# Kernel Methods
## SC4/SM4 Data Mining and Machine Learning, Hilary Term 2017
**Dino Sejdinovic**

### 6.1 Feature Maps and Feature Spaces

Kernel methods are a versatile algorithmic framework which allows construction of nonlinear machine learning algorithms (for a variety of both supervised and unsupervised learning tasks: clustering, dimensionality reduction, classification, regression) by employing linear tools in a nonlinearly transformed feature space. Let us first recall the definition of an abstract inner product, which is central to kernel methods.

**Definition 6.1.** [Inner product] Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be *an inner product* on $\mathcal{H}$ if

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

We can define a norm using the inner product as $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$. A *Hilbert space* is a vector space on which an inner product is defined, along with an additional technical condition.[1] We are now ready to define the notion of a *kernel*.

**Definition 6.2.** Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *kernel* if there exists a Hilbert space and a map $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \left\langle \varphi(x), \varphi(x') \right\rangle_{\mathcal{H}}.$$

We will call such $\mathcal{H}$ of kernel $k$ a *feature space* and the map $\varphi$ will be called a *feature map*. Note that we imposed almost no conditions on $\mathcal{X}$: in particular, we do not require there to be an inner product defined on the elements of $\mathcal{X}$. The case of text documents is an instructive example: one cannot take an inner product between two books, but can take an inner product between features of the text in those books.

Clearly, a single kernel can correspond to multiple pairs of underlying feature maps and feature spaces. For a simple example, consider $\mathcal{X} := \mathbb{R}^p$:

$$\phi_1(x) = x \qquad \text{and} \qquad \phi_2(x) = \left[ \frac{x_1}{\sqrt{2}}, \cdots, \frac{x_p}{\sqrt{2}}, \frac{x_1}{\sqrt{2}}, \cdots, \frac{x_p}{\sqrt{2}} \right]^\top.$$

Both $\phi_1$ and $\phi_2$ are valid feature maps (with feature spaces $\mathcal{H}_1 = \mathbb{R}^p$ and $\mathcal{H}_2 = \mathbb{R}^{2p}$) of kernel $k(x, x') = x^\top x'$.

[1] Specifically, a Hilbert space must be *complete*, i.e. it must contain the limits of all Cauchy sequences with respect to the norm defined by its inner product.

## 6.2 Positive definiteness of an inner product in a Hilbert space

It turns out that all kernel functions (defined as inner products between some features) are *positive definite*.

**Definition 6.3.** [Positive definite functions] A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1$, $\forall (a_1, \dots a_n) \in \mathbb{R}^n$, $\forall (x_1, \dots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

The function $k(\cdot, \cdot)$ is *strictly* positive definite if for mutually distinct $x_i$, the equality holds only when all the $a_i$ are zero.[2]

Every inner product is a positive definite function, and more generally:

**Lemma 6.1.** Let $\mathcal{H}$ be any Hilbert space (not necessarily an RKHS), $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. Then $k(x, y) := \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ is a positive definite function.

*Proof.*

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0. \end{aligned}$$

□

## 6.3 Reproducing Kernel Hilbert Spaces

We have introduced the notation of feature spaces, and kernels on these feature spaces. What's more, we've determined that these kernels are positive definite. In this section, we use these kernels to define *functions* on $\mathcal{X}$. The space of such functions is known as a reproducing kernel Hilbert space (RKHS).

**Definition 6.4.** [Reproducing kernel] Let $\mathcal{H}$ be a *Hilbert space of functions* $f \colon \mathcal{X} \to \mathbb{R}$ defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if it satisfies

- $\forall x \in \mathcal{X}$, $k_x = k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (*the reproducing property*).

If $\mathcal{H}$ has a reproducing kernel, it is called a reproducing kernel Hilbert space (RKHS).

[2] The corresponding terminology used for matrices is "positive semi-definite" vs "positive definite".

