

# Perceptron and Online Perceptron

Daniel Hsu (COMS 4771)

## Margins

Let  $S$  be a collection of labeled examples from  $\mathbb{R}^d \times \{-1, +1\}$ . We say  $S$  is *linearly separable* if there exists  $w \in \mathbb{R}^d$  such that

$$\min_{(x,y) \in S} y \langle w, x \rangle > 0,$$

and we call  $w$  a *linear separator* for  $S$ .

The (*minimum*) *margin* of a linear separator  $w$  for  $S$  is the minimum distance from  $x$  to the hyperplane orthogonal to  $w$ , among all  $(x, y) \in S$ . Note that this notion of margin is invariant to positive scaling of  $w$ . If we rescale  $w$  so that

$$\min_{(x,y) \in S} y \langle w, x \rangle = 1,$$

then this minimum distance is  $1/\|w\|_2$ . Therefore, the linear separator with the largest minimum margin is described by the following maximization problem:

$$\max_w \min_{(x,y) \in S} y \langle w, x \rangle$$

$$y \langle w, x \rangle \geq 1, \quad (x, y) \in S.$$

## Perceptron algorithm

The Perceptron algorithm is given as follows. The input to the algorithm is a collection  $S$  of labeled examples from  $\mathbb{R}^d \times \{-1, +1\}$ .

- Begin with  $\hat{w}_1 := 0 \in \mathbb{R}^d$ .
- For  $t = 1, 2, \dots$ :
  - If there is a labeled example in  $S$  (call it  $(x_t, y_t)$ ) such that  $y_t \langle \hat{w}_t, x_t \rangle \leq 0$ , then set  $\hat{w}_{t+1} := \hat{w}_t + y_t x_t$ .
  - Else, return  $\hat{w}_t$ .

**Theorem.** Let  $S$  be a collection of labeled examples from  $\mathbb{R}^d \times \{-1, +1\}$ . Suppose there exists a vector  $w_\star \in \mathbb{R}^d$  such that

$$\min_{(x,y) \in S} y \langle w_\star, x \rangle = 1.$$

Then Perceptron on input  $S$  halts after at most  $\|w_\star\|_2^2 L^2$  loop iterations, where  $L := \max_{(x,y) \in S} \|x\|_2$ .

*Proof.* Suppose Perceptron does not exit the loop in the  $t$ -th iteration. Then there is a labeled example  $(x_t, y_t) \in S$  such that

$$y_t \langle w_\star, x_t \rangle \geq 1,$$

$$y_t \langle \hat{w}_t, x_t \rangle \leq 0.$$

We bound  $\langle w_\star, w_{t+1} \rangle$  from above and below to deduce a bound on the number of loop iterations. First, we bound  $\langle w_\star, \hat{w}_t \rangle$  from below:

$$\langle w_\star, \hat{w}_{t+1} \rangle = \langle w_\star, \hat{w}_t \rangle + y_t \langle w_\star, x_t \rangle \geq \langle w_\star, \hat{w}_t \rangle + 1.$$

Since  $\hat{w}_1 = 0$ , we have

$$\langle w_*, \hat{w}_{t+1} \rangle \geq t.$$

We now bound  $\langle w_*, \hat{w}_{t+1} \rangle$  from above. By Cauchy-Schwarz,

$$\langle w_*, \hat{w}_{t+1} \rangle \leq \|w_*\|_2 \|\hat{w}_{t+1}\|_2.$$

Also,

$$\|\hat{w}_{t+1}\|_2^2 = \|\hat{w}_t\|_2^2 + 2y_t \langle \hat{w}_t, x_t \rangle + y_t^2 \|x_t\|_2^2 \leq \|\hat{w}_t\|_2^2 + L^2.$$

Since  $\|\hat{w}_1\|_2 = 0$ , we have

$$\|\hat{w}_{t+1}\|_2^2 \leq L^2 t,$$

so

$$\langle w_*, \hat{w}_{t+1} \rangle \leq \|w_*\|_2 L \sqrt{t}.$$

Combining the upper and lower bounds on  $\langle w_*, \hat{w}_{t+1} \rangle$  shows that

$$t \leq \langle w_*, \hat{w}_{t+1} \rangle \leq \|w_*\|_2 L \sqrt{t},$$

which in turn implies the inequality  $t \leq \|w_*\|_2^2 L^2$ . ■

## Online Perceptron algorithm

The Online Perceptron algorithm is given as follows. The input to the algorithm is a sequence  $(x_1, y_1), (x_2, y_2), \dots$  of labeled examples from  $\mathbb{R}^d \times \{-1, +1\}$ .

- Begin with  $\hat{w}_1 := 0$
- For  $t = 1, 2, \dots$ :
  - If  $y_t \langle \hat{w}_t, x_t \rangle \leq 0$
  - Else,  $\hat{w}_{t+1} := \hat{w}_t + y_t x_t$

We say that Online Perceptron makes a *mistake* in round  $t$

**Theorem.** Let  $(x_1, y_1), (x_2, y_2), \dots$  be a sequence of labeled examples from  $\mathbb{R}^d \times \{-1, +1\}$  such that there exists a vector  $w_* \in \mathbb{R}^d$  satisfying

$$\min_{t=1,2,\dots} y_t \langle w_*, x_t \rangle = 1.$$

Then Online Perceptron on input  $(x_1, y_1), (x_2, y_2), \dots$  makes at most  $\|w_*\|_2^2 L^2$  mistakes, where  $L := \max_{t=1,2,\dots} \|x_t\|_2$ .

