Support Vector Machines – Part 2

Scribes: Alexander Zhiliakov* and Sulaimon Oyeleye**

1 Non-linear SVMs

Recall that the standard linear SVM problem reads as follows.

Find $(\mathbf{w}^\star, b^\star) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$\begin{aligned} &(\mathbf{w}^\star, b^\star) = \arg\min_{(\mathbf{w},b)\in\mathbb{R}^d\times\mathbb{R}} \frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle \\ &\text{subject to} \quad y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geq 1. \end{aligned} \tag{1}$$

We refer to the constrained optimization (1) as a *linear* SVM problem. The decision function associated with this problem is

$$f(\mathbf{x}) := \operatorname{sgn}(\langle \mathbf{w}^\star, \mathbf{x}\rangle + b^\star) \tag{2}$$

and it is designed to find the maximum-margin hyperplane $\{\mathbf{x} : \langle \mathbf{w}^\star, \mathbf{x}\rangle + b^\star = 0\}$ separating a set of training points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.

There are two major issues with this classification approach [3, Chapter 1.3]:

1. The linear form of (2) may not be suitable for a classification task, i.e., the training set is *not* linearly separable. In this case $(\mathbf{w}^\star, b^\star)$ simply does not exist.
2. Overfitting may be a serious problem for $d \geq n$ and we need to somehow misclassify some training points in order to avoid overfitting in the presence of noise.

* Department of Mathematics, University of Houston, Houston, Texas 77204 (alex@math.uh.edu).

** Department of Mathematics, University of Houston, Houston, Texas 77204 (soyeleye@math.uh.edu).

1.1 Hard and soft margin SVM approaches and the 'kernel trick'

In order to resolve the first issue, we consider a *feature map* Φ that maps the input data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbf{X}$ to some Hilbert space $\mathscr{H}$ called *feature space*:

$$\Phi : \mathbf{X} \to \mathscr{H} \tag{3}$$

The feature map Φ is typically nonlinear and $\mathscr{H}$ may be infinite dimensional.

Using a feature map Φ, one can build analogous problem to (1) by considering the mapped data $\Phi(\mathbf{x}_1)$, $\Phi(\mathbf{x}_2)$, $\ldots$, $\Phi(\mathbf{x}_n)$ and then solving the *nonlinear* SVM problem in the feature space $\mathscr{H}$ as follows. Find $(\mathbf{w}^\star, b^\star) \in \mathscr{H} \times \mathbb{R} \times \mathbb{R}^n$ such that

$$\begin{aligned} &(\mathbf{w}^\star, b^\star) = \arg\min_{(\mathbf{w},b)\in\mathscr{H}\times\mathbb{R}} \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle_{\mathscr{H}} \\ &\text{subject to} \quad y_i \left(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b \right) \geq 1. \end{aligned} \tag{4}$$

This approach is called *hard margin SVM* approach, and initially was proposed by Boser et al. [1].

To deal with the second issue, the so called *soft margin SVM* technique was introduced by Cortes and Vapnik [2]. While the constraints in (1) force the data set to be divided by a hyperplane exactly, the soft margin approach[3] introduces a slack variables $\xi \in \mathbb{R}^n$ to relax this constraint leading to the following *nonlinear* SVM optimization problem. Find $(\mathbf{w}^\star, b^\star, \xi^\star) \in \mathscr{H} \times \mathbb{R} \times \mathbb{R}^n$ such that

$$\begin{aligned} &(\mathbf{w}^\star, b^\star, \xi^\star) = \arg\min_{(\mathbf{w},b)\in\mathscr{H}\times\mathbb{R}} \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle_{\mathscr{H}} \\ &\text{subject to} \quad y_i \left(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0. \end{aligned} \tag{5}$$

Assume that there is a *kernel* function $k : \mathbf{X} \times \mathbf{X} \to \mathbb{K}$ on the input space[4] satisfying

$$k(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \right\rangle_{\mathscr{H}}. \tag{6}$$

Given (6), we can then formulate the SVM problem in the dual form as

$$\begin{aligned} &\max_{\alpha\in\mathbb{R}^n} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \, \alpha_j \, y_i \, y_j \, k(\mathbf{x}_i, \mathbf{x}_j) \right) \\ &\text{subject to} \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i \, y_i = 0, \end{aligned} \tag{7}$$

and correspondingly write the decision function for (5) as

[3] The original approach by Cortes and Vapnik also includes a regularization of the objective functional to deal with overrelaxation of the constraints. We omit it here for simplicity.

[4] Here and further we will use $\mathbb{K}$ for a field (either real $\mathbb{R}$ or complex $\mathbb{C}$).

