A Counterexample

A Candidate for a Kernel

\[
k(x, x') = \begin{cases} 
1 & \text{if } \|x - x'\| \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

This is symmetric and gives us some information about the proximity of points, yet it is not a proper kernel . . .

Explicit Counterexample

We use three points, \(x_1 = 1, x_2 = 2, x_3 = 3\) and compute the resulting “kernel matrix” \(K\). This yields

\[
K = \begin{bmatrix} 
1 & 1 & 0 \\
1 & 1 & 1 \\
0 & 1 & 1 \\
\end{bmatrix}
\]

and

\[
\begin{bmatrix} 
\frac{1}{\sqrt{2} - 1} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & 0 \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\end{bmatrix}
\]

as eigensystem. Clearly this is not what we want since \(K\) must have nonnegative eigenvalues to be a kernel matrix. Hence \(k\) is not a kernel.

Interpretation, Part II

Integral Operator

A useful trick is to consider the integral operator \(T_k\) associated with \(k\) via

\[
T_k : L_2(\mathbb{X}) \to L_2(\mathbb{X}) \text{ where } (T_kf)(x) = \int_{\mathbb{X}} k(x, x') f(x') dx'
\]

Eigensystem of Operators

In this case Mercer’s condition reads as

\[
\int_{\mathbb{X} \times \mathbb{X}} k(x, x') f(x) f(x') dx dx' = \int_{\mathbb{X}} f(x) \langle T_k f(x) \rangle dx = \langle f, T_k f \rangle \geq 0
\]

In other words, \(T_k\) has to be an operator with nonnegative eigenvalues. There the \(\lambda_i, \phi_i(x)\) are the eigenvalues and eigenfunctions of \(T_k\).

This means that we replaced the condition that all the eigenvalues of a matrix be nonnegative by the requirement that all the eigenvalues of an operator be nonnegative.

Mercer’s Theorem

The Theorem

For any symmetric function \(k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}\) which is square integrable in \(\mathbb{X} \times \mathbb{X}\) and which satisfies

\[
\int_{\mathbb{X} \times \mathbb{X}} k(x, x') f(x) f(x') dx dx' \geq 0 \text{ for all } f \in L_2(\mathbb{X})
\]

there exist functions \(\phi_i : \mathbb{X} \to \mathbb{R}\) and numbers \(\lambda_i \geq 0\) such that

\[
k(x, x') = \sum_i \lambda_i \phi_i(x) \phi_i(x') \text{ for all } x, x' \in \mathbb{X}.
\]

Interpretation

Effectively the double integral is the continuous version of a vector-matrix-vector multiplication. Recall that for positive semidefinite matrices we had

\[
\sum_i \sum_j k(x_i, x_j) \alpha_i \alpha_j \geq 0
\]

Gaussian RBF Kernels

Radial Basis Function Kernels

The polynomial kernels so far were of the type \(\kappa(\langle x, x' \rangle)\). Quite often we would, however, prefer a kernel which depends on the distance between points. This can be achieved by

\[
k(x, x') = \kappa(x - x') \text{ such as } k(x, x') = \exp \left( -\frac{1}{2\sigma^2} \|x - x'\|^2 \right).
\]

Properties

Typically we set \(\kappa(0) = 1\). This means that for all \(x\) we have

\[
\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle = k(x, x) = \kappa(\|x - x\|) = 1.
\]

In other words, all observations are mapped onto the unit sphere in the feature space given by \(\Phi\).

As we shall see, the Fourier transform of \(\kappa\) tells us about how smooth the features that we are extracting.
When are RBF Kernels OK?

Problem
Not all RBF kernels are admissible. Recall the indicator function kernel with the negative eigenvalues in $K$.

Goal
We need a simple criterion to figure out whether some $k$ satisfies Mercer’s condition and therefore corresponds to a dot product in some feature space.

Idea
Maybe, applying the Fourier transformation to the integral condition will help. In the RBF case $k(x, x') = \kappa(x - x')$ and therefore Mercer’s condition reads as

$$\int_{\mathcal{X} \times \mathcal{X}} k(x, x') f(x) f(x') dx dx'$$

This looks like a dot product and a convolution with $\kappa$ . . .

Fourier Transform and Fourier Plancherel

Fourier Transform
For a square integrable function $f : \mathbb{R}^n \to \mathbb{R}$ the Fourier Transform $\tilde{f}$ is given by

$$\tilde{f}(\omega) := (2\pi)^{-\frac{n}{2}} \int \exp(-i(\omega, x)) f(x) dx.$$ 

Fourier Plancherel
The power in the time domain and in frequency domain are the same. More formally this means

$$||f||^2 = \int |f(x)|^2 dx = \int |\tilde{f}(\omega)|^2 d\omega = ||\tilde{f}||^2$$

However, due to the polarization inequality this also holds for dot products between functions, i.e.

$$\langle f, g \rangle = \langle \tilde{f}, \tilde{g} \rangle$$.

