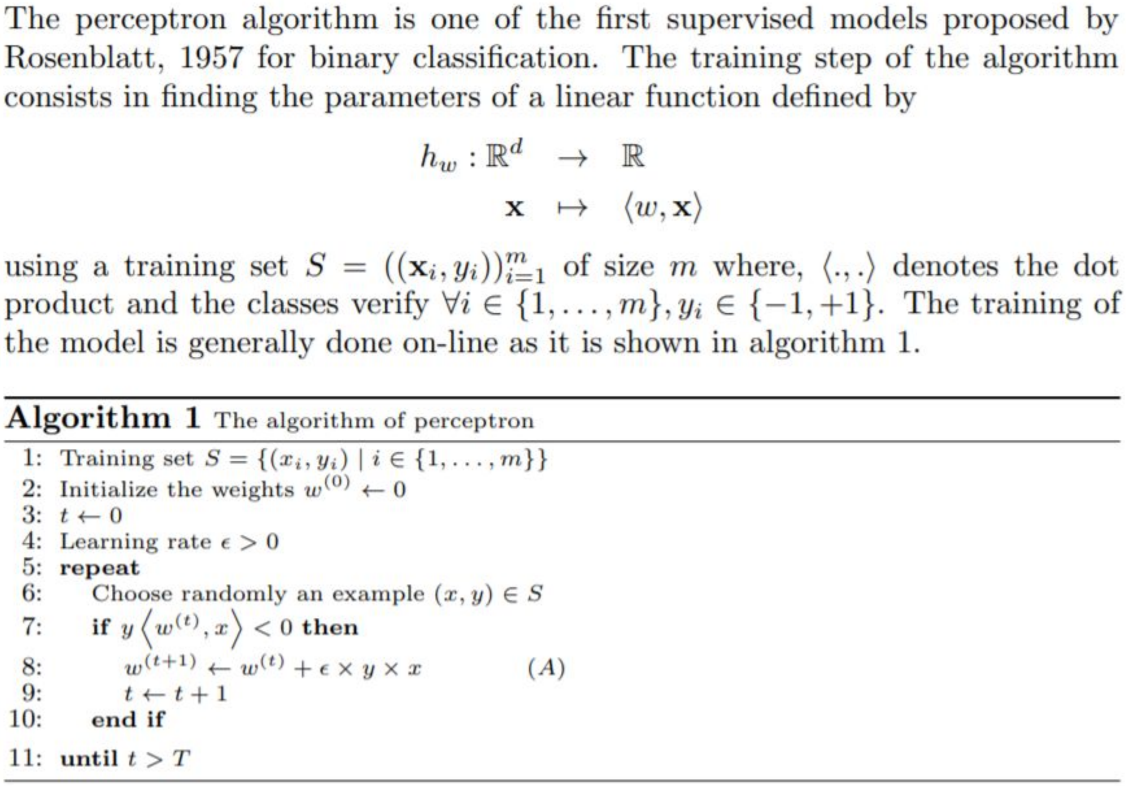
Machine Learning Fundamentals - Homework 1

Milena Markovic (milena.markovic@grenoble-inp.org)

An analysis of the perceptron algorithm



The perceptron 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

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

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

Algorithm 1 The algorithm of perceptron

```
Training set $S = \{(x_i, y_i) \mid i \in \{1, \dots, m\}\}$
Initialize the weights $w^{(0)} \leftarrow 0$
$t \leftarrow 0$
Learning rate $\epsilon > 0$
repeat
    Choose randomly an example $(x, y) \in S$
    if $y \left\langle w^{(t)}, x \right\rangle < 0$ then
        $w^{(t+1)} \leftarrow w^{(t)} + \epsilon \times y \times x$        $(A)$
        $t \leftarrow t + 1$
    end if
until $t > T$
```

Question 1: Explain the algorithm.

