Discussiones Mathematicae
Probability and Statistics 30 (2010) 179–201

SOME METHODS OF CONSTRUCTING KERNELS IN STATISTICAL LEARNING

Tomasz Górecki

Faculty of Mathematics and Computer Science
Adam Mickiewicz University
Umultowska 87, 61–614 Poznań, Poland

e-mail: drizzt@amu.edu.pl

AND

Maciej Łuczak

Department of Civil and Environmental Engineering
Koszalin University of Technology
Śniadeckich 2, 75–453 Koszalin, Poland

e-mail: mluczak@wbiis.tu.koszalin.pl

Abstract

This paper is a collection of numerous methods and results concerning a design of kernel functions. It gives a short overview of methods of building kernels in metric spaces, especially R^n and S^n. However we also present a new theory. Introducing kernels was motivated by searching for non-linear patterns by using linear functions in a feature space created using a non-linear feature map.

Keywords: positive definite kernel, dot product kernel, statistical kernel, SVM, kPCA.

2010 Mathematics Subject Classification: 62H30, 68T10.

1. Introduction

The mid-1990's yielded major advance in machine learning: the support vector machine (SVM). The fundamental idea beyond this method is

especially far-reaching. SVM utilizes the positive definite kernels. What does it mean? The most of machine learning methods is very well developed in the linear case. However in practice we have real data and we need non-linear methods to detect the kind of dependencies that enable us to predict successfully the properties of interest. The kernel corresponds to a dot product in a feature space (usually high-dimensional, even infinite-dimensional for Gaussian kernel). In this space, our methods are linear, but as long as we can formulate everything in terms of kernel evaluations, we will never have to compute explicitly in the high-dimensional feature space. If we can show that a linear algorithm is depended on the data matrix $\boldsymbol{X}$ only by

$$\boldsymbol{K} \equiv \boldsymbol{X}\boldsymbol{X}^T,$$

then it can be easily "kernelized". In general, this procedure is known as the *kernel trick*. The kernel trick transforms any algorithm that solely depends on the dot product. Wherever a dot product is used, it is replaced with the kernel function. Thus, a linear algorithm can easily be transformed into a non-linear algorithm. The algorithms capable of operating with kernels are (apart from SVM) Fisher's linear discriminant analysis (LDA), principal components analysis (PCA), canonical correlation analysis (CCA), ridge regression, spectral clustering, and many others. A full review of kernel methods was presented in (Hoffman, Schölkopf, Smola, 2008).

In this article we focus one's attention on constructing kernels. The modularity is important advantage of kernel methods. To solve a different problem, we should use a different kernel function. Hence, it is essential to have as many as possible kernel functions, because we never know which kernel will be the best (to be effective in practice, obviously we should use the correct kernel function and with right parameters (Zu, 2008)).

In our paper first we present the main ideas beyond kernels methods in machine learning (Section 2). We review especially SVM and kernel PCA as the members of big family of "kernelized" methods. Then we describe basic kernels and main methods of constructing kernels (Section 3). In Section 4 we show how kernels are constructed using superposition of kernels with other functions, namely functions with "good" Taylor or Legendre polynomials series expansion. We pay attention particularly to sigmoidal kernels (Corollary 2, Example 4–7). In the same section we propose superposition of kernels with maps from $\mathbb{R}^n$ to unit sphere (Corollary 3–6). Section 5 concerns with the special case of the method proposed in Section 4.

We construct the inverse of the stereographic projection. This enables us to construct new kernels as a superposition of this projection and kernels on sphere. In Section 6 we generalize results from Section 4 to the case of multivariable functions.

2. Kernels in machine learning

Suppose we are given empirical data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{\pm 1\}$. Here, the domain $\mathcal{X}$ is some nonempty set from which the patterns x_i are taken; the y_i are called class labels. In order to study the problem of learning, we need additional structure. In learning, we want to generalize unseen data points. This means in the case of pattern recognition, that given some new pattern $x \in \mathcal{X}$, we want to predict the corresponding $y \in \{\pm 1\}$. Although the most of kernels methods can manage with multi-class classification problems as well we are limited mostly to the two-class classification problem. We choose y such as (x, y) is in some way similar to the training examples. So we need similarity measures in $\mathcal{X}$. We require a similarity measure

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

i.e., given two examples x_1 and x_2, a function returns a real number characterizing their similarity. The function k is called a *kernel*.

A type of similarity measure being of particular mathematical appeal are *dot products*. In order to use a dot product as a similarity measure, we first need to embed them into some dot product space F, which may not be identical to $\mathbb{R}^n$. To this end, we use a map

$$\Phi : \mathcal{X} \to F.$$

The space F is called a *feature space*. Embedding the data into F has some benefits:

- It lets us define a similarity measure from the dot product in F:

$$k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle.$$

- It allows us to deal with the patterns geometrically, and thus lets us study learning algorithm using linear algebra and analytic geometry. The geometrical interpretation of dot product means computing the cosine of the angle between the vectors, provided they are normalized to length 1. Moreover, it allows computation of the length of a vector, and the distance between two vectors as the length of their difference.

- Ability to choose the mapping Φ will enable us to design a large variety of learning algorithms.

Now we give two basic "kernelization" examples of well known linear method.