Convolutions

Definition
The convolution of two functions $f, g : \mathcal{X} \to \mathbb{R}$ is defined as

$$(f \star g)(x) := \int_{\mathcal{X}} f(x') g(x - x') dx'$$

Symmetry

$$f \star g := \int_{\mathcal{X}} f(x') g(x - x') dx = \int_{\mathcal{X}} f(x - \tau) g(\tau) d\tau = g \ast f$$

Here we used the variable substitution $\tau = x - x'$.

Convolutions and Fourier Transform
The Fourier transform of a convolution is the product of the Fourier transforms of the arguments and vice versa, i.e.

$$f \star g = (2\pi)^{\frac{n}{2}} \tilde{f} \cdot \tilde{g}$$

Recall linear filters where the final signal was a convolution in time domain and a multiplication in frequency domain.

Proof of Convolution Property

Time to Frequency
We ignore all integrability considerations (or divergence thereof) and simply write out the equations (don’t do that at home).

$$f \star g = (2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \exp(-i(\omega, x)) \int_{\mathbb{R}^n} f(x') g(x - x') dx' dx$$

$$= (2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \exp(-i(\omega, x')) \exp(-i(\omega, x')) f(x') g(x - x') dx' dx'$$

$$= (2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \exp(-i(\omega, \tau)) g(\tau) \exp(-i(\omega, x)) f(x') dx' d\tau$$

$$= (2\pi)^{\frac{n}{2}} \int_{\mathbb{R}^n} \left(2\pi\right)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \exp(-i(\omega, \tau)) g(\tau) d\tau \left(2\pi\right)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \exp(-i(\omega, x)) f(x') dx'$$

$$= (2\pi)^{\frac{n}{2}} \tilde{f} \cdot \tilde{g}$$

Time to Frequency
The same reasoning as above, again we have to swap the order of integration.
Proof for RBF Kernels

Rewriting Mercer’s Condition

\[
\begin{align*}
\int_X \int_X f(x')k(x - x')f(x)dx dx' \\
= \int_X (f \circ k)(x)f(x)dx \\
= (f \circ k, f) \\
= (2\pi)^2 \int_X f(x)k(x)dx \\
= (2\pi)^2 \int_X \hat{f}(\omega)\hat{k}(\omega)d\omega
\end{align*}
\]

Positivity Condition

The integral is exactly then always nonnegative if

\[\hat{k}(\omega) \geq 0 \text{ for all } \omega \in \mathcal{X}\]

This means that Mercer’s condition is easy to check — simply compute \(\hat{k}(\omega)\) and check its sign.

Examples

Gaussian Kernels

Now we finally can check whether \(k(x, x') = \exp \left(-\frac{1}{2\sigma^2}||x - x'||\right)\) is positive semidefinite. We know that the Fourier transform of a Gaussian is a Gaussian, hence never negative. That’s sufficient.

Laplacian Kernel

For the kernel \(k(x, x') = \exp(-||x - x'||)\) things are a bit trickier, since there the Fourier transform depends on the dimensionality of \(\mathcal{X}\). For \(\mathcal{X} = \mathbb{R}\) we have

\[
\begin{align*}
\tilde{k}(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-|x|} e^{-i\omega x} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-(1+i\omega)x} + e^{-(1-i\omega)x} dx \\
&= \frac{1}{\sqrt{2\pi}} \left( \frac{1}{1 + i\omega} + \frac{1}{1 - i\omega} \right) = \sqrt{\frac{2}{\pi}} \frac{1}{1 + \omega^2} \geq 0
\end{align*}
\]

Linear Regression in Feature Space

Regression Problem

- Patterns \(x_1, \ldots, x_m\) together with target values \(y_1, \ldots, y_m\).
- Quadratic loss function \(c(x, y, f(x)) = \frac{1}{2}(y - f(x))^2\).
- Linear model in feature space \(f(x) = \langle w, \Phi(x) \rangle\), hence \(\Phi\) and kernel \(k\).
- Quadratic regularizer of the form \(\Omega(f) = \frac{1}{2}\|w\|^2\).
- Regularization constant \(\lambda\).

Goal

Minimize the regularized risk functional

\[R_{reg}[f] = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}(y_i - f(x_i))^2 + \frac{1}{2}\|w\|^2\]
Solving the Expansion

It follows that $\alpha_i$ is given by

$$\alpha_i = \frac{1}{m\lambda} (y_i - f(x_i)) = \frac{1}{m\lambda} \left( y_i - \sum_{j=1}^{m} \alpha_j k(x_j, x_i) \right)$$

In vector notation this reads as

$$\alpha = \frac{1}{\lambda m} (y - K\alpha)$$

and therefore $\alpha = (K + \lambda m 1)^{-1} y$

Interpretation

This equation resembles the one obtained in the linear case, only that now we replaced $XX^T$, the outer product between the observations with the kernel matrix.

Important Observation

This estimator is one of the currently best regression estimators available. In doubt, use it rather than Neural Networks.