

**Solutions to**  
**Understanding Machine Learning**  
**by Shai Shalev-Shwartz and Shai Ben-David**

**Chapter 9 (Linear Predictors)**

**Zahra Taheri<sup>1</sup>**

Apr 2020

**Exercise 9.1**

Firstly, we prove that for all  $c \in \mathbb{R}$ ,  $|c| = \min_{a \geq 0} a$  such that  $-a \leq c \leq a$ . If  $c \geq 0$  then obviously  $c = |c| = \min_{a \geq 0} a$  such that  $0 \leq c \leq a$ . If  $c < 0$  then  $-c > 0$  and with the latter discussion,  $-c = |c| = |-c| = \min_{a \geq 0} a$  such that  $0 \leq -c \leq a$ . Therefore, for all  $c \in \mathbb{R}$ ,  $|c| = \min_{a \geq 0} a$  such that  $-a \leq c \leq a$ . So, for all  $i \in [m]$ , if we consider  $a_i = |\langle w, x_i \rangle - y_i|$ , then we have

$$-a_i \leq \langle w, x_i \rangle - y_i \leq a_i.$$

Therefore, we want to write the following problem as a linear program:

$$\min \sum_{i=1}^m a_i \text{ s.t. } \langle w, x_i \rangle - a_i \leq y_i \text{ and } -\langle w, x_i \rangle - a_i \leq -y_i, \text{ for all } i \in [m]. \quad (1)$$

Let

$$A = \begin{pmatrix} x_1^T & & \\ & \vdots & -I_m \\ x_m^T & & \\ -x_1^T & & \\ & \vdots & -I_m \\ -x_m^T & & \end{pmatrix}, w' = (w_1, \dots, w_d, a_1, \dots, a_m)^T, y = (y_1, \dots, y_m, -y_1, \dots, -y_m)^T.$$

Then by (1) and with the above notations we have

$$Aw' \leq y.$$

---

<sup>1</sup><https://github.com/zahta/Exercises-Understanding-Machine-Learning>

Therefore, we can write the ERM problem of linear regression with respect to the absolute value loss function as the following linear program:

$$\min aw'Aw' \leq y,$$

where  $a = (\overbrace{0, \dots, 0}^d, \overbrace{1, \dots, 1}^m)^T$ .

### Exercise 9.3

Let  $m$  be an arbitrary positive integer. Suppose that  $d = m$  and for all  $i \in [m]$ ,  $x_i = e_i$  and  $y_i = 1$ . Then  $R = \max_i \|x_i\| = 1$ . If  $w^* = (1, \dots, 1)^T$ , then for all  $i \in [m]$ ,  $y_i \langle w^*, x_i \rangle = 1$  and  $\|w^*\|^2 = m$ , and so, using the notation in Theorem 9.1, we have  $B = \min\{\|w\| : \forall i \in [m], y_i \langle w, x_i \rangle \geq 1\} = \sqrt{m}$ . Therefore,  $(RB)^2 = m$  and by Theorem 9.1, the Perceptron algorithm stops after at most  $m$  iterations.

On the other hand by the Perceptron algorithm, we have

$$w^{(1)} = 0w^{(2)} = e_1w^{(3)} = e_1 + e_2 \dots w^{(m)} = \sum_{i=1}^{m-1} e_i w^{(m+1)} = (1, \dots, 1)^T = w^*.$$

If we assume that  $\text{sign}(0) = -1$  then for all  $i \in [m]$ ,  $\langle w^{(i)}, x_i \rangle = 0$  and so  $\text{sign}(\langle w^{(i)}, x_i \rangle) \neq y_i$ . Also by Theorem 9.1, when the Perceptron algorithm stops it holds that for all  $i \in [m]$ ,  $y_i \langle w^{(t)}, x_i \rangle > 0$ . So the Perceptron algorithm stops after  $m$  iterations and then we obtain  $w^* = (1, \dots, 1)^T$  with the desired condition  $y_i \langle w^*, x_i \rangle = 1$ , for all  $i \in [m]$ .

### Exercise 9.4

Suppose that  $x_1, \dots, x_m$  are all positive examples such that for all  $i \in [m]$ ,  $x_i = (a, b, 1)$  and  $\|x_i\|^2 = R^2$ . Since the examples have the same labels,  $(x_1, 1), \dots, (x_m, 1)$  is separable. Let  $w^* = (0, 0, 1)$ . Then obviously  $y_i \langle w^*, x_i \rangle \geq 1$ , for all  $i \in [m]$  and we have  $B = \|w^*\| = 1$ . Then, the Perceptron algorithm stops after at most  $(RB)^2 = R^2$  iterations, and when it stops it holds that for all  $i \in [m]$ ,  $\langle w^{(t)}, x_i \rangle > 0$ .

Now we want to construct  $m = R^2$  positive examples on which the upper bound of Theorem 9.1 equals  $R^2$  and the perceptron algorithm is bound to make  $R^2$  mistakes. Let  $x_1 := (a_1, 0, 1)$  such that  $a_1 = \sqrt{R^2 - 1}$ . Suppose that in iteration  $t$  of the Perceptron algorithm, the new example  $x_t = (a, b, 1)$  be such that

- (1)  $\|x_t\|^2 = R^2$ , and
- (2)  $\langle w^{(t)}, x_t \rangle = 0$ .

We want to prove that as long as  $t \leq R^2$ , we can meet the above two desired conditions.