The algorithm finds a hyper-plane that splits linearly separable datasets. The hyper-plane of dimension *d* is defined by the equation:

$$\omega_0 + \omega_1 x_1 + \dots + \omega_d x_d = 0$$

where $\omega_n,\ n \, \varepsilon \, [0, d]$ are weights and $x_n,\ n \, \varepsilon \, [1, d]$ are inputs to the perceptron. The goal of the algorithm is to find the appropriate values of $\omega_n,\ n \, \varepsilon \, [0, d]$ for the given dataset *S*. Since the algorithm is a supervised learning classifier, the dataset consists of multiple pairs of ($\vec{x}, y$) where $\vec{x}$ is a vector of input values of dimension *d* and *y* is the classification of the given input, $y \, \varepsilon \, \{-1, \ +1\}$. The last factors in the algorithm are the learning rate $\varepsilon > 0$ which controls the rate at which the algorithm accepts the changes to the weights and *t* which measures the number of updates.

Initially, all weights are set to zero and the learning rate is adjusted according to the dataset. After that, in a predefined amount of updates, examples from the dataset S are selected at random. Each example goes through the classifier and depending on whether the output is the expected value, the weights are adjusted so that the new weights are:

$$\omega_i^{(t+1)} = \omega_i^{(t)} + \varepsilon * y * x_i$$

If an adjustment is made in an interaction, the time counter t is incremented and the weight correction step is repeated until the counter t reaches the predefined limit T.

Question 2: How is called the update rule (eq. (A)), and what does it do?

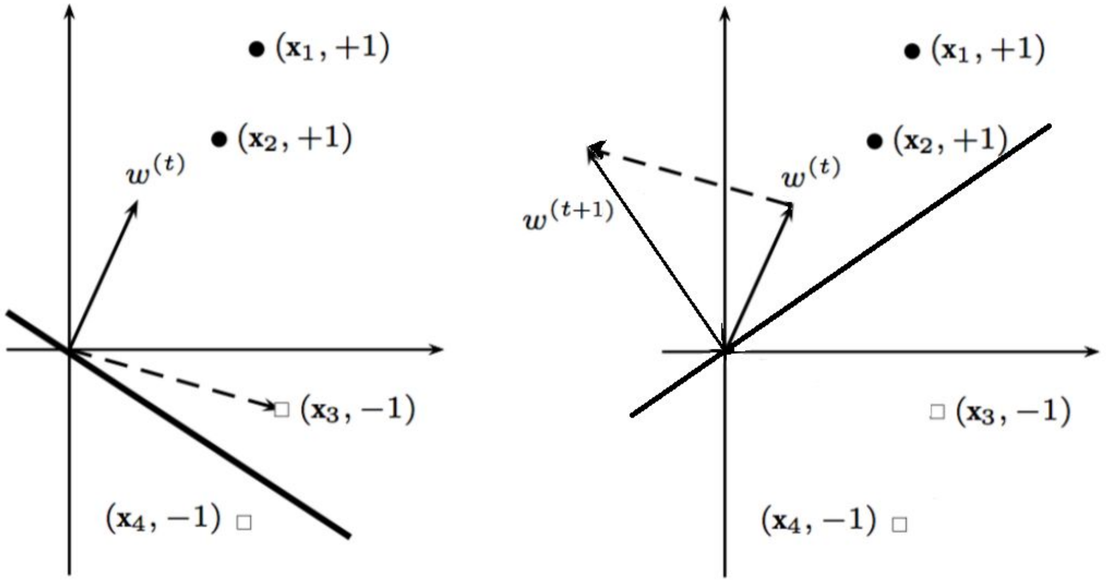
The update rule is an equation that represents the stochastic gradient descent algorithm.

$$\vec{\omega}^{(t+1)} = \vec{\omega}^{(t)} + \varepsilon * y * \vec{x}$$

This is an iterative algorithm that optimizes parameters of the perceptron by gradually adjusting them until all of the data from the set is correctly classified:

- If a record is classified correctly, then weight vector $\vec{\omega}$ remains unchanged
- If it is not classified correctly, we add the vector $\vec{x}$to $\vec{\omega}$ when $y = 1$ and otherwise, subtract it when $y = -1$.

