Machine Learning
Kernels II

Prof. Matthias Hein

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

Lecture 14, 11.12.2013

Learning with kernels:

- As hypothesis space we use the RKHS $\mathcal{H}_k$ associated to the kernel k,
- As regularization functional we use: $\Omega(f) = \|f\|^2_{\mathcal{H}_k}$ (or more generally a strictly monotonically increasing function of $\|f\|_{\mathcal{H}_k}$)

Regularized empirical risk minimization problem with a RKHS as hypothesis space:

$$f^* = \operatorname*{arg\,min}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\, \Omega\Big(\|f\|^2_{\mathcal{H}_k} \Big),$$

Problems !?

- The RKHS has often very high dimension or is even infinite dimensional. This means we have a very high dimensional hypothesis space

$\Longrightarrow$ **Danger of overfitting !**

Well we use regularization... and the following representer theorem saves the day !

$\boxed{\textbf{Effectively we are working in an } n\textbf{-dimensional subspace of } \mathcal{H}_k \textbf{ !}}$

Representer Theorem II

Theorem (Representer Theorem)

Denote by $\Omega : [0, \infty) \to \mathbb{R}$ a strictly monotonically increasing function. Let $\mathcal{X}$ be the input space, $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ an arbitrary loss function and $\mathcal{H}_k$ the reproducing kernel Hilbert space associated to the kernel k. Then each minimizer $f^* \in \mathcal{H}_k$ of the regularized empirical risk

$$f^* = \underset{f \in \mathcal{H}_k}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \Omega\Big(\|f\|^2_{\mathcal{H}_k} \Big),$$

admits a representation as

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

(and we have for this function $\|f\|^2_{\mathcal{H}_k} = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$).

Representer Theorem III

Proof:

- $\mathcal{G} = \mathrm{Span}\{k(x_i, \cdot) \,|\, i = 1, \ldots, n\}$ is the finite dimensional subspace of $\mathcal{H}_k$ spanned by the data.
- Decompose any $f \in \mathcal{H}_k$ into $f^{\parallel} \in \mathcal{G}$ and the orthogonal part $f^{\perp} \in \mathcal{G}^{\perp}$.
  Then $f(x) = f^{\parallel}(x) + f^{\perp}(x) = \sum_{i=1}^n \alpha_i k(x_i, x) + f^{\perp}(x)$.
- Note that since $k(x_i, \cdot) \in \mathcal{G}$ and $f^{\perp} \in \mathcal{G}^{\perp}$ we have,
  $0 = \langle f^{\perp}, k(x_i, \cdot)\rangle_{\mathcal{H}_k} = f^{\perp}(x_i)$, for all $i = 1, \ldots, n$. Therefore,

$$f(x_j) = \sum_{i=1}^n \alpha_i k(x_i, x_j) + f^{\perp}(x_j) = \sum_{i=1}^n \alpha_i k(x_i, x_j).$$

Moreover,

$$\Omega\Big(\|f\|^2_{\mathcal{H}_k} \Big) = \Omega\Big(\left\|f^{\parallel}\right\|^2_{\mathcal{H}_k} + \left\|f^{\perp}\right\|^2_{\mathcal{H}_k} \Big) \geq \Omega\Big(\left\|f^{\parallel}\right\|^2_{\mathcal{H}_k} \Big)$$

Kernelization of algorithms

Which learning methods can be used with kernels ?

- Any regularized empirical risk minimization problem of the form,

$$f^* = \mathop{\arg\min}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\, \Omega\Big(\|f\|^2_{\mathcal{H}_k} \Big).$$

- Any method which can be formulated only using inner products (usually inner product in $\mathbb{R}^d$)

Replace inner product with kernel ! (or equivalently)

- Use the representer theorem - final function: $f(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$.
- Use the representer theorem - regularizer:

$$\|f\|^2_{\mathcal{H}_k} = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j).$$

