i)}$ (error)

Output: $\mathbf{w}^{(T+1)}$, best estimate of the separating hyperplane

---

We say we make an error in the i'th step if we reach the else statement (line 6) in the i'th iteration of the for loop above.

Theorem 1 Let

$$\gamma = \min_{1\leq i\leq T} \left| \frac{\mathbf{w}^* \cdot z^{(i)}}{||\mathbf{w}^*||} \right| = \min_{1\leq i\leq T} \left| \mathbf{w}^* \cdot z^{(i)} \right| \text{ (since } ||\mathbf{w}^*|| = 1). \tag{1}$$

Then the Perceptron Training Algorithm makes $\leq \frac{1}{\gamma^2}$ errors.

We prove this theorem using the following steps:

Lemma 2 For every example, $\text{sign}(\mathbf{w}^* \cdot \mathbf{z}^{(i)})y_i > 0$. In every step that we make an error (namely, reach the else statement in the algorithm), we will have $\text{sign}(\mathbf{w}^{(i)} \cdot \mathbf{z}^{(i)})y_i < 0$.
Proof Clear from the definitions of $\mathbf{w}^*$ and because we enter the else statement only when $\text{sign}(\mathbf{w}^{(i)} \cdot \mathbf{z}^{(i)}) \neq y_i$.

Lemma 3 In step i, if we make an error, then

$$\mathbf{w}^{(i+1)} \cdot \mathbf{w}^* \leq \mathbf{w}^{(i)} \cdot \mathbf{w}^* + \gamma.$$

Proof

$$\begin{aligned}
\mathbf{w}^{(i+1)} \cdot \mathbf{w}^* &\overset{(a)}{=} (\mathbf{w}^{(i)} + y_i z^{(i)}) \cdot \mathbf{w}^* \\
&\overset{(b)}{=} \mathbf{w}^{(i)} \cdot \mathbf{w}^* + y_i \mathbf{w}^* \cdot \mathbf{z}^{(i)} \\
&\overset{(c)}{\geq} \mathbf{w}^{(i)} \cdot \mathbf{w}^* + \gamma
\end{aligned}$$

where (a) follows because we reach the else line if there is an error in our prediction during step i, (b) because the dot product distributes over addition, and (c) from Lemma 2 and from Equation (1).

Lemma 4 In step i, if we make an error, then

$$||\mathbf{w}^{(\mathrm{i+1})}||^2 \leq ||\mathbf{w}^{(\mathrm{i})}||^2 + 1.$$

Proof

$$\begin{aligned}
||\mathbf{w}^{(i+1)}||^2 &\stackrel{(a)}{=} ||\mathbf{w}^{i)} + y_i\mathbf{z}^{(i)}||^2 \\
&\stackrel{(b)}{=} (\mathbf{w}^{i)} + y_i\mathbf{z}^{(i)}) \cdot (\mathbf{w}^{i)} + y_i\mathbf{z}^{(i)}) \\
&\stackrel{(c)}{=} \mathbf{w}^{(i)} \cdot \mathbf{w}^{(i)} + 2y_i\mathbf{w}^{(i)} \cdot \mathbf{z}^{(i)} + (y_i\mathbf{z}^{(i)}) \cdot (y_i\mathbf{z}^{(i)}) \\
&= ||\mathbf{w}^{(i)}||^2 + 2y_i\mathbf{w}^{(i)} \cdot \mathbf{z}^{(i)} + ||y_i\mathbf{z}^{(i)})||^2 \\
&\stackrel{(d)}{\leq} ||\mathbf{w}^{(i)}||^2 + ||y_i\mathbf{z}^{(i)})||^2 \\
&= ||\mathbf{w}^{(i)}||^2 + 1
\end{aligned}$$

where (a) follows because we reach the else statement of the algorithm if there is an error in the i'th step, (b) because the length of a vector squared equals the dot product of the vector with itself, (c) because the dot product distributes over addition, (d) from Lemma 2 which asserts $y_i\mathbf{w}^{(i)} \cdot z^{(i)} < 0$ whenever there is an error in the i'th step and the last step because the vector $\mathbf{z}^{(i)}$ has length 1, and y_i is just ± 1.

Proof of the Theorem. We put all the Lemmas together.

Suppose we make M errors in the T iterations of the for loop. Using Lemma 3 and because $\mathbf{w}^{(1)} = 0$, we have

$$\mathbf{w}^{(T+1)} \cdot \mathbf{w}^* \geq M\gamma. \tag{2}$$

Using Lemma 4 and because $\mathbf{w}^{(1)} = 0$, we have

$$||\mathbf{w}^{(T+1)}||^2 \leq M. \tag{3}$$

But we know

$$\mathbf{w}^{(T+1)} \cdot \mathbf{w}^* = ||\mathbf{w}^{(T+1)}|| ||\mathbf{w}^*|| \cos\theta,$$

where θ is the angle between $\mathbf{w}^{(T+1)}$ and $\mathbf{w}^*$. Because $\cos\theta \leq 1$ no matter what θ is, we have

$$\mathbf{w}^{(T+1)} \cdot \mathbf{w}^* \leq ||\mathbf{w}^{(T+1)}|| ||\mathbf{w}^*||. \tag{4}$$

The equation above is very important and is an instance of the Cauchy Schwartz inequality. But combining Equations (2), (3) and (4), we have

$$M\gamma \leq \mathbf{w}^{(T+1)} \cdot \mathbf{w}^* \leq ||\mathbf{w}^{(T+1)}|| ||\mathbf{w}^*|| \leq \sqrt{M} \cdot 1,$$

or that $M\gamma \leq \sqrt{M}$. This of course implies

$$M \leq \frac{1}{\gamma^2}$$

as we wanted to show.