Machine Learning
Kernels I

Prof. Matthias Hein

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Lecture 12, 29.11.2013

Motivation for using kernel methods ?

- The approach using kernels includes the previous basis function/feature map approach,
- easier intuition of kernels in terms of a similarity function, $\Longrightarrow$ kernels on structured domains !
- direct penalization of functional properties (smoothness) instead of indirect penalization of weights.

Program for today:

- Basic notions of functional analysis,
- From basis functions to kernels - Motivation
- Positive definite kernels and reproducing kernel Hilbert spaces

Definition

A **metric space** is a set $\mathcal{X}$ with a distance function $d : \mathcal{X} \times X \to \mathbb{R}$ such that:

- $d(x, y) \geq 0$,
- $d(x, y) = 0$ if and only if $x = y$,
- $d(x, y) = d(y, x)$, (symmetry)
- $d(x, y) \leq d(x, z) + d(z, y)$. (triangle inequality)

It is denoted as $(\mathcal{X}, d)$. For a **semi-metric** $d(x, y) = 0$ does not imply $x = y$.

Remark: any semi-metric space $(\mathcal{X}, d)$ can be turned into a metric space by identifying points which have zero distance.

Convergence and Cauchy sequences:

Definition

A sequence of elements $\{x_n\}_{n\in\mathbb{N}}$ of a metric space $(\mathcal{X}, d)$ is said to **converge** to an element $x \in \mathcal{X}$ if $\lim_{n\to\infty} d(x, x_n) = 0$. We will denote this either as $x_n \xrightarrow{d} x$ or $\lim_{n\to\infty} x_n = x$.

Definition

A sequence of elements $\{x_n\}$ of a metric space $(\mathcal{X}, d)$ is called a **Cauchy sequence** if $\forall\, \epsilon > 0, \quad \exists N$ such that $d(x_n, x_m) < \epsilon, \;\forall\, n, m > N$.

Proposition: Every convergent sequence is a Cauchy sequence.

Definition

A metric space in which all Cauchy sequences converge is called **complete**.

Function spaces as vector spaces

Sets of functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ as vector spaces - apply vector axioms pointwise.

Three functions $f, g, h \in \mathcal{F}$, $\alpha, \beta \in \mathbb{R}$,

$$(f+g)(x) := f(x) + g(x), \quad \forall x \in \mathcal{X},$$
$$(\alpha f)(x) := \alpha\, f(x), \quad \forall x \in \mathcal{X}.$$

Associativity	$\big(f(x)+g(x)\big)+h(x) = f(x) + \big(g(x)+h(x)\big)$,
Commutativity	$f(x)+g(x) = g(x)+f(x)$,
Identity (addition)	$f(x)+0 = f(x)$, $\Rightarrow$ zero function $h(x) = 0$, $\forall x \in \mathcal{X}$,
Distributivity I	$(\alpha+\beta)f(x) = \alpha f(x) + \beta f(x)$,
Distributivity II	$\alpha(f(x)+g(x)) = \alpha f(x) + \alpha g(x)$,
Compatibility	$(\alpha\beta)f(x) = \alpha(\beta f(x))$,
Identity (multiplication)	$(1f(x)) = f(x)$.

Examples of function spaces

Sets of functions as vector spaces:

- all linear functions (finite dimensional),
- all polynomials (infinite dimensional),
- given a set of functions $\{\phi_1, \ldots, \phi_D\}$, they generate an D-dimensional vector space by taking all linear combinations:

$$\mathcal{F} = \mathrm{span}\{\phi_1, \ldots, \phi_D\} := \Big\{ \sum_{i=1}^{D} \alpha_i \phi_i \,\Big|\, \alpha_i \in \mathbb{R}, \quad i = 1, \ldots, D \Big\}.$$

$\Longrightarrow$ given that the functions are linearly independent,

$$\sum_{i=1}^{D} c_i \phi_i(x) = 0, \quad \forall\, x \in \mathcal{X}, \quad \Longrightarrow \quad c_i = 0, \quad i = 1, \ldots, D,$$

$\{\phi_1, \ldots, \phi_D\}$ is then also a basis of $\mathcal{F}$ (by definition).

