The Online Perceptron Algorithm And Linear Support Vector Machine

Scribe: Vivien Conti

November 16, 2023

1 Linear Discrimination

1.1 Formulation

Let $D = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1}^n$ be a set of labeled points. We want to construct from D a function $f : \mathcal{X} \to \{-1, 1\}$ or $f : \mathcal{X} \to \mathbb{R}$ that predicts the class -1 or 1 of a point $x \in \mathcal{X}$.

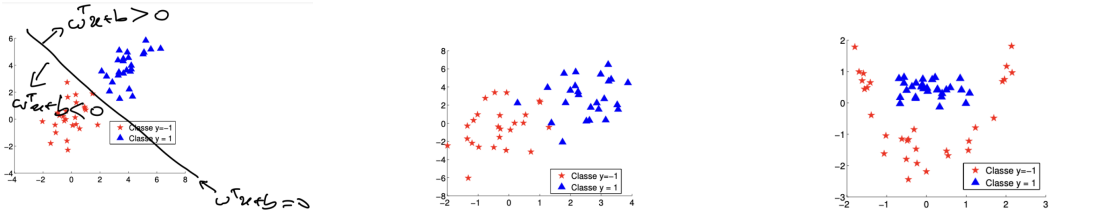
Let the input space be $X = \mathbb{R}^d$. We can construct a scoring function: $f : \mathbb{R}^d \to \mathbb{R}$ such that:

$$f(x) = \begin{cases} f(x) < 0 & \text{assign } x \text{ to class } -1 \\ f(x) > 0 & \text{assign } x \text{ to class } 1 \end{cases}$$

This leads to the introduction of linear scoring function: $f(x) = w^T x + b$, where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Definition 1.1 *Linearly Separable Problem*
The points $\{(x_i, y_i)\}$ are linearly separable if there exists a hyperplane that correctly discriminates the entire set of data. Otherwise, we refer to them as linearly non-separable examples.

Some examples are shown on the figure 1.



(a) Example of separability (b) Example of no separability (c) Example of no separability

Figure 1: Examples of separable and no separable problems

1.2 Linear Separator and Maximization of the Margin

Proposition 1.0.1 *Distance from a Point to the Decision Boundary*
Let $H(w, b) = \{z \in \mathbb{R}^d \mid f(z) = w^T z + b = 0\}$ be a hyperplane, and let $x \in \mathbb{R}^d$. The distance from the point x to the hyperplane H is $d(x, H) = \frac{|w^T x + b|}{\|w\|} = \frac{|f(x)|}{\|w\|}$.

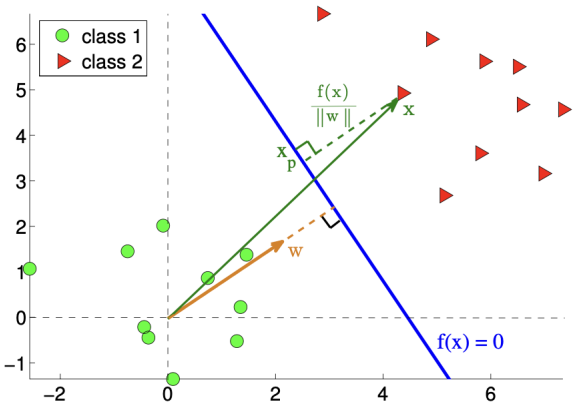


Figure 2: Distance from a Point to the Decision Boundary

Proof 1.0.1 *As we can see on the figure 2:*

$$
\begin{aligned}
x &= x_p + \frac{w}{\|w\|} \times d(x, H) \\
w^T x &= w^T x_p + w^T \frac{w}{\|w\|} \times d(x, H) \\
\|w\| \times d(x, H) &= w^T x - w^T x_p \\
\|w\| \times d(x, H) &= (w^T x + b) - (w^T x_p + b) \\
d(x, H) &= \frac{|w^T x + b|}{\|w\|}
\end{aligned}
$$

The term $(w^T x_p + b) = 0$ *because the point* x_p *is on the hyperplane.*

Definition 1.2 *Canonical Hyperplane*
An hyperplane is said to be canonical with respect to the data $\{x_1, \ldots, x_N\}$ *if* $\min_i |w^T x_i + b| = 1$.

