The Online Perceptron Algorithm and Linear SupportVector Machines

Lecturer: Yann Chevaleyre
Scribe: Thomas Boudras

Lecture nº
14/11/2023

Contents

1 Problem Definition

1.1 Linear Separators

A linear separator is a function $f : \mathcal{X} \to \mathbb{R}$ or $f : \mathcal{X} \to \{-1, 1\}$ that, based on a set $D = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1,\dots,n}$, predicts the class -1 or 1 for a point $x \in \mathcal{X}$.
Here, we consider $\mathcal{X} = \mathbb{R}^d$, and define the following decision function:

$$f(x) < 0 \quad \text{, assign } x \text{ to class -1}$$
$$f(x) > 0 \quad \text{, assign } x \text{ to class 1}$$

with $f(x) = w^T x + b$, where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ in the case of a linear decision function.

1.2 Linearly Separable Problem

The points $\{(x_i, y_i)\}$ are linearly separable if there is a hyperplane that correctly discriminates all the data. Otherwise, we refer to them as linearly non-separable examples.

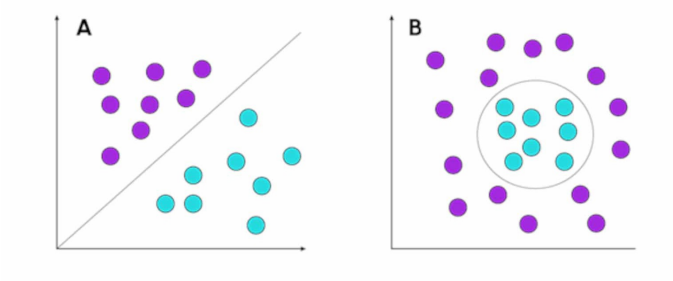

Figure 1: Difference between linearly separable and non-linearly separable

We establish that the distance from a point x to the hyperplane $H(w, b) = \{z \in \mathbb{R}^d \mid f(z) = w^T z + b = 0\}$ of the decision boundary is given by:

$$d(x, H) = \frac{|w^T x + b|}{\|w\|}$$

Proof:

Let x_p be the projection of x onto the hyperplane, and d be the distance separating x from x_p.
We consider the case where x is on the positive side, i.e., $f(x) > 0$. By definition, we have:
$x = x_p + \frac{w}{\|w\|} \times d$
Multiplying both sides by w^T on the left, we get:
$w^T x = w^T x_p + w^T \frac{w}{\|w\|} \times d$
$w^T x - w^T x_p = \|w\| d$
Expressing b:
$\|w\| d = (w^T x + b) - (w^T x_p + b)$
Using the definition of x_p, we have $w^T x_p + b = 0$, so:
$d = \frac{w^T x + b}{\|w\|}$

For the case where x is on the negative side, all we have to do is place a minus in front of the w in the definition of x at the beginning of the proof and an almost similar demonstration yields :
$d = \frac{-(w^T x + b)}{\|w\|}$

Thus, in the general case: $d = \frac{|w^T x + b|}{\|w\|}$

We define the margin as twice the distance from the closest point to the hyperplane.

1.3 Choice of the Decision Function

For a linearly separable problem, there can be multiple possible decision functions. We choose the one with the maximum margin.

As for any non-zero λ, the pair (w, b) and $(w', b') = (\lambda w, \lambda b)$ generates the same hyperplane equation.

$$\forall x, w^T x + b = 0 \iff \forall x, \forall \lambda \in \mathbb{R}^*, \lambda w^T x + \lambda b = 0$$

We then define the canonical hyperplane with respect to the data $\{x_1, \ldots, x_N\}$ as the hyperplane that satisfies $\min_{x_i} |w^T x_i + b| = 1$.

To obtain it, for any hyperplane $H(w, b)$, we simply divide w and b by $\min_{x_i} |w^T x_i + b|$.

The margin associated with this hyperplane is then called the geometric margin and is equal to $M = \frac{2}{\|w\|}$ (by definition of the margin).

Since $\min_{x_i} |w^T x_i + b| = 1$ correctly classifies as $y_i f(x_i) > 1$ (Conquequence of $y_i f(x_i) > 0$ by condition of good classification, $|f(x_i)| \geq 1$ because it is the canonical hyperplane and $y_i \in \{-1, 1\}$.