Definition

A real vector space V is called an **inner product space** if there is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the following four conditions $\forall x, y, z \in V$ and $\forall \alpha \in \mathbb{R}$:

1. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$,
2. $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$,
3. $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle$,
4. $\langle x, y \rangle = \langle y, x \rangle$.

The function $\langle \cdot, \cdot \rangle$ is called **inner product**.

Every inner product defines a norm, $\|x\| := \sqrt{\langle x, x \rangle}$, and a metric, $d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$.
On inner product spaces we have the **Cauchy-Schwarz inequality:**

$$|\langle x, y \rangle| \leq \|x\| \, \|y\| .$$

Example - Inner product spaces

An inner product on functions: Let $f, g : \mathcal{X} \to \mathbb{R}$,

$$\langle f, g \rangle := \int_{\mathcal{X}} f(x)g(x)dx.$$

We obtain:

- $\langle f, f \rangle = \int_{\mathcal{X}} \big(f(x)\big)^2 dx \geq 0$,
- $\langle f, f \rangle = 0$ if and only if $f = 0$ (almost everywhere),
- $\langle f, g + h \rangle = \int_{\mathcal{X}} f(x)\big(g(x) + h(x)\big)dx = \int_{\mathcal{X}} f(x)g(x)dx + \int_{\mathcal{X}} f(x)h(x)dx$,
- $\langle f, \alpha g \rangle = \int_{\mathcal{X}} f(x)(\alpha\, g)(x)dx = \alpha \int_{\mathcal{X}} f(x)g(x)dx = \alpha \, \langle f, g \rangle$,
- $\langle f, g \rangle = \int_{\mathcal{X}} f(x)g(x)dx = \int_{\mathcal{X}} g(x)f(x)dx = \langle g, f \rangle$.

Induced norm

$$\|f\| = \sqrt{\langle f, f \rangle} = \Big(\int_{\mathcal{X}} \big(f(x)\big)^2 dx \Big)^{\frac{1}{2}}.$$

Hilbert spaces

Definition

A complete, inner product space is called a **Hilbert space**.

Definition

If S is an orthonormal set in a Hilbert space $\mathcal{H}$ and no other orthonormal set contains S as a proper subset, then S is called an **orthonormal basis** (or a **complete orthonormal system**) for $\mathcal{H}$.

Theorem

Let $\mathcal{H}$ be a Hilbert space and $S = \{x_\alpha\}_{\alpha \in A}$ an orthonormal basis. Then for each $y \in \mathcal{H}$,

$$y = \sum_{\alpha \in A} \langle x_\alpha, y \rangle \, x_\alpha, \qquad \|y\|^2 = \sum_{\alpha \in A} |\, \langle x_\alpha, y \rangle \,|^2.$$

Example - Hilbert Space

The space $L_2(\mathcal{X})$ of square-integrable functions:

$$L_2(\mathcal{X}) := \Big\{ f : \mathcal{X} \to \mathbb{R} \,\Big|\, \int_{\mathcal{X}} \big(f(x)\big)^2 dx < \infty \Big\},$$

is a Hilbert space together with the inner product,

$$\langle f, g \rangle := \int_{\mathcal{X}} f(x) g(x) dx.$$

but !

- $\|f\| = 0$ if and only if f is zero almost everywhere !
- Functions which agree almost everywhere are identified (the relation equal almost everywhere defines equivalence classes of functions),
- $L_2(\mathcal{X})$ is not a space of pointwise defined functions !

In machine learning we need a space of pointwise defined functions, since we have to do predictions at each point !

