Problem set 1

Machine Learning

Lucas Emanuel Resck Domingues
Professor: Rodrigo Targino

School of Applied Mathematics
Getulio Vargas Foundation

February 28, 2021

Exercise 1.3

The weight update rule in (1.3) [1] has the nice interpretation that it moves in the direction of classifying $\mathbf{x}(t)$ correctly.

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

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

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

(a) Because $\mathbf{x}(t)$ is misclassified, $\mathbf{w}^\intercal(t)\mathbf{x}(t)$ has the opposite sign of $y(t)$.

(b) Note that, because $x_0 = 1$, we have $\mathbf{x}(t) \neq 0$. So:

$$\begin{aligned} y(t)\mathbf{w}^\intercal(t+1)\mathbf{x}(t) &= y(t)(\mathbf{w}(t) + y(t)\mathbf{x}(t))^\intercal\mathbf{x}(t) \\ &= y(t)(\mathbf{w}^\intercal(t) + y(t)\mathbf{x}^\intercal(t))\mathbf{x}(t) \\ &= y(t)\mathbf{w}^\intercal(t)\mathbf{x}(t) + [y(t)]^2\mathbf{x}^\intercal(t)\mathbf{x}(t) \\ &= y(t)\mathbf{w}^\intercal(t)\mathbf{x}(t) + [y(t)\|\mathbf{x}\|]^2 \\ &> y(t)\mathbf{w}^\intercal(t)\mathbf{x}(t) \end{aligned}$$

(c) We would like to have $y(t)\mathbf{w}^\intercal(t)\mathbf{x}(t) > 0$, if the classification of $\mathbf{x}(t)$ was correct. So we take a greater $y(t)\mathbf{w}^\intercal(t+1)\mathbf{x}(t)$, with the hope that it will be positive. If it is not the case, we can try again.

Exercise 1.4

Let us create our own target function f and data set $\mathcal{D}$ and see how the perceptron learning algorithm works. Take $d = 2$ so you can visualize the problem, and choose a random line in the plane as your target function, where one side of the line maps to $+1$ and the other maps to -1. Choose the inputs $\mathbf{x}_n$ of the data set as random points in the plane, and evaluate the target function on each $\mathbf{x}_n$ to get the corresponding output $\mathbf{y}_n$. Now, generate a data set of size 20. Try the perceptron learning algorithm on your data set and see how long it takes to converge and how well the final hypothesis g matches your target f. You can find other ways to play with this experiment in Problem 1.4.

The code for this experiment can be found in GitHub.

We choose a random line in the plane, as it can be seen in the Figure 1.

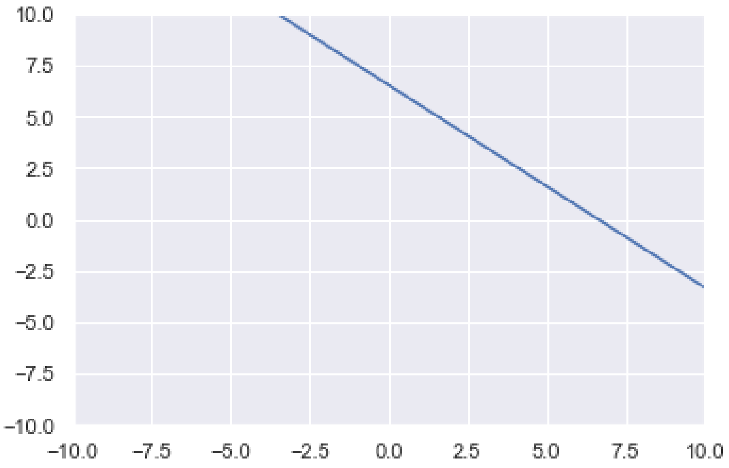

Figure 1: Random line in the plane.

We generate a data set of $n = 20$ random points in the plane, and we evaluate the target function, as shown in Figure 2. Note that the target function evaluation, in the Figure, is the color of the points.