In particular, note that for any $x, y \in \mathcal{X}$, reproducing kernel satisfies $k(x,y) = \langle k(\cdot, y), k(\cdot, x)\rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}}$. Thus, reproducing kernel is clearly a kernel, i.e. an inner product between features with a feature space $\mathcal{H}$ and a feature map $\phi\colon x \mapsto k(\cdot, x)$. This way of writing feature mapping is called the *canonical feature map.* Note that these features are not specified explicitly in a vector form, but rather as functions on $\mathcal{X}$.

We have seen that any reproducing kernel is a kernel and that every kernel is a positive definite function. Remarkably, *Moore-Aronszajn theorem* [1] shows that *for every positive definite function $k$, there exists a unique RKHS with kernel $k$.* The theorem is outside of the scope of this course, but it provides an insight into the structure of the RKHS corresponding to $k$. It turns out RKHS can be written as $\overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$, i.e. the space of all linear combinations of canonical features, completed with respect to an inner product on these linear combinations defined as

$$\left\langle \sum_{i=1}^{r} \alpha_i k(\cdot, x_i), \sum_{j=1}^{s} \beta_j k(\cdot, y_j) \right\rangle := \sum_{i=1}^{r}\sum_{j=1}^{s} \alpha_i \beta_j k(x_i, y_j).$$

Thus, all three notions: (1) reproducing kernel, (2) kernel as inner product between features and (3) positive definite function, are equivalent. Recall that the feature space of a kernel is not unique - but RKHS (feature space as a space of functions) is. For example, for the *linear kernel* $k(x,y) = x^\top y$ considered earlier, many possible feature representations exist but the canonical feature representation that associates to each $x$ the function $k(\cdot, x)\colon y \mapsto x^\top y$ is what determines the structure of its RKHS. In particular, linear kernel $k(x,y) = x^\top y$ corresponds to the space of all linear functions $f(x) = w^\top x$ (*why*?).

### 6.4 Representer Theorem

Now that we have defined an RKHS, we can consider it as a hypothesis class for empirical risk minimisation (noneERM). In particular, we are looking for the function $f^*$ in the RKHS $\mathcal{H}$ which solves the regularised ERM problem

$$\min_{f\in\mathcal{H}} \hat{R}(f) + \Omega\left(\|f\|_{\mathcal{H}}^2\right),$$

for empirical risk $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i), x_i)$, a loss function $L\colon \mathcal{Y}\times\mathcal{Y}\times\mathcal{X} \to \mathbb{R}_+$ and any non-decreasing function $\Omega$.

**Theorem 6.1** (Representer Theorem)**.** There is a solution to

$$\min_{f\in\mathcal{H}} \hat{R}(f) + \Omega\left(\|f\|_{\mathcal{H}}^2\right) \tag{6.1}$$

that takes the form $f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$. If $\Omega$ is strictly increasing, all solutions have this form.

*Proof.* Let $f$ be any minimiser of (6.1). Denote by $f_s$ the projection of $f$ onto the subspace

$$\text{span}\{k(\cdot, x_i) :\ i = 1, \ldots, n\}$$

such that

$$f = f_s + f_\perp,$$

where $f_s = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ and $f_\perp$ is orthogonal to span $\{k(\cdot, x_i) : i = 1, \ldots, n\}$.

Since

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2 ,$$

we have

$$\Omega\left(\|f\|_{\mathcal{H}}^2\right) \geq \Omega\left(\|f_s\|_{\mathcal{H}}^2\right).$$

On the other hand, the individual terms $f(x_i)$ in the loss are given by

$$f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_s + f_\perp, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_s, k(\cdot, x_i)\rangle_{\mathcal{H}} = f_s(x_i),$$

so

$$L(y_i, f(x_i), x_i) = L(y_i, f_s(x_i), x_i) \quad \forall i = 1, \ldots, n.$$

and thus empirical risks must be the same: $\hat{R}(f) = \hat{R}(f_s)$. Thus $f_s$ is also a minimiser of (6.1) and if $\Omega$ is strictly increasing, it must be that $f_\perp = 0$.