*Proof.* The proof of this theorem is essentially the same as the proof of the iteration bound for Perceptron. ■

Online Perceptron may be applied to a collection of labeled examples  $S$  by considering the labeled examples in  $S$  in any (e.g., random) order. If  $S$  is linearly separable, then the number of mistakes made by Online Perceptron can be bounded using the theorem.

However, Online Perceptron is also useful when  $S$  is not linearly separable. This is especially notable in comparison to Perceptron, which never terminates if  $S$  is not linearly separable.

**Theorem.** Let  $(x_1, y_1), (x_2, y_2), \dots$  be a sequence of labeled examples from  $\mathbb{R}^d \times \{-1, +1\}$ . Online Perceptron on input  $(x_1, y_1), (x_2, y_2), \dots$  makes at most

$$\min_{w_* \in \mathbb{R}^d} \left[ \|w_*\|_2^2 L^2 + \|w_*\|_2 L \sqrt{\sum_{t \in \mathcal{M}} \ell(\langle w_*, x_t \rangle, y_t)} + \sum_{t \in \mathcal{M}} \ell(\langle w_*, x_t \rangle, y_t) \right]$$

mistakes, where  $L := \max_{t=1,2,\dots} \|x_t\|_2$ ,  $\mathcal{M}$  is the set of rounds on which Online Perceptron makes a mistake, and  $\ell(\hat{y}, y) := [1 - \hat{y}y]_+ = \max\{0, 1 - \hat{y}y\}$  is the *hinge loss* of  $\hat{y}$  when  $y$  is the correct label.

*Proof.* Fix any  $w_\star \in \mathbb{R}^d$ . Consider any round  $t$  in which Online Perceptron makes a mistake. Let  $\mathcal{M}_t := \{1, \dots, t\} \cap \mathcal{M}$  and  $M_t := |\mathcal{M}_t|$ . We will bound  $\langle w_\star, \hat{w}_{t+1} \rangle$  from above and below to deduce a bound on  $M_t$ , the number of mistakes made by Online Perceptron through the first  $t$  rounds. First we bound  $\langle w_\star, \hat{w}_{t+1} \rangle$  from above. By Cauchy-Schwarz,

$$\langle w_\star, \hat{w}_{t+1} \rangle \leq \|w_\star\|_2 \|\hat{w}_{t+1}\|_2.$$

Moreover,

$$\|\hat{w}_{t+1}\|_2^2 = \|\hat{w}_t\|_2^2 + 2y_t \langle \hat{w}_t, x_t \rangle + y_t^2 \|x_t\|_2^2 \leq \|\hat{w}_t\|_2^2 + L^2.$$

Since  $\hat{w}_1 = 0$ , we have

$$\|\hat{w}_{t+1}\|_2^2 \leq L^2 M_t,$$

and thus

$$\langle w_\star, \hat{w}_{t+1} \rangle \leq \|w_\star\|_2 L \sqrt{M_t}.$$

We now bound  $\langle w_\star, \hat{w}_{t+1} \rangle$  from below:

$$\begin{aligned} \langle w_\star, \hat{w}_{t+1} \rangle &= \langle w_\star, \hat{w}_t \rangle + 1 - [1 - y_t \langle w_\star, x_t \rangle] \\ &\geq \langle w_\star, \hat{w}_t \rangle + 1 - [1 - y_t \langle w_\star, x_t \rangle]_+ \\ &= \langle w_\star, \hat{w}_t \rangle + 1 - \ell(\langle w_\star, x_t \rangle, y_t), \end{aligned}$$

Since  $\hat{w}_1 = 0$ ,

where  $\langle w_\star, \hat{w}_{t+1} \rangle \geq M_t - H_t$ , **Assignment Project Exam Help**

Combining the upper and lower

**<https://eduassistpro.github.io/>**  
i.e., **Add WeChat edu\_assist\_pro**

This inequality is quadratic in  $\sqrt{M_t}$ . By solving it<sup>1</sup>, we deduce the bound

$$M_t \leq \frac{1}{2} \|w_\star\|_2^2 L^2 + \frac{1}{2} \|w_\star\|_2 L \sqrt{\|w_\star\|_2^2 L^2 + 4H_t} + H_t,$$

which can be further loosened to the following (slightly more interpretable) bound:

$$M_t \leq \|w_\star\|_2^2 L^2 + \|w_\star\|_2 L \sqrt{H_t} + H_t.$$

The claim follows. ■

---

<sup>1</sup>The inequality is of the form  $x^2 - bx - c \leq 0$  for some non-negative  $b$  and  $c$ . This implies that  $x \leq (b + \sqrt{b^2 + 4c})/2$ , and hence  $x^2 \leq (b^2 + 2b\sqrt{b^2 + 4c} + 4c + b^2 + 4c)/4$ . We can then use the fact that  $\sqrt{A+B} \leq \sqrt{A} + \sqrt{B}$  for any non-negative  $A$  and  $B$  to deduce  $x^2 \leq b^2 + b\sqrt{c} + c$ .