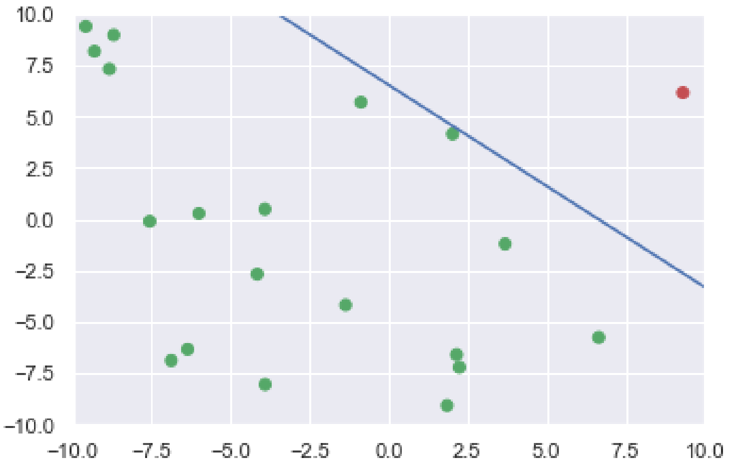

Figure 2: Random data set and random line in the plane.

Now we run the perceptron learning algorithm, with 23 iterations, and the result is shown in Figure 3. We see that, even though the result is the same, the two lines are different.

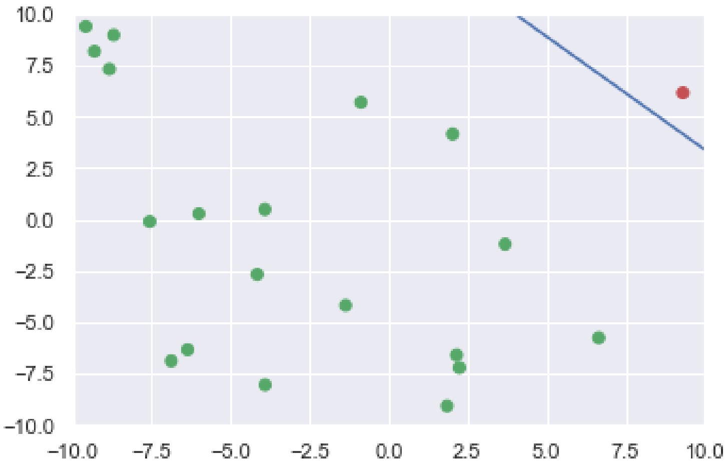

Figure 3: Line result of perceptron learning algorithm.

Exercise 1.8

If $\mu = 0.9$, what is the probability that a sample of 10 marbles will have $\nu \leq 0.1$? [Hints: 1. Use binomial distribution. 2. The answer is a very small number.]

Consider X the number of red marbles. Because every marble has the same probability of being red, this is a binomial distribution with parameters $n = N = 10$ and $p = \mu = 0.9$. The probability calculation is as follows:

$$\begin{aligned}\mathbb{P}(\nu \leq 0.1) &= \mathbb{P}\left(\frac{X}{10} \leq 0.1\right) \\ &= \mathbb{P}(X \leq 1) \\ &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1) \\ &= \binom{10}{0} \cdot 0.9^0 \cdot 0.1^{10} + \binom{10}{1} \cdot 0.9^1 \cdot 0.1^9 \\ &= 9.1 \cdot 10^{-9}\end{aligned}$$

Exercise 1.9

If $\mu = 0.9$, use the Hoeffding Inequality to bound the probability that a sample of 10 marbles will have $\nu \leq 0.1$ and compare the answer to the previous exercise.

It is asked to bound $\mathbb{P}(\nu \leq 0.1)$. It is as follows:

$$\begin{aligned}\mathbb{P}(\nu \leq 0.1) &= \mathbb{P}(\nu - \mu \leq -0.8) \\ &= \mathbb{P}(\mu - \nu \geq 0.8) \\ &= \mathbb{P}(|\mu - \nu| \geq 0.8) + \mathbb{P}(\mu - \nu \leq -0.8)\end{aligned}$$

Note that $\mathbb{P}(\mu - \nu \leq -0.8) = \mathbb{P}(\nu \geq 1.7) = 0$, so

$$\begin{aligned}\mathbb{P}(\nu \leq 0.1) &= \mathbb{P}(|\mu - \nu| \geq 0.8)\\ &\leq 2e^{-2 \cdot 0.8^2 \cdot 10}\\ &\simeq 5.5 \cdot 10^{-6}\end{aligned}$$

Of course, the bound is greater than the probability found in the previous exercise.

References

[1] Yaser S. Abu-Mostafa, M. Magdon-Ismail, and H.T. Lin. *Learning from Data: A Short Course.* AMLBook.com, 2012.