Perceptron

Course: Optmization Methods for Machine Learning

Laura Palagi

http://www.dis.uniroma1.it/~palagi

Dipartimento di Ingegneria informatica automatica e gestionale A. Ruberti
Sapienza Università di Roma

Via Ariosto 25

October 5, 2016

Perceptron

$$f_{w,b}(x) = sgn(w^T x + b) = \begin{cases} 1 & \text{if } w^T x + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

- Linear separating hyperplanes
- Weighted linear combination
- Nonlinear decision function
- Linear offset (bias)
- Goal of Learning: *w* and *b*

SAPIENZA UNIVERSITÀ DI ROMA

Perceptron algorithm

Data. Input x^i, con $\|x^i\| \leq R$, Target y^i, $i = 1, \dots, \ell$.

Inizialization. Set $w^0 = 0, b^0 = 0$, $k = 0$, $\sharp cc = 0$.

While $\sharp cc < \ell$ **do**

For $i = 1, \dots, \ell$ **do**

If $y^i \cdot ({w^k}^T x^i + b^k) \leq 0$ **then**

$w^{k+1} \leftarrow w^k + y^i x^i$ and

$b^{k+1} \leftarrow b^k + y^i$

$k = k + 1$

else $\sharp cc = \sharp cc + 1$

End For

If $\sharp cc < \ell$ **then set** $\sharp cc = 0$

end While

Perceptron algorithm

- It is error driven. No update if input is already classified correctly
- Weight vector is linear combination $w = \sum_{p=1}^{\ell} y^p x^p$
- Classifier is linear combination of the inner products $x^T x^p$

$$f_{w,b}(x) = sign(\sum_{p=1}^{\ell} y^p x^T x^p + b)$$

- Find w, b such that Empirical Risk $R_{emp} = 0$

Epoch

- It is an *On line algorithm*: it does not consider the entire data set at the same time, it only ever looks at one example. Training examples appear sequentially (internal FOR cycle).
- The number of passes to make over the full training data (external While cycle) is called *epoch*
- It may be useful to have a control on the maximum number of epochs *MaxIter* allowed

  The value *MaxIter* is the only **hyperparameter** of the perceptron algorithm.
- If we make many many passes over the training data, then the algorithm is likely to overfit.
- On the other hand, going over the data only one time might lead to underfitting.

Convergence theorem

If there exists some vector $\begin{pmatrix} \bar{w} \\ \bar{b} \end{pmatrix}$ with unit norm such that $y^i \cdot (\bar{w}^T x^i + \bar{b}) > 0$ for all $i = 1 \ldots, \ell$ then the perceptron converges to a linear separator after a number of steps bounded by

$$\frac{R^2}{\rho^2}$$

where $R = \max_i\{\|x^i\|\}$ and $\rho = \min_i\{y^i \cdot (\bar{w}^T z^i)\} > 0$.

SAPIENZA
UNIVERSITÀ DI ROMA

Pros-cons

- Dimensionality independent
- Convergence for linearly separable sets
- Order dependent (it may depend on the order in which data are presented): random reordering is useful
- Scales with ”difficulty” of problem
- For non linearly separable points the algorithm will never converge

Convergence proof

For the sake of simplicity we consider the unbias case. Namely w.l.g we set

$$w \leftarrow \begin{pmatrix} w \\ b \end{pmatrix} \qquad z^i \leftarrow \begin{pmatrix} x^i \\ 1 \end{pmatrix}$$

and all the vectors corresponding, so that we assume that $\exists \bar{w}$ with $\|\bar{w}\| = 1$ which classifies correctly, i.e. $y^i \cdot (\bar{w}^T z^i) > 0$. Assume that the algorithm does not stop, so that for each k there is at least a p such that $w^{k+1} = w^k + y^p z^p$. By the assumption on $\bar{w}$ we can write for all k

$$\begin{aligned} \bar{w}^T w^{k+1} &= \bar{w}^T w^k + y^p \bar{w}^T z^p \\ &\geq \bar{w}^T w^k + \rho, \end{aligned}$$

Alignment $\bar{w}^T w^{k+1}$ increases with number of errors.