Kernelization of algorithms II

- **Optimization point of view:** Transformation of any regularized empirical risk minimization problem of the form,

$$f^* = \operatorname*{arg\,min}_{f \in \mathcal{H}_k} \frac{1}{n}\sum_{i=1}^n L\big(y_i, f(x_i)\big) + \lambda\,\Omega\Big(\,\|f\|^2_{\mathcal{H}_k}\,\Big)$$

$$\Downarrow$$

$$\alpha^* = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n L\Big(y_i, \sum_{j=1}^n \alpha_j k(x_j, x_i)\Big) + \lambda\,\Omega\Big(\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)\Big)$$

  and $f^*(x) = \sum_{i=1}^n \alpha_i^* k(x_i, x)$.
- **Geometric point of view:**
  - map data to high-dimensional feature space: $\phi : \mathcal{X} \to \mathcal{H}_k$
  - apply linear algorithm in $\mathcal{H}_k$. Equivalently: Replace inner product with kernel function,

$$\langle x, y\rangle_{\mathbb{R}^d} \quad \Longrightarrow \quad k(x,y) = \langle \Phi_x, \Phi_y\rangle_{\mathcal{H}_k}.$$

Kernel Ridge Regression

Kernel Ridge Regression: Ridge Regression over $\mathcal{H}_k$,

$$f^* = \operatorname*{arg\,min}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \left\| f \right\|_{\mathcal{H}_k}^2 .$$

Representer Theorem: we do not have to optimize over whole $\mathcal{H}_k$

$$\alpha^* = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \sum_{j=1}^{n} \alpha_j k(x_j, x_i)\big)^2 + \lambda \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j),$$

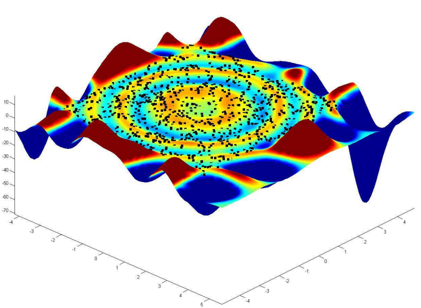
$\Rightarrow$ final function is given as $f^* = \sum_{j=1}^{n} \alpha_j^* k(x_j, \cdot)$.

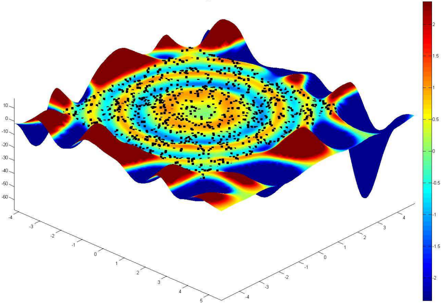
Derivation of optimal α^*:

$$\operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \left\| Y - K\alpha \right\|_2^2 + \lambda \left\langle \alpha, K\alpha \right\rangle$$

has solution $\alpha^* = (K^T K + n\lambda K)^{-1} K^T Y$ which can be written as $\alpha^* = (K + n\lambda \mathbb{1})^{-1} Y$ if K has full rank.

Example: Ridge versus Kernel ridge regression





- input: unif. on $[-\frac{7}{2}, \frac{7}{2}]^2$, output: $Y = \sin(\|X\|^2) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \frac{4}{100})$
- regularization parameter λ chosen by optimizing on test set,
- MSE for ridge regression was 0.121 and for kernel ridge regression 0.109,
- basis functions: $\phi_i(x) = e^{-\|x-x_i\|^2}$ and the Gaussian kernel, $\Longrightarrow$ solutions f^* have the expansion: $f^*(x) = \sum_{i=1}^n \alpha_i e^{-\|x-x_i\|^2}$,

Kernel Least Squares

Kernel Least Squares: Hypothesis space $\mathcal{H}_k$,

$$f^* = \operatorname*{arg\,min}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2.$$

Representer Theorem does not hold ! But any solution has the form

$$f^* = \sum_{j=1}^{n} \alpha_j^* k(x_j, \cdot) + f^{\perp}, \qquad \text{where } f^{\perp} \in \operatorname{span}\{k(x_i, \cdot) \,|\, i = 1, \ldots, n\}^{\perp},$$

Reminder: $f^{\perp}(x_i) = 0$, $i = 1, \ldots, n$, and

$$\alpha^* = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \sum_{j=1}^{n} \alpha_j k(x_j, x_i)\big)^2,$$

which can be computed via

$$K^T K \alpha^* = K^T y,$$

or $K\alpha^* = y$ if K has full rank.