2.1. SVM

In the case of support vector machines, a data point is viewed as a p-dimensional vector. We want to know whether we can separate such points with a $p-1$-dimensional hyperplane. This is called a *linear classifier*. There are many hyperplanes that might classify the data. However, we are additionally interested in finding out if we can achieve maximum separation (*margin*) between the two classes. By this we mean that we choose the hyperplane such as the distance from the hyperplane to the nearest data point is maximized. In other words the nearest distance between a point in one separated hyperplane and a point in the other separated hyperplane is maximized. Now, if such hyperplane exists, it is clearly of interest and is known as the *maximum-margin hyperplane*. Furthermore such linear classifier is known as a *maximum margin classifier*. The samples on the margin are called the *support vectors* (maximum margin hyperplane and hence the classification task is only a function of the support vectors). To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets.

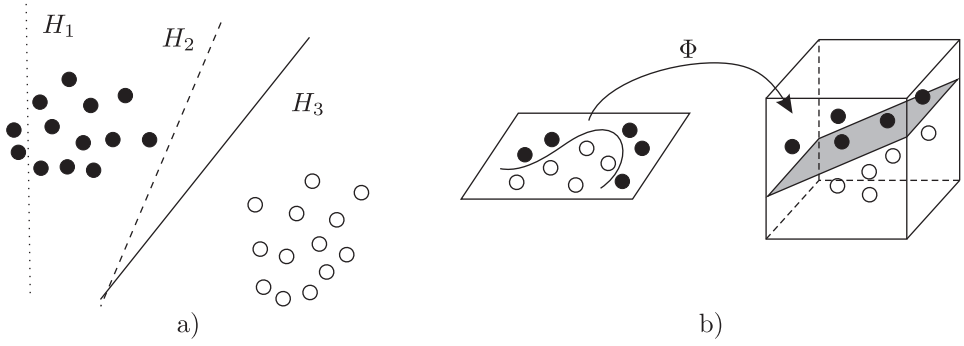

Figure 1. (a) H1 does not separate the 2 classes, H2 does, with a small margin and H3 with the maximum margin. b) Correct map (kernel) changes non-linear classifier into linear in higher dimensional space.

If two classes are perfectly separable, then there exist an infinite number of separating hyperplanes. SVM method is based on the fact that the perfect hyperplane for separating two classes is the one that is the farthest away from the training points.

Intuitively, a good separation is achieved by the hyperplane of the largest distance to the neighboring datapoints of both classes. In general the larger the margin is, the better the generalization error of the classifier (Figure 1a).

Cortes and Vapnik (1995) suggested a modified maximum margin idea (*soft margin*) providing mislabeled examples (generally we can not assume that the two classes are always perfectly separable). If there exists no hyperplane splitting examples, the method will choose a hyperplane that splits the examples as cleanly as possible, still maximizing the distance to the nearest cleanly split examples. The method introduces *slack variables*, $\xi_i \geq 0$, which measure the degree of misclassification of the datum x_i. Then the objective function is increased by a function which penalizes non-zero ξ_i. Additionally the optimization becomes a trade off between a large margin, and a small error penalty.

However, one can not possibly expect a linear classifier to succeed in general situations, no matter how optimal the hyperplane is. Boser, Guyon and Vapnik (1992) suggested a way to create non-linear classifiers by applying the kernel trick to maximum-margin hyperplanes. The resulting algorithm is formally similar, except for every dot product replacing by a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in the transformed feature space. The transformation may be non-linear and the transformed space might be high dimensional. Although the classifier is a hyperplane in the high-dimensional feature space, it may be non-linear in the original input space (Figure 1b).

SVMs belong to a family of generalized linear classifiers. They can also be considered as a special case of Tikhonov regularization (most commonly used method of regularization of ill-posed problems, in statistics the method is also known as ridge regression – Tarantola, 2004).

2.2. Kernel PCA

Principal Component Analysis (PCA) is a vector space transformation often used to reduce multidimensional data sets to lower dimensions for analysis.

PCA is an orthogonal linear transformation of the coordinate system in which we describe our data such, that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. Unfortunately we have to assume that observed data set to be linear combinations of certain basis. Kernel principal component analysis (kPCA) is an extension of principal component analysis (PCA) using techniques of kernel methods (without assuming linearity). Using a kernel, the originally linear operations of PCA are done in a *reproducing kernel Hilbert space* with a non-linear mapping (Fugure 2). It is a successful example of "kernelizing a well-known linear algorithm. Schölkopf *et al.* (1998) showed that finding and projecting onto principal components depend on just the inner-product and kernel trick could be use. There are several important points to note about the behavior of the kPCA components, which should be contrasted with the behavior of classic PCA:

- The maximum number of components is determined not by the dimensionality of the input data, but by the number of input data points.
- Not all sets of values of the components correspond to an actual point in input space.

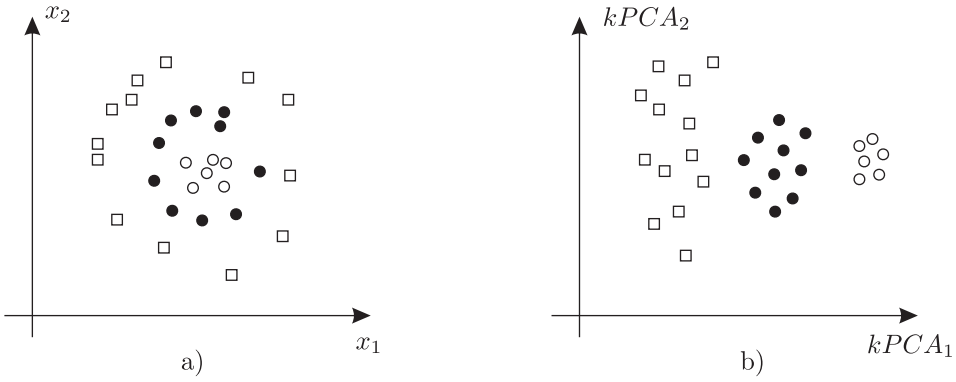

Figure 2. a) Input points before kernel PCA. b) Output after kernel PCA. The three groups are distinguishable using the first component only.

