Machine Learning Fundamentals Homework 1

Predrag Pilipovic (predrag.pilipovic@grenoble-inp.org)

January 12, 2020

An analysis of the perceptorn algorithm

The perceptorn algorithm is one of the first supervised models proposed by Rosenblatt, 1957 for binary classification. The training step of the algorithm consists in finding the parameters of a linear function defined by

$$\begin{aligned} h_{\mathbf{w}} : \mathbb{R}^d &\to \mathbb{R} \\ \mathbf{x} &\mapsto \langle \mathbf{w}, \mathbf{x} \rangle \end{aligned}$$

using a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ of size m, where $\langle \cdot, \cdot \rangle$ denotes the dot product and the classes verify for $i \in \{1, 2, ..., m\}$, and $y_i \in \{-1, +1\}$. The training of the model is generally done on-line as shown in algorithm 1.

Algorithm 1 The algorithm of perceptron

```
Training set S = {(x_i, y_i) | i = 1, 2, ..., m}
Initialize the weights w^(0) ← 0
t ← 0
Learning rate ε > 0
repeat
    Choose randomly an example (x, y) ∈ S
    if y ⟨w^(t), x⟩ < 0 then
        w^(t+1) ← w^(t) + ε · y · x        (A)
        t ← t + 1
    end if
until t > T
```

Question 1.

Explain the algorithm.

The goal is to find a hyper-plane passing through the origin, which will classify data based on the training set S. Hyper-plane is given by the equation $\langle \mathbf{w}, \mathbf{x} \rangle = 0$, i.e.

$$\omega_1 x_1 + \omega_2 x_2 + ... + \omega_d x_d = 0.$$

In other words, we need to find weights $\mathbf{w}$. So, we have the training set S consisting of vectors $\mathbf{x_i}$ and their classification $y_i = 1$ or $y_i = -1$, for two classes. Now we set starting weights $\mathbf{w}^{(0)}$ at zero and the beginning time t also at zero. We then chose the learning rate $\varepsilon > 0$ which controls how much we change the weights. In other words, when we are trying to reach minimum and choosing ε helps us determine how fast we can reach that minimum. Nevertheless, for the perception model learning rate is not a necessity because this model guarantees to find a solution, if such exists, in

final number of steps. Now, we will repeat the algorithm until given number of iteration T. At time t we have our hyper-plane

$$\omega_1^{(t)} x_1 + \omega_2^{(t)} x_2 + ... + \omega_d^{(t)} x_d = 0.$$

Firstly, we choose randomly an example $(\mathbf{x}, y)$ from set S. Then we need to see if this element of ours is rightly classified. We do that by computing

$$y \left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle.$$

If the element $(\mathbf{x}, y)$ is rightly classified, the value of the last expression will always be grater or equal to zero. The reason for this is due to fact that sign of $\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle$ depends of the side of the hyper-plane in which is $\mathbf{x}$ is located. So, if our current hyper-plane is good for element $(\mathbf{x}, y)$, we have positive sign in the previous expression and we should not change weights. On the other hand, if the sign is negative it means that our $\mathbf{x}$ is misclassified by the hyper-plane determined by $\mathbf{w}^{(t)}$. So, we need to change our weights. We do it by next expression

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \varepsilon \cdot y \cdot \mathbf{x}.$$

So basically we added or subtracted scaled vector $\mathbf{x}$ from $\mathbf{w}^{(t)}$. At the end we increment t and repeat the whole process.

Question 2.

How is called the update rule (eq. (A)), and what does it do?

The equation (A), i.e.

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \varepsilon \cdot y \cdot \mathbf{x}$$

is called *stochastic gradient descent*. This equation changes the weights and thus hyper-plane. Let's discuss why it works. Without loss of generality we can assume that $\mathbf{x}$ should be in positive class, i.e. $y = 1$. Then, we know from before that when $\mathbf{x}$ is rightly classified then $\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle > 0$. We also know the angle between $\mathbf{w}^{(t)}$ and $\mathbf{x}$, let's call it α_t, is determined by

$$\cos \alpha_t = \frac{\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle}{\left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}||} > 0.$$

This means that $\alpha < 90°$. So, whatever the $\mathbf{w}$ vector may be, as long as it makes an angle less then $90°$ with the positive example vector $\mathbf{x}$, the hyper-plane is good. We can now show why the stochastic gradient descent works. For the simplicity of the proof let's assume that $\varepsilon = 1$. Now, we have

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \mathbf{x},$$

which leads us to

$$\begin{aligned}
\cos \alpha_{t+1} = \frac{\left\langle \mathbf{w}^{(t+1)}, \mathbf{x} \right\rangle}{\left|\left|\mathbf{w}^{(t+1)}\right|\right| \cdot ||\mathbf{x}||} &= \frac{\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle + ||\mathbf{x}||^2}{\left|\left|\mathbf{w}^{(t)} + \mathbf{x}\right|\right| \cdot ||\mathbf{x}||} \\
&\geqslant \frac{\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle + ||\mathbf{x}||^2}{\left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}|| + ||\mathbf{x}||^2} \\
&\geqslant \frac{\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle}{\left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}||} = \cos \alpha_t.
\end{aligned}$$