1.4 The perceptron algo (online)

The perceptron algo (online)

For homogenous linear classifier $f(x) = w^T x$ (no b).

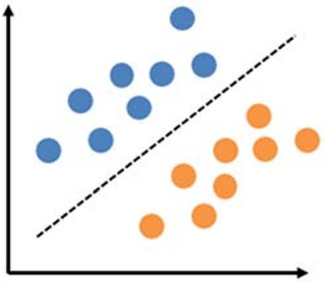

Figure 2: $w^T x = 0$

Perceptron Algo :

$t \leftarrow 0$

$w_0 \leftarrow 0$

Repeat :

- receive x_t
- predict $\hat{y}_t = sign(w_t^T x_t)$
- receive $y_t \in \{-1, 1\}$
- if $\hat{y}_t \neq y_t$ then
  - $w_{t+1} \leftarrow w_t + y_t x_t$
- else
  - $w_{t+1} \leftarrow w_t$

The algorithm oscile towards the right answer because each time it changes, it gradually reduces the number of misclassified points and therefore converges more and more towards the right solution.

1.5 Theorem Block-Novikoff

Assume $\|x_t\| < R, \forall t, y_t \in \{-1, 1\}$
Assume there exist canonical hyperplane w^* passing through the origin with a half margin $\rho = \frac{1}{\|w^*\|}$
The error number of the perceptron is at most $\frac{R^2}{\rho^2}$.

Proof :

Step 1

After an update (erroneous prediction) w_{t+1} is "more aligned" to w^* (with w^* the perfect canonical classifier). So $\langle w_{t+1}, w^* \rangle$ must increase with each iteration.

$$\begin{aligned}\langle w_{t+1}, w^* \rangle &= \langle w_t + y_t x_t, w^* \rangle \\ &= \langle w_t, w^* \rangle + \underbrace{y_t \langle x_t, w^* \rangle}_{\geq 1 \text{ by } w^* \text{ definition}}\end{aligned}$$

$\langle w_{t+1}, w^* \rangle \geq \langle w_t, w^* \rangle + 1$
Unrolling, we get $\langle w_t, w^* \rangle \geq t$ because $w_0 = 0$ (with t number of errors)

Step 2

After an update (classification error) we get:

$$\begin{aligned}\|w_{t+1}\|^2 &= \langle w_{t+1} + y_t x_t, w_t + y_t x_t \rangle \\ &= \|w_t\|^2 + \underbrace{2 y_t \langle w_t, y_t \rangle}_{\leq 0 \text{ because it's an error}} + \underbrace{\|y_t x_t\|^2}_{\leq R^2} \\ &\leq \|w_t\|^2 + R^2\end{aligned}$$

$\|w_t\|^2 \leq tR^2$ by unrolling.

Step 3

Using step 1 and 2, we get :

$$t < \underbrace{\langle w_t, w^* \rangle \leq \|w_t\| \|w^*\|}_{\text{Cauchy-Schwarz}} \leq \underbrace{\sqrt{t} R \|w^*\|}_{= \frac{\sqrt{t} R}{\rho}}$$

whence $t \leq \sqrt{t} \frac{R}{\rho}$
whence $t \leq \frac{R^2}{\rho^2}$

Exercice:
Rewrite perceptron algo for non homogenous hyper plan

We have : $w^T x + b = 0 \iff w' x' = 0$ with $w' = \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ b \end{pmatrix}$ et $x' = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ 1 \end{pmatrix}$

So, rewrite the perceptron agloritme means replacing $\left\{ w'_{t+1} \leftarrow w'_t + y_t x'_t \right.$
by $\left\{ \begin{array}{l} w_{t+1} \leftarrow w_t + y_t x_t \\ b_{t+1} \leftarrow b_t + y_t \end{array} \right.$

2 The perceptron Algo as an SGD online learner

Perceptron update

if $y_t \neq \hat{y}_t$ then

$w_{t+1} \leftarrow w_t + y_t x_t$

else

$w_{t+1} \leftarrow w_t$

SGD update

$$w_{t+1} \leftarrow w_t - \alpha \nabla_w l(\underbrace{w_t^T x}_{\text{predicted score of } x_t}, y_t)$$