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n} y_i\, \alpha_i \, \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i)\rangle_{\mathscr{H}} + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{n} y_i\, \alpha_i\, k(\mathbf{x}, \mathbf{x}_i) + b\right).$$

To conclude, the ‘kernel trick’ makes it possible to achieve nonlinear separation in the input space by implicitly mapping the input space into a feature space where features are linearly separable; see Figure 1. These observations motivates us to study kernels and their properties and this will be the topic of the following lectures.

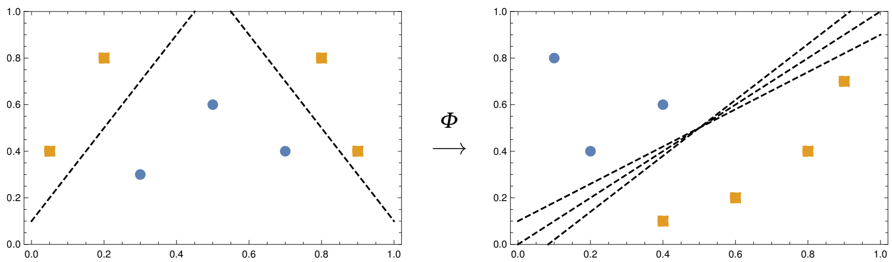

Fig. 1: Illustration of the “kernel trick”. Left: Initial input data $\mathbf{x}_1, \ldots, \mathbf{x}_7 \in \mathbf{X} = \mathbb{R}^2$ is *not* linearly separable. Right: Mapped data $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_7) \in \mathscr{H}$ is separable in $\mathscr{H} = \Phi(\mathbf{X})$.

2 Kernels and Reproducing Kernel Hilbert Spaces (RKHS)

Definition 1. Let $\mathbf{X} \neq \emptyset$ be a set. A function $k : \mathbf{X} \times \mathbf{X} \to \mathbb{K}$ is called a *kernel* on $\mathbf{X}$ iff there is a $\mathbb{K}$-Hilbert space $\mathscr{H}$ and a feature map $\Phi : \mathbf{X} \to \mathscr{H}$ such that for any $\mathbf{x}, \mathbf{x}' \in \mathbf{X}$

$$k(\mathbf{x}, \mathbf{x}') = \left\langle \Phi(\mathbf{x}'), \Phi(\mathbf{x}) \right\rangle_{\mathscr{H}}$$

holds.

Given a kernel k, neither Φ nor $\mathscr{H}$ are uniquely determined.

Example 1. Let $\mathbf{X} := \mathbb{R}$ and $k(x, x') := x'x$. Obviously, k is a kernel on $\mathbf{X}$ with $\Phi_1(x) := x$ being the identity map and $\mathscr{H}_1 := \mathbb{R}$.
Consider $\Phi_2 : \mathbf{X} \to \mathbb{R}^2 =: \mathscr{H}_2$ given by

$$\Phi_2(x) := \frac{1}{\sqrt{2}}(x, x).$$

We have

$$\left\langle \Phi_2(x'), \Phi_2(x) \right\rangle_{\mathbb{R}^2} = \frac{x'x}{\sqrt{2}} + \frac{x'x}{\sqrt{2}} = x'x =: k(x, x'),$$

and hence k is a kernel on $\mathbf{X}$ also for Φ_2 and $\mathscr{H}_2$.

Next we present one of the commonly used kernels that has series representation.

Example 2. Let $\mathbf{X} \neq \emptyset$ and $\{f_n\}_{n=1}^{\infty}$ be a set of functions $f_n : \mathbf{X} \to \mathbb{K}$ with the property that $f_n(\mathbf{x}) \in \ell^2$ for any $\mathbf{x} \in \mathbf{X}$. Then

$$k(\mathbf{x}, \mathbf{x}') := \sum_{i=1}^{\infty} f_n(\mathbf{x}) \overline{f_n(\mathbf{x}')}$$

is a kernel on $\mathbf{X}$ with $\Phi(\mathbf{x}) := \overline{f_n(\mathbf{x})}$, $\Phi : \mathbf{X} \to \ell^2$, i.e., the sum

$$\left\langle \Phi(\mathbf{x}'), \Phi(\mathbf{x}) \right\rangle_{\ell^2} = \sum_{i=1}^{\infty} f_n(\mathbf{x}) \overline{f_n(\mathbf{x}')} =: k(\mathbf{x}, \mathbf{x}')$$

is well defined since $f_n(\mathbf{x}) \in \ell^2$ for any $x \in \mathbf{X}$ by Hölder's inequality.