□

We see that the key parts of the theorem are the fact that the empirical risk only depends on the components of $f$ lying in the subspace spanned by the canonical features and that the regulariser $\Omega(\ldots)$ is minimised when $f = f_s$ (adding additional orthogonal components to the function makes it more complex but does not change the empirical risk). Moreover, if $\Omega$ is strictly increasing, then $\|f_\perp\|_{\mathcal{H}} = 0$ is required at the minimum.

## 6.5 Operations with Kernels

Kernels can be combined and modified to get new kernels. For example,

**Lemma 6.2.** [Sums of kernels are kernels] Given $\alpha > 0$ and $k$, $k_1$ and $k_2$ all kernels on $\mathcal{X}$, then $\alpha k$ and $k_1 + k_2$ are kernels on $\mathcal{X}$.

To prove the above, just check *positive definiteness*. Note that a difference between two kernels need not be a kernel: if $k_1(x, x) - k_2(x, x) < 0$, then condition 3 of inner product definition 6.1 may be violated.

**Lemma 6.3.** [Mappings between spaces] Let $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ be non-empty sets, and define a map $A : \mathcal{X} \to \widetilde{\mathcal{X}}$. Define the kernel $k$ on $\widetilde{\mathcal{X}}$. Then $k(A(x), A(x'))$ is a kernel on $\mathcal{X}$.

**Lemma 6.4.** [Products of kernels are kernels] Given $k$ on $\mathcal{X}$ and $l$ on $\mathcal{Y}$, then

$$\kappa\left((x, y), (x', y')\right) = k\left(x, x'\right) l\left(y, y'\right)$$

is a kernel on $\mathcal{X} \times \mathcal{Y}$. Moreover, if $\mathcal{X} = \mathcal{Y}$, then

$$\kappa\left(x, x'\right) = k\left(x, x'\right) l\left(x, x'\right)$$

is a kernel on $\mathcal{X}$.

The general proof would require some technical details about Hilbert space tensor products, but the main idea can be understood with some simple linear algebra. We consider the case where $\mathcal{H}$ corresponding to $k$ is $\mathbb{R}^M$, and $\mathcal{G}$ corresponding to $l$ is $\mathbb{R}^N$. Write $k\left(x, x'\right) = \varphi(x)^\top \varphi(x')$ and $l\left(y, y'\right) = \psi(y)^\top \psi(y')$. We will use that a notion of inner product between matrices $A \in \mathbb{R}^{M \times N}$ and $B \in \mathbb{R}^{M \times N}$ is given by

$$\langle A, B \rangle = \operatorname{trace}(A^\top B). \tag{6.2}$$

Then

$$\begin{aligned} k\left(x, x'\right) l\left(y, y'\right) &= \varphi(x)^\top \varphi(x') \psi(y')^\top \psi(y) \\ &= \operatorname{tr}(\psi(y) \varphi(x)^\top \varphi(x') \psi(y')^\top) \\ &= \left\langle \varphi(x)\psi(y)^\top, \varphi(x')\psi(y')^\top \right\rangle, \end{aligned}$$

thus we can define features $A(x, y) = \varphi(x)\psi(y)^\top$ of the product kernel.

The sum and product rules allow us to define a huge variety of kernels.

**Lemma 6.5.** [Polynomial kernels] Let $x, x' \in \mathbb{R}^p$ for $p \geq 1$, and let $m \geq 1$ be an integer and $c \geq 0$. Then

$$k(x, x') := \left(\left\langle x, x' \right\rangle + c\right)^m$$

is a valid kernel.

To prove: expand out this expression into a sum (with non-negative scalars) of kernels $\langle x, x' \rangle$ raised to integer powers. These individual terms are valid kernels by the product rule.

