Perceptron:

update algo: $w(t) = w(t-1) + y(t-1)x(t-1)$

Exercise 1.3

The weight update rule in (1.3) has the nice interpretation that it moves in the direction of classifying $\mathbf{x}(t)$ correctly.

(a) Show that $y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) < 0$. *[Hint: $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$.]*

(b) Show that $y(t)\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)$. *[Hint: Use (1.3).]*

(c) As far as classifying $\mathbf{x}(t)$ is concerned, argue that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move 'in the right direction'.

a. $\left.\begin{matrix} w^T(t)\,x(t) \\ y(t) \end{matrix}\right\} \rightarrow$ both are of different signs

b. $w^T(t+1) = w^T(t) + y(t)\,x^T(t)$

$y(t)\,w^T(t+1)\,x(t) = y(t)\,w^T(t)\,x(t) + \underbrace{y(t)\,y(t)}_{>0}\,\underbrace{x^T(t)\,x(t)}_{>0}$

$\rightarrow y(t)\,w^T(t+1)\,x(t) > y(t)\,w^T(t)\,x(t)$

c. let's say actual label is '-1'. & $w^T(t)\,x(t) > 0$

from 'b', we see that the value of $w^T(t+1)\,x(t) < w^T(t)\,x(t)$, that means we are approaching to the negative side by updating w.

Ily we can argue if actual label is +1.

Problem 1.3 Prove that the PLA eventually converges to a linear separator for separable data. The following steps will guide you through the proof. Let $\mathbf{w}^*$ be an optimal set of weights (one which separates the data). The essential idea in this proof is to show that the PLA weights $\mathbf{w}(t)$ get "more aligned" with $\mathbf{w}^*$ with every iteration. For simplicity, assume that $\mathbf{w}(0) = \mathbf{0}$.

(a) Let $\rho = \min_{1 \le n \le N} y_n(\mathbf{w}^{*\mathsf{T}}\mathbf{x}_n)$. Show that $\rho > 0$.

(b) Show that $\mathbf{w}^{\mathsf{T}}(t)\mathbf{w}^* \ge \mathbf{w}^{\mathsf{T}}(t-1)\mathbf{w}^* + \rho$, and conclude that $\mathbf{w}^{\mathsf{T}}(t)\mathbf{w}^* \ge t\rho$. *[Hint: Use induction.]*

(c) Show that $\|\mathbf{w}(t)\|^2 \le \|\mathbf{w}(t-1)\|^2 + \|\mathbf{x}(t-1)\|^2$.

[Hint: $y(t-1) \cdot (\mathbf{w}^{\mathsf{T}}(t-1)\mathbf{x}(t-1)) \le 0$ because $\mathbf{x}(t-1)$ was misclassified by $\mathbf{w}(t-1)$.]

(d) Show by induction that $\|\mathbf{w}(t)\|^2 \le tR^2$, where $R = \max_{1 \le n \le N} \|\mathbf{x}_n\|$.

(e) Using (b) and (d), show that

$$\frac{\mathbf{w}^{\mathsf{T}}(t)}{\|\mathbf{w}(t)\|}\mathbf{w}^* \ge \sqrt{t} \cdot \frac{\rho}{R},$$

and hence prove that

$$t \le \frac{R^2\|\mathbf{w}^*\|^2}{\rho^2}.$$

$\left[\textit{Hint: } \frac{\mathbf{w}^{\mathsf{T}}(t)\mathbf{w}^*}{\|\mathbf{w}(t)\|\,\|\mathbf{w}^*\|} \le 1. \textit{ Why?}\right]$

In practice, PLA converges more quickly than the bound $\frac{R^2\|\mathbf{w}^*\|^2}{\rho^2}$ suggests. Nevertheless, because we do not know ρ in advance, we can't determine the number of iterations to convergence, which does pose a problem if the data is non-separable.

a. since w^* classifies all the instances correctly

$\Rightarrow y_n w^{*T} x_n \geq 0 \quad \forall n \in \{1, \dots N\}$

$\therefore \rho = \min_{1 \leq n \leq N} y_n (w^{*T} x_n)$

$w(t) = w(t-1) + y(t-1) x(t-1)$

$w^T(t) = w^T(t-1) + y(t-1) x^T(t-1)$

$w^T(t) w^* = w^T(t-1) w^* + \underbrace{y(t-1) x^T(t-1) w^*}_{\geq \rho}$ (from a)

$\therefore w^T(t) w^* \geq w^T(t-1) w^* + \rho$

Proof for $w^T(t) w^* \geq t\rho$:

$w(0) \geq 0 \quad [w(0) = 0]$

let

$w^T(t-1) w^* \geq (t-1)\rho$

$w^T(t) = w^T(t-1) + y(t-1) x(t-1)$

$w^T(t) w^* = \underbrace{w^T(t-1) w^*}_{\geq (t-1)\rho} + \underbrace{y(t-1) x(t-1) w^*}_{\geq \rho}$

$\Rightarrow w^T(t) w^* \geq t\rho$

c. $w(t) = w(t-1) + y(t-1)\,x(t-1)$

$$\|w(t)\|^2 = \|w(t-1)\|^2 + \|x(t-1)\|^2$$
$$+ \ y(t-1)\left(w^T(t-1)\,x(t-1)\right)$$
$$+ \ y(t-1)\cdot\left(x^T(t-1)\,w(t-1)\right)$$
$$< 0$$

$$\Rightarrow \|w(t)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2$$

d.

$$\|w(1)\|^2 \leq 0 + \|x(0)\|^2$$

induction:

$$\|w(1)\|^2 \leq R^2 \qquad [R = \max \|x_n\|]$$

let's assume

$$\|w(t-1)\|^2 \leq (t-1)R^2$$

$$\|w(t)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2$$

$$\|w(t)\|^2 \leq (t-1)R^2 + \underbrace{\|x(t-1)\|^2}_{\leq R^2}$$

$$\|w(t)\|^2 \le tR^2$$

2.

To prove: $$\frac{w^T(t)}{\|w(t)\|} w^* \ge \sqrt{t} \cdot \frac{\rho}{R}$$

$$\|w(t)\| \le \sqrt{t} \cdot R$$

$$w^T(t)\, w^* \ge t \cdot \rho$$

$$\Rightarrow \frac{w^T(t) w^*}{\|w(t)\|} \ge \sqrt{t} \cdot \frac{\rho}{R}$$

$$\frac{(w^T(t) w^*)^T (w^T(t) w^*)}{\|w(t)\|^2} \ge t \frac{\rho^2}{R^2}$$

$$w^{*T} \frac{\cancel{w(t)\, w^T(t)}}{\cancel{\|w(t)\|^2}} w^* \geq \frac{t\rho^2}{R^2}$$

$$\|w^*\|^2 \geq \frac{t\rho^2}{R^2}$$

$$\boxed{t \leq \frac{R^2}{\rho^2} \|w^*\|^2}$$