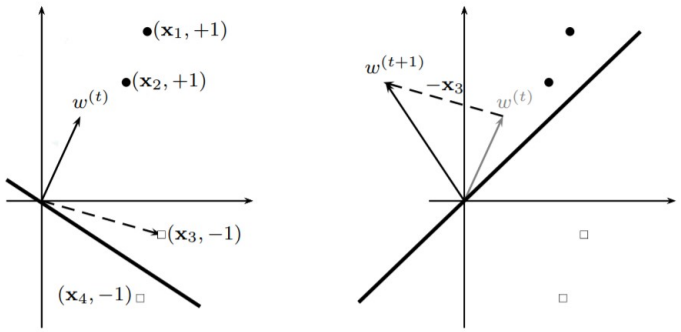
The first inequality is just the triangle inequality, and the reason for the last inequality comes from the fact that $\cos\alpha_t \leqslant 1$, so it is $\left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}|| \geqslant \left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle$. From previous we have

$$\frac{\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle + ||\mathbf{x}||^2}{\left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}|| + ||\mathbf{x}||^2} - \frac{\left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle}{\left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}||}$$
$$= \frac{\left(\left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}|| - \left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle\right) \cdot ||\mathbf{x}||^2}{\left(\left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}|| + ||\mathbf{x}||^2\right) \cdot \left|\left|\mathbf{w}^{(t)}\right|\right| \cdot ||\mathbf{x}||} \geqslant 0.$$

Finally, we proved that $\cos\alpha_{t+1} \geqslant \cos\alpha_t$ and we know that $\alpha_t < 90°$, so we can conclude that $\alpha_t > \alpha_{t+1}$. To conclude, with stochastic gradient descent we change the hyper-plane such that for every new positive vector $\mathbf{x}$, the angle is smaller then the previous one. Similarly, for every negative vector we have $\cos\alpha_t < 0$, which means that $\alpha_t > 90°$. So whatever $\mathbf{w}$ is, as long as it makes an angle more than 90° with the negative example data vector $\mathbf{x}$, the hyper-plane is good. And in the same way we can prove that when $\mathbf{x}$ is in negative class, i.e. $y = -1$, the stochastic gradient descent will decrease the $\cos\alpha_t$, which means, it will increase the α_t, which is what we wanted from the beginning.

Question 3.

Consider the following classification problem in a two dimensional space. Suppose that the chosen example is $\mathbf{x}_3$, what will be the new weight vector using the update rule of the perceptron if $\varepsilon = 1$? Draw the weight vector by reproducing the figure in your sheet.



Question 4.

We are now interested to demonstrate the convergence of the algorithm in a finite number of iterations and in the case where there exists a weight vector $\mathbf{w}^*$ such that for all $(\mathbf{x}_i, y_i) \in S$

$$y_i \left\langle \mathbf{w}^*, \mathbf{x}_i \right\rangle > 0.$$

What is the meaning of the condition $y_i \left\langle \mathbf{w}^*, \mathbf{x}_i \right\rangle > 0$?

As it is discussed in previous questions the mentioned condition is sufficient for having all data in right classes.

Question 5.

We suppose that there exists $\mathbf{w}^*$ such that for all $(\mathbf{x}_i, y_i) \in S$

$$y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle > 0,$$

and we define

$$\rho = \min_{i=1,2,\dots,m} y_i \left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{x}_i \right\rangle.$$

What does ρ represent? Explain why it is a strictly positive real value?

We suppose that $\mathbf{w}^*$ is a weight vector which provides the hyper-plane that separates all data rightly. Then, ρ represents tho shortest distance from hyper-plane to data set S, in other words ρ corresponds to closest $\mathbf{x}$ to the hyper-plane given by $\mathbf{w}^*$. If we suppose that $\lambda = y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle > 0$, for all $(\mathbf{x}_i, y_i) \in S$, then it follows

$$y_i \left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{x}_i \right\rangle = \frac{\lambda}{||\mathbf{w}^*||} > 0,$$

for all $(\mathbf{x}_i, y_i) \in S$. So it must be $\rho > 0$. Observe that the condition $\rho > 0$ provides that we can find a hyper-plane such that none of the data points actually lie on the hyper-plane. If the separating hyper-plane must necessarily pass through some of the data points, note that perceptron's predictions for these points would be an arbitrary choice.

Question 6.