3. Basic kernels

According to (Schölkopf and Smola, 2002) we introduce basic definitions.

Definition 1. Given a function $k : \mathcal{X} \to \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the $n \times n$ matrix K with elements

$$K_{ij} = k(x_i, x_j)$$

is called the *Gram matrix* (or *kernel matrix*) of k with respect to $x_1, \ldots, x_n \in \mathcal{X}$.

Definition 2. A real symmetric $n \times n$ matrix K satisfying

$$\sum_{i,j} c_i c_j K_{ij} \geq 0$$

for all $c_i, c_j \in \mathbb{R}$ is called *positive definite* *.

Definition 3. Let $\mathcal{X}$ be a nonempty set. A function k on $\mathcal{X} \times \mathcal{X}$ which for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathcal{X}$ gives rise to a positive definite Gram matrix is called a *positive definite kernel* or in short form *kernel*.

The key idea of the kernel technique is to invert the chain of arguments. i.e., choose a kernel k rather than a mapping before applying a learning algorithm. Not every symmetric function can be a kernel. The necessary and sufficient condition for this are given by Mercer's theorem.

Theorem 1 (Mercer's theorem)**.** *Suppose $k \in L_\infty(\mathcal{X} \times \mathcal{X})$ is a symmetric function, such that the integral operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ given by*

$$T_k f(\cdot) = \int_{\mathcal{X}} k(\cdot, x) f(x) dx$$

is positive definite, that is,

$$\int_{\mathcal{X}} \int_{\mathcal{X}} k(x_1, x_2) f(x_1) f(x_2) dx_1 dx_2 \geq 0,$$

*some authors call this nonnegative definite

for all $f \in L_2(\mathcal{X})$. *Such kernel we call* Mercer kernel.

Let $\psi_i \in L_2(\mathcal{X})$ *be the eigenfunction of* T_k *associated with the eigenvalue* $\lambda_i \geq 0$ *and normalized such that* $||\psi_i||_2 = \int_{\mathcal{X}} \psi_i^2(x)dx = 1$, *i.e,*

$$\forall x \in \mathcal{X} : \int_{\mathcal{X}} k(x_1, x_2)\psi_i(x_2)dx_2 = \lambda_i \psi_i(x_1).$$

Then

1. $(\lambda_i)_{i\in\mathbb{N}} \in l_1$,
2. $\psi_i \in L_\infty(\mathcal{X})$,
3. k *can be expanded in a uniformly convergent series, i.e.,*

$$k(x_1, x_2) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x_1)\psi_i(x_2)$$

holds for all $x_1, x_2 \in \mathcal{X}$.

Proposition 1 (Herbrich, 2002)**.** *The function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is a Mercer kernel if and only if is a kernel in sense of Definition* 3 *(for almost all* x*).*

If we have a positive definite kernel such as in Definition 3, which, however, is not in $L_\infty(\mathcal{X})$, to use Mercer's theorem we can take any compact subset of $\mathcal{X}$ containing all observations.

Example 1. A few simple examples of functions that are kernels or not.

- Kernels: $\langle x_1, x_2\rangle$, $e^{-\|x_1-x_2\|^2}$, $e^{\langle x_1,x_2\rangle}$;
- Not kernels: $\|x_1 - x_2\|^2$, $-\|x_1 - x_2\|^2$, $-\langle x_1, x_2\rangle$, $e^{\|x_1-x_2\|^2}$, $e^{-\langle x_1,x_2\rangle}$.

Example 2. Lets show that $k(x_1, x_2) = \langle x_1, x_2\rangle$, $x_1, x_2 \in \mathcal{X} \subset \mathbb{R}^n$ is a kernel in sense of Definition 3 and Theorem 1.

For $x_i = (x_i^1, \dots, x_i^n) \in \mathcal{X}$, $c_i \in \mathbb{R}$, $i = 1, \dots, m$ we have

$$\begin{aligned}\sum_{i,j} c_i c_j \langle x_i, x_j \rangle &= \sum_{i,j} c_i c_j \sum_k x_i^k x_j^k \\ &= \sum_k \sum_{i,j} c_i x_i^k \, c_j x_j^k \\ &= \sum_k \left[\left(\sum_i c_i x_i^k \right) \left(\sum_j c_j x_j^k \right) \right] \\ &= \sum_k \left(\sum_i c_i x_i^k \right)^2 \geq 0.\end{aligned}$$

For $f \in L_\infty(\mathcal{X})$ we have

$$\begin{aligned}\int_{\mathcal{X}} \int_{\mathcal{X}} \langle x_1, x_2 \rangle f(x_1) f(x_2) \, dx_1 dx_2 &= \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\sum_i x_1^i x_2^i \right) f(x_1) f(x_2) dx_1 dx_2 \\ &= \sum_i \int_{\mathcal{X}} \int_{\mathcal{X}} x_1^i f(x_1) \, x_2^i f(x_2) \, dx_1 dx_2 \\ &= \sum_i \left(\int_{\mathcal{X}} x_1^i f(x_1) \, dx_1 \int_{\mathcal{X}} x_2^i f(x_2) \, dx_2 \right) \\ &= \sum_i \left(\int_{\mathcal{X}} x^i f(x) \, dx \right)^2 \geq 0.\end{aligned}$$

Example 3. We can show that function $k(x_1, x_2) = \|x_1 - x_2\|^2$ is not a kernel.