Definition 1.3 *Geometric Margin*
The geometric margin is $M = \frac{2}{\|w\|}$

Definition 1.4 *Optimal Canonical Hyperplane*
The optimal canonical hyperplane respects these two properties (cf figure 3:

- *It maximizes the margin*
- *It correctly classifies each point:* $\forall i, y_i f(x_i) \geq 1$

1.3 Perceptron Algortihm

We are in the case of homogeneous linear classifiers which means: $f(x) = w^T x$. The algorithm is the following:

Theorem 1.1 *Block Norikoff*
Assume $\|x_t\| < R$ *for all* t *and* $y_t \in \{-1, 1\}$. *Assume there exists a canonical hyperplane* $w^\star$ *passing through the origin with half a margin* $\rho = \frac{1}{\|w^\star\|}$.
Then, then number of mistakes of perceptron is at most of $\frac{R^2}{\rho^2}$.

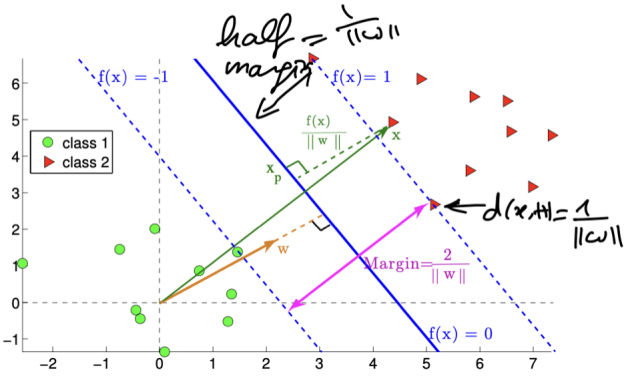


Figure 3: Example of an optimal canonical hyperplane

Algorithm 1 Perceptron algorithm

$w_0 \leftarrow 0$
for $t = 1$ `to` T **do**
  `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
    $w_{t+1} \leftarrow w_t$
  end if
end for

Proof 1.1.1
Step 1)
After an update (a predicttion error), we can say that $_{t+1}$ *'is more aligned' with* $w^\star$*:*

$$\langle w_{t+1}, w^\star \rangle = \langle w_t + y_t x_t, w^\star \rangle$$
$$\langle w_{t+1}, w^\star \rangle = \langle w_t, w^\star \rangle + y_t \langle x_t, w^\star \rangle$$

Because of $w^\star$ *is a canonical hyperplane,* $y_t \langle x_t, w^\star \rangle \geq 1$*. So we have:*

$$\langle w_{t+1}, w^\star \rangle = \langle w_t, w^\star \rangle + 1$$

By unrolling we get: $\langle w_t, w^\star \rangle \geq t_e$ *with* t_e *the number of errors.*

Step 2)
After an update we can write:

$$\|w_{t+1}\|^2 = \langle w_t + y_t x_t, w_t + y_t x_t \rangle$$
$$\|w_{t+1}\|^2 = \|w_t\|^2 + 2 y_t \langle w_t, y_t \rangle + \|y_t x_t\|^2$$

Because of we made an error we have $2 y_t \langle w_t, y_t \rangle \leq 0$*. So we can write:*

$$\|w_{t+1}\|^2 \leq \|w_t\|^2 + R^2$$

By unrolling we get: $\|w_t\|^2 \leq t_e R^2$

Step 3)
By Cauchy-Scharwtz we can write:

$$t \leq \langle w_t, w^\star \rangle \leq \|w_t\| \|w^\star\| \leq \sqrt{t_e} R \|w^\star\|$$
$$\implies \sqrt{t_e} \leq \frac{R}{\rho}$$
$$\implies t \leq \frac{R^2}{\rho^2}$$

1.3.1 Perceptron as a 'SGD' online learner

We can remark that there are not a big difference between perceptron and SGD algorithms. As a reminder, the update in the SGD algorithm is $w_{t+1} \leftarrow w_t - \alpha \nabla_w l(w_t^T x, y)$. Can we designed a loss function to retrieve perceptron algorithm?
Let $S_t = w_t^T x_t$:

$$l(s_t, y_t) = \begin{cases} 0 & \text{if } y_t s_t \geq 0 \\ -y_t s_t & \text{otherwise} \end{cases}$$

Applying SGD algorithm here gives:

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

Which is equivalent to the perceptron algorithm and $l^{perceptron}(s_t, y) = \max(0, 1 - y s_t)$ (cf figure 4).