Question 3: Consider the following classification problem in a two dimensional space. Suppose that the chosen example is 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 $\omega *$ such that $\forall (x_i , y_i) \in S;\ y \times \langle \omega *, x \rangle > 0$. What is the meaning of the condition $y \times \langle \omega *, x \rangle > 0$?

The condition $y \times \langle \omega *, x \rangle > 0$ means that for a weight vector values $\omega *$, the classification is correct for all pairs (x_i, y_i) of the dataset S. If a weight vector with such property exists, the dataset is convergent.

Question 5: We suppose that there exists $\omega *$ such that $\forall (x_i , y_i) \in S;\ y \times \langle \omega *, x \rangle > 0$ and we define $\rho = min_{i \in \{1,\ldots,m\}}\ (y_i \langle \frac{\omega *}{||\omega *||}, x_i \rangle)$. What does ρ represent? Explain why it is a strictly positive real value?

The provided condition establishes that there is a value of $\omega *$ that separates the dataset classes correctly. For the hyper-plane defined by such weight vector, this value ρ represents the distance from the hyper-plane to the closest dataset member.
The value has to be strictly positive because of the given condition $y \times \langle \omega *, x \rangle > 0$, and since the vector norm is positive by definition, ρ therefore has to be positive as well.

Question 6: We suppose that all the examples in the training set are within a hypersphere of radius R (*i.e.* $\forall x_i \in S,\ ||x_i|| \le R$). Further, we initialise the weight vector to be the null vector (i.e. $w(0) = 0$) as well as the learning rate $\varepsilon = 1$. Show that after *t* updates, the norme of the current weight vector satisfies :

$$||\omega^{(t)}||^2 \le t \times R^2 \qquad (1)$$

hint : You can consider $||\omega^{(t)}||^2$ as $||\omega^{(t)} - \omega^{(0)}||^2$

$$\left\|\omega^{(t)}\right\|^2 = \left\|\omega^{(t-1)} + y\varepsilon x\right\|^2 = \left\|\omega^{(t-1)}\right\|^2 + 2yx\left\|\omega^{(t-1)}\right\| + y^2x^2,\ y < 0 \Rightarrow 2yx\left\|\omega^{(t-1)}\right\| \le 0$$

$$= \left\|\omega^{(t-2)}\right\|^2 + 2yx\left\|\omega^{(t-2)}\right\| + y^2x^2 + 2yx\left\|\omega^{(t-1)}\right\| + y^2x^2$$

$$= \left\|\omega^{(0)}\right\|^2 + 2yx\left\|\omega^{(1)}\right\| + \ldots + 2yx\left\|\omega^{(t-1)}\right\| + t\,y^2x^2,\ y^2 = 1$$

$$\left\|\omega^{(t)}\right\|^2 = \left\|\omega^{(0)}\right\|^2 + 2yx\left\|\omega^{(1)}\right\| + \ldots + 2yx\left\|\omega^{(t-1)}\right\| + t\,||x||^2$$

$$||x|| \le R \quad \wedge \quad A = 2yx\left\|\omega^{(1)}\right\| + \ldots + 2yx\left\|\omega^{(t-1)}\right\| \le 0 \quad \wedge \quad \left\|\omega^{(0)}\right\| = 0$$

$$\Rightarrow \left\|\omega^{(t)}\right\|^2 = t \times R^2 - A,\ A \ge 0 \ \Rightarrow \left\|\omega^{(t)}\right\|^2 \le t \times R^2$$

