Machine Learning Theory
Tübingen University, WS 2016/2017
Lecture 12

Tolstikhin Ilya

Abstract

In this lecture we will consider Support Vector Machines (SVM). We will start with linearly separable case and the corresponding *hard margin SVM* algorithm. Then we will relax the separability assumption and introduce the *soft margin SVM* algorithm. We will provide dual problems for both of the algorithms, using representer's theorem. Finally, we will extend these algorithms to the non-linear case by means of a kernel trick. Blue colour will be used to highlight parts appearing in the upcoming homework assignments.

1 Hard margin SVM

We will be solving a binary classification problem with $\mathcal{Y} = \{-1, +1\}$ and $\mathcal{X} = \mathbb{R}^d$. In this section we will make the following assumption on our training set $S_n = \{(X_i, Y_i)\}_{i=1}^n$: *there is a vector* $w^* \in \mathbb{R}^d$ *such that*

$$Y_i \cdot \langle X_i, w^* \rangle_{\mathbb{R}^d} > 0, \quad i = 1, \ldots, n. \tag{1}$$

What this assumption says is that the training set is *linearly separable*, i.e. there is a linear classifier making no mistakes on S_n. Indeed, $Y_i \cdot \langle X_i, w^* \rangle_{\mathbb{R}^d} > 0$ means that the classifier $\operatorname{sgn}\langle X_i, w^* \rangle_{\mathbb{R}^d}$ does not make a mistake on the point X_i. Moreover, by demanding that all the margins are *strictly positive*, we ruled out the cases when, while there are no mistakes, some of the training points will rest on the line $\langle X_i, w^* \rangle_{\mathbb{R}^d} = 0$. This could happen, for instance, if $\mathcal{X} = \mathbb{R}^2$ and the training set consists of only 3 points: $X_1 = (0,0), Y_1 = +1$, $X_2 = (0,1), Y_2 = -1$, $X_3 = (0,2), Y_3 = +1$.

One more comment is in place here. Note that conditions (1) state that there is a *homogeneous* linear separator, perfectly separating S_n, i.e. a classifier with hyperplane passing through the origin. We use this assumption as it will simplify computations a bit. If we would still want to allow for hyperplanes not passing through the origin, we could (a) add one more dimension to the input points, (b) put constant 1 in $d+1$-st coordinate for all the points, and (c) repeat everything we did for $\mathbb{R}^{d+1}$.

Large margin Now note that, because we demanded margins in (1) to be strictly positive, there are infinitely many hyperplanes perfectly separating the training set S_n. Every one of them is ERM. Which one should we choose? A rather natural choice is to say *among all the hyperplanes perfectly separating* S_n*, let's choose the one maximizing the distance to the closest point in* S_n. This is indeed reasonable, as intuitively this should be the most *robust choice*, i.e. if we observe another i.i.d. training set S'_n from the same distribution and the points will be perturbed, we expect the performance of a maximum margin hyperplane to be affected the least.

Note that for any homogeneous hyperplane $\langle x, w\rangle_{\mathbb{R}^d}$ and a point x_0 the distance from x_0 to the hyperplane is measured by

$$\frac{|\langle x_0, w\rangle_{\mathbb{R}^d}|}{\|w\|}.$$

In other words, in order to find the maximum margin hyperplane we need to solve the following constrained optimization problem:

$$\max_{w\in\mathbb{R}^d} \min_{i=1,\dots,n} \frac{|\langle X_i, w\rangle_{\mathbb{R}^d}|}{\|w\|}; \quad \text{s.t. } Y_i\langle X_i, w\rangle_{\mathbb{R}^d} > 0. \tag{2}$$

Because of the assumption we made, we are sure that all the constraints can be actually satisfied. Now, take any vector w satisfying the constraints. Note that multiplying w by any positive constant $c > 0$ does not change anything: neither the value of the objective in the problem (2), nor the predictions of the associated classifier $\operatorname{sgn}\langle x, c\cdot w\rangle_{\mathbb{R}^d}$. The same holds also for the vector $w^\star$ solving (2). In other words, this is a degree of freedom, which we need to resolve. We will now announce that while solving (2) we will consider only vectors $w\in\mathbb{R}^d$ for which

$$\min_{i=1,\dots,n} |\langle X_i, w\rangle_{\mathbb{R}^d}| = 1.$$

Once again note that, whatever vector w perfectly separates the training set, we can always choose c so as to guarantee $\min_{i=1,\dots,n} |\langle X_i, c\cdot w\rangle_{\mathbb{R}^d}| = 1$. Finally, we may rewrite the original problem in the following way:

$$\max_{w\in\mathbb{R}^d} \frac{1}{\|w\|}; \quad \text{s.t. } Y_i\langle X_i, w\rangle_{\mathbb{R}^d} \geq 1$$

or equivalently

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w\|^2; \quad \text{s.t. } Y_i\langle X_i, w\rangle_{\mathbb{R}^d} \geq 1. \tag{3}$$