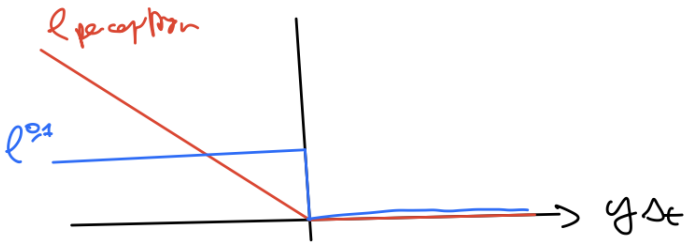


Figure 4: Perceptron loss vs 0/1 loss

1.3.2 Margin and Generalization Bound

If we focus on the risk over a function class $\mathcal{H}$, we can write with probability $1 - \delta$:

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

Where D is the VC dimension of $\mathcal{H}$. Moreover, if we consider $\mathcal{H}$ as the class function $f(x) = w^T x + b$ with a margin ρ we can write:

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

Where R is the radius of a ball containing the training data. The idea here is to see that reducing D allows to reduce the risk of h. And increasing the margin allows to reduce D. So a large margin is a good way to prevent from overfitting.

2 Solving the SVM problem

2.1 Primal Problem and Lagrangian

Definition 2.1 *Large Margin Separator (SVM): Formulation*

- $D = \{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}\}_{i=1}^n$: *linearly separable points*
- *Objective: Find a decision function* $f(x) = w^T x + b$ *that maximizes the margin and correctly discriminates the points in* D.

Definition 2.2 *Primal Problem of SVM*
This is an optimization problem defined as the following:

$$\begin{aligned} \min_{w \in \mathbb{R}, b \in \mathbb{R}} \quad & \frac{1}{2}\|w\|^2 \\ s.t. \quad & y_i(w^T x_i + b) \geq 1 \ \forall i = 1, ..., n \end{aligned}$$

To resolve this we need the Lagrangian and the n Lagrange multipliers $\alpha_i \geq 0$ that correspond to the n inequality constraints. The Lagrangian is written as:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(w^T x_i + b) - 1)$$

2.2 Dual Problem of SVM

Definition 2.3 *Dual Problem of SVM*
The stationary conditions of the primal problem of SVM are:

- $\frac{\partial L(w,b,\alpha)}{\partial b} = 0$
- $\frac{\partial L(w,b,\alpha)}{\partial w} = 0$

wich can be written as:

- $\sum_{i=1}^n \alpha_i y_i = 0$
- $w = \sum_{i=1}^n \alpha_i y_i x_i$

By substituting into the Lagrangian, the dual problem is written as:

$$\begin{aligned} \max_{\{\alpha_i\}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ s.t. \quad & \alpha_i \geq 0 \ \forall i = 1, ..., n \\ & \sum_{i,j=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

The condition of complementary slackness is written as:

$$\alpha_i (y_i(w^T x_i + b) - 1) = 0$$

By solving the dual problem to find the n parameters $\{\alpha_i\}$, two cases are obtained:

- For a point x_j, if $y_j(w^T x_j + b) > 1$, then $\alpha_j = 0$
- For a point x_i, if $y_i(w^T x_i + b) = 1$, then $\alpha_i \geq 0$

Hence, the solution $w = \sum_{i=1}^n \alpha_i y_i x_i$ is uniquely defined by points such as $y_i(w^T x_i + b) = 1$. This is what we called the **support vectors**. In other words, the hyperplane is entirely defined by a linear combination of support vectors (cf figure 5)

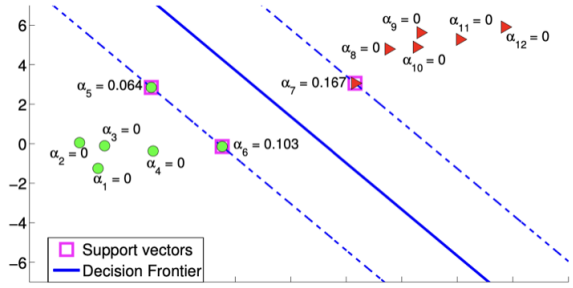


Figure 5: Example of the result of the SVM problem. Here the hyperplane is defined by this linear combination: $w = 0.167x_7 - 0.064x_5 - 0.103x_6$. Here, x_5, x_6 and x_7 are the support vectors.

