Kernel Methods and SVMs: A convex optimization point of view

Gael Lederrey   Corentin Tallec

June 1, 2016

Abstract

Kernel methods, and more specifically support vectors machines, are well known, well theorized and efficient machine learning tools. For years, they were presenting state of the art performances in many machine learning fields, before being overtaken by deep learning methods. They have the enormous advantage of being well understood theoretically, and most interestingly to belong to the class of convex optimization problems, which is clearly not the case of the latter.

In this document, we aim at presenting the general framework of Kernel methods, to expose how they relate to convex optimization, and how their optimization can be undertaken in practice. Besides, we intend to give a more precise treatment of the support vectors machines special case.

Most theoretical results used here are presented, explained and proved more thoroughly in the "Master vision et apprentissage" of the "Ecole normale superieure de Cachan" course on kernels.

1 Kernel methods

1.1 Generalities

Kernel methods are tools broadly used in various fields, but most notably in machine learning and pattern discrimination. Among this class of methods, support vector machines are probably the best known kernel method. This class of algorithm is used to find and study some types of relations in datasets, *e.g.* classifications of the data, correlations between them, etc. Many machine learning algorithms rely on the idea of embedding the data at hand into a vector space that makes it easier analyse. Kernel methods provide a quite general framework for doing so, and provide tools to transport data into possibly infinite dimensional spaces. Kernel methods rely on a measure of similarity between input data known as a **kernel**, which has to be selected beforehand.

A typical case of application for a kernel method is the following. Define $\mathcal{X}$ (resp. $\mathcal{Y}$) as a set of datapoints (resp. a set of labels). Given a training set, that is a set of couples $(x_i, y_i)_{i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$, we want to find an $f : \mathcal{X} \mapsto \mathcal{Y}$ in a particular set of functions $\mathcal{F}$ such that f predicts accurately labels in the training set, and is able to generalize to unseen datapoints. One way of doing so is to solve the following optimization problem:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \Omega(f) \tag{1}$$

where L is a cost function, that measure how close $f(x_i)$ is to y_i for each i, λ is a positive real number and Ω is a regularizing term, that aims at describing the complexity of a certain function f. In the general case, this optimization problem is not required to be convex. Kernel methods provide a set of function $\mathcal{F}$ and a regularizing term $\Omega(f)$ such that when the cost function is assumed to be convex in its second argument, the entire problem is convex, and can be easily solved.

In the following sections, kernel functions are defined, as well as their relation to the idea of feature mapping, and the general principle of kernel methods is exposed.

1.1.1 Positive Definite Kernel

Denote by $\mathcal{X}$ an arbitrary input set (notably $\mathcal{X}$ is not required to be a vector space). A kernel can be seen as a measure of similarity between two elements of $\mathcal{X}$. Formally

Definition 1 *A positive definite (p.d.) kernel on $\mathcal{X}$ is a function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ that is symmetric and that satisfies for all $N \in \mathbb{N}$, $(x_1, \ldots, x_n) \in \mathcal{X}^N$, $(a_1, \ldots, a_N) \in \mathbb{R}^{\mathbb{N}}$:*

$$\sum_{1 \leq i,j \leq N} a_i a_j K(x_i, x_j) \geq 0. \tag{2}$$

If $\mathcal{X} = R^d$, the simplest kernel that can be thought of is the canonical inner product. It obviously is symmetric, and positivity is easily verified. Under the same assumptions, it can quite easily be shown that $K(x, x') = (\langle x, x' \rangle_{\mathbb{R}^d})^p$ is a kernel too (known as the polynomial kernel). Another well known kernel is the gaussian kernel, defined as $K(x, x') = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$.

1.1.2 Feature mapping

As seen in Subsection 1.1.1, the canonical inner product of a vector space is a kernel. Mercer's theorem provides a kind of reciprocal statement: any kernel can be viewed as an inner product in a certain hilbert space which is a functional space on $\mathcal{X}$.

Theorem 1 (Mercer's theorem) *The function K is a positive definite kernel on $\mathcal{X}$ if and only if there exists a Hilbert space $\mathcal{H}$ with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a mapping*

$$\phi : \mathcal{X} \mapsto \mathcal{H} \tag{3}$$

such that, for any x, x' in $\mathcal{X}$:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} \tag{4}$$

Using a kernel can thus somehow be viewed as embedding the datapoints in a (larger) space. One benefit of doing so can be, for example, to simplify regression or classification problems, as these problems tend to be easier in high dimensional space than in low dimensional ones.

1.1.3 Kernel methods as convex optimization problems

Recall the kind of problem exposed in Sec 1.1, by eq 1. Now, once a kernel K is defined, it implicitly defines a functional space $\mathcal{H}$ on $\mathcal{X}$, as explained in Sec 1.1.2. This space can be used as the prediction function space, that is $\mathcal{F} = \mathcal{H}$, and the regularization term can be replaced by the squared norm in $\mathcal{H}$.

This gives the following minimization problem:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \left\| f \right\|_{\mathcal{H}}^2 . \tag{5}$$

When L is a convex function over its second argument, it is quite clear that the resulting minimization problem is a convex optimization problem: the squared norm is a convex function of f, $f \mapsto f(x)$ is a linear function for any $x \in \mathcal{X}$, thus $L(y_i, f(x_i))$ is convex for all i, which lead to the global convexity. However, the space $\mathcal{H}$ considered is, in the general case, infinite, and this makes convex optimization unapplicable.