Question 7: Using the the same condition than in the previous question, show that after t updates of the weight vector we have

$$\langle \frac{\omega*}{\|\omega*\|}, \omega^{(t)} \rangle \geq t \times \rho \qquad (2)$$

$$\rho \; = \; min_{i \in \{1,\dots,m\}} \; (y_i \langle \; \frac{\omega*}{\|\omega*\|} \; , \; x_i \rangle), \; \varepsilon = 1$$

$$\langle \frac{\omega*}{\|\omega*\|}, \omega^{(t)} \rangle = \langle \frac{\omega*}{\|\omega*\|}, \omega^{(t-1)} + y\varepsilon x \rangle = \langle \frac{\omega*}{\|\omega*\|}, \omega^{(t-1)} \rangle + y \langle \frac{\omega*}{\|\omega*\|}, x \rangle$$

$$= \langle \frac{\omega*}{\|\omega*\|}, \omega^{(t-2)} + y\varepsilon x \rangle + y \langle \frac{\omega*}{\|\omega*\|}, x \rangle = \langle \frac{\omega*}{\|\omega*\|}, \omega^{(t-2)} \rangle + 2y \langle \frac{\omega*}{\|\omega*\|}, x \rangle$$

$$= \dots = \langle \frac{\omega*}{\|\omega*\|}, \omega^{(0)} \rangle + ty \langle \frac{\omega*}{\|\omega*\|}, x \rangle$$

$$y \langle \frac{\omega*}{\|\omega*\|}, x \rangle \geq \rho \Rightarrow \langle \frac{\omega*}{\|\omega*\|}, \omega^{(0)} \rangle + ty \langle \frac{\omega*}{\|\omega*\|}, x \rangle \geq \langle \frac{\omega*}{\|\omega*\|}, \omega^{(0)} \rangle + t\rho$$

$$\langle \frac{\omega*}{\|\omega*\|}, \omega^{(t)} \rangle \geq \langle \frac{\omega*}{\|\omega*\|}, \omega^{(0)} \rangle + t\rho$$

$$\langle \frac{\omega*}{\|\omega*\|}, \omega^{(0)} \rangle = 0 \Rightarrow \langle \frac{\omega*}{\|\omega*\|}, \omega^{(0)} \rangle \geq t\rho$$

Question 8: Deduce from equations (1) and (2) that the number of iterations t is bounded by

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

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

$$\langle \frac{\omega*}{\|\omega*\|}, t \rangle \geq t\rho$$

$$\left\| \omega^{(t)} \right\|^2 \leq \; t \; \times R^2$$

$$\langle \frac{\omega*}{\|\omega*\|}, \omega^{(0)} \rangle = \left\| \frac{\omega*}{\|\omega*\|} \right\| \left\| \omega^{(t)} \right\| cos(\frac{\omega*}{\|\omega*\|}, \omega^{(0)})$$

$$cos(\frac{\omega*}{\|\omega*\|}, \omega^{(0)}) \leq 1, \left\| \frac{\omega*}{\|\omega*\|} \right\| \; = \; 1 \Rightarrow t\rho \leq \langle \frac{\omega*}{\|\omega*\|}, t \rangle \leq \left\| \omega^{(t)} \right\|$$

$$t\rho \leq \left\| \omega^{(t)} \right\|, \quad \left\| \omega^{(t)} \right\|^2 \leq \; t \; \times R^2$$

$$t^2 \rho^2 \leq \; \left\| \omega^{(t)} \right\|^2 \leq \; t \; \times R^2$$

$$t^2 \rho^2 \leq \; t \; \times R^2$$

$$t \, \rho^2 \leq R^2$$

$$t \leq \left(\frac{R}{\rho} \right)^2$$

Question 9. Explain the previous result.

Questions 7 and 8 explore the proof of the Novikoff 1962 theorem of convergence of perceptrons for linearly separable datasets that consist of data contained within a hypersphere of a limited radius. The theorem asserts that the weight vector should converge before reaching a certain amount of misclassifications (t) that is in correlation with the radius of the hypersphere (R) - as the result of question 8 shows, this limit of errors is $\left(\frac{R}{\rho}\right)^2$.