Definition

Positive Definite Kernels:

Definition

A real-valued symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **positive definite (PD) kernel** if for all $m \geq 1$, $x_1, \ldots, x_m \in \mathcal{X}$, $c_1, \ldots, c_m \in \mathbb{R}$

$$\sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \geq 0$$

The set of all real-valued positive definite kernels on $\mathcal{X}$ is denoted $\mathbb{R}_+^{\mathcal{X} \times \mathcal{X}}$.

Remark:

- In this lecture a kernel is always positive definite if not stated otherwise.
- Note that $\mathcal{X}$ is a general set $\Longrightarrow$ later on we will define kernels on structured domains (graphs, histograms, etc.).

Basis functions and positive definite kernels

Using so called **basis functions** or **feature maps** $\phi_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, D$ we created nonlinear functions

$$f(x) = \sum_{i=1}^{D} w_i \phi_i(x).$$

How to choose a set of basis functions:

- prior knowledge about the true underlying function,
- subset of the basis functions of a complete basis of the function space (e.g. Fourier basis, Wavelet basis etc.)
- local functions (e.g. Gaussians) centered on the data points.

In the following we assume $D \gg n$.

Problem: A large set of basis functions can lead to overfitting !

New Parameters:

- replace original parameters w_i by new parameters α_j, $j = 1, \ldots, n$,
- one parameter per data point instead of one per basis function.

Definition

$$w_i = \sum_{j=1}^{n} \alpha_j \phi_i(x_j) \quad \Longleftrightarrow \quad w = \Phi^T \alpha,$$

where $\Phi \in \mathbb{R}^{n \times D}$ is the design matrix introduced in the last chapter.

Problem: if $D \gg n$ there exists for general w no solution α.

Solution: w_i determined by data $\rightarrow$ n degrees of freedom. Given that the map "data" $\mapsto$ "weights" is linear, weights fill n dimensional subspace.

Basis functions and positive definite kernels III

With the new parameters α we can write the function f as

$$f(x) = \sum_{i=1}^{D} w_i \phi_i(x) = \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{D} \phi_i(x_j)\phi_i(x).$$

Defining the function $k(x, y) = \sum_{i=1}^{D} \phi_i(x)\phi_i(y)$ we can write $f(x)$ as

$$f(x) = \sum_{i=1}^{n} \alpha_i \, k(x_i, x) \qquad \text{(kernel expansion of } f\text{)}.$$

Note that $k(x, y)$ is **symmetric and positive definite**,

$$\sum_{r,s=1}^{m} c_r c_s k(x_r, x_s) = \sum_{r,s=1}^{m} c_r c_s \sum_{i=1}^{D} \phi_i(x_r)\phi_i(x_s) = \sum_{i=1}^{D} \sum_{r=1}^{m} c_r \phi_i(x_r) \sum_{s=1}^{m} c_s \phi_i(x_s)$$

$$= \sum_{i=1}^{D} \Big(\sum_{r=1}^{m} c_r \phi_i(x_r) \Big)^2 \geq 0.$$

Basis functions and positive definite kernels IV

In the case where $D \gg n$ the kernel expansion of f

$$f(x) = \sum_{i=1}^{n} \alpha_i \, k(x_i, x) \qquad \text{(kernel expansion of } f\text{)}.$$

can be efficiently computed if we have a closed form expression of $k(x, y)$.
Example: countably infinite many feature maps $\phi_r : \mathbb{R} \to \mathbb{R}$ with

$$\phi_r(x) = \frac{1}{\sqrt{r!}} x^r, \quad r = 0, 1, \ldots$$

$$\Longrightarrow \qquad k(x, y) = \sum_{r=0}^{\infty} \phi_r(x)\phi_r(y) = \sum_{r=0}^{\infty} \frac{x^r}{\sqrt{r!}} \frac{y^r}{\sqrt{r!}} = \sum_{r=0}^{\infty} \frac{(xy)^r}{r!} = e^{xy}.$$

- every kernel admits a feature map expansion but it is not necessarily unique.
- The kernel function itself is more important than the feature maps.