1.2 From infinite to finite

Happily enough, the representer theorem stated below turns this infinite dimensional convex optimization problem into your gentle everyday n-dimensional convex optimization problem:

Theorem 2 (Representer theorem) *Using the notations previously defined, let $\Psi : \mathbb{R}^{n+1} \mapsto \mathbb{R}$ be a function strictly increasing in the last variable. Then any solution of*

$$\min_{f \in \mathcal{H}} \Psi(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}}) \tag{6}$$

admits a representation of the form:

$$\forall x \in \mathcal{X}, f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x). \tag{7}$$

The equation 3 clearly falls in the field of application of the representer theorem. The optimization problem can thus be rewritten in term of α's as:

$$\min_{(\alpha_1,\ldots,\alpha_n)\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} L(y_i, \sum_{j=1}^{n} \alpha_j K(x_j, x_i)) + \lambda \sum_{1\leq i,j\leq n} \alpha_i \alpha_j K(x_i, x_j) \tag{8}$$

which is a finite dimensional convex optimization problem.

2 Support vector machines: a special case

Support vector machines are kernel methods used in the field of binary classification, and that can be extended (quite painfully, unfortunately) to n-ary classification. Among kernel methods, they are probably the best known. This might be explained by the fact that they enforce sparsity in the best fit representation. Formally, they ensure that in eq 5, a varying number of α's will be exactly zero. Sparsity often brings along a lot of nice properties. For SVMs, most notably, they ensure a better regularization (intuitively, a fit with less term is less complex than a fit with many terms, and the number of term is directly related to the number of non-zero α's), and they induce smaller computational cost. Indeed, the best fit represented in eq 5 has as many terms as there are datapoints in the train dataset. This means that the computational cost at test time will depend directly on the number of datapoints in the training set. This might prove computationally costly. SVMs partly solve this problem by ensuring that some (most) α's are zero, and thus that the best fit will only depend on a few training datapoints.

2.1 Generalities

Given an input set $\mathcal{X}$, a p.s.d kernel on $\mathcal{X}$, K and a training dataset, $((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \{\pm 1\})^n$, SVMs solve the kernel optimization problem eq 3 with the particular choice of loss $L(x, y) = \max(0, 1 - x^\top y)$, which is known as the Hinge Loss.

A few remarks might be of use at this point:

- The Hinge Loss is clearly convex in x. It is a maximum over two affine functions of x, and is thus convex.

- Datapoints are classified simply by looking at the sign of $f(x)$ where f is the fit. If $f(x) > 0$, x is classified as $+1$, else x is classified ad -1. The specificity of the Hinge Loss is to try and drive $f(x)$ towards $+1$

when x is in the $+1$ class, and toward -1 when x is in the -1 class, but not to do any further effort.

Rewritting eq 3 in the SVM case, and introducing additionnal slack variables, the optimization problem we intend to solve takes the form:

$$\min_{f \in \mathcal{H}, \xi \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \|f\|_{\mathcal{H}}^2 \tag{9}$$

subject to:

$$\begin{cases} \xi_i \geq 1 - y_i f(x_i) \\ \xi_i \geq 0 \end{cases} \tag{10}$$

Using the representer theorem, this can be transformed in

$$\min_{\alpha \in \mathbb{R}^n, \xi \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^\top \mathbf{K} \alpha \tag{11}$$

subject to:

$$\begin{cases} y_i \sum_{j=1}^{n} \alpha_j K(x_i, x_j) + \xi_i - 1 \geq 0 \\ \xi_i \geq 0 \end{cases} \tag{12}$$

where $\mathbf{K}$ is the p.s.d. matrix associated to the kernel K on the dataset, that is $\mathbf{K}_{i,j} = K(x_i, x_j)$. This is a quadratic program.

The dual problem can easily be expressed:

$$\max_{0 \leq \mu \leq 1/n} \sum_{i=1}^{n} \mu_i - \frac{1}{4\lambda} \sum_{1 \leq i,j \leq n} y_i y_j \mu_i \mu_j K(x_i, x_j). \tag{13}$$

Now expressing the relation between primal and dual solutions:

$$\alpha = \operatorname{diag}(y) \mu / 2\lambda \tag{14}$$

(this is done by rewriting the optimization condition on the Lagrangian with respect to the α variables) and injecting this relation into the dual, we get the following easier problem:

$$2\alpha^\top y - \alpha^\top \mathbf{K} \alpha \tag{15}$$

subject to:

$$0 \leq y_i \alpha_i \leq \frac{1}{2\lambda n} \tag{16}$$

(The crucial point in deriving this new optimization problem is to notice that $y_i^2 = 1$, since $y_i \in \{\pm 1\}$).

On this very simple optimization problem, the KKT conditions can be easily expressed, and they lead to the following complementary slackness equations:

$$\begin{cases} \alpha_i \left[y_i f(x_i) + \xi_i - 1 \right] = 0 \\ \left[\alpha_i - \frac{y_i}{2\lambda n} \right] \xi_i = 0. \end{cases} \tag{17}$$

With a bit of analysis with these equations and the constraints of the primal, it can be easily shown that the only terms for which $\alpha_i \neq 0$ are the examples "hard to classify" that is for which $y_i f(x_i) \leq 1$ (those examples can be well classified, but their classifiction margin is not large). All in all, this tends to express the fact that only a few datapoints are involved in the best fit expression. This has a tremendous impact on further computations involving the best fit.