Let $s_t = w_t^T x$, we get : $y_t \neq \hat{y}_t \iff y_t.s_t < 0$

if $y_t s_t < 0$ then

$w_{t+1} \leftarrow w_t + y_t x_t$

else:

$w_{t+1} \leftarrow w_t$

$$\iff \boxed{w_{t+1} \leftarrow w_t - \alpha \begin{cases} 0 \text{ if } y_t s_t \geq 0 \\ -y_t x_t \text{ otherwise} \end{cases}}$$

So, we introduce : $l^{\text{perceptron}}(s_t, y) = \begin{cases} 0 & \text{if } y_t s_t \geq 0 \\ -y_t s_t & \text{otherwise (because if it's an error, we need } \nabla_w l = -y_t x_t) \end{cases}$

Finally we get : $l^{perceptron}(s_t, y_t) = \max(0, -y_t s_t)$

and we obtain : $w_{t+1} \leftarrow w_t - \nabla_w l^{perceptron}(s_t, y_t)$

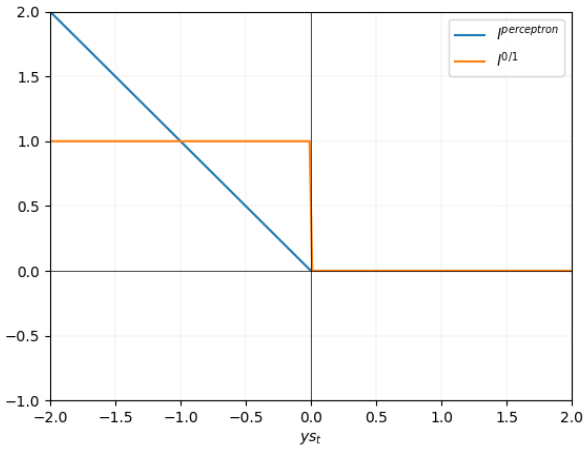

Figure 3: Plot of $l^{perceptron}$ and $l^{0/1}$.

3 Margin and Generalization Bound

3.1 VC Bound

The risk on a function class H with probability $1 - \delta$ is defined as:

$$R(h) \leq R_{\text{emp}}(h) + C\sqrt{\frac{D(log(2N/D) + 1 + log(4\delta)}{N}}$$

where D is the VC dimension of H.

3.2 VC Dimension of the Class of Linear Functions with Margin ρ

Let H be the class of functions $f(x) = w^T x + b$ with a margin ρ from the training examples. Then,

$$D \leq 1 + \min\left(d, \frac{R^2}{\rho^2}\right)$$

where R is the radius of a ball containing the training data.

4 Formulation of the max margin problem

The aim of SVM is :

$$\max p = \frac{1}{\|w\|} \iff \min\|w\| \iff \min \frac{\|w\|^2}{2}$$

subject to the constraints

$$\forall i, y_i(w^T x_i + b) \geq 1$$

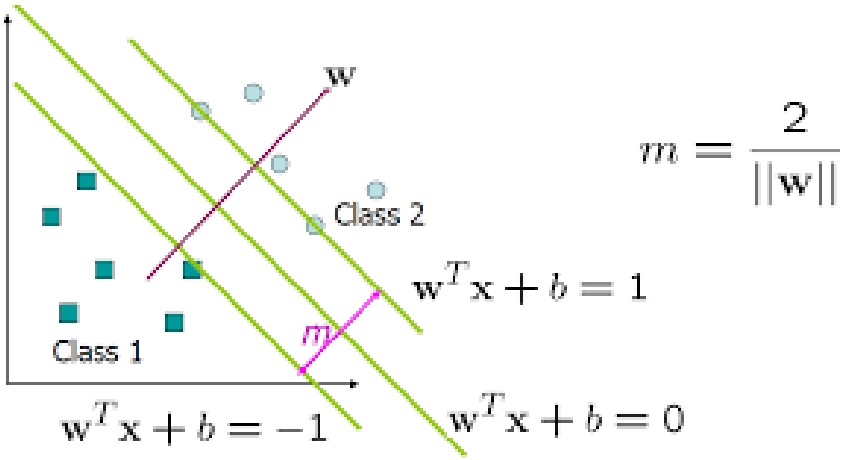

Figure 4: Visualization of the max margin problem

4.1 The SVM Lagrangian

In the general case, the Lagrangian is commonly employed in optimization problems of the form:

$$\min f(w) \quad \text{subject to the constraints} \quad h_i(w) \leq 0$$

The problem is then equivalent to :

$$\max_{\alpha} L(w) = f(w) + \sum \alpha_i h_i(w)$$

4.1.1 Separable Case in SVM

In our case, we get:

$$\max_{\alpha} \frac{1}{2}\|w\|^2 - \sum \alpha_i (y_i(w^T x_i + b) - 1)$$

We have stationarity conditions : $\frac{\partial h}{\partial b} = 0$ and $\frac{\partial h}{\partial w} = 0$

So, we get : $\sum \alpha_i y_i = 0$ et $w = \sum \alpha_i y_i x_i$

So we get : $max_\alpha L = \frac{1}{2}\langle \sum_i \alpha_i y_i x_i, \sum_i \alpha_i y_i x_i \rangle - \sum_i \alpha_i (y_i (\overbrace{\sum_j \alpha_j y_j x_j^T x_i}^{\text{same term as in the other sum}} + \underbrace{b}_{\substack{\text{disappears} \\ \text{because } \sum \alpha_i y_i = 0}}) - 1)$

Whence the problem become :

$$max_\alpha L = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j^T \text{ with } \alpha_i \geq 0$$

However, the complementary stackness condition imposes :
$\alpha_i(y_i(w^T x_i + b) - 1) = 0$ implies either $\alpha_i = 0$ or $y_i(w^T x_i + b) = 1$

Thus, $w = \sum \alpha_i x_i y_i$ corresponds to a sum only for points on the margin line.

4.1.2 Linearly Separable SVM in Practice

Calculating w

Utilize the dataset $D = \{(x_i, y_i)\}_{i=1}^n$ to solve the dual problem! This yields the parameters $\{\alpha_i\}_{i=1}^n$. Consequently, the solution w is given by $w = \sum_{i=1}^n \alpha_i y_i x_i$.

Calculating b

The $\alpha_i > 0$ correspond to the support points that satisfy the relation $y_i(w^T x_i + b) = 1$. Infer the value of b from these support points.

Decision Function

The decision function is given by $f(x) = w^T x + b = \sum_{i=1}^n \alpha_i y_i x_i^T x + b$.

4.2 Non-Separable Case in SVM

4.2.1 Redefining the problem

In the Non-Separable case we can relax the constraints by introducing the concept of error ε.

Relaxing the Constraints

Relax the constraint $y_i(w \cdot x_i + b) \geq 1$

Accepting with an Error Margin

introducing the concept of error (slack variable) ε_i, where $\varepsilon_i \geq 0$.

$$y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i$$

Error Term

Include the sum of errors ($\sum_{i=1}^{n} \varepsilon_i$) in the SVM optimization problem. The objective is to minimize both the norm of the weight vector w and the sum of errors:

$$\min_{w,b,\varepsilon} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \varepsilon_i$$

Subject to the constraints:

$$y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i$$
$$\varepsilon_i \geq 0$$

Here, C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing errors. **It is fixed bu the user**.

4.2.2 Dual Problem

The Lagrangian for the SVM optimization problem is given by:

$$\mathcal{L}(w, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\nu}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i(w \cdot x_i + b) + \xi_i) + \sum_{i=1}^{n} \nu_i \xi_i$$

where $\boldsymbol{\nu_i} \geq 0$, $\boldsymbol{\alpha_i} \geq 0$ for all $i = 1, \ldots, n$.