Kernels as (dis)similarity measures

The kernel function as similarity measure of points x and y:
This interpretation is motivated by the fact that the kernel function $k(x, y)$ can be seen as an inner product

$$k(x, y) = \langle \Psi(x), \Psi(y) \rangle \,,$$

where $\Psi : \mathcal{X} \to \mathcal{H}$ and $\mathcal{H}$ is a **Hilbert space**.

We will show later that the new kernel $\tilde{k}$ defined as,

$$\tilde{k}(x, y) = \frac{k(x, y)}{\sqrt{k(x, x)k(y, y)}} = \frac{\langle \Psi(x), \Psi(y) \rangle}{\|\Psi(x)\| \, \|\Psi(y)\|} = \cos(\angle(\Psi(x), \Psi(y))),$$

is again a positive definite kernel.

- The cosine is a common similarity measure (text classification).
- $|\tilde{k}(x, y)| \leq 1$ and $\tilde{k}(x, y) = 1$ if and only if $x = y$.

Dissimilarity measure:
The kernel induces a distance function (semi-metric):

$$\begin{aligned} d(x,y) &= \|\Psi(x)-\Psi(y)\| = \sqrt{\|\Psi(x)-\Psi(y)\|^2} \\ &= \sqrt{k(x,x)+k(y,y)-2k(x,y)} \end{aligned}$$

Why are (dis)similarity measures useful for learning ?

- one needs only to define the similarity $k(x,y)$ of two points x and y instead of a set of functions $\phi_i(x)$ in the feature maps approach.
- construction of a similarity measure between structured objects e.g. graphs is conceptually much easier than defining a certain set of feature maps on these structured objects.

$\Longrightarrow$ One of the main reasons why learning methods based on kernels are very popular in machine learning.

We will show that we have the relationship

$$\text{positive definite kernel} \quad \Longleftrightarrow \quad \text{Reproducing Kernel Hilbert Space } \mathcal{H}_k.$$

The function space $\mathcal{H}_k$ will be our hypothesis class for learning.
The norm of $\mathcal{H}_k$ will be used as regularization functional $\Omega(f) = \|f\|^2_{\mathcal{H}_k}$.

Advantages of a Reproducing Kernel Hilbert space

- In the basis function approach, $f(x) = \sum_j w_j \phi_j(x)$, we used $\Omega(f) = \sum_j w_j^2$,
  $\Rightarrow$ no direct penalization of properties (in particular smoothness) of the function but only indirectly via the weights.
- The RKHS norm $\|f\|^2_{\mathcal{H}_k}$ can often be directly connected to smoothness properties of the function.

Example for a regularization functional of a RKHS:

The Gaussian kernel on $\mathbb{R}$ is defined as

$$k(x,y) = e^{-\frac{(x-y)^2}{2\sigma^2}},$$

It can be shown that the induced norm of the RKHS is given as

$$\|f\|_{\mathcal{H}_k}^2 = \int_{\mathbb{R}} \sum_{s=0}^{\infty} \frac{\sigma^{2s}}{s!2^s} \big((\partial_x^s f)(x)\big)^2 \, dx.$$

- The RKHS consists only of smooth functions ($f \in C^\infty(\mathbb{R})$ for every $f \in \mathcal{H}_k$) ,
- In the norm the integral of squared s-th derivative of the function $f \in \mathcal{H}_k$ is penalized with a weight $\frac{\sigma^{2s}}{s!2^s}$ and then the contributions of all derivatives are summed.

Properties of kernels

Definition

A real-valued symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **positive definite (PD) kernel** if for all $m \geq 1$, $x_1, \ldots, x_m \in \mathcal{X}$, $c_1, \ldots, c_m \in \mathbb{R}$,

$$\sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \geq 0$$

A kernel k is **strictly positive definite** if strict inequality holds for any distinct $x_1, \ldots, x_m$ and $c \neq 0$.