Example: Support-Vector-Machine I

The soft margin SVM is formulated using **slack variables** $\xi_i \geq 0$.

$$\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R},\, \xi\in\mathbb{R}^n} \frac{1}{2}\left\|w\right\|_2^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$\text{subject to: } y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \quad \forall\, i = 1,\ldots,n, \qquad \xi_i \geq 0,$$

- the geometric margin is given by $\frac{2}{\left\|w\right\|_2}$,
- maximizing the margin corresponds to minimizing $\left\|w\right\|_2$,
- slack variables allow points to get inside the margin - soft margin

SVM = RERM with Hinge loss and squared regularizer:

$$\min_{w\in\mathbb{R}^d,\ b\in\mathbb{R}} C\,\frac{1}{n}\sum_{i=1}^{n}\max\big(0, 1-y_i(\langle w, x_i\rangle + b)\big) + \|w\|_2^2\,,$$

- error parameter C is inverse to the regularization parameter $\lambda = \frac{1}{C}$.

Dual problem:

$$\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle\,,$$

$$\text{subject to: } 0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1,\ldots,n, \quad \sum_{i=1}^{n} y_i\alpha_i = 0.$$

SVM with kernels

SVM = RERM with Hinge loss and squared regularizer:

$$\min_{f \in \mathcal{H}_k,\, b \in \mathbb{R}} C \,\frac{1}{n} \sum_{i=1}^{n} \max\big(0, 1 - y_i(f(x_i) + b)\big) + \|f\|_{\mathcal{H}_k}^2 ,$$

becomes with the representer theorem,

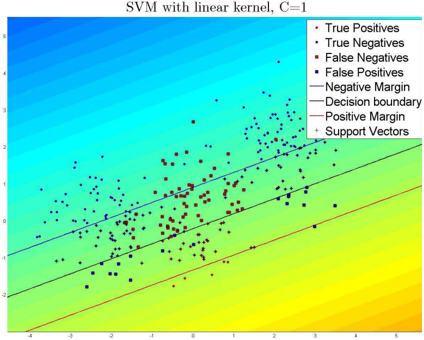
$$\min_{\alpha \in \mathbb{R}^n,\, b \in \mathbb{R}} C \,\frac{1}{n} \sum_{i=1}^{n} \max\Big(0, 1 - y_i\Big(\sum_{j=1}^{n} \alpha_j k(x_j, x_i) + b\Big)\Big) + \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j),$$

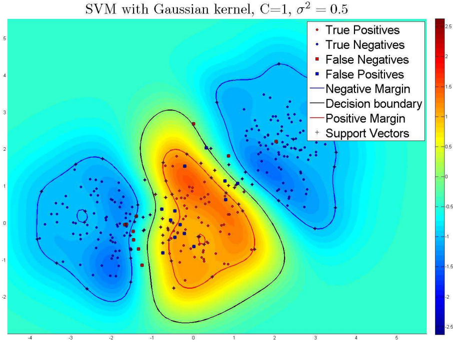
The dual problem:

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, k(x_i, x_j),$$

$$\text{subject to: } 0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n} y_i \alpha_i = 0.$$

Demo: Support-Vector-Machine with kernels II





Left: the result of the linear SVM with error parameter C - clearly no linear hyperplane can solve this problem. **Right:** the result of the SVM with a Gaussian kernel with $\sigma^2 = \frac{1}{2}$ and $C = 1$. We observe that the Gaussian kernel can nicely identify the class structure.

Replace inner products with kernels:

- any linear method can be kernelized,
- often the dual formulation is more easily accessible and better suited for optimization,
- Kernel Logistic Regression, Kernel Fisher Discriminant Analysis, Kernel PCA, Kernel Perceptron, ...

What is the purpose of regularization ?

- penalize functions which are not smooth and penalize slowly varying functions less.
- regularization functional should measure complexity of the function.

How can we measure smoothness of a function ?

- penalize the derivatives of a function e.g. $\Omega(f) = \int_{\mathbb{R}^d} \|\nabla f\|_2^2 \, dx$.
- How can we achieve that using a RKHS ? Can we see directly from a kernel what kind of regularization functional it induces ?

Translation invariant kernels in $\mathbb{R}^d$

$$k(x, y) = k(x - y).$$

What does translation invariant mean ?