Notice that two last optimization problems have exactly the same solutions. The fact that we chose to minimize $\frac{1}{2}\|w\|^2$ instead of $\|w\|$ or any other monotonic function of the norm is dictated purely by convenience, as it will simplify certain computations below. This form of the hard-margin SVM is often called *primal* form. Problem (3) is a quadratic program which can be solved efficiently using great number of available software packages.

Representer theorem and a dual form Consider the following optimization problem:

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^n 1\{Y_i\langle X_i, w\rangle_{\mathbb{R}^d} < 1\}, \tag{4}$$

where $C > 0$ is any constant. Note that we can equivalently rewrite this problem in the following way:

$$\min_{f\in\mathcal{H}_k} \frac{1}{2}\|f\|^2_{\mathcal{H}_k} + \frac{C}{n}\sum_{i=1}^n 1\{Y_i f(X_i) < 1\},$$

where k is a linear kernel on $\mathbb{R}^d$ with the corresponding RKHS $\mathcal{H}_k$. We may now apply the representer theorem, which tells us that we may seek for solution of the last problem in the form

$$f^*_C(X) = \sum_{i=1}^n \alpha_{i,C} k(X_i, X) = \sum_{i=1}^n \alpha_{i,C}\langle X_i, X\rangle_{\mathbb{R}^d} = \left\langle \sum_{i=1}^n \alpha_{i,C} X_i, X\right\rangle_{\mathbb{R}^d}.$$

Going back to the original w variables we conclude that the problem (4) has a solution of the form

$$w^*_C = \sum_{i=1}^n \alpha_{i,C} X_i.$$

Why did we start considering problem (4) in the first place? The thing is, if we take $C \to \infty$ then, under the linear separability assumption, problem (4) becomes equivalent to (3). Indeed, unless w satisfies all the margin constraints, the second term in (4) will be non-zero and explode because $C \to \infty$. Thus, w will have to make sure the second term is zeroed out, i.e. all the margin constraints are met. This results precisely in (3).

To summarize, we just showed that the solution of (3) can be searched in the form

$$w^* = \sum_{i=1}^n \alpha_i X_i.$$

I.e., we may instead solve the following problem:

$$\min_{\alpha_1,\dots,\alpha_n\in\mathbb{R}} \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\langle X_i, X_j\rangle_{\mathbb{R}^d}; \quad \text{s.t. } Y_i\sum_{j=1}^n \alpha_j\langle X_i, X_j\rangle_{\mathbb{R}^d} \geq 1. \tag{5}$$

This form of the hard-margin SVM is often called *the dual form.* Note that this is again a quadratic optimization problem. Interesting thing to notice is that in (5) we are tuning n parameters, while in (3) we have d parameters. Depending on whether d is larger than n or not, we may prefer one of these two forms.

2 Soft margin SVM

In previous section we assumed the linear separability of the training set S_n. However, this may be too strong an assumption to make. How could we relax this assumption?

Let's try to mimic the previous algorithm, but this time allow the margin conditions in (3) to be partially violated:

$$\min_{\substack{w\in\mathbb{R}^d\\ \xi_1,\dots,\xi_n\in\mathbb{R}}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i; \quad \text{s.t. } Y_i\langle X_i, w\rangle_{\mathbb{R}^d} \geq 1-\xi_i, \quad \xi_i \geq 0, \tag{6}$$

where C is a positive constant and the *slack variables* ξ measure how much are we violating the margin conditions. Note that if $\xi_i > 1$ we even allow the classifier to make a mistake on X_i. However, we want to punish these violations, and this is why we add the sum of slacks to the objective function. Now we see that C controls the balance between making sure that w corresponds to the large margin and making as few margin violations as possible. In extreme case when $C \to \infty$, if the linear separability assumption holds, we get back to the hard margin SVM (3). Finally, it is important to notice that we decided to penalize violations by adding the sum of slacks, while we could consider adding any other transformations, for instance $\sum_{i=1}^n \xi_i^2$. However, this would lead to a different algorithm with another solution.