Stationarity Conditions

The optimality conditions for stationarity are given by:

$$\frac{\partial \mathcal{L}(w, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\nu})}{\partial b} = 0$$
$$\frac{\partial \mathcal{L}(w, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\nu})}{\partial w} = 0$$
$$\frac{\partial \mathcal{L}(w, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\nu})}{\partial \xi_k} = 0$$

which yields the following equations:

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$
$$C - \alpha_i - \nu_i = 0, \quad \forall i = 1, \ldots, n$$

The dual problem for a linear SVM in the non-separable case is so given by:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

4.2.3 Theorem [Solution of a Linear SVM: Non-Separable Case]

Consider a non-separable linear SVM problem with the decision function $f(x) = w^T x + b$. The vector w is defined as $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, where the coefficients α_i are solutions to the dual problem described above.

Compare to separable case, nothing changes except the constraints on α_i, which are now $0 \leq \alpha_i \leq C$.

The choice of C influences the solution: with a small C the margin is large, with a large C the margin is small.

4.2.4 Linearly Non-separable SVM in Practice

The methodology for implementing a linear SVM in the non-separable case is as follows :

1. **Center the Data:** $\{x_i\}_{i=1} \rightarrow \{x_i = x_i - \bar{x}\}_{i=1}$
2. **Set the Parameter $C > 0$ for the SVM.**
3. **Use a Solver to Solve the Dual Problem:** Obtain $\boldsymbol{\alpha}_i \neq 0$, corresponding support vectors $\mathbf{x}_i$, and the bias b, as we did in the separable case.
4. **Derive the Decision Function:**
$$f(x) = \sum_{i \in \text{SV}} \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b$$
5. **Evaluate the Generalization Error of the SVM Obtained:** (Cross-validation, etc.) If unsatisfactory, return to Step 2.

4.2.5 Model Selection: Tuning C for Linear SVM

```
function [bestC] = tuneC(X, Y, options)
    (Xa, Ya, Xv, Yv) = SplitData(X, Y, options);

    for each C
        (w, b) = TrainLinearSVM(Xa, Ya, C, options);
        error = EvaluateError(Xv, Yv, w, b);

    bestC = arg min error;

end
```

5 Relation between soft-SVM, Hinge loss and Hinge-loss perceptron

For soft SVM (SVM with stack variable) we get :
$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum \xi_i \\ y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

For ξ_i we have the followings constraints :
$$\begin{cases} \xi_i \geq 0 \\ \xi_i \geq 1 - y_i(\langle w, x_i\rangle + b) = 1 - y_i s_i \text{ avec } s_i = \langle w, x_i\rangle + b \end{cases}$$

$$\iff \xi_i \geq \max(0, 1 - y_i s_i)$$

Consider this optimisation sub problem :
$$\begin{cases} \min \sum \xi_i \\ \text{s.t. } \xi_i \geq \max(0, 1 - y_i s_i) \end{cases}$$

$\rightarrow$ Solution : $\xi_i = \max(0, 1 - y_i s_i)$

The problem SVM become

$$\begin{aligned} & \min \frac{1}{2}\|w\|^2 + C\sum \max(0, 1 - y_i(\langle w, x_i\rangle + b)) \text{ without constraint now} \\ \iff & \min \left(\frac{1}{2}\|w\|^2 + \sum l^{hinge}(\langle w, x_i\rangle + b, y_i)\right) \\ & \text{where } l^{hinge} = max(0, 1 - y_i s_i \end{aligned}$$

With this loss, we can rewrite the SGD objectif function of SVM :

$$\nabla_w(\frac{1}{2C}\|w\|^2 + \sum_i l^{hinge}(s_i, y_i)) = \frac{w}{C} + \sum_i \begin{cases} 0 & \text{if } y_i s_i > 1, \\ -y_i x_i & \text{else.} \end{cases}$$

So, we obtain the following SGD of soft SVM

if $y_t s_t < 1$ then
$\quad w_{t+1} \leftarrow w_t + \alpha y_t x_t - \frac{\alpha}{C} w_t$
else:
$\quad w_{t+1} \leftarrow w_t - \frac{\alpha}{C} w_t$

ATTENTION : $l^{hing} \geq l^{0/1}$

6 Conclusions

Construction of an optimal hyperplane in terms of maximizing the margin:
An in-depth theoretical analysis shows that maximizing the margin is equivalent to minimizing an upper bound on the generalization error.
The non-linear case (where a non-linear decision function is sought) can be addressed through the use of kernels.
Possible generalization to cases where multiple classes are present.
A widely used classification algorithm in practice.