Convergence proof

By Schwarz inequality $\bar{w}^T w^{k+1} \leq \|\bar{w}\|\|w^{k+1}\|$ and applying induction starting with $w^0 = 0$ we get

$$\begin{aligned} \|w^{k+1}\| = \quad & \|\bar{w}\|\|w^{k+1}\| \\ & \geq \bar{w}^T w^{k+1} \geq \bar{w}^T w^k + \rho \\ & \geq \bar{w}^T w^{k-1} + 2\rho \\ & \geq (k+1)\rho \end{aligned}$$

Convergence proof

On the other hand, since z^p satisfies $y^p {w^k}^T z^p \leq 0$ we get

$$\begin{aligned} \|w^{k+1}\|^2 &= \|w^k\|^2 + 2y^p {w^k}^T z^p + \|y^p z^p\|^2 \\ &\leq \|w^k\|^2 + |y^p|\|z^p\|^2 \\ &\leq \|w^k\|^2 + R^2 \\ &\leq \|w^0\|^2 + R^2(k+1) \end{aligned}$$

where the last inequality is obtained by induction.
Combination of these two steps

$$(k+1)^2\rho^2 \leq \|w^{k+1}\|^2 \leq R^2(k+1)$$

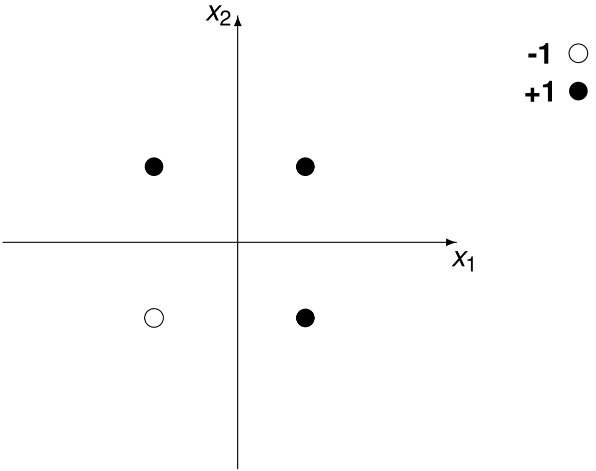
$$k+1 \leq R^2/\rho^2$$

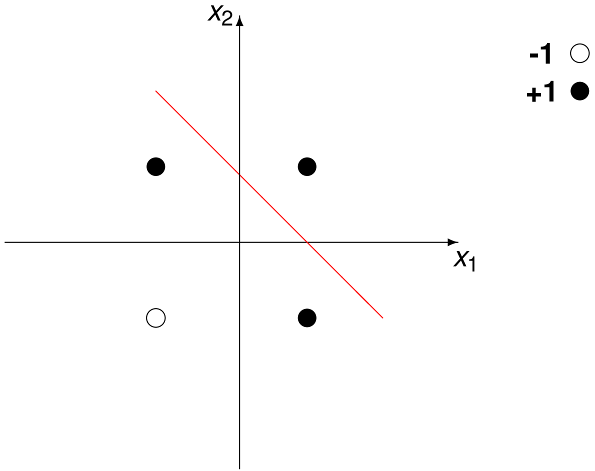
An Example: OR

Training Set (x^i, y^i), $i = 1, \ldots, 4$ with $x^i \in R^2$ and $y^i \in \{-1, 1\}$