- translating all feature vectors by a constant vector $c \in \mathbb{R}^d$, $x \mapsto x + c$, does not change the kernel.
- $k\big(x + c, y + c\big) = k\big((x + c) - (y + c)\big) = k\big(x + c - y - c\big) = k\big(x - y\big) = k\big(x, y\big).$

Why translation invariant ?

- use if only **relative** properties of the features are important, but not **absolute** ones.

Fourier transform

Fourier transform in $\mathbb{R}$ $\left(e^{-i\,x\,\omega} = \cos(\omega\,x) - i\sin(\omega\,x)\right)$:

$$f(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x)\, e^{-i\,x\omega}\, dx.$$

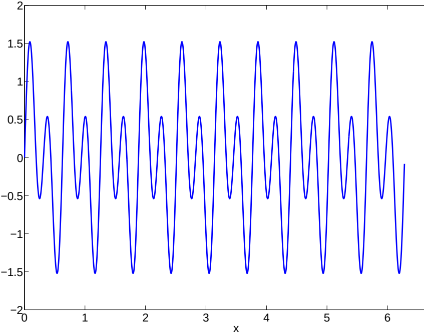
Inverse Fourier transform:

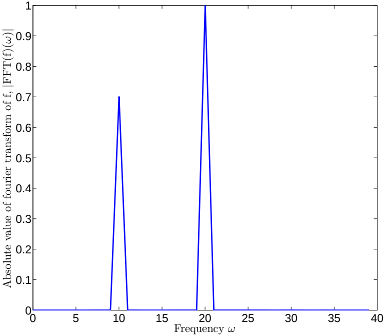
$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(\omega)\, e^{i\,x\omega}\, d\omega.$$

What is the interpretation of the Fourier transform ?

- decomposition of $f(x)$ into its **oscillation components or harmonics**,
- $f(\omega)$ is called the **spectrum of f**, $f(\omega)$ is complex, $f(\omega) = A(\omega)e^{i\phi(\omega)}$ ($A(\omega)$: **amplitude**, $\phi(\omega)$: **phase**).
- $|f(\omega)|^2$ is called the **power spectrum** of f.

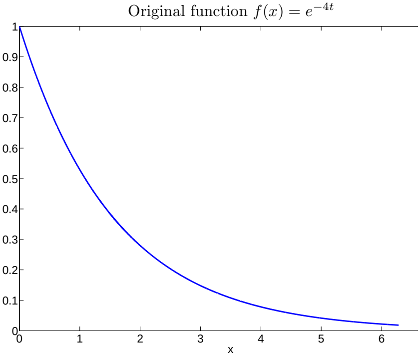
Fourier transform: Mixture of sinusoids

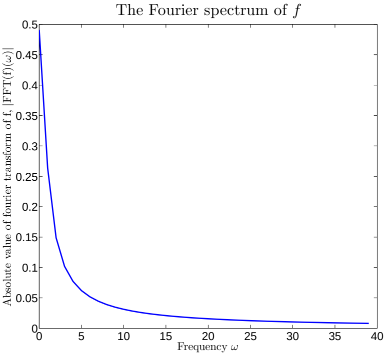




Left: The original function $f(x) = \frac{7}{10}\sin(10x) + \cos(20x)$, **Right:** The amplitude spectrum of its fourier transform.

Fourier transform of the exponential





Left: The original function $f(x) = \exp(-4t)$, **Right:** The amplitude spectrum of its fourier transform.

Fourier transform II

Derivative becomes multiplication in the Fourier domain:

$$\frac{d}{dx} f(x) = \frac{d}{dx}\left(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(\omega)\, e^{i\,x\omega}\, d\omega\right) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} i\,\omega\, f(\omega)\, e^{i\,x\omega}\, d\omega.$$

Fourier transform of $\frac{d}{dx} f$ is $i\omega f(\omega)$ (multiplication in Fourier domain).

General k-th derivatives:

$$\frac{d^k}{dx^k} f(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} (i\omega)^k\, f(\omega)\, e^{ix\omega}\, d\omega.$$

Moreover, we have **Plancherel's theorem:**

$$\int_{\mathbb{R}} |f(x)|^2 dx = \int_{\mathbb{R}} |f(\omega)|^2 d\omega.$$

Thus, $\int_{\mathbb{R}} |\frac{d}{dx} f|^2 = \int_{\mathbb{R}} |\omega|^2 |f(\omega)|^2 d\omega$.