Remark: In mathematics one usually calls a matrix A with

- $\langle w, Aw \rangle \geq 0$ for all $w \neq 0$ $\quad\Longrightarrow\quad$ A is **positive semi-definite**
- $\langle w, Aw \rangle > 0$ for all $w \neq 0$ $\quad\Longrightarrow\quad$ A is **positive definite**.

Mathematics	$\Longleftrightarrow$	Machine Learning
positive semi-definite	$\Longleftrightarrow$	positive definite
positive definite	$\Longleftrightarrow$	strictly positive definite.

Notation

Definition

Given a kernel k and a set of n points $x_1, \ldots, x_n \in \mathcal{X}$ the $n \times n$ matrix

$$K = (k(x_i, x_j))_{ij},$$

is called the **kernel matrix** (or **Gram Matrix**) K of the kernel k with respect to $x_1, \ldots, x_n$.

Properties of kernels

Properties of kernels: Transformations of kernels which preserve the property of positive definiteness are important for

1. the construction of new kernels,
2. the verification that a given function $k(x, y)$ is positive definite
   $\Rightarrow$ people often have a similarity measure (biology) they would like to use.

Proposition

Let k, k_1 and k_2 be positive definite kernels on $\mathcal{X} \times \mathcal{X}$.

i) *for any $\alpha \geq 0$, $k(x, y) = \alpha\, k_1(x, y)$ is positive definite,*
ii) *$k(x, y) = k_1(x, y) + k_2(x, y)$ is positive definite (**pointwise addition**),*
iii) *$k(x, y) = k_1(x, y)k_2(x, y)$ is positive definite (**pointwise multiplication**),*
iv) *the **pointwise limit** k of a sequence of positive definite kernels k_n on $\mathcal{X} \times \mathcal{X}$ is positive definite,*
v) *for any $f : \mathcal{X} \to \mathbb{R}$, $k'(x, y) = f(x)f(y)k(x, y)$ is positive definite, especially $k(x, y) = f(x)f(y)$,*
vi) *for any $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is a dot product space, $k(x, y) = \langle \phi(x), \phi(y) \rangle$ is positive definite,*

iii) Let $K^{(1)}$ and $K^{(2)}$ be the $n \times n$ kernel matrices with respect to n points. Since $K^{(2)}$ is positive definite it allows a decomposition $K^{(2)}_{ij} = \sum_{m=1}^{n} L_{im} L_{jm}$ where L is the square root of $K^{(2)}$. Then

$$\sum_{i,j=1}^{n} c_i c_j K^{(1)}_{ij} K^{(2)}_{ij} = \sum_{m=1}^{n} \sum_{i,j=1}^{n} c_i L_{im} c_j L_{jm} K^{(1)}_{ij} \geq 0,$$

For every m define $d_i^{(m)} = c_i L_{im} \Longrightarrow \sum_{m=1}^{n} \sum_{i,j=1}^{n} d_i^{(m)} d_j^{(m)} K^{(1)}_{ij} \geq 0$.

iv) Since one can exchange finite sums and limits, we have

$$\sum_{i,j=1}^{n} c_i c_j \lim_{l \to \infty} k_l(x_i, x_j) = \lim_{l \to \infty} \sum_{i,j=1}^{n} c_i c_j k_l(x_i, x_j) \geq 0,$$

v) $\sum_{i,j=1}^{n} c_i c_j f(x_i) f(x_j) k(x_i, x_j) = \sum_{i,j=1}^{n} (c_i f(x_i))(c_j f(x_j))\, k(x_i, x_j) \geq 0$.

vi) $\sum_{i,j=1}^{n} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle = \sum_{i,j=1}^{n} \langle c_i \phi(x_i), c_j \phi(x_j) \rangle =$ $\left\| \sum_{i=1}^{n} c_i \phi(x_i) \right\|^2 \geq 0$.