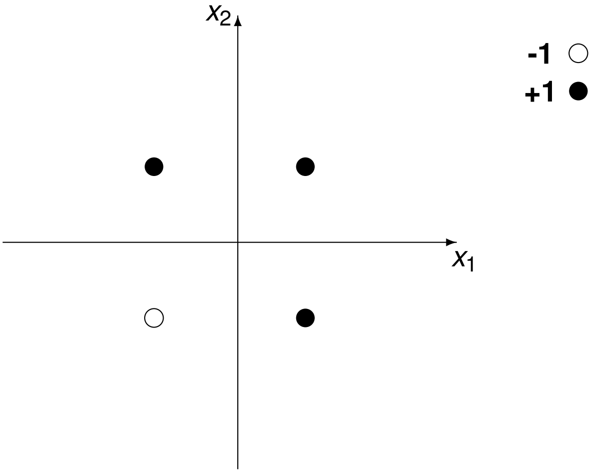
x_1	x_2	y
-1	1	1
1	-1	1
1	1	1
-1	-1	-1

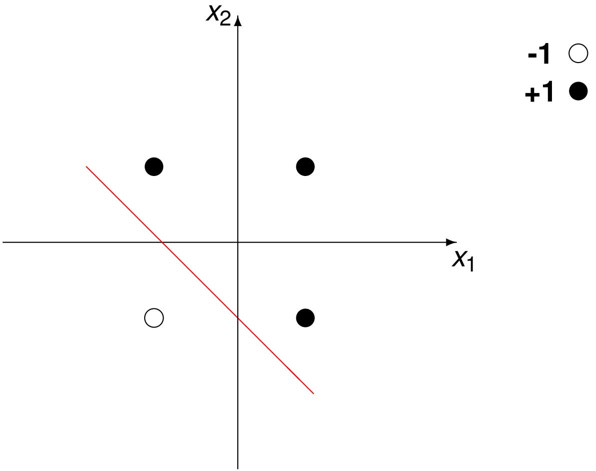
$$\begin{array}{lll} \textit{Iniz.} & w^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; & b^0 = 0 \\ i = 1 & w^1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}; & b^1 = 1 \\ i = 2 & w^2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; & b^2 = 2 \\ i = 3 & \text{point correctly classified} & \\ i = 4 & w^3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}; & b^4 = 1 \end{array}$$


x_2
x_1
-1
+1



x_2
x_1
-1
+1


x_2
x_1
-1
+1

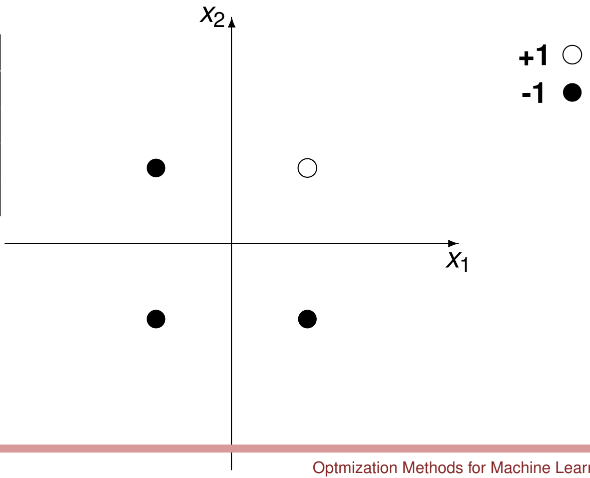


x_2
x_1
-1
+1

An Example: AND

AND

x_1	x_2	y
-1	1	-1
1	-1	-1
1	1	1
-1	-1	-1

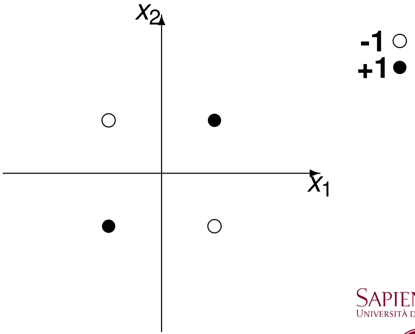


Sets not linearly separable

What if sets aren't separable?

XOR

x_1	x_2	y
-1	1	-1
1	-1	-1
1	1	1
-1	-1	1



Then perceptron doesn't work

Beyond Perceptron

- Voting Perceptron
- Average Perceptron
- ....