Bochner's theorem: A real-valued function $k(x-y)$ is positive definite if and only if it is the Fourier transform of a symmetric, positive function $v(\omega)$.

$$k(x-y) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(\omega)\, e^{-i\,(x-y)\omega}\, d\omega.$$

$\Longrightarrow$ Important theorem for building kernels in $\mathbb{R}^d$.

One direction is easy:

$$\begin{aligned}
\sum_{r,s=1}^{m} c_i c_j k(x_r - x_s) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(\omega) \sum_{r,s=1}^{m} c_r\, c_s\, e^{-i\,(x_r - x_s)\omega}\, d\omega \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(\omega) \sum_{r=1}^{m} c_r e^{-i x_r \omega} \sum_{s=1}^{m} c_s e^{i x_s \omega}\, d\omega \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(\omega) \Big| \sum_{r=1}^{m} c_r e^{i x_r \omega} \Big|^2 d\omega \geq 0.
\end{aligned}$$

RKHS norm of this kernel: One can show that

$$\|f\|^2_{\mathcal{H}_k} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{|f(\omega)|^2}{v(\omega)}, d\omega.$$

Example: Using the previous results we get for $v(\omega) = \sqrt{\frac{2}{\pi}\frac{1}{1+\omega^2}}$:

$$\|f\|^2_{\mathcal{H}_k} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |f(\omega)|^2 \sqrt{\frac{2}{\pi}}(1+\omega^2)\, d\omega = \frac{1}{\pi}\int_{\mathbb{R}} f(x)^2 + \Big(\frac{df}{dx}\Big)^2 dx.$$

Associated kernel function:

$$k(x-y) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-i\,(x-y)\omega}\, \sqrt{\frac{2}{\pi}}\frac{1}{1+\omega^2} d\omega = e^{-|x-y|}.$$

This is the so called Laplace kernel.

Fourier transform V

Interpretation of the norm in the frequency domain:

- functions which have high-frequency components are heavily penalized,
- saturation at zero - extremely low frequency components (constant functions) are also penalized.

Generalization to higher order: Let $v(\omega) = \frac{1}{\sum_{j=0}^{s} \alpha_j \, \omega^{2j}}$. Then the induced norm is given by

$$\|f\|_{\mathcal{H}_k}^2 = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \sum_{j=0}^{s} \alpha_j \left(\frac{d^j f}{dx^j}\right)^2 dx.$$

Penalization of all derivatives:

The **Gaussian kernel**

$$k(x-y) = \exp\Big(-\frac{(x-y)^2}{2\sigma^2}\Big)$$

has the Fourier-transform

$$v(w) = \sigma \exp\Big(-\sigma^2\omega^2/2\Big).$$

Thus we can argue (the rigorous mathematics is quite tricky)

$$\|f\|^2_{\mathcal{H}_k} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \sum_{j=0}^{\infty} \frac{\sigma^{2j}}{j!2^j} \Big(\frac{d^j f}{dx^j}\Big)^2 dx.$$

A **translation and rotation invariant kernel** has the form

$$k(x, y) = \phi(\|x - y\|^2).$$

Such kernels are called **radial**.

What means rotational invariance ?

Let R be an orthogonal matrix, that is $RR^T = R^T R = \mathbb{1}$, then

$$\begin{aligned} k(Rx, Ry) &= \phi\big(\, \|Rx - Ry\|^2 \,\big) = \phi\big(\, \langle R(x - y), R(x - y)\rangle \,\big) \\ &= \phi\Big(\, \Big\langle (x - y), R^T R(x - y) \Big\rangle \,\Big) = \phi\big(\, \langle x - y, x - y\rangle \,\big) = \phi\big(\, \|x - y\| \\ &= k(x, y). \end{aligned}$$

Applying a rotation on the whole space does not change the kernel.

Translation and rotation invariant kernels II

Theorem of Schoenberg for radial kernels $k(x, y) = \phi(\|x - y\|)$

A continuous function $\phi : [0, \infty) \to \mathbb{R}$ is positive definite on $\mathbb{R}^d$ if and only if it is the **Bessel transform** of a finite, nonnegative measure μ on $[0, \infty)$.

$$\phi(r) = \int_0^\infty \Omega_d(rt)\, d\mu(t),$$

where

$$\Omega_d(r) = \begin{cases} \cos(r) & d = 1 \\ \Gamma(\frac{d}{2})\left(\frac{2}{r}\right)^{(d-2)/2} J_{(d-2)/2}(r) & d \geq 2 \end{cases}$$

and J_d is the Bessel function of first kind.

