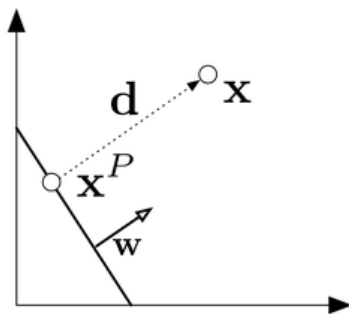


# SVM notes

Matteo Lugli

October 22, 2023

## 1 Margins



Considering the above figure,  $x^p = x - d$  and  $d = w\alpha$ . But  $x^p$  is also a point on the hyperplane, so  $w^t x^p + b = 0$ . By substituting we get  $w^t(x - \alpha w) + b = 0$ .

$$w^t - w^t \alpha w + b = 0$$

$$\frac{w^t x + b}{w^t w} = \alpha$$

$$d = \frac{w^t x + b}{w^t w} w$$

Now we want to calculate the magnitude of the distance  $d$ , which is

$$\begin{aligned} \|d\|_2 &= \sqrt{d^t d} = \sqrt{\alpha^2 w^t w} = \sqrt{\alpha^2 w^t w} = \alpha \sqrt{w^t w} \\ &= \frac{w^t x + b}{w^t w} \sqrt{w^t w} = \frac{w^t x + b}{\sqrt{w^t w}} = \frac{w^t x + b}{\|w\|_2} \end{aligned}$$

So the margin is defined as follows:

$$\gamma(w, b) = \min_{x \in D} \left( \frac{w^t x + b}{\|w\|_2} \right) \quad (1)$$

In the SVM problem we want to find the separating hyperplane that maximizes the margin:

$$\begin{aligned} \max_{w, b} (\gamma(w, b)) &= \max_{w, b} \left( \min_{x \in D} \left( \frac{w^t x + b}{\|w\|_2} \right) \right) \\ &= \max_{w, b} \left( \frac{1}{\|w\|_2} \min_{x \in D} (w^t x + b) \right) \end{aligned}$$

If we consider the hyperplane  $w^t x + b$  it is scale invariant, meaning that we can multiply  $w$  and  $b$  for whatever number  $\beta$  and nothing will change. Intuitively, you can think of it this way: imagine that the weights of your classifier are  $[0.1, 0.2, 0.3]$ . It means that feature 1 of your vector has importance 0.1 on the overall classification, feature 2 has importance 0.2, etc. . .

The important thing here is that feature 2 is twice more important than feature 1, so if you multiply everything by the same amount nothing will change. It means that we can set  $\beta$  so that  $w^t + b = 1$ , obtaining the following objective function:

$$\max_{w, b} \frac{1}{\|w\|_2} (1) = \min_{w, b} \|w\| \quad (2)$$

Given that  $f(z) = z^2$  is a monotonic function, we can write our full optimization problem like this:

$$\begin{aligned} &\min_{w, b} \|w\|^2 \\ \text{s.t.} \quad &\forall_i \quad y_i(w^t x_i + b) \geq 0 \\ &\min_i \quad w^t x_i + b = 1 \end{aligned}$$

We can intuitively combine the two constraints in a single one and get the final formulation:

$$\begin{aligned} &\min_{w, b} \|w\|^2 \\ \text{s.t.} \quad &\forall_i \quad y_i(w^t x_i + b) \geq 1 \end{aligned} \quad (3)$$

This problem can be solved by using a quadratic programming solver, and it's totally fine. Moreover there are many ways to re-formulate and solve the problem, which are particularly useful when our data is not linearly separable.

## 2 Lagrangian

In machine learning we usually deal with problems in which points are not linearly separable. The previous problem can be re-written in a way that allows us to use the so called **Kernel Trick**. To do that we need to introduce the *Lagrangian formulation*. It is basically a way to reduce a constrained optimization problem into a single equation.

$$\begin{aligned}
 & \min f(x) \\
 & s.t \quad h_i(x) \geq 0 \quad \forall i = 1 \dots m \\
 & \Downarrow \\
 & L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i h_i(x)
 \end{aligned} \tag{4}$$

where  $\lambda_i$  are called **lagrangian multipliers** and are all  $\geq 0$ .