2.3 SVM for linearly no-separable problems

How can we adapt what we saw in the case of tha data are not linearly separable? In fact, our constraints are too strong and so we need to relax the constraints. That is to say:

- Relax x_i, if $y_i(w^T x_i + b) = 1$
- Accept that x_i, if $y_i(w^T x_i + b) \geq 1 + \epsilon_i$ with ϵ_i the error term.
- Include $\sum_{i=1}^{n} \epsilon_i$ in the SVM problem.

Definition 2.4 *Primal Problem of no-separable SVM*
This optimization problem is defined as the following:

$$\begin{aligned} \min_{w \in \mathbb{R}, b \in \mathbb{R}, \{\epsilon_i\}} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \epsilon_i \\ s.t. \quad & y_i(w^T x_i + b) \geq 1 - \epsilon_i \ \forall i = 1, ..., n \\ & \epsilon_i \geq 0 \ \forall i = 1, ..., n \end{aligned}$$

C is positive and has to be set up by the user. It represents the regularization parameter that indicates how tolerant we are.

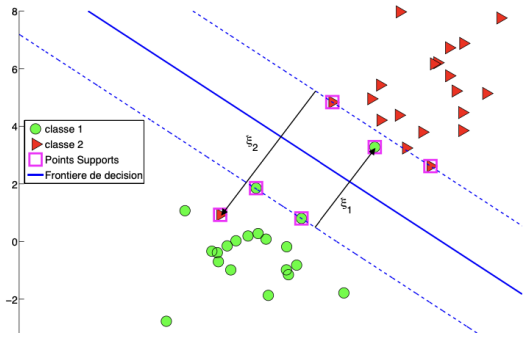


Figure 6: Example of a no-separable SVM problem. The support vectors are indicates by purple bounding boxes.

Definition 2.5 *Dual Problem of no-separable SVM*
For the no-separable problem, the Lagrangian becomes:

$$L(w, b, \epsilon, \alpha, \nu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \epsilon_i - \sum_{i=1}^{n} \alpha_i(y_i(w^T x_i + b) - 1 + \epsilon_i) - \sum_{i=1}^{n} \nu_i \epsilon_i$$

Where α_i, $\nu_i \geq 0$ $\forall i = 1, ..., n$. *The stationary conditions are now:*

- $\frac{\partial L(w,b,\epsilon_i,\alpha)}{\partial b} = 0$
- $\frac{\partial L(w,b,\epsilon_i,\alpha)}{\partial w} = 0$
- $\frac{\partial L(w,b,\epsilon_i,\alpha)}{\partial \epsilon_k} = 0$

wich can be written as:

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

By substituting into the Lagrangian, the dual problem is written as:

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

Theorem 2.1 *Solution of linear SVM: no-separable case*
Consider a linear non-separable SVM problem with a 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 the solutions of the dual problem above.*

Compared to the previous separable case, very few things have changed. The condition on α_i is now different since we have $0 \leq \alpha_i \leq C$. The influence of the C parameter is shown on the figure 7.

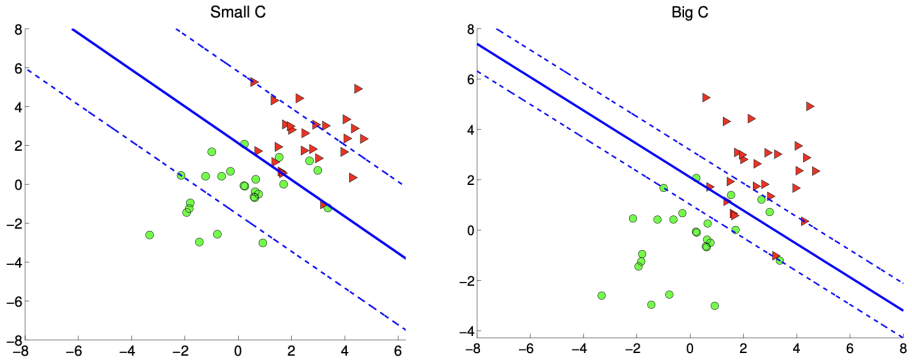


Figure 7: Example of the influence of C parameter. If C is small (left) then the margin is big and we accept a lot of errors. If C is big (right) then the margin is small and we accept a small amount of errors.