- property of being positive definite depends on the dimension d.
- there exists **no radial kernels of compact support**, that means $\phi(\|x - y\|) = 0$ if $\|x - y\| \geq r$, for **any** dimension.

Translation and rotation invariant kernels III

Why is that interesting ?

- Your desired kernel might not be a kernel for the number of features d you are using $\rightarrow$ representation theorem for radial kernels valid for all dimensions exists.
- We cannot hope for general purpose (good for all dimensions) sparse methods with kernels.
  $\Rightarrow$ complexity of kernel methods is often $O(n^3)$
- sparse radial kernels are rare.

Standard radial kernels:

$$\begin{aligned} \text{Gaussian kernel:} \qquad & k(x,y) = \exp\Big(-\frac{\|x-y\|^2}{2\sigma^2}\Big), \\ \text{Laplace kernel:} \qquad & k(x,y) = \exp\Big(-\lambda\,\|x-y\|\Big). \end{aligned}$$

Embedding induced by the kernel

Function space induced by the kernel:

- Kernel function $k(x, \cdot)$ for all $x \in \mathcal{X}$,
- RKHS = Hilbert space of functions,
- For $\mathcal{X} = \mathbb{R}^d$: $\|f\|_{\mathcal{H}_k}$ measures smoothness of f.

Vector space point of view:

- embedding $\Phi : \mathcal{X} \to \mathcal{H}$, with $x \to \Phi(x) \in \mathcal{H}$ of the input space into a Hilbert space = **feature space**,
- the embedding is not unique given a kernel but there exists one and they are all isometric isomorphic, the easiest is

$$\Phi : x \to k(x, \cdot) \in \text{ RKHS } \mathcal{H}_k.$$

- kernel as inner product of embedded vectors: $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$,
- $w = \sum_{i=1}^n \alpha_i \Phi(x_i)$ is a vector in the $\mathcal{H}$.

Vector space structure in $\mathcal{H}$

Norm of a feature vector $\Phi(x)$

$$\|\Phi(x)\|_{\mathcal{H}} = \sqrt{\|\Phi(x)\|_{\mathcal{H}}^2} = \sqrt{k(x,x)}.$$

Distance/Metric:

$$\begin{aligned} d(x,y) &= \sqrt{\|\Phi(x)-\Phi(y)\|_{\mathcal{H}}^2} = \sqrt{\|\Phi(x)\|_{\mathcal{H}}^2 + \|\Phi(y)\|_{\mathcal{H}}^2 - 2\left\langle \Phi(x), \Phi(y)\right\rangle_{\mathcal{H}}} \\ &= \sqrt{k(x,x)+k(y,y)-2k(x,y)}. \end{aligned}$$

Angle:

$$\cos\left(\angle(\Phi(x),\Phi(y))\right) = \frac{\left\langle \Phi(x), \Phi(y)\right\rangle}{\|\Phi(x)\|\,\|\Phi(y)\|} = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}.$$

Center of mass/Centroid/Mean vector:

$$\Phi_m = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i).$$

Distance of $\Phi(x)$ to the center of mass:

$$\begin{aligned}\|\Phi(x)-\Phi_m\|^2 &= \langle\Phi(x),\Phi(x)\rangle + \langle\Phi_m,\Phi_m\rangle - 2\,\langle\Phi(x),\Phi_m\rangle\\ &= k(x,x) + \frac{1}{n^2}\sum_{i,j=1}^{n}k(x_i,x_j) - \frac{2}{n}\sum_{i=1}^{n}k(x,x_i).\end{aligned}$$

Centering of datapoints in the feature space:

$$\tilde{\Phi}(x) = \Phi(x) - \Phi_m = \Phi(x) - \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i).$$

$$\left\langle\tilde{\Phi}(x),\tilde{\Phi}(z)\right\rangle = k(x,z) - \frac{1}{n}\sum_{i=1}^{n}k(x,x_i) - \frac{1}{n}\sum_{i=1}^{n}k(z,x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n}k(x_i,x_j).$$