2.1 Properties of kernels

1. Let k be a kernel on $\mathbf{X}$ and A be a map, $A : \bar{\underline{\mathbf{Y}}} \to \mathbf{X}$, where $\bar{\underline{\mathbf{Y}}}$ is another set. Then, $\overline{k}(x, x') := k(A(x), A(x')),\ \ x, x' \in \mathbf{X}$ is a kernel on $\bar{\underline{\mathbf{Y}}}$. This include the special case where A is a restriction map. Hence, if $\bar{\underline{\mathbf{Y}}} \subset \mathbf{X}$, then $k_{|\bar{\underline{\mathbf{Y}}} \times \mathbf{X}}$ is a kernel.

2. If k_1, k_2 are kernels then $k_1 + k_2$ is a kernel.

3. If $\alpha \geq 0$ and k is a kernel, then αk is a kernel.

   Remark: The space of kernels forms a cone but not a vector space.
   Let k_1, k_2 be kernels on $\mathbf{X}$ such that, for some $x \in \mathbf{X}$,

   $$k_1(x,x) - k_2(x,x) < 0.$$

   If $k_1 - k_2$ is kernel, then there exist a map $\Phi : \mathbf{X} \to H$ such that

   $$0 \leq \langle \Phi(x), \Phi(x) \rangle = k_1(x,x) - k_2(x,x) < 0,$$

   giving a contradiction. So $k_1 - k_2$ is not a kernel.

4. If k_1 is a kernel on $\mathbf{X}_1$ and k_2 is a kernel on $\mathbf{X}_2$, then $k_1.k_2$ is a kernel on the tensor space $\mathbf{X}_1 \times \mathbf{X}_2$. In particular, if $\mathbf{X}_1 = \mathbf{X}_2 = \mathbf{X}$, then $k(x, x') := k_1(x, x')k_2(x, x'),\ x, x' \in \mathbf{X}$ defines a kernel on $\mathbf{X}$.

Example 3. For any $n > 0$, the map $k_n(x, x') := (xx')^n$, where $x, x' \in \mathbf{X}$ is a kernel. Hence, if $p : \mathbf{X} \to \mathbb{R}$ is of the form,

$$p(t) = a_n t^n + ... + a_1 t + a_0$$

with non-negative coefficients a_i, then $k(x,x') = p(xx')$, with $x,x' \in \mathbf{X}$ is a kernel.
In general, the function: $k(z,z_1) = (\langle z,z' \rangle + c)^m$ with $z,z' \in \mathbb{C}^d, c \geq 0$, is a polynomial kernel on $\mathbb{C}^d$.

Lemma (Taylor type kernels). Let $B_{\mathbb{C}},\ B_{\mathbb{C}^d}$ be the open unit ball in $\mathbb{C}, \mathbb{C}^d$ respectively. Let $r > 0$ and $f : rB_{\mathbb{C}} \to \mathbb{C}$ be a holomorphic function with Taylor series expansion;

$$f(z) = \sum_{n=0}^{\infty} a_n z^n; \quad z \in rB_{\mathbb{C}}$$

If $a_n \geq 0$ for all $n \in \mathbb{N}$, then

$$k(z,z') := f(\langle z,z' \rangle)_{\mathbb{C}^d} = \sum_{n=0}^{\infty} a_n \langle z,z' \rangle^n_{\mathbb{C}^d}, \qquad z,z' \in \sqrt{r} B_{\mathbb{C}^d}$$

defines a kernel on $\sqrt{r} B_{\mathbb{C}^d}$.
It follows that the restriction to $\mathbf{X} := \{x \in \mathbb{R}^d : ||x||_2 < \sqrt{r}\}$ is a real-valued kernel.

Example 4. For $d \in \mathbb{N},\ x,x' \in \mathbb{R}^d,\ \ k(x,x') = exp(\langle x,x' \rangle)$ is a $\mathbb{K}$ - valued kernel on $\mathbb{R}^d$.

Example 5. (Exponential kernel). Let $d \in \mathbb{N},\ \gamma > 0,\ z = (z_1, ..., z_d),\ z' = (z'_1, ..., z'_d) \in \mathbb{C}^d$. It follows from the lemma above that

$$k^{(z,z')}_{\gamma,\mathbb{C}^d} := exp(-\gamma^{-2} \sum_{j=1}^{d} (z_j, -\bar{z}'_j)^2)$$

is a kernel on $\mathbb{C}^d$. Its restriction $k_\gamma := exp(-\frac{||x-x'||_2^2}{\gamma^2})$, for $x,x' \in \mathbb{R}^d$, is a kernel on $\mathbb{R}^d$.