3 Relation Between soft SVM, Hinge-loss and Hinge-loss Perceptron

The soft-SVM problem is considering the no separable case. The optimization problem is the following :

$$\begin{aligned} \min_{w \in \mathbb{R}, b \in \mathbb{R}, \{\epsilon_i\}} \quad & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \epsilon_i \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \epsilon_i \ \forall i = 1, ..., n \\ & \epsilon_i \geq 0 \ \forall i = 1, ..., n \end{aligned}$$

The constraint on ϵ_i give the following:

$$\begin{aligned}
&\epsilon_i \geq 0 \\
&\epsilon_i \geq 1 - y_i(\langle w, x_i \rangle + b = 1 - y_i s_i \\
&\epsilon_i \geq \max(0, 1 - y_i s_i)
\end{aligned}$$

We can hence considering this optimisation sub-problem:

$$\begin{aligned}
\min_{\{\epsilon_i\}} \quad & \sum_{i=1}^{n} \epsilon_i \\
\text{s.t.} \quad & \epsilon_i \geq \max(0, 1 - y_i s_i) \ \forall i = 1, ..., n
\end{aligned}$$

The solution are $\epsilon_i = \max(0, 1 - y_i s_i)$. This is also the solution on ϵ_i to the original problem!

The soft SVM problem becomes:

$$\min_{w \in \mathbb{R}, b \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(\langle w, x_i \rangle + b))$$

We remark that there are no more constraints anymore it's a convex optimization problem:

$$\min_{w \in \mathbb{R}, b \in \mathbb{R}} \quad \frac{1}{2C}\|w\|^2 + \sum_{i=1}^{n} l^{\text{hindge}}(\langle w, x_i \rangle + b, y_i)$$

where $l^{\text{hindge}}(s_i, y_i) = \max(0, 1 - y_i s_i)$.

We can visually compare the difference between l^{hindge}, $l^{0,1}$ and $l^{\text{perceptron}}(s_t, y_t) = \max(0, -y_t s_t)$ in Figure 8. Notice that $l^{\text{hindge}} \geq l^{0,1}$.

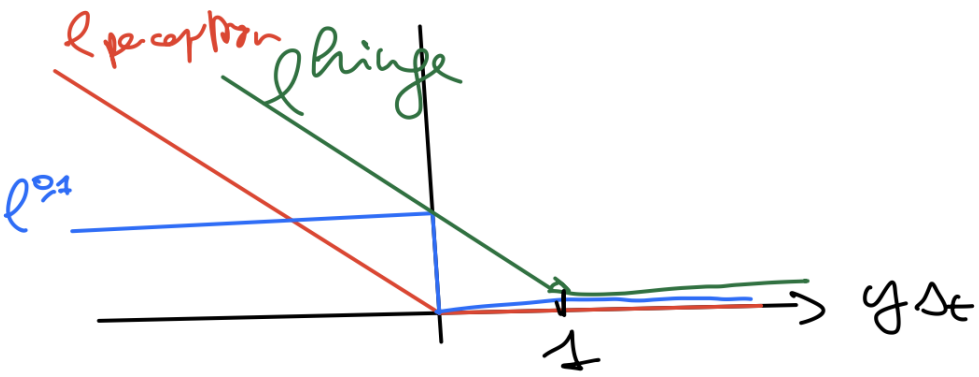


Figure 8: Losses comparison

If we look at the SGD on the objective function we get:

$$\begin{aligned}
&\nabla(\frac{1}{2C}\|w\|^2 + \sum_{i=1}^{n} l^{\text{hindge}}(\langle w, x_i \rangle + b, y_i)) \\
&= \frac{w}{c} + \sum_{i=1}^{n} \begin{cases} 0 & \text{if } y_i s_i \geq 1 \\ -y_i x_i & \text{otherwise} \end{cases}
\end{aligned}$$

Finally the SGD algorithm for SVM becomes:

Algorithm 2 SGD algorithm for SVM

if $y_t\langle w_t, x_t\rangle \leq 1$ **then**
  $w_{t+1} \leftarrow w_t + y_t x_t - \frac{\alpha}{c} w_t$
else
  $w_{t+1} \leftarrow w_t - \frac{\alpha}{c} w_t$
end if

The condition $y_t\langle w_t, x_t\rangle \leq 1$ allows to optimze if a point is far from the margin. Also, the term $\frac{\alpha}{c} w_t$ could be considering as a weight decay for the regularization.