Simple classification method

Center of mass/Centroid/Mean vector:

$$\Phi_m^+ = \frac{1}{n_+}\sum_{i=1}^{n_+}\Phi(x_i), \quad \Phi_m^- = \frac{1}{n_-}\sum_{j=1}^{n_-}\Phi(z_j).$$

Classify by assigning point to class of closest centroid:

$$\begin{aligned} f(x) &= \mathrm{sign}\left(\left\|\Phi(x) - \Phi_m^-\right\|^2 - \left\|\Phi(x) - \Phi_m^+\right\|^2 \right) \\ &= \mathrm{sign}\left(\frac{1}{n_+}\sum_{i=1}^{n_+} k(x, x_i) - \frac{1}{n_-}\sum_{j=1}^{n_-} k(x, z_j) + b\right) \end{aligned}$$

where

$$b = \frac{1}{n_-^2}\sum_{i,j=1}^{n_-} k(z_i, z_j) - \frac{1}{n_+^2}\sum_{i,j=1}^{n_+} k(x_i, x_j).$$

This is a so called **Parzen window classifier** (with offset).

Kernels on structured domains

Kernels can be defined on arbitrary sets !

Not any positive definite kernel is useful !

$$k(x,y) = c, \quad c \geq 0, \quad \forall\, x, y \in \mathcal{X},$$

$$k(x,y) = \begin{cases} 1 & \text{if} \quad x = y \\ 0 & \text{else} \end{cases}.$$

$\Longrightarrow$ no generalization possible.

Kernels on structured domains II

How we should we construct kernels (on structured domains) ?

- the kernel function $k(x, y)$ should be a natural similarity measure. In particular, objects

$$\text{for all } y \sim x \text{ then } k(x, y) \geq k(x, z) \text{ where } z \nsim x.$$

- distance function $d(x, y)$ induced by the kernel should be a natural dissimilarity measure.
- the evaluation of the kernel function should include less computations than an explicit feature mapping.

General scheme: compare objects by comparing substructures !

Kernels on sets

Application scenario:
each object is described by a set of features where the cardinality of the set can differ between objects.

Prominent examples:

- **computer vision:** extract features (image patches, gradients, histograms,...) at interesting points (variation of location and scale). Then the image is summarized by the set of extracted features.
- **natural language processing:** neglecting semantic information a text document simply consists of a set of words or sentences.

Kernels on sets II

Two approaches:

- directly compare two sets using a kernel defined on the components of the sets,
- count the number of occurrences of elements and compare the counts

bag-of-words representation

Kernels on sets III

Reminder: $2^{\mathcal{X}}$ is the powerset of $\mathcal{X}$, the set of all finite subsets of $\mathcal{X}$.

Proposition

Let $\mathcal{X}$ be a set and $k' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive definite kernel on $\mathcal{X}$, then a kernel on finite subsets of $\mathcal{X}$, the **set kernel**, *$k : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \mathbb{R}$, is given by*

$$\forall\, A, B \in 2^{\mathcal{X}}, \qquad k(A, B) = \sum_{a \in A} \sum_{b \in B} k'(a, b).$$

Proof: Let $\Phi : \mathcal{X} \to \mathcal{H}_{k'}$ be the feature mapping associated to the kernel k'. Then using the linear mapping $\Phi_{2^{\mathcal{X}}} : 2^{\mathcal{X}} \to \mathcal{H}_{k'}$ defined as $A \to \Phi_{2^{\mathcal{X}}}(A) = \sum_{a \in A} k'(a, \cdot)$ we get

$$\begin{aligned}
\langle \Phi_{2^{\mathcal{X}}}(A), \Phi_{2^{\mathcal{X}}}(B) \rangle_{\mathcal{H}_{k'}} &= \left\langle \sum_{a \in A} k'(a, \cdot), \sum_{b \in B} k'(b, \cdot) \right\rangle_{\mathcal{H}_{k'}} \\
&= \sum_{a \in A} \sum_{b \in B} \langle k'(a, \cdot), k'(b, \cdot) \rangle_{\mathcal{H}_{k'}} = \sum_{a \in A} \sum_{b \in B} k'(a, b) = k(A, B).
\end{aligned}$$