2.2 Characterization of kernels

Definition: A function $k : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ is *positive definite* if for all $n \in \mathbb{N}$, $\alpha_1, ..., \alpha_n \in \mathbb{R}$, and all $x_1, ..., x_n \in \mathbf{X}$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

Furthermore, it is *strictly positive definite* if for mutually distinct $x_1, ..., x_n \in \mathbf{X}$, equality only occurs when $\alpha_1 = ... = \alpha_n = 0$. k is symmetric if $k(x,x') = k(x',x),\ for\ all\ \ x,x' \in$

X.

NOTE: $K = (k(x_i, x_j))_{i,j}$ is the *Gram matrix.*

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \iff K \text{ is positive definite.}$$

Theorem 1. *A function $k : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite*

Proof. ($\Longrightarrow$) If k is a $\mathbb{R}$-kernel, then $k(x,x') = \langle \Phi(x), \Phi(x') \rangle = \langle \Phi(x'), \Phi(x) \rangle = k(x', x)$ is symmetric.
Also, for any $n \in \mathbb{N},\ \alpha_1, ..., \alpha_n \in \mathbb{R},\ x_1, ..., x_n \in \mathbf{X}$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) = \left\langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \sum_{j=1}^{m} \alpha_j \Phi(x_j) \right\rangle = || \sum_{i=1}^{n} \alpha_i \Phi(x_i) ||^2 \geq 0$$

Hence, k is positive definite.
($\Longleftarrow$) Assume $k : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ is symmetric and positive definite.
Define

$$\mathscr{H}_{pre} := \left\{ \sum_{i=1}^{n} \alpha_i k(.\ ,\ x_i) : n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathbf{X} \right\}.$$

For any $f = \sum_{i=1}^{n} \alpha_i k(.\ ,\ x_i),\ g = \sum_{j=1}^{m} \alpha_j k(.\ ,\ x'_j) \in H_p$, set

$$\langle f, g \rangle := \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(x'_j, x_j).$$

We want to show that this operation defines an inner product on $\mathscr{H}_{pre}$, hence we will show that $\langle ., . \rangle$ is bilinear, symmetric and positive definite.

First we observe that, for any $x'_j \in \mathbf{X}$, we have $f(x'_j) = \sum_{i=1}^{n} \alpha_i k(x'_j, x_i)$, hence we can write $\langle f, g \rangle = \sum_{j=1}^{m} \beta_j f(x'_j)$. Similarly, we can write $\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i g(x_i)$. This shows that $\langle f, g \rangle$ is independent of the representation of f and g.

By the assumption on k, it is straightforward tp verify that $\langle f, g \rangle$ is symmetric, bilinear and positive, that is $\langle f, f \rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0$ for any $\alpha_1, ..., \alpha_n, x_1, ..., x_n, f \in \mathscr{H}_{pre}$. We remark that these properties also imply the Cauchy-Schwartz inequality, $|\langle f, g \rangle|^2 \leq \langle f, f \rangle \langle g, g \rangle$ for all $f, g, \in \mathscr{H}_{pre}$.

It is also clear that if $f = 0$, then $\langle f, f \rangle = 0$. It remains to show that n$\langle f, f \rangle$ implies $f = 0$. WE observe that $\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i g(x_i)$, then $\sum_{i=1}^{n} \alpha_i k(x, x_i) = \langle f, k(x, x_i) \rangle \leq k(.\ ,\ x) k(.\ ,\ x) > \langle f, f \rangle$.

Using this observation and Cauchy-Schwartz inequality, for any $x \in \mathbf{X}$ we have

$$|f(x)|^2 = |\sum_{i=1}^{n} \alpha_i k(x, x_i)|^2 = |\langle f, k(., x) \rangle|^2 \leq \langle k(.\ ,\ x), k(.\ ,\ x) \rangle . \langle f, f \rangle = 0$$

Thus, $f(x) = 0$ for any $x \in \mathbf{X}$ hence $f = 0$.

Let $\mathscr{H}$ be a completion of $\mathscr{H}_{pre}$ and $T : \mathscr{H}_{pre} \to H$ be the corresponding isometric embedding. Thus $\mathscr{H}$ is a Hilbert space and for any $x \in \mathbf{X}$

$$\left\langle Tk(.\ ,\ x'), Tk(.\ ,\ x)\right\rangle_H = \left\langle k(.\ ,\ x'), k(.\ ,\ x)\right\rangle_{\mathscr{H}_{pre}} = k(x, x')$$

The map $x \mapsto Tk(.\ ,\ x)$ for $x \in \mathbf{X}$ defines a feature map of k, hence k is a kernel.

References

1. Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.
2. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
3. Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.