Intelligence Artificial and the Web Exam

MOSIG-2013

Massih-Reza Amini (Part 1), Ahlame Douzal (Part 2)
Duration: 3 hours, Documents: authorized

Part 1: An analysis of the perceptron algorithm (10 Pts)

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

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

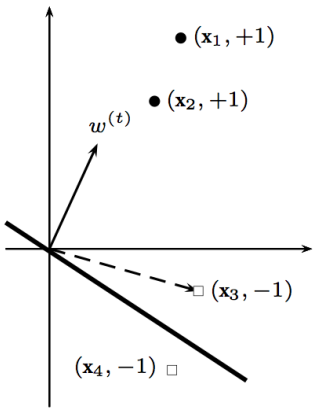
using a training set $S = ((\mathbf{x}_i, y_i))_{i=1}^m$ of size m where, $\langle ., . \rangle$ denotes the dot product and the classes verify $\forall i \in \{1, \ldots, m\}, y_i \in \{-1, +1\}$. The training of the model is generally done on-line as it is shown in algorithm 1.

Algorithm 1 The algorithm of perceptron

Training set $S = \{(x_i, y_i) \mid i \in \{1, \ldots, m\}\}$
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**
    $w^{(t+1)} \leftarrow w^{(t)} + \epsilon \times y \times x$ $\quad (A)$
    $t \leftarrow t + 1$
  end if
until $t > T$

1. Explain the algorithm.
2. How is called the update rule (eq. (A)), and what does it do?
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 $\epsilon = 1$? Draw the weight vector by reproducing the figure in your sheet.



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 w^* such that $\forall(\mathbf{x}_i, y_i) \in S; y_i \times \langle w^*, \mathbf{x}_i \rangle > 0$. What is the meaning of the condition $y \times \langle w^*, \mathbf{x} \rangle > 0$?

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

6. We suppose that all the examples in the training set are within a hypersphere of radius R (i.e. $\forall \mathbf{x}_i \in S, ||\mathbf{x}_i|| \leq R$). Further, we initialise the weight vector to be the null vector (i.e. $w^{(0)} = 0$) as well as the learning rate $\epsilon = 1$. Show that after t updates, the norme of the current weight vector satisfies :
$$||w^{(t)}||^2 \leq t \times R^2 \tag{1}$$
hint : You can consider $||w^{(t)}||^2$ as $||w^{(t)} - w^{(0)}||^2$

7. Using the the same condition than in the previous question, show that after t updates of the weight vector we have
$$\left\langle \frac{w^*}{||w^*||}, w^{(t)} \right\rangle \geq t \times \rho \tag{2}$$

8. Deduce from equations (1) and (2) that the number of iterations t is bounded by
$$t \leq \left\lfloor \left(\frac{R}{\rho} \right)^2 \right\rfloor$$
where $\lfloor x \rfloor$ represents the floor function (This result is due to Novikoff, 1966).

9. Explain the previous result.

Part 2: Temporal Data Analysis (10 Pts)

2.1 Time series alignment

We define the alignment π_i of length $|\pi_i| = m$ between the time series x_1, x_2 of length T as the set of m couples of aligned instants:

$$\pi_i = ((\pi_1(1), \pi_2(1)), (\pi_1(2), \pi_2(2)), ..., (\pi_1(m), \pi_2(m)))$$

where the applications π_1 and π_2 defined from $\{1, ..., m\}$ to $\{1, .., T\}$ obey to the following boundary and monotonicity conditions:

$$1 = \pi_1(1) \leq \pi_1(2) \leq ... \leq \pi_1(m) = T$$
$$1 = \pi_2(1) \leq \pi_2(2) \leq ... \leq \pi_2(m) = T$$

whereas $\forall\, j \in \{1, .., m\}$,

$$\pi_1(j+1) \leq \pi_1(j) + 1 \text{ and } \pi_2(j+1) \leq \pi_2(j) + 1$$

Let $C(\pi_i)$ be the cost of π_i defined as follows:

$$C(\pi_i) = \frac{1}{|\pi_i|} \sum_{k=1}^{|\pi_i|} \varphi(x_{1\,\pi_1(k)}, x_{2,\pi_2(k)}) \tag{3}$$

1. For $x_1 = (1, 0, 2, 3, 2)$, $x_2 = (0, 3, 1, 2, 3)$ and a divergence $\varphi(x_{1\,\pi_1(k)}, x_{2,\pi_2(k)}) = \|x_{1\,\pi_1(k)} - x_{2,\pi_2(k)}\|_2$, determine the alignement π_* that minimizes the cost C (Eq. 3).
2. Evaluate the DTW and the Fréchet distance between x_1 and x_2.

2.2 About temporal correlation

Let us define the temporal correlation or order r between two time series x, y, as:

$$Cort(x, y) \quad = \quad \frac{\sum_{i,i'} m_{ii'}(x_i - x_{i'})(y_i - y_{i'})}{\sqrt{\sum_{i,i'} m_{ii'}(x_i - x_{i'})^2}\sqrt{\sum_{i,i'} m_{ii'}(y_i - y_{i'})^2}} \tag{4}$$

with $m_{ii'} = 1$ if $|i' - i| \leq r$, 0 otherwise; $r \geq 1$.

1. Explain the role of the parameter r.
2. For x_1, x_2 given above, evaluate $Cort(x_1, x_2)$ for $r = 1$, interpret the obtained measure.
3. Discuss the effect of the parameter r and indicate how it can be suitably fixed in each of the following cases:
   - Time series with strong tendency effects,
   - Noisy time series

2.3 Support Vector Machine

Let $(x_1, y_1), ..., (x_m, y_m)$ be m points, $x_i \in H$. Assume a binary classification of linearly separable points. Let HP be a separable hyperplan of direction w, with $y_i = +1$ (vs. $y_i = -1$) for points belonging to the side of direction w (vs. opposite direction to w). We recall, the primal ν-Support Vector Machine formalization:

$$\begin{aligned} \min_{w \in H, \xi \in \mathbb{R}^m, b \in \mathbb{R}, \varphi \in \mathbb{R}} \quad & \frac{1}{2}|w|^2 \; - \; \nu \, \rho \; + \; \frac{1}{m} \sum_{i=1}^{m} \xi_i \\ s.t. \quad & y_i(< x_i, w > + b) \geq \rho - \xi_i \quad \forall i = 1, ..., m \\ & \xi_i \geq 0 \quad \forall i = 1, ..., m \\ & \rho \geq 0 \end{aligned} \tag{5}$$

1. Explain the main differences between the hard, soft and ν-SVM.
2. Explain the concept of sparsity that characterizes the SVM.
3. In the equation 5, indicate the role of the right term $\frac{1}{m} \sum_{i=1}^{m} \xi_i$.
4. For which samples ξ_i is canceled ?
5. Discuss the relationship between ν and respectively: the sparsity, precision, margin and the model complexity.