The lagrangian dual function is defined like this:

$$g(\lambda) = \min_x (L(x, \lambda)) \tag{5}$$

According to the **strong duality** we can write that

$$d^* = \max_{\lambda \geq 0} g(\lambda) = \min f(x) = p^* \tag{6}$$

where  $d^*$  is the solution of the dual and  $p^*$  is the solution of the primal. So to find the equation of the dual, we must set  $\frac{dL}{dx_1} = 0, \frac{dL}{dx_2} = 0, \dots, \frac{dL}{dx_n} = 0$ . If we plug what we obtain in L we get a function of  $\lambda$  only, which is what we need. So let's try to do that for our specific minimization problem and see what we get. Remember that we have a constraint for each one of the points in our dataset, so  $x^1 \dots x^i$

$$L(w, b, \lambda) = \frac{1}{2} w^t w + \sum_{i=1}^m \lambda_i (1 - y_i (w^t x^{(i)} + b)) \tag{7}$$

So if we simply compute the derivatives  $\frac{dL}{dw} = 0$  and  $\frac{dL}{db} = 0$  we get:

$$\begin{aligned}
 \frac{dL}{db} = 0 & \Rightarrow \sum_{i=1}^m \lambda_i y_i = 0 \\
 \frac{dL}{dw} = 0 & \Rightarrow w = \sum_{i=1}^m \lambda_i y_i x_i
 \end{aligned} \tag{8}$$

Let's plug the easiest of the two conditions (8) in the lagrangian and see what we get:

$$\frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x^{(i)t} x^{(j)} + \sum_i \lambda_i - \sum_{i,j} \lambda_i y_i \lambda_j y_j x^{(i)t} x^{(j)} \quad (9)$$

This sum is composed by three terms: if you look carefully the first and the last one are the same (except for the  $\frac{1}{2}$ ), so we can simplify:

$$\sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x^{(i)t} x^{(j)}$$

let's write the formulation of the new problem:

$$\begin{aligned} \max_{\lambda \geq 0} \quad & \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \langle x^{(i)} x^{(j)} \rangle \\ \text{s.t.} \quad & \sum_{i=1}^m \lambda_i y_i = 0 \end{aligned} \quad (10)$$

Now we have a new objective function that only depends on  $\lambda$ . It also contains a inner product on which we can apply a **Kernel** (explained later).

### 3 KKT conditions

There is a series of equations called KKT conditions that summarize the relation of the primal and the dual problem. If  $x^*$  is a solution for the primal problem,  $f$  is convex and all of the  $h_i(x)$  are linear constraints, then there exist some multipliers  $\lambda_1 \dots \lambda_m$  such that:

$$\nabla f(x^*) - \lambda^t \nabla h(x^*) = 0 \quad (11)$$

$$h(x^*) \geq 0 \quad (12)$$

$$\lambda_i \geq 0 \quad \forall i = 1 \dots m \quad (13)$$

$$\lambda^t h(x^*) = 0 \quad (14)$$

where  $\lambda^t$  is the solution of the dual! The first 3 conditions are quite intuitive, the last one is the tricky one. Let's write the demonstration for that. Let  $x^*$  be the solution of the primal and  $\lambda^*$  be the solution of the dual. Because of the strong duality, we can write the following:

$$\begin{aligned} f(x^*) = g(\lambda^*) &= \min_x \left( f(x) + \sum_{i=1}^m \lambda_i^* h_i(x) \right) \\ &\geq f(x^*) + \sum_{i=1}^m \underbrace{\lambda_i^*}_{\geq 0} \underbrace{h_i(x^*)}_{\geq 0} \\ &\geq f(x^*) \end{aligned}$$

which means that  $\sum_{i=1}^m \lambda_i^* h_i(x^*) = 0 \Rightarrow \lambda^{*T}(x^*) = 0$ .

This means that if we have a lagrangian multiplier  $\lambda_i \geq 0$ , then the relative constraint should be active (it is satisfied with equality), so  $h_i(x^*) = 0$ .