The set kernel:

- adds up all similarities between elements of the sets.
- problems if cardinality varies very much $\Longrightarrow$ sets with large number of elements will be similar to every other set $\Longrightarrow$ normalization necessary,

$$\tilde{k}(A,B) := \frac{k(A,B)}{\sqrt{k(A,A)k(B,B)}} = \frac{\sum_{a\in A}\sum_{b\in B} k'(a,b)}{\sqrt{\sum_{a,a'\in A} k'(a,a')\sum_{b,b'\in B} k(b,b')}},$$

or

$$\tilde{k}(A,B) := \frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} k'(a,b),$$

- **Advantage:** two disjoint sets A and B ($A \cap B = \emptyset$) can have a non-zero similarity value,
- the set kernel can be used for arbitrary sets not only subsets of $\mathcal{X}$.

Invariances via sets:

- classifier should be invariant under small transformations of the data (small rotations/translations in the case of handwritten digit recognition.
- add to each training object all its small transformations

  new object = old object + all transformations (set of objects)
- apply set kernel to this set.

Kernels on sets VI

A simple set kernel not taking into account any structure of $\mathcal{X}$:

Proposition

Let $\mathcal{X}$ be some set. Then a kernel on finite subsets of $\mathcal{X}$, the **intersection kernel**, $k : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \mathbb{R}$, *is given by*

$$\forall\, A, B \in 2^{\mathcal{X}}, \qquad k(A, B) = |A \cap B|.$$

Proof: One can show that $\min\{x, y\}$ is a kernel on $\mathbb{R}_+$. For a finite set $\mathcal{X}$ one has

$$|A \cap B| = \sum_{x \in \mathcal{X}} \min\{A(x), B(x)\},$$

where $A(x)$ denotes the number of elements of type x in the set A. This finishes the proof since we add up valid kernels and the index set of the sum is **fixed**.

Taking into account both aspects ($M(\mathcal{X})$ denotes arbitrary sets consisting of elements in $\mathcal{X}$):

Proposition

Let $\mathcal{X}$ be a finite set and

- $k' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *a positive definite kernel on* $\mathcal{X}$,
- $\overline{k} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ *a positive definite kernel on* $\mathbb{R}_+$.

Then the **general set kernel** *between arbitrary sets consisting of elements in $\mathcal{X}$, $k : M(\mathcal{X}) \times M(\mathcal{X}) \to \mathbb{R}$, is given by*

$$k(A, B) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} k'(x, y)\overline{k}(A(x), B(y)),$$

where $A(x)$ is the number of times the element x is contained in set A.

Kernels on sets VII

Properties of the general set kernel:

- comparison of arbitrary sets (the standard form is a histogram),
- integration of a complex weighting scheme depending on the similarity of the frequency of occurrence via $\overline{k}(A(x), B(y))$,
- integration of a given similarity measure on $\mathcal{X}$. This can be e.g. used to integrate semantic similarity when comparing texts.

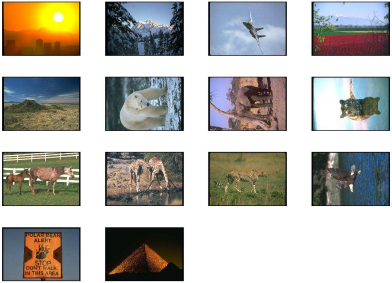
Normalization of the kernel or normalization of the counts $A(x)$ might be useful.

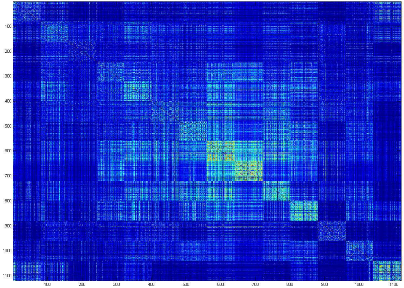
Kernels on sets: Example

Problem:

- 14 categories of images (different animals, landscapes, airplanes, mountains),
- image representation: color histogram (set of colors !)
  (each channel in RGB is quantized into 16 levels - yielding a 4096 dimensional histogram).
- bag-of-colors representation.

Kernels on sets: Example II





- good block-diagonal structure of the kernel matrix,
- 10.4% error for a 14-class problem.