Let $m = 2$, $x_1, x_2 \in \mathcal{X}$, $c_1, c_2 \in \mathbb{R}$. Then

$$\sum_{i,j \in \{1,2\}} c_i c_j \|x_i - x_j\|^2 = 2 c_1 c_2 \|x_1 - x_2\|^2 < 0$$

for $x_1 \neq x_2$, $c_1 c_2 < 0$.

4. Functions of kernels

Following facts show that we can create new kernels from existing kernels using a number of simple operations. In this way we can create complex kernels by basic operations that combine simpler kernels.

Theorem 2 (Herbrich, 2002)**.** *Let $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be any kernels. Then, the functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by*

1. $k(x_1, x_2) = k_1(x_1, x_2) + k_2(x_1, x_2)$,
2. $k(x_1, x_2) = ck_1(x_1, x_2)$ *for all* $c \in \mathbb{R}^+$,
3. $k(x_1, x_2) = k_1(x_1, x_2)k_2(x_1, x_2)$,
4. $k(x_1, x_2) = f(x_1)f(x_2)$ *for any function* $f : \mathcal{X} \to \mathbb{R}$,
5. $k(x_1, x_2) = x_1' B x_2$ *for any symmetric positive definite* B *matrix,*
6. $k(x_1, x_2) = \lim_{i \to \infty} k_i(x_1, x_2)$, *if the limit exists*

are also kernels.

The combination of kernels given in part (3) is often refereed to as the *Schur product*. We can easily decompose any kernel into the Schur product of its normalisation and the 1-dimensional kernel of part (4) with $f(x) = \sqrt{k(x, x)}$.

Corollary 1 (Herbrich, 2002)**.** *Let $k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel. Then, the functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by*

1. $k(x_1, x_2) = (k_1(x_1, x_2) + \theta_1)^{\theta_2}$ *for all* $\theta_1 \in \mathbb{R}^+$ *and* $\theta_2 \in \mathbb{N}$,
2. $k(x_1, x_2) = \exp\left(\frac{k_1(x_1, x_2)}{\sigma^2}\right)$ *for all* $\sigma \in \mathbb{R}^+$,
3. $k(x_1, x_2) = \exp\left(-\frac{k_1(x_1, x_1) - 2k_1(x_1, x_2) + k_1(x_2, x_2)}{2\sigma^2}\right)$ *for all* $\sigma \in \mathbb{R}^+$,
4. $k(x_1, x_2) = \frac{k_1(x_1, x_2)}{\sqrt{k_1(x_1, x_1)k_1(x_2, x_2)}}$,

are also kernels.

From Theorem 2 we see that the following theorems are true.

Theorem 3. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel. Let P be a polynomial of degree n with nonnegative coefficients:*

$$P(t) = \sum_{i=0}^{n} a_i t^i \qquad (a_i \geq 0).$$

Then the function $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$\tilde{k}(x_1, x_2) := P(k(x_1, x_2))$$

is a kernel.

Theorem 4. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel. Let $f : k(\mathcal{X}, \mathcal{X}) \to \mathbb{R}$ be a function which Taylor expansion has only nonnegative coefficients:*

$$f(t) = \sum_{i=0}^{\infty} a_i t^i \qquad (a_i \geq 0).$$

Then the function $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$\tilde{k}(x_1, x_2) := f(k(x_1, x_2))$$

is a kernel.

For example, functions with "good" (nonnegative coefficients) Taylor expansion: e^x, $\arcsin x$, $\sinh x$, $\cosh x$, $\tan x$, $\operatorname{arctanh} x$. Functions with "bad" Taylor expansion: $\sin x$, $\cos x$, $\arccos x$, $\arctan x$, $\operatorname{arcsinh} x$, $\tanh x$.

Corollary 2. *Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function and let $f\colon k(\mathcal{X}, \mathcal{X}) \to \mathbb{R}$ be a function for which there is an inverse f^{-1} with nonnegative Taylor expansion coefficients. Then k is a kernel if the superposition $f \circ k$ is a kernel.*

Example 4. If k is not a kernel, then functions

$$k_1(x_1, x_2) = \tanh(k(x_1, x_2)),$$

$$k_2(x_1, x_2) = \arctan(k(x_1, x_2))$$

are not kernels.

Indeed, functions arctanh and tan have "good" Taylor expansions.

Next theorems concern dot product kernels.

Definition 4 (Abramowitz and Stegun, 1972)**.** The solutions of Legendre's differential equation

$$\frac{d}{dx}\left[(1-x^2)\frac{d}{dx}P(x)\right] + n(n+1)P(x) = 0$$

are called *Legendre functions*. When n is an integer, the solution $P_n(x)$ is a polynomial. These solutions for $n = 0, 1, \ldots$ (with the normalization $P_n(1) = 1$) form a orthogonal polynomials called the *Legendre polynomials*. Each Legendre polynomial $P_n(x)$ is an nth-degree polynomial. It may be expressed using Rodrigues formula:

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left[(x^2-1)^n\right].$$

Associated Legendre functions are the canonical solutions of the general Legendre equation

$$(1-x^2)\,y'' - 2xy' + \left(n[n+1] - \frac{m^2}{1-x^2}\right)y = 0.$$

This equation has solutions that are nonsingular on $[-1, 1]$ only if n and m are integers with $0 \le m \le n$. When in addition m is even, the function is a polynomial. When m is zero and n integer, these functions are identical to the Legendre polynomials. These functions are denoted $P_n^m(x)$. Their most straightforward definition is in terms of derivatives of ordinary Legendre polynomials ($m \ge 0$):

$$P_n^m(x) = (-1)^m\,(1-x^2)^{m/2}\,\frac{d^m}{dx^m}\left(P_n(x)\right).$$

Since, by Rodrigues formula one obtains

$$P_n^m(x) = \frac{(-1)^m}{2^n n!}(1-x^2)^{m/2}\frac{d^{n+m}}{dx^{n+m}}(x^2-1)^n.$$

Coefficients c_n and c_n^m of expansion function $f(x)$ to series of Legendre and associated Legendre polynomials we calculate:

$$c_n = \frac{(2n+1)}{2} \int_{-1}^{1} f(x) P_n(x) dx,$$

$$c_n^m = \frac{(2n+1)(n-m)!}{2(n+m)!} \int_{-1}^{1} f(x) P_n^m(x) dx.$$

Theorem 5 (Shoenberg, 1942)**.** *Let $k(x_1, x_2) = f(\langle x_1, x_2 \rangle)$ be a function defined on $S \times S \subset \mathbb{R}^{m+3} \times \mathbb{R}^{m+3}$, where S is the unit sphere, and $f\colon [-1,1] \to \mathbb{R}$ is a function with expansion into Legendre polynomials P_n^m*

$$f(t) = \sum_{n=0}^{\infty} a_n P_n^m(t).$$

Then k is a kernel if and only if all $a_n \geq 0$.

Theorem 6 (Shoenberg 1942)**.** *Let $k(x_1, x_2) = f(\langle x_1, x_2 \rangle)$ be a function defined on $S \times S \subset H \times H$, where S is the unit sphere in an infinite dimensional Hilbert space H, and $f\colon [-1,1] \to \mathbb{R}$ is a function with a power series expansion*

$$f(t) = \sum_{n=0}^{\infty} a_n t^n.$$

Then k is a kernel if and only if all $a_n \geq 0$.

Remark 1. In order to prove positive definiteness for arbitrary dimensions it suffices to show that the Taylor expansion contains only positive coefficients. On the other hand, in order to prove that a candidate for a kernel function will never be positive definite, it is sufficient to show this for the Legendre Polynomials P_n.

Example 5. The function

$$k(x_1, x_2) = \tanh(a\langle x_1, x_2 \rangle + b)$$

is not a kernel for any $a, b \in \mathbb{R}$, $a \neq 0$. We have to show that the kernel does not satisfy the conditions of Theorem 5. Since this is very technical we refer the reader to work of Ovari (2000) for details, and explain how the method works in the simpler case of Theorem 6. The Taylor series of $\tanh(at+b)$ is equal

$$\tanh b + (1 - \tanh^2 b)at + (\tanh^3 b - \tanh b)a^2t^2 + \dots$$

Since the coefficients have to be nonnegative we have $\tanh b \geq 0$, $\tanh^3 b - \tanh b \geq 0$. Hence $b \geq 0$ and if $b \neq 0$ then $\tanh^2 b \geq 1$ — contradiction. If $b = 0$ the expansion is equal $at - \dfrac{a^3t^3}{3} + \dots$, and then $a = 0$ — contradiction.

Example 6. By computer computations we obtain that for parameters $a, b \in \{-3, -2, \dots, 2, 3\}$, $a \neq 0$ any function

$$f(x) := \frac{1}{1 + \exp(ax + b)}$$

has a negative coefficient in its expansion into Legendre polynomial series, therefore $k(x_1, x_2) = f(\langle x_1, x_2 \rangle)$ is not a kernel.

Example 7. Consider a function

$$f(x) := \frac{1}{1 - \exp(ax - b)}.$$

The function f is well definite for $0 < a < b$, $x \in [-1, 1]$ and its Taylor series is equal

$$\frac{e^b}{e^b - 1} + \frac{e^b\, a\, x}{(e^b)^2 - 2\, e^b + 1} + \frac{\left((e^b)^2 + e^b\right) a^2 x^2}{2\,(e^b)^3 - 6\,(e^b)^2 + 6\, e^b - 2} + \dots$$

$$= \sum_{n=0}^{\infty} c_n \frac{a^n\, x^n}{(e^b - 1)^{n+1}}, \qquad c_n > 0.$$

Thus all coefficients of the series are nonnegative and $k(x_1, x_2) = f(\langle x_1, x_2\rangle)$ is a kernel on the product of the unit spheres.

The next corollaries are simple consequence of the definition of kernels and Remark 1.

Corollary 3. *Let $k\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a map and let $T\colon \mathcal{X} \to \mathcal{Y}$ be a map such that $\mathcal{Y} = T(\mathcal{X})$. Then the map*

$$\tilde{k}(x_1, x_2) := k\big(T(x_1), T(x_2)\big)$$

is a kernel on $\mathcal{X} \times \mathcal{X}$ if and only if k is a kernel.