Reproducing kernel Hilbert spaces

Definition

A **reproducing kernel Hilbert space (RKHS)** $\mathcal{H}$ on $\mathcal{X}$ is a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ with a reproducing kernel $k(x, y)$ on $\mathcal{X} \times \mathcal{X}$ such that

$$\begin{aligned} &\forall\, x \in \mathcal{X}, \quad k(x, \cdot) \in \mathcal{H} \\ &\forall\, f \in \mathcal{H}, \quad \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x), \quad \text{(reproducing property)} \end{aligned}$$

Steps for the construction of a RKHS:

- consider the set of all finite linear combinations of the kernel:

$$\mathcal{G} = \mathrm{Span}\{k(x,.) : x \in \mathcal{X}\}$$

- Let $f(x) = \sum_i a_i k(x_i, x)$ and $g(x) = \sum_j b_j k(z_j, x)$. Then

$$\Big\langle \sum_i a_i k(x_i,.), \sum_j b_j k(z_j,.) \Big\rangle_{\mathcal{G}} := \sum_{i,j} a_i\, b_j\, k(x_i, z_j)\,.$$

- check that $\langle \cdot, \cdot \rangle$ is well-defined.

$$\sum_{i,j} a_i\, b_j\, k(x_i, z_j) = \sum_i a_i g(x_i) = \sum_j b_j f(z_j)$$

  The value of the inner product does not depend on the expansion of f or g $\Rightarrow$ (semi)-inner product with the reproducing property on $\mathcal{G}$.

Construction of Reproducing kernel Hilbert spaces II

Steps for the construction of a RKHS:

- construct the **semi-norm** associated to this inner product,

$$\|f\|_{\mathcal{G}}^2 = \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j).$$

  The Cauchy-Schwarz inequality holds also on semi-inner product spaces,

$$|f(x)| = |\langle f, k(x,.)\rangle_{\mathcal{G}}| \leq \|f\|_{\mathcal{G}} \|k(x,.)\|_{\mathcal{G}} = \|f\|_{\mathcal{G}} \sqrt{k(x,x)}.$$

  $\|f\|_{\mathcal{G}} = 0$ implies $f \equiv 0 \Rightarrow$ inner product on $\mathcal{G}$ and $\mathcal{G}$ is an **inner product space**.
- Standard completion by adding all limits of Cauchy sequences in $\mathcal{G}$
  - one has to check that the inner product as well as the reproducing property carries over to the limit elements.

Theorem (Moore)

If k is a positive definite kernel then there exists a **unique** *reproducing kernel Hilbert space $\mathcal{H}$ whose kernel is k.*

$\Rightarrow$ there is a **one-to-one** relation between reproducing kernel Hilbert spaces and positive definite kernels.

Characterization of Reproducing kernel Hilbert spaces II

Theorem (Aronszajn)

Let $\mathcal{H}$ be a Hilbert space of function from $\mathcal{X}$ to $\mathbb{R}$, then $\mathcal{H}$ is a reproducing kernel Hilbert space if and only if **all evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x(f) = f(x)$ are continuous**, *equivalently for all $x \in \mathcal{X}$, there exists a $M_x < \infty$ such that*

$$\forall f \in \mathcal{H}, \qquad |f(x)| \leq M_x \left\| f \right\|_{\mathcal{H}} .$$

- The RKHS is an Hilbert space of pointwise defined functions $\left\| f \right\|_{\mathcal{H}_k} = 0 \quad \Longrightarrow \quad f(x) = 0, \ \forall x \in \mathcal{X}$.
- The set of square integrable functions $L_2(\mathcal{X}) = \{ f \mid \int_{\mathcal{X}} f(x)^2 dx < \infty \}$ is not a Hilbert space of pointwise defined functions, $\int_{\mathcal{X}} f(x)^2 dx = 0 \quad \not\Longrightarrow \quad f(x) = 0, \ \forall x \in \mathcal{X}$.

In learning we need pointwise-defined functions since we want to make predictions at every point of the space.