Let  $t \leq R^2$ . It is easy to see that if the above two conditions hold, for some  $\alpha, \beta \in \mathbb{R}$  we have

$$w^{(t)} = (\alpha, \beta, t-1). \quad (1)$$

Also by condition  $\langle w^{(t)}, x_t \rangle = 0$ , the inequality (9.4) in the proof of Theorem 9.1 holds with equality and we have

$$\|w^{(t)}\|^2 = \|w^{(t-1)}\|^2 + R^2.$$

Now since  $\|w^{(1)}\|^2 = 0$ , if we use above equation recursively for  $t$  iterations, we obtain that

$$\|w^{(t)}\|^2 = (t-1)R^2. \quad (2)$$

Therefore, by equations (1) and (2) we have  $\alpha^2 + \beta^2 + (t-1)^2 = (t-1)R^2$ . If  $\beta \neq 0$ , then without loss of generality we may rotate  $w^{(t)}$  with respect to the  $z$  axis and consider  $w^{(t)} = (\alpha, 0, t-1)$ . Therefore, with the discussion above we have  $\alpha = \sqrt{(t-1)R^2 - (t-1)^2}$  and so  $w^{(t)} = (\sqrt{(t-1)R^2 - (t-1)^2}, 0, t-1)$ . Now let  $a := -\frac{t-1}{\alpha}$  and  $x_t := (a, b, 1)$ , where  $b \in \mathbb{R}$ . Then it is easy to see that  $\langle w^{(t)}, x_t \rangle = 0$ . Therefore, if  $b := \sqrt{R^2 - a^2 - 1}$ , then  $\|x_t\|^2 = R^2$  and so, the two desired conditions hold. We just need to prove that  $R^2 - a^2 - 1 \geq 0$  or equivalently  $a^2 + 1 \leq R^2$ . Since  $t \leq R^2$ , we have  $R^2 - t + 1 \geq 1$  and so

$$a^2 + 1 = \frac{(t-1)^2}{\alpha^2} + 1 = \frac{(t-1)R^2}{(t-1)R^2 - (t-1)^2} = \frac{R^2}{R^2 - t + 1} \leq R^2.$$

This completes the proof.

### Exercise 9.6

1. Suppose that  $\mathcal{B}_d = \{B_{v,r} : v \in \mathbb{R}^d \text{ and } r > 0\}$  such that  $B_{v,r}$  is a closed ball of center  $v$  and radius  $r$  in  $\mathbb{R}^d$ , i.e.,

$$B_{v,r}(x) = \begin{cases} 1 & \text{if } \|x - v\| \leq r \\ 0 & \text{otherwise} \end{cases}$$

Also suppose that  $C = \{x_1, \dots, x_m\}$  is shattered by  $\mathcal{B}_d$ . If  $y = (y_1, \dots, y_m)$  is an arbitrary vector of labels of  $C$ , then there exists  $B_{v,r} \in \mathcal{B}_d$  such that for all  $i \in [m]$  we have  $B_{v,r}(x_i) = y_i$ . By definition of

$B_{v,r}$ ,  $y_i = 1$  for some  $i \in [m]$ , if and only if  $\|x_i - v\| \leq r$  if and only if

$$\sum_{j=1}^d x_{ij}^2 - 2 \sum_{j=1}^d x_{ij} v_j + \sum_{j=1}^d v_j^2 - r^2 \leq 0,$$

where  $x_i = (x_{i1}, \dots, x_{id})^T$  and  $v = (v_1, \dots, v_d)^T$ , or equivalently

$$\langle w, x \rangle + b \leq 0,$$

where  $w := (-2v_1, \dots, -2v_d, 1)^T$ ,  $x := (x_{i1}, \dots, x_{id}, \|x_i\|^2)^T$  and  $b := \sum_{j=1}^d v_j^2 - r^2$ .

Let  $\mathcal{L}_{d+1}$  be the class of halfspaces in  $\mathbb{R}^{d+1}$  and consider the mapping  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d+1}$  defined by  $\phi(a) = (a_1, \dots, a_d, \|a\|^2)^T$ . With the discussion above and the assumption  $\text{sign}(0) = 1$ , if  $x_1, \dots, x_m$  are shattered by  $\mathcal{B}_d$  then  $\phi(x_1), \dots, \phi(x_m)$  are shattered by  $\mathcal{L}_{d+1}$ . Therefore,

$$VCdim(\mathcal{B}_d) \leq VCdim(\mathcal{L}_{d+1}) = d + 2.$$

2. Let  $C = \{0, e_1, \dots, e_d\}$ . We want to show that if  $\emptyset \neq C' \subseteq C$ , then there exists  $B_{v,r} \in \mathcal{B}_d$  such that for all  $x \in C'$ ,  $B_{v,r}(x) = 1$ , and for all  $x \in C \setminus C'$ ,  $B_{v,r}(x) = 0$ .

Let  $v := \sum_{x \in C'} x$ . Therefore,

$$\|x - v\| = \begin{cases} \sqrt{|C'| - 1} & \text{if } x \in C' \text{ and } x \neq 0 \\ \sqrt{|C'|} & \text{if } x = 0 \\ \sqrt{|C'| + 1} & \text{if } x \in C \setminus C' \text{ and } x \neq 0 \end{cases}$$

If  $0 \notin C'$ , then let  $r := \sqrt{|C'| - 1}$ . Otherwise, let  $r := \sqrt{|C'|}$ . With the discussion above, it is easy to see that  $C$  is shattered by  $\mathcal{B}_d$  and so  $d + 1 \leq VCdim(\mathcal{B}_d)$ .

Therefore by parts 1 and 2 we have

$$d + 1 \leq VCdim(\mathcal{B}_d) \leq d + 2.$$