Can we extend this combination of sum and product rule to sums with infinitely many terms? Consider for example the exponential function applied to an inner product $k(x, x') = \exp\left(\langle x, x' \rangle\right)$. Since addition and multiplication preserve positive definiteness and since all the coefficients in the Taylor series expansion of the exponential function are nonnegative, $k_m(x, x') = \sum_{r=1}^{m} \frac{\langle x, x' \rangle^r}{r!}$ is a valid kernel $\forall m \in \mathbb{N}$. Fix some $\{\alpha_i\}$ and $\{x_i\}$. Then $A_m = \sum_{i,j} \alpha_i \alpha_j k_m(x_i, x_j) \geq 0$ $\forall m$ since $k_m$ is positive definite. But $A_m \to \sum_{i,j} \alpha_i \alpha_j \exp\left(\langle x_i, x_j \rangle\right)$ as $m \to \infty$, so $\sum_{i,j} \alpha_i \alpha_j \exp\left(\langle x_i, x_j \rangle\right) \geq 0$ as well. Thus, $\exp\left(\langle x, x' \rangle\right)$ is also a valid kernel (it is called *exponential kernel*). We may combine all the results above (*exercise*) to show that the following in practice widely used kernel, known under various names: *Gaussian, Gaussian RBF, squared exponential* or *exponentiated quadratic* is a valid kernel on $\mathbb{R}^p$:

$$k(x, x') := \exp\left(-\frac{1}{2\sigma^2} \left\|x - x'\right\|^2\right).$$

The RKHS of this kernel is infinite-dimensional.

## 6.6 Kernel PCA

Kernel PCA is a popular nonlinear dimensionality reduction technique [2]. Assume we have a dataset $\{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$. Consider an explicit feature transformation $x \mapsto \varphi(x) \in \mathcal{H}$, and assume that we are interested in performing PCA in the feature space $\mathcal{H}$. Assume that the features $\{\varphi(x_i)\}_{i=1}^n$ are centred. Assume for the moment that the feature space is finite-dimensional, i.e. $\mathcal{H} = \mathbb{R}^M$. Then the $M \times M$ sample covariance matrix in the feature space is given by

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n \varphi(x_i)\varphi(x_i)^\top = \frac{1}{n-1} \Phi^\top \Phi,$$

where $\Phi \in \mathbb{R}^{n \times M}$ is the feature representation of the data. To perform PCA, recall that we are interested in solving the eigenvalue problem $\mathbf{S}v_m = \lambda_m v_m$, $m = 1, \ldots, M$, and we need the top $k \ll \min\{n, M\}$ eigenvectors $v_m$, $m = 1, \ldots, k$, to construct the PC projections $z_i^{(m)} = v_m^\top \varphi(x_i)$. A property analogous to the representer theorem holds here: whenever $\lambda_m > 0$, the eigenvectors lie in the linear span of feature vectors span $\{\varphi(x_i) : i = 1, \ldots, n\}$, i.e.

$$v_m = \sum_{i=1}^n a_{mi} \varphi(x_i) \tag{6.3}$$

for some scalars $a_{mi}$. To see this, note that

$$\lambda_m v_m = \mathbf{S}v_m = \frac{1}{n-1} \sum_{i=1}^n \varphi(x_i) \left( \varphi(x_i)^\top v_m \right)$$

and since $\lambda_m > 0$, it suffices to take $a_{mi} = \frac{1}{\lambda_m (n-1)} \left( \varphi(x_i)^\top v_m \right)$ and clearly $v_m$ has form (6.3). Thus eigenvectors can also be recovered in the *dual space.* Consider now the $n \times n$ kernel matrix $\mathbf{K}$ with $\mathbf{K}_{ij} = k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j)$. By substituting $v_m = \sum_{i=1}^n a_{mi} \varphi(x_i)$ back into the eigenvalue problem, we have:

$$\mathbf{S}v_m = \frac{1}{n-1} \sum_{i=1}^n \varphi(x_i) \sum_{\ell=1}^n a_{m\ell} k(x_i, x_\ell) = \lambda_m \sum_{i=1}^n a_{mi} \varphi(x_i).$$

To express the above in terms of the kernel matrix, we project both sides onto $\varphi(x_j)$, for each $j = 1, \ldots, n$. This gives

$$\frac{1}{n-1} \sum_{i=1}^n k(x_j, x_i) \sum_{\ell=1}^n a_{m\ell} k(x_i, x_\ell) = \lambda_m \sum_{i=1}^n a_{mi} k(x_j, x_i), \quad j = 1, \ldots, n,$$

which in matrix notation can be written as

$$\mathbf{K}^2 a_m = \lambda_m (n-1) \mathbf{K} a_m.$$

Assuming that $\mathbf{K}$ is invertible, $a_m$ vectors can be found as the eigenvectors of the kernel matrix $\mathbf{K}$ with corresponding eigenvalues given by $\lambda_m(n-1)$.

