Support Vector Machines - Passerini

Mattia Carolo - @Carolino96

Support Vector Machines

Support Vector Machine, SVM from now on, are linear classifiers that separate using a **large margin classifier** which solution depends only on a small subset of the traing examples called **support vector**. It's very important to note that it has a sound generalization theory (not to study) and they can be easily extended to non linear separation retaining the separation properties thanks to *kernel machines*.

Maximum margin Classifier

Let's try to formalize the margin. We already know that $yf(x)$ is the confidence on the correct prediction, if negative the prediction is wrong otherwise if positive correct and the value is the confidence on the prediction. Now suppose we have a classifier that correctly separates with no training errors. If this is the case the minimum value among the training examples is called *confidence margin* and it's written like

$$\rho = \min_{(\mathbf{x},y)\in D} yf(\mathbf{x})$$

Since it depends on w we can compute the distance from the minimal distance to our classifier and it's called **geometric margin** which is formalized like

$$\frac{\rho}{\|\mathbf{w}\|} = \min_{(\mathbf{x},y)\in D} \frac{yf(\mathbf{x})}{\|\mathbf{w}\|}$$

Ideally we want to maximize the last formula in order to get w in order to maximize the margin. However if we put in an optimization problem we have actually one degree of freedom that is being removed. Suppose we have a solution where

$$\mathbf{w}^T\mathbf{x} + w_0 = 0$$

now if we want to characterize further the plane we can, for example, multiply the terms with an $\alpha \neq 0$ and still we will return to a formula that look like before since we can incorporate the α in our formalization. This is because there is an infinite number of equivalent formulation for the same hyperplane even with different parameters.

We can counter this problem through the introduction of the *canonical hyperplane* in which we set the constraint that ρ must be equal to a number given a priori (in our case we take 1) in order to get:

$$\rho = \min_{(\mathbf{x},y)\in D} yf(\mathbf{x}) = 1$$

and it's geometric margin will be $\dfrac{\rho}{\|\mathbf{w}\|} = \dfrac{1}{\|\mathbf{w}\|}$

the numerical value in the geometric margin must match

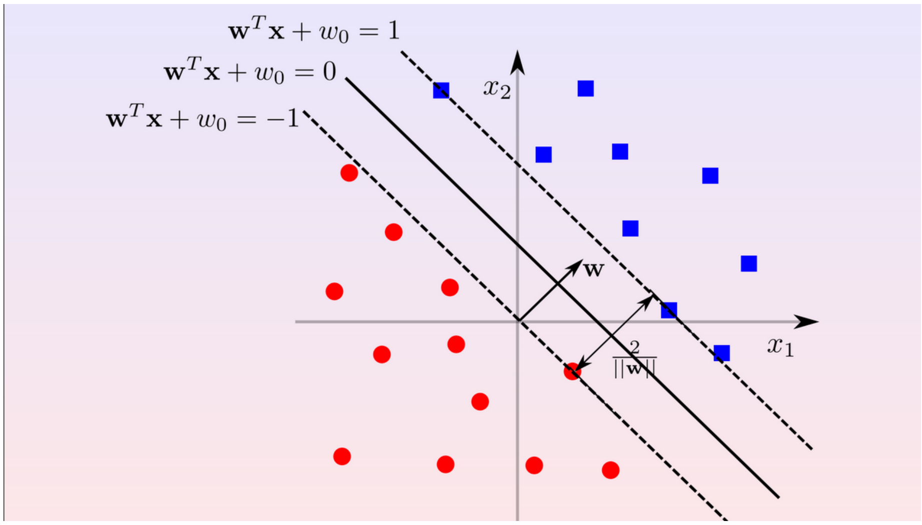
$\mathbf{w}^T\mathbf{x} + w_0 = 1$
$\mathbf{w}^T\mathbf{x} + w_0 = 0$
$\mathbf{w}^T\mathbf{x} + w_0 = -1$
x_2
$\mathbf{w}$
x_1
$\frac{2}{||\mathbf{w}||}$

Figure 1: MLClassifier.png

As we can see from the image above the two dotted lines are the two canonical hyperplanes with their ρ set to 1 so summing their respective geometric margin we get that the total geometric margin is equal to $\dfrac{2}{||\mathbf{w}||}$.