We suppose that all the examples in the training set are within a hyper-sphere of radius R, i.e. for all $\mathbf{x}_i \in S$,

$$||\mathbf{x}_i|| \leqslant R.$$

Further, we initialize the weight vector to be the null vector, i.e. $\mathbf{w}^{(0)} = \mathbf{0}$, as well as the learning rate $\varepsilon = 1$. Show that after t updates, the norm of the current weight vector satisfies

$$\left|\left|\mathbf{w}^{(t)}\right|\right|^2 \leqslant t \cdot R^2. \tag{1}$$

Hint: You can consider $\left|\left|\mathbf{w}^{(t)}\right|\right|^2$ as $\left|\left|\mathbf{w}^{(t)} - \mathbf{w}^{(0)}\right|\right|^2$.

We have that

$$\begin{aligned}
\left|\left|\mathbf{w}^{(t)}\right|\right|^2 &= \left|\left|\mathbf{w}^{(t-1)} + \cancel{\varepsilon}^{1} \cdot y \cdot \mathbf{x}\right|\right|^2 = \left\langle \mathbf{w}^{(t-1)} + y \cdot \mathbf{x}, \mathbf{w}^{(t-1)} + y \cdot \mathbf{x} \right\rangle \\
&= \left|\left|\mathbf{w}^{(t-1)}\right|\right|^2 + 2 \underbrace{y \left\langle \mathbf{w}^{(t-1)}, \mathbf{x} \right\rangle}_{\leqslant 0} + \cancel{y^2}^{1} \cdot ||\mathbf{x}||^2 \\
&\leqslant \left|\left|\mathbf{w}^{(t-1)}\right|\right|^2 + ||x||^2 \leqslant \left|\left|\mathbf{w}^{(t-1)}\right|\right|^2 + R^2 \\
&\leqslant \left|\left|\mathbf{w}^{(t-2)}\right|\right|^2 + 2R^2 \leqslant \dots \leqslant \cancel{\left|\left|\mathbf{w}^{(0)}\right|\right|^2}^{0} + t \cdot R^2 = t \cdot R^2.
\end{aligned}$$

Note that $y \left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle$ is negative or zero, otherwise we would have not corrected $\mathbf{w}^{(t)}$.

Question 7.

Using the the same condition than in the previous question, show that after t updates of the weight vector we have

$$\left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{w}^{(t)} \right\rangle \geqslant t \cdot \rho. \tag{2}$$

Similarly, we have

$$\begin{aligned}
\left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{w}^{(t)} \right\rangle &= \left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{w}^{(t-1)} + y \cdot \mathbf{x} \right\rangle \\
&= \left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{w}^{(t-1)} \right\rangle + \underbrace{y \left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{x} \right\rangle}_{\geqslant \rho} \\
&\geqslant \left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{w}^{(t-1)} \right\rangle + \rho \geqslant \left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{w}^{(t-2)} \right\rangle + 2\rho \\
&\vdots \\
&\geqslant \cancelto{0}{\left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{w}^{(0)} \right\rangle} + t \cdot \rho = t \cdot \rho.
\end{aligned}$$

Question 8.

Deduce from equations (1) and (2) that the number of iterations t is bounded by

$$t \leqslant \left\lfloor \left(\frac{R}{\rho} \right)^2 \right\rfloor,$$

where $\lfloor x \rfloor$ represents the floor function. (This result is due to Novikoff, 1966)

Using (2) we have

$$t \cdot \rho \leqslant \left\langle \frac{\mathbf{w}^*}{||\mathbf{w}^*||}, \mathbf{w}^{(t)} \right\rangle = \cancelto{1}{\left|\left| \frac{\mathbf{w}^*}{||\mathbf{w}^*||} \right|\right|} \cdot \left|\left| \mathbf{w}^{(t)} \right|\right| \cdot \underbrace{\cos \angle \left(\mathbf{w}^*, \mathbf{w}^{(t)} \right)}_{\leqslant 1} \leqslant \left|\left| \mathbf{w}^{(t)} \right|\right|.$$

Now, from previous result and from (1) we can conclude

$$t^2 \cdot \rho^2 \leqslant \left|\left| \mathbf{w}^{(t)} \right|\right|^2 \leqslant t \cdot R^2,$$

so it must be

$$t \leqslant \frac{R^2}{\rho^2}.$$

Considering the fact that $t \in \mathbb{N}$, we can say

$$t \leqslant \left\lfloor \left(\frac{R}{\rho} \right)^2 \right\rfloor.$$

Question 9.

Explain the previous result.

Assume that our input data points are linearly separable, i.e.

$$y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geqslant \rho > 0,$$

for all $(\mathbf{x}_i, y_i) \in S$, and for some vector $\mathbf{w}^*$. Moreover, we assumed not only that input data points are linearly separable, but the hyper-plane should pass through the origin. The last condition can be avoided but we used it for the simplicity of the proof. Finally, last assumption is that the distance between the input points and the origin is bounded by R. Then, the perceptron algorithm makes at most $\frac{R^2}{\rho^2}$ errors, where this means that the error occurs whenever we have misclassified data, after which it returns a separating hyper-plane. Which means we proved the convergence for this algorithm.