Machine Learning Theory
Lecture 7

Nicholas Harvey

October 9, 2018

1 Basic Perceptron

Algorithm 1 Perceptron algorithm.

> **procedure** PERCEPTRON($(x_1, y_1), ..., (x_m, y_m)$)
>   Initialize $w_0 = 0$ and $t = 0$
>   **repeat**
>   **if** there exists i with $y_i \neq \operatorname{sign}(\langle\, w_t,\, x_i \,\rangle)$ **then**
> $w_{t+1} \leftarrow w_t + y_i x_i$
> $t \leftarrow t + 1$
>   **end if**
>   **until** no such i exists
>   **return** w_t

Let's define margin(w) to be $\min_i |\langle\, w,\, x_i \,\rangle| / \|w\|$. Note that $\langle\, w,\, x_i \,\rangle$ is the length of the orthogonal projection of x_i onto the subspace $\{\, x \,:\, \langle\, w,\, x \,\rangle = 0 \,\}$. Alternatively, it is the cosine of the angle between w and x_i.

Theorem 1.1. Let w^* be a consistent linear classifier with $\|w^*\| = 1$ such that $\gamma := \operatorname{margin}(w^*)$ is maximized. Then Algorithm 1 terminates after at most $1/\gamma^2$ iterations.

This is a nice result because it doesn't depend on the dimensionality of the data, just on the geometric margin properties of the data.

Why are the Perceptron updates a good idea? Suppose $y_i = 1$ but $\langle\, w_t,\, x_i \,\rangle < 0$. Then

$$\langle\, w_{t+1},\, x_i \,\rangle \;=\; \langle\, w_t + x_i,\, x_i \,\rangle \;=\; \langle\, w_t,\, x_i \,\rangle + \underbrace{\langle\, x_i,\, x_i \,\rangle}_{=1}$$

So the inner product between the solution and this example improves by 1, which seems good.

The formal analysis of the algorithm argues that w_t gets "closer" to w^*.

Claim 1.2. $\langle\, w_{t+1},\, w^* \,\rangle \geq \langle\, w_t,\, w^* \,\rangle + \gamma$.

Proof. Suppose $y_i = 1$ but $\langle\, w_t,\, x_i \,\rangle < 0$. Then

$$\langle\, w_{t+1},\, w^* \,\rangle \;=\; \langle\, w_t + x_i,\, w^* \,\rangle \;=\; \langle\, w_t,\, w^* \,\rangle + \underbrace{\langle\, x_i,\, w^* \,\rangle}_{\geq\gamma}.$$

The argument is similar if $y_i = -1$. ■

But this does not really show that w_{t+1} gets closer to w^*. w_{t+1} could be "cheating" by just increasing its norm. The next claim rules out excessive cheating.

Claim 1.3. $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$.

Proof. Again, suppose $y_i = 1$ but $\langle\, w_t,\, x_i \,\rangle < 0$. Then

$$\langle\, w_{t+1},\, w_{t+1} \,\rangle \;=\; \langle\, w_t + x_i,\, w_t + x_i \,\rangle \;=\; \langle\, w_t,\, w_t \,\rangle + 2\underbrace{\langle\, w_t,\, x_i \,\rangle}_{<0} + \underbrace{\langle\, x_i,\, x_i \,\rangle}_{=1} \;\leq\; \langle\, w_t,\, w_t \,\rangle + 1.$$

■

Proof (of Theorem 1.1). By induction, Claim 1.2 gives that $\langle\, w_t,\, w^* \,\rangle \geq t \cdot \gamma$. By induction, Claim 1.3 gives that $\|w_t\| \leq \sqrt{t}$. How can we combine these two? How can we relate inner products and norms? Cauchy-Schwarz of course.

$$\begin{aligned} & t \cdot \gamma \;\leq\; \langle\, w_t,\, w^* \,\rangle \;\leq\; \|w_t\|\,\|w^*\| \;\leq\; \sqrt{t} \\ \Longrightarrow\quad & \sqrt{t} \;\leq\; 1/\gamma \\ \Longrightarrow\quad & t \;\leq\; 1/\gamma^2. \end{aligned}$$

■

2 Margin Perceptron

The analysis of Algorithm 1 is elegant, but unsatisfying in one way. The hypothesis of the theorem is that there is a hypothesis with large margin. The hypothesis output by the algorithm is guaranteed to correctly classify all points, but there are no guarantees about its margin.

It turns out that we can modify the algorithm in a simple way so that we can analyze the margin too. See Algorithm 2. Roughly, any point that has small margin with respect to the current hypothesis is treated the same as a misclassified point.

Theorem 2.1. Suppose there is a hypothesis w^* with margin at least γ. Then Algorithm 2 outputs a classifier with margin at least $\gamma/3$ after at most $3/\gamma^2$ iterations.

As before, we may assume that $\|w^*\| = 1$.

Claim 2.2. $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 3$.

Proof. Again assume $y_i = 1$. As before,

$$\|w_{t+1}\|^2 \;=\; \|w_t\|^2 + 2\langle\, w_t,\, x_i \,\rangle + \|x_i\|^2$$

Now, either x_i was a misclassification, in which case $\langle\, w_t,\, x_i \,\rangle < 0$, or it had poor margin, in which case $\langle\, w_t,\, x_i \,\rangle \leq 1$. In either case, we have $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 3$. ■

Algorithm 2 The Margin-Perceptron algorithm.

```
procedure MarginPerceptron((x_1, y_1), ..., (x_m, y_m))
    Initialize w_1 = 0 and t = 1
    repeat
        Find any i with
            (Misclassification)   y_i ≠ sign(⟨ w_t, x_i ⟩)
            (Poor margin)         |⟨ w_t, x_i ⟩| ≤ 1
        if such an i exists then
            w_{t+1} ← w_t + y_i x_i
            t ← t + 1
        end if
    until no such i exists
    return w_t
```

Thus, by induction, the cumulative increase in the norm of w_t is

$$\|w_t\|^2 \leq 3t \qquad \Longrightarrow \qquad \|w_t\| \leq \sqrt{3t}.$$

Proof (of Theorem 2.1).

Number of iterations. Claim 1.2 holds without change so we have $\langle\, w_t,\, w^* \,\rangle \geq t \cdot \gamma$, as in Theorem 1.1. The bound on the number of iterations is similar:

$$\begin{aligned} & t \cdot \gamma \;\leq\; \langle\, w_t,\, w^* \,\rangle \;\leq\; \|w_t\|\,\|w^*\| \;\leq\; \sqrt{3t} \\ \Longrightarrow \quad & \sqrt{t} \;\leq\; \sqrt{3}/\gamma \\ \Longrightarrow \quad & t \;\leq\; 3/\gamma^2. \end{aligned}$$

Margin. The output classifier w has $|\langle\, w,\, x_i \,\rangle| > 1$ for each i. So

$$\text{margin}(w) \;=\; \min_i \frac{|\langle\, w,\, x_i \,\rangle|}{\|w\|} \;>\; \frac{1}{\|w\|} \;\geq\; \frac{1}{\sqrt{3t}} \;\geq\; \frac{1}{\sqrt{3 \cdot (3/\gamma^2)}} \;=\; \gamma/3.$$

■