We can take this and convert it to an optimization problem.

First of all we want to maximize the margin so $\dfrac{2}{||\mathbf{w}||}$ and we want to do it by enforcing all examples to stay on the correct part of the hyperplane for both canonical ones. Formalized will be

$$\max \frac{2}{||\mathbf{w}||} \text{s.t.} \forall x_i : y_i = 1 \Rightarrow \mathbf{w}^T x_i + w_0 \geq 1 \;\; \& \; \forall x_i : y_i = -1 \Rightarrow \mathbf{w}^T x_i + w_0 \leq -1$$

and the term to maximize can be inverted in order to get

$$\min\frac{||\mathbf{w}||}{2} = \frac{\sqrt{w^T w}}{2}$$

which is not a quadratic function but it's monotonic so if we found a maximum it will be the same even squared so we can minimize doing $\min\frac{||\mathbf{w}||^2}{2}$ and to summarize the constraints we can just say $yf(x) \geq 1$ since it's our confidence

Margin Error Bound (just a citation not study material)

Margin Error Bound: $\nu + \sqrt{\frac{c}{m}(\frac{R^2 \bigwedge^2}{\rho^2}\ln^2 m + \ln(\frac{1}{\delta}))}$

The probability of test error so depends on:

- ν is number of margin errors (samples that are outside the confidence margin, correcly classified samples with low confidence)
- m training example in the $\sqrt{\frac{\ln^2 m}{m}}$ so the result goes down if m goes up
- R is the radius of the space containing all the samples
- larger the margin ρ, the smaller test error (so we want the margin $\frac{2}{\|\mathbf{w}\|}$ to be large)

  if ρ is fixed to 1, maximizing margin corresponds to minimizing $||\mathbf{w}||$
- c is a constant

  it makes an upper bound of the generalization error (?)

The name **hard margin** is because we require all examples to be at confidence margin at least one.

Learning Problem

The learning problem is formalized like $\min\frac{||\mathbf{w}||^2}{2}$ with linear constraints in w $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1, \forall(\mathbf{x}_i, y_i) \in D$. Still this is a quadratic optimization problem which means that is convex and it has only one global optimum. Problem now is that we need to minimize respect to the constraints and one way to do this is the **KKT approach**

Karush-Kuhn-Tucker (KKT) approach

With this approach basically we turn a *constrained problem* into an *uncostrained* one with the same solution. To do this suppose we have $f(z)$ to minimize with some constraints like $g_i(z) \geq 0 \forall \mathbf{i}$. Now how can we het rid of the constraints? To do so we introduce a non negative variable called **Lagrange multiplier** noted with $\alpha_i \geq 0$ for each constraint and we rewrite the optimization problem as a **Lagrangian**:

$$\min_{z} \max_{\alpha \geq 0} f(z) - \sum_{i} \alpha_i g_i(z)$$

If we find an optimum of this lagrangian called z^* it's still an optimum for the original constrained problem. That's because suppose we find a solution called z' than:

- if at least one constraint is not satisfied ($\exists i \mid g_i(z') < 0$), maximizing over α_i leads to an infinite value;
- if all constraints are satisfied, maximizing over α sets all elements in the sum to zero so that z' is a solution for $\min\limits_{z} f(z)$.

Applying the approach to our learning problem we will get that

$$\min_{\mathbf{w}, w_0} \frac{1}{2} ||\mathbf{w}||^2$$

subject to:

$$y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1$$
$$\forall (\mathbf{x}_i, \mathbf{y_i}) \in D$$