Corollary 4. *Let $T\colon \mathcal{X} \to H$ be a map such that $T(\mathcal{X}) \subset S$, where S is the unit sphere in Hilbert space H (finite ($\mathbb{R}^n$) or infinite-dimensional). Let $f\colon [-1, 1] \to \mathbb{R}$ be a function with Taylor expansion*

$$f(t) = \sum_{i=0}^{\infty} a_i t^i.$$

If all $a_i \geq 0$ then the map

$$k(x_1, x_2) := f\big(\langle T(x_1), T(x_2)\rangle\big)$$

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

Corollary 5. *Let $T\colon \mathcal{X} \to H$ be a map such that $S \subset T(\mathcal{X})$, where S is the unit sphere in Hilbert space H (finite ($\mathbb{R}^n$) or infinite-dimensional). Let $f\colon [-1, 1] \to \mathbb{R}$ be a function with expansion into Legendre polynomials*

$$f(t) = \sum_{i=0}^{\infty} a_i P_i(t).$$

If some $a_i < 0$ then the map

$$k(x_1, x_2) := f\big(\langle T(x_1), T(x_2)\rangle\big)$$

is not a kernel on $\mathcal{X} \times \mathcal{X}$.

Example 8. For any transformation T from X onto the unit sphere S the function

$$k(x_1, x_2) := \tanh\big(a\langle T(x_1), T(x_2)\rangle + b\big)$$

in not a kernel for any parameters a, b.

Corollary 6. *Let $f\colon D \subset \mathbb{R} \to \mathbb{R}$ be a function which can be written as $f(t) = g(at + b)$, where g — some function, and Legendre polynomial expansion of f has some negative coefficient for any a, b. Then for any kernel $k_0(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$ such that $S(0, r) \subset \Phi(X)$ ($S(0, r)$ — a sphere with a radius of $r > 0$ and the center in 0), the function*

$$k(x_1, x_2) := f\big(k_0(x_1, x_2)\big) = g\big(ak_0(x_1, x_2) + b\big)$$

in not a kernel for any parameters a, b.

Proof. Since $S(0, r) \subset \Phi(X)$, $S(0, 1) \subset \frac{1}{r}\Phi(X)$. Then, by Corollary 5,

$$g\big(r^2 a\langle \tfrac{1}{r}\Phi(x_1), \tfrac{1}{r}\Phi(x_2)\rangle + b\big) = g\big(a\langle \Phi(x_1), \Phi(x_2)\rangle + b\big) = k(x_1, x_2)$$

is not a kernel. ■

Example 9. For any kernel $k_0(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$ such that $S(0, r) \subset \Phi(\mathcal{X})$, the function

$$k(x_1, x_2) := \tanh\big(ak_0(x_1, x_2) + b\big)$$

in not a kernel for any parameters a, b.

Theorem 7 (Burges, 1999)**.** *Let $k(x_1, x_2) = f(\langle x_1, x_2 \rangle)$ be a dot product kernel, where $f\colon \mathbb{R} \to \mathbb{R}$ is a differentiable function. Then*

$$f(t) \geq 0, \quad f'(t) \geq 0, \quad f'(t) + tf''(t) \geq 0$$

for any $t \geq 0$.

Example 10. Let $k(x_1, x_2) = \exp\big(-\langle x_1, x_2 \rangle\big)$. Then $f(t) = e^{-t}$, $f'(t) = -e^{-t} < 0$, thus k is not a kernel.

5. Inverse of the stereographic projection

We see that we have many methods of checking of kernel on the sphere. So if we have observations in $\mathbb{R}^n$ we can use inverse of the stereographic projection T into $S \subset \mathbb{R}^{n+1}$ (see below) and then the superposition of T and any kernel k on sphere will be a kernel. Similarly, we can use this technique if we have a kernel on sphere which is not a kernel on a whole space.

Example 11. Let take the second Legendre polynomial $P_2(t) = -\frac{1}{2} + \frac{3}{2}t^2$. The function

$$k(x_1, x_2) := P_2(\langle x_1, x_2\rangle) = -\tfrac{1}{2} + \tfrac{3}{2}\langle x_1, x_2\rangle^2$$

is, by Theorem 5, a kernel on the unit spheres $S \subset \mathbb{R}^3$. But k is not a kernel on any subset of $\mathbb{R}^2$ including zero. Indeed, for $c \neq 0$, we have $c^2 k(0,0) < 0$ so, by Definition 3, k is not a kernel. Even if we exclude zero from the subset, k is not a kernel because $\lim\limits_{x\to 0} k(x,x) = -\frac{1}{2}$.

We construct inverse of the stereographic projection and introduce a new metric on $\mathbb{R}^n$ induced from Euclidean metric on the unit sphere in $\mathbb{R}^{n+1}$. This metric could be used not only to constructing kernels but also directly in, for example, classification methods.

We define a map $h\colon \mathbb{R}^n \cup \{\infty\} \to \mathbb{R}^{n+1}$,

$$\mathbb{R}^n \ni x = (x^1, \ldots, x^n) \mapsto y = (y^1, \ldots, y^n, y^{n+1}),$$

$$\infty \mapsto (0, \ldots, 0, 1).$$

To find $y = h(x)$ for $x \in \mathbb{R}^n$ we take an n-dimensional sphere $S_{(\frac{1}{2},\frac{1}{2})}$ in $\mathbb{R}^{n+1}$ with center in the point $(0, \ldots, 0, \frac{1}{2})$ and a radius of $\frac{1}{2}$. We draw a line through the points $(x^1, \ldots, x^n, 0)$ and $(0, \ldots, 0, 1)$. The intersection of the line and the sphere (another than $(0, \ldots, 0, 1)$) is the result point $y = h(x)$.