But if we simply perform the eigendecomposition of $\mathbf{K}$, we will obtain $n$-dimensional eigenvectors of unit norm, and we are after the $M$-dimensional eigenvectors $v_m$ of $\mathbf{S}$ which have unit norm. We

see that $1 = v_m^\top v_m = a_m^\top \mathbf{K} a_m = \lambda_m(n-1) a_m^\top a_m$. Thus, if $u_m$ denotes the $m$-th eigenvector of $\mathbf{K}$ with unit norm, to ensure that $v_m$ has unit norm, we need to rescale $a_m = u_m/\sqrt{\lambda_m(n-1)}$. Now, we have an implicit representation of eigenvectors in terms of their dual coefficients. The PC projections are

$$z_i^{(m)} = v_m^\top \varphi(x_i) = \left( \sum_{j=1}^n a_{mj} \varphi(x_j) \right)^\top \varphi(x_i) = \sum_{j=1}^n a_{mj} k(x_j, x_i),$$

or equivalently, the $m$-th dimension of the PC projections is given by

$$\mathbf{z}^{(m)} \quad = \quad \mathbf{K} a_m = \lambda_m(n-1) a_m = \sqrt{\lambda_m(n-1)}\, u_m. \tag{6.4}$$

We have seen this before! Note that PC projections can be discovered from the SVD $\Phi = UDV^\top$ as either $\mathbf{Z} = \Phi V$ or $\mathbf{Z} = UD$. The latter expression is exactly (6.4), since $u_m$ are the eigenvectors of kernel matrix $\mathbf{K}$ (i.e. the left singular vectors of the feature matrix $\Phi$) and $D_{mm} = \sqrt{\lambda_m(n-1)}$ (*why*?). But note that the eigendecomposition of $\mathbf{K}$ and these projections do not require explicit feature transformations - thus, all the computation is happening in the dual representation and $\varphi(x_i)$ need not be computed, only the kernel matrix $\mathbf{K}$ with $\mathbf{K}_{ij} = k(x_i, x_j)$. The kernel formalism also allows us to compute the projection $v_m^\top \varphi(\tilde{x})$ of a new (previously unseen) data vector $\tilde{x} \in \mathbb{R}^p$ to the $m$-th kernel principal component using

$$\left( \sum_{i=1}^n a_{mi} \varphi(x_i) \right)^\top \varphi(\tilde{x}) = \sum_{i=1}^n a_{mi} k(x_i, \tilde{x}) = a_m^\top \mathbf{k}_{\tilde{x}},$$

where $\mathbf{k}_{\tilde{x}} = [k(x_1, \tilde{x}), \ldots, k(x_n, \tilde{x})]^\top$, so again no explicit feature transformations are needed.

Recall that the above all assumes that the features are centred, i.e. that $\frac{1}{n}\sum_{i=1}^n \varphi(x_i) = 0$, but if we are just given a kernel function $k(x, x')$, there is no reason to believe that the features would be centred. Fortunately, it is straightforward to transform *any* kernel matrix into a centred form. Recall that the squared distance matrix in the feature space, i.e. matrix $\mathbf{D}$ for which $\mathbf{D}_{ij} = \|\varphi(x_i) - \varphi(x_j)\|_{\mathcal{H}}^2$ can be recovered from the Gram/kernel matrix (Notes 1, page 6). But distances are invariant to centering and the Gram matrix corresponding to centred features can then be recovered from the distance matrix (Q6 on Problem Sheet 1).

## References


[1] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[2] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.