If w is fixed, it is easy to see that the optimal slacks can be computed in the following form:

$$\xi_i^* = (1 - Y_i\langle X_i, w\rangle_{\mathbb{R}^d})_+ := \max\{0, 1 - Y_i\langle X_i, w\rangle_{\mathbb{R}^d}\}.$$

Plugging this back to (6) we obtain the following equivalent unconstrained optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n (1 - Y_i \langle X_i, w\rangle_{\mathbb{R}^d})_+. \tag{7}$$

Using representer theorem once again we conclude that the solution of (6) and (7) can be searched in the form

$$w^* = \sum_{i=1}^n \alpha_i X_i.$$

I.e. we need to solve the following:

$$\min_{\alpha_1,\ldots,\alpha_n \in \mathbb{R}} \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle X_i, X_j\rangle_{\mathbb{R}^d} + C\sum_{i=1}^n \left(1 - Y_i \sum_{j=1}^n \alpha_j \langle X_i, X_j\rangle_{\mathbb{R}^d}\right)_+. \tag{8}$$

This is a convex problem and can be solved using gradient descend methods. Again note that we ended up with two problems, one optimizing over n parameters and the other over d parameters.

3 Kernel trick

Finally, what if we want to use non-linear methods? Good news are, we can use precisely the same algorithms (hard and soft margin SVM) to obtain a non-linear separating boundaries. Note that the dual forms of SVMs, i.e. problems (5) and (8) depend on points only through the inner products $\langle X_i, X_j\rangle_{\mathbb{R}^d}$. So does the final classifier: $f(x) = \operatorname{sgn}\sum_{i=1}^n \alpha_i \langle X, X_i\rangle_{\mathbb{R}^d}$.

Now, remember how we saw that once we have a kernel k, we also get a Hilbert space (RKHS) $\mathcal{H}_k$, and also a way to map all the inputs $X \in \mathcal{X}$ to $\mathcal{H}_k$ via $X \mapsto k(X, \cdot) \in \mathcal{H}_k$. Imagine running the same hard-margin or soft-margin SVM algorithms, but replacing all the occurrences of $\langle X', X''\rangle_{\mathbb{R}^d}$ with $k(X', X'')$. This will implicitly mean that we are (a) first mapping our input space $\mathcal{X}$ to $\mathcal{H}_k$ and then (b) running hard (soft) margin SVM in the feature space $\mathcal{H}_k$. One delicate thing to note here is that we are mapping input points X to the *functions* $k(X, \cdot)$. Thus, the linear separator in $\mathcal{H}_k$ takes *a function* $f \in \mathcal{H}_k$ and returns a label (or a number). This kind of mappings are generally called *functionals.*

Now, we already saw that generally $\mathcal{H}_k$ can be of much higher dimensionality than $\mathcal{X}$. For instance recall the polynomial kernel, or even Gaussian kernel corresponding to the feature space $\mathcal{H}_k$ of infinite dimensions. This would mean that we are running SVM in $\mathbb{R}^\ell$ with huge or even infinite ℓ. Luckily, we have the dual equivalent forms (5) and (8), which reduce this hard problem to the one of tuning only n real parameters! Moreover, the linear separator (linear functional) on $\mathcal{H}_k$ may correspond to a non-linear one in the original space $\mathcal{X}$! Assume we already ran the dual SVM algorithm in the feature space $\mathcal{H}_k$ and obtained weights $\alpha_1^*, \ldots, \alpha_n^*$. Now we receive the new test point $X \in \mathcal{X}$ and need to classify it. For this we first map X to $k(X, \cdot) \in \mathcal{H}_k$ and apply our classifier, which will result in:

$$\operatorname{sgn}\sum_{i=1}^n \alpha_i^* \langle k(X_i, \cdot), k(X, \cdot)\rangle_{\mathcal{H}_k} = \operatorname{sgn}\sum_{i=1}^n \alpha_i^* k(X, X_i).$$

Surprisingly, note that the function inside the sgn is *nonlinear*! For instance, for a polynomial kernel of degree 2 it will be a curve of degree 2.