The equation of the sphere is $(y^1)^2 + \cdots + (y^n)^2 + (y^{n+1} - \frac{1}{2})^2 = \frac{1}{4}$ and the parametrical equations of the line are

$$y^1 = x^1 t, \quad \ldots, \quad y^n = x^n t, \quad y^{n+1} = 1 - t, \qquad t \in \mathbb{R}.$$

Then the intersection point $y = (y^1, \ldots, y^n, y^{n+1})$ has the coordinates

$$y^1 = \frac{x^1}{\|x\|^2 + 1}, \quad \ldots, \quad y^n = \frac{x^n}{\|x\|^2 + 1}, \quad y^{n+1} = \frac{\|x\|^2}{\|x\|^2 + 1}.$$

Now, we define a metric d on $\mathbb{R}^n$ by

$$d(x_1, x_2) := d^{(n+1)}\big(h(x_1), h(x_2)\big),$$

where $d^{(n+1)}$ is the usual Euclidean metric on $\mathbb{R}^{n+1}$. We have

$$d(x_1, x_2) = \frac{\|x_1 - x_2\|}{\sqrt{\|x_1\|^2 + 1}\sqrt{\|x_2\|^2 + 1}}.$$

The metric d has the following properties

$$d(x_1, x_2) < 1, \quad d(x, \infty) = \frac{1}{\sqrt{\|x\|^2 + 1}}, \quad d(0, \infty) = 1$$

for $x_1, x_2 \in \mathbb{R}^n$.

If we need a map onto the unit sphere $S = S_{(0,1)}$, we can define

$$\tilde{h}(x) := 2\big[h(x) - (0, \ldots, 0, \tfrac{1}{2})\big].$$

Then we have $\tilde{d}(x_1, x_2) = 2d(x_1, x_2)$.

If we need to transform the closed ball $\bar{B} \subset \mathbb{R}^n$ (with the center in x_0 and a radius of $]r > 0$) onto a sphere $S \in \mathbb{R}^{n+1}$ we can take a transformation $T\colon \bar{B} \subset \mathbb{R}^n \to \mathbb{R}^n \cup \{\infty\}$ defined

$$T(x) := \begin{cases} g\left(\dfrac{\|x - x_0\|}{r}\right)(x - x_0) & \text{for } x \in B, \\ \infty & \text{for } x \in \bar{B} \setminus B, \end{cases}$$

where g is some function which maps $[0, 1]$ onto $[0, \infty)$, for example: $\operatorname{arctanh}(t)$ or $\tan(\frac{\pi}{2}t)$. Then we define $\hat{h}\colon \bar{B} \subset \mathbb{R}^n \to S \subset \mathbb{R}^{n+1}$ as $\hat{h} = h \circ T$ or $\hat{h} = \tilde{h} \circ T$. Note that all points from boundary of $\bar{B}$ are mapped on the same point in $\mathbb{R}^{n+1}$.

Example 12. If we have all data in the unit ball $B \subset \mathbb{R}^n$, we can transform the ball into the unit sphere $S \subset \mathbb{R}^{n+1}$. We have

$$T(x) = g(\|x\|)\, x,$$

where g is a function mentioned above. Then

$$\langle y_1, y_2\rangle = \frac{4\tilde{g}(x_1,x_2) + (\tilde{g}(x_1)-1)(\tilde{g}(x_2)-1)}{(\tilde{g}(x_1)+1)(\tilde{g}(x_2)+1)},$$

where

$$\tilde{g}(x_1,x_2) = g(\|x_1\|)g(\|x_2\|)\,\langle x_1,x_2\rangle$$

$$\tilde{g}(x) = \tilde{g}(x,x).$$

Thus we have a kernel on $B\times B$

$$k(x_1,x_2) := f\big(\langle y_1,y_2\rangle\big)$$

for any function f which satisfies conditions of Corollary 4.

6. Multivariable functions of kernels

We can generalize the method of superposition functions of one variable with kernels (Section 4) to the case of multivariable functions. If we have a function of n variables with "good" Taylor expansion then its superposition with n kernels is a kernel. This enables us to construct new kernels and simplifies checking that a function is a kernel.

Next theorems concern multivariable functions and directly follow Theorem 2.

Theorem 8. *Let $k_i\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ $(i=1,\dots,n)$ be kernels. Let $P\colon \mathbb{R}^n\to\mathbb{R}$ be a several variable polynomial with nonnegative coefficients:*

$$P(t_1,\dots,t_n) = \sum_{i=1}^{m} a_i\, t_1^{i_1}\dots t_n^{i_n}, \qquad a_i\geq 0.$$

Then the function $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ defined by

$$k(x_1,x_2) := P\big(k_1(x_1,x_2),\dots,k_n(x_1,x_2)\big)$$

is a kernel.

Theorem 9. *Let $k_i\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ $(i=1,\dots,n)$ be kernels. Let $f\colon k_1(\mathcal{X},\mathcal{X})\times\cdots\times k_n(\mathcal{X},\mathcal{X})\to\mathbb{R}$ be a several variable function which Taylor expansion has only nonnegative coefficients:*

$$f(t_1,\dots,t_n)=\sum_{i=1}^{\infty} a_i\, t_1^{i_1}\dots t_n^{i_n}, \qquad a_i\geq 0.$$

Then the function $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ defined by

$$k(x_1,x_2):=f\big(k_1(x_1,x_2),\dots,k_n(x_1,x_2)\big)$$

is a kernel.

Example 13. Let $n\in\mathbb{N}$. If $k_i\colon \mathcal{X}\times\mathcal{X}\to(-1,1)$, $i=1,\dots,n$ are kernels then the following functions are kernels:

$$K_1(x_1,x_2):=\prod_{i=1}^{n}\frac{1}{1-k_i(x_1,x_2)},$$

$$K_2(x_1,x_2):=\prod_{i=1}^{n}\frac{1+k_i(x_1,x_2)}{1-k_i(x_1,x_2)}.$$

For $t_i\in(-1,1)$, $i=1,\dots,n$ we have

$$\sum_{\alpha_1,\dots,\alpha_n\in\mathbb{N}\cup\{0\}} t_1^{\alpha_1}\dots t_n^{\alpha_n}=\prod_{i=1}^{n}\frac{1}{1-t_i},$$

$$\sum_{\alpha_1,\dots,\alpha_n\in\mathbb{Z}} t_1^{|\alpha_1|}\dots t_n^{|\alpha_n|}=\prod_{i=1}^{n}\frac{1+t_i}{1-t_i}.$$

Therefore, by Theorem 9, K_1, K_2 are kernels.

Example 14. If k_1, k_2 are kernels and $a, b, d \geq 0$, $c > 0$, $m, n \in \mathbb{N}$ then

$$k(x_1, x_2) := \frac{\big(a + bk_1(x_1, x_2)\big)^m}{\big(c - dk_2(x_1, x_2)\big)^n}$$

is a kernel.

Indeed, we have $k(x_1, x_2) = f\big(k_1(x_1, x_2), k_2(x_1, x_2)\big)$, where

$$f(t_1, t_2) = \frac{(a + bt_1)^m}{(c - dt_2)^n}.$$

The partial derivatives are

$$\frac{\partial^k}{\partial t_1^k}(a + bt_1)^m = \begin{cases} \dfrac{m!}{(m-k)!} b^k (a + bt_1)^{m-k} & \text{for } k \leq m \\ 0 & \text{for } k > m \end{cases}$$

$$\frac{\partial^l}{\partial t_2^l}(c - dt_2)^{-n} = \frac{(n-1+l)!}{(n-1)!} d^l (c - dt_2)^{-(n+l)} \quad \text{for } l \in \mathbb{N}$$

and

$$\frac{\partial^{k+l} f}{\partial t_1^k \partial t_2^l}(0,0) = \begin{cases} \dfrac{m!\,(n-1+l)!}{(m-k)!\,(n-1)!} \dfrac{a^{m-k}\, b^k\, d^l}{c^{n+l}} & \text{for } k \leq m \\ 0 & \text{for } k > m. \end{cases}$$

Hence all coefficients in Taylor expansion of function f are nonnegative and, by Theorem 9, k is a kernel.

Example 15. Let k_{1i}, k_{2i} be kernels and $a_i, b_i, d_i \geq 0$, $c_i > 0$, $m_i, n_i \in \mathbb{N}$, $i = 1, \ldots, n$. Then

$$k(x_1, x_2) := \prod_{i=1}^{n} \frac{\big(a_i + b_i k_{1i}(x_1, x_2)\big)^{m_i}}{\big(c_i - d_i k_{2i}(x_1, x_2)\big)^{n_i}}$$

is a kernel and a generalization of kernels from Examples 13–14.

In all above examples we have to note that the Taylor expansions are convergent to appropriate functions only if the kernels are well definite. In this example it has to hold $|k_{2i}(x_1, x_2)| < \frac{c_i}{d_i}$ $(d_i \neq 0)$ for $i = 1, \ldots, n$.

7. Conclusion

We showed a few method of constructing kernels on metric spaces. We hope that this could be useful for researchers using kernel methods. The choice of proper kernel is very difficult and corresponds to:

- choosing a similarity measure for the data,
- choosing a linear representation of the data,
- choosing a regularization functional,
- choosing a covariance function for correlated observations.

Therefore, this choice should reflect prior knowledge about the problem at hand.

References

[1] M. Abramowitz and I.A. Stegun, Chs. Legendre functions and orthogonal polynomials in Handbook of mathematical functions, Dover Publications, New York 1972.

[2] B.E. Boser, I.M. Guyon and V.N. Guyon, *A training algorithm for optimal margin classifiers*, in D. Haussler, eds. 5th Annual ACM Workshop on COLT. ACM Press, Pittsburgh (1992), 144–152.

[3] C.J.C. Burges, *Geometry and invariance in kernel based methods* in: Schölkopf, B. Burges, C.J.C. Smola, A.J. eds. Advances in kernel methods — support vector learning. MIT Press, Cambridge (1999), 89–116.

[4] C. Cortes and V. Vapnik, *Support-Vector Networks*, Machine Learning **20** (1995), 273–297.

[5] R. Herbrich, *Learning Kernel Classifiers*, MIT Press, London 2002.

[6] T. Hofmann, B. Schölkopf and A.J. Smola, *Kernels methods in machine learning*, Annals of Statistics **36** (2008), 1171–1220.

[7] Z. Ovari, Kernels, eigenvalues and support vector machines, Honours thesis, Australian National University, Canberra 2000.

[8] B. Schölkopf and A.J. Smola, *Learning with Kernels*, MIT Press, London 2002.

[9] B. Schölkopf, A.J. Smola and K.R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural Computation **10** (1998), 1299–1319.

[10] I.J. Schoenberg, *Positive definite functions on spheres*, Duke Mathematical Journal **9** (1942), 96–108.

[11] A. Tarantola, Inverse problem theory and methods for model paramenter estimation, SIAM, Philadelphia 2005.

[12] M. Zu, *Kernels and ensembles: perspective on statistical learning*, American Statistician **62** (2008), 97–109.

Received 8 March 2010