A Counterexample

A Candidate for a Kernel

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

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 "kernelmatrix" K. This yields

$$K = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \text{ and } \left(\frac{1}{\sqrt{2}-1}, \begin{bmatrix} \frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{2} \end{bmatrix} \right), \left(1, \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix} \right), \left(1-\sqrt{2}, \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{2} \end{bmatrix} \right)$$

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.

Mercer's Theorem

The Theorem

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

$$\int_{\mathcal{X} \times \mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0 \text{ for all } f \in L_2(\mathcal{X})$$

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

$$k(\mathbf{x}, \mathbf{x}') = \sum_i \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}') \text{ for all } \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

Interpretation

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

$$\sum_i \sum_j k(\mathbf{x}_i, \mathbf{x}_j) \alpha_i \alpha_j \geq 0$$

Interpretation, Part II

THE AUSTRALIAN NATIONAL UNIVERSITY

Integral Operator

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

$$T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X}) \text{ where } (T_k f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') d\mathbf{x}'$$

Eigensystem of Operators

In this case Mercer's condition reads as

$$\int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' = \int_{\mathcal{X}} f(\mathbf{x})(T_k f)(\mathbf{x}) d\mathbf{x} = \langle f, T_k f \rangle \geq 0$$

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

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

Gaussian RBF Kernels

Radial Basis Function Kernels

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

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

Properties

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

$$\|\Phi(\mathbf{x})\|^2 = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle = k(\mathbf{x}, \mathbf{x}) = \kappa(\|\mathbf{x} - \mathbf{x}\|) = 1.$$

In other words, all observations are mapped onto the **unit sphere** in the feature space given by Φ.

As we shall see, the Fourier transform of κ tells us about how smooth the features that we are extracting.

When are RBF Kernels OK?

THE AUSTRALIAN NATIONAL UNIVERSITY

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 $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ and therefore Mercer's condition reads as

$$\int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x})\kappa(\mathbf{x} - \mathbf{x}')f(\mathbf{x}')d\mathbf{x}d\mathbf{x}'$$

This looks like a dot product and a convolution with κ ...

Fourier Transform and Fourier Plancherel

THE AUSTRALIAN NATIONAL UNIVERSITY

Fourier Transform

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

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

Fourier Plancherel

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

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

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

$$(f \circ g)(\mathbf{x}) := \int_{\mathcal{X}} f(\mathbf{x}')g(\mathbf{x} - \mathbf{x}')d\mathbf{x}'$$

Symmetry

$$f \circ g := \int_{\mathcal{X}} f(\mathbf{x}')g(\mathbf{x} - \mathbf{x}')d\mathbf{x} = \int_{\mathcal{X}} f(\mathbf{x} - \tau)g(\tau)d\tau = g \circ f$$

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

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.

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

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).

$$
\begin{aligned}
& f \,\tilde{\circ}\, g \\
= \; & (2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \exp(-i\langle \omega, \mathbf{x} \rangle) \int_{\mathbb{R}^n} f(\mathbf{x}') g(\mathbf{x} - \mathbf{x}') d\mathbf{x}' d\mathbf{x} \\
= \; & (2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \exp(-i\langle \omega, \mathbf{x} - \mathbf{x}' \rangle) \exp(-i\langle \omega, \mathbf{x}' \rangle) f(\mathbf{x}') g(\mathbf{x} - \mathbf{x}') d\mathbf{x} d\mathbf{x}' \\
= \; & (2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \exp(-i\langle \omega, \tau \rangle) g(\tau) \exp(-i\langle \omega, \mathbf{x} \rangle) f(\mathbf{x}') d\tau d\mathbf{x}' \\
= \; & (2\pi)^{\frac{n}{2}} \left((2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \exp(-i\langle \omega, \tau \rangle) g(\tau) d\tau \right) \left((2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} \exp(-i\langle \omega, \mathbf{x} \rangle) f(\mathbf{x}') d\mathbf{x}' \right) \\
= \; & (2\pi)^{\frac{n}{2}} \tilde{f} \cdot \tilde{g}
\end{aligned}
$$

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

$$\begin{aligned}
&\int_{\mathcal{X}} \int_{\mathcal{X}} f(\mathbf{x}')k(\mathbf{x}-\mathbf{x}')f(\mathbf{x})d\mathbf{x}d\mathbf{x}' \\
= &\int_{\mathcal{X}} (f \circ k)(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\
= &\langle (f \circ k), f \rangle \\
= &(2\pi)^{\frac{n}{2}} \langle \tilde{f} \cdot \tilde{k}, \tilde{f} \rangle \\
= &(2\pi)^{\frac{n}{2}} \int_{\mathcal{X}} |\tilde{f}(\omega)|^2 \tilde{k}(\omega) d\omega
\end{aligned}$$

Positivity Condition

The integral is exactly then always nonnegative if

$$\tilde{k}(\omega) \geq 0 \text{ for all } \omega \in \mathcal{X}$$

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

Examples

Gaussian Kernels

Now we finally can check whether $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}'\|\right)$ 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 $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|)$ things are a bit trickier, since there the Fourier transform depends on the dimensionality of $\mathcal{X}$. For $\mathcal{X} = \mathbb{R}$ we have

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

Linear Regression in Feature Space

THE AUSTRALIAN NATIONAL UNIVERSITY

Regression Problem

- Patterns $\mathbf{x}_1, \ldots, \mathbf{x}_m$ together with target values $y_1, \ldots, y_m$.
- Quadratic loss function $c(\mathbf{x}, y, f(\mathbf{x})) = \frac{1}{2}(y - f(\mathbf{x}))^2$.
- Linear model in feature space $f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle$, hence Φ and kernel k.
- Quadratic regularizer of the form $\Omega[f] = \frac{1}{2}\|\mathbf{w}\|^2$.
- Regularization constant λ.

Goal

Minimize the regularized risk functional

$$R_{\text{reg}}[f] = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}(y_i - f(\mathbf{x}_i))^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Linear Regression in Feature Space, II

THE AUSTRALIAN NATIONAL UNIVERSITY

Regularized Risk

$$R_{\text{reg}}[f] = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(y_i - \langle \mathbf{w}, \Phi(\mathbf{x})\rangle)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

We compute the derivative with respect to $\mathbf{w}$. For optimality we need

$$\partial_{\mathbf{w}} R_{\text{reg}}[f] = \frac{1}{m}\sum_{i=1}^{m}(f(\mathbf{x}_i) - y_i)\Phi(\mathbf{x}_i) + \lambda\mathbf{w} = 0$$

Kernel Expansion

The above equation shows that $\mathbf{w}$ can be expanded in terms of $\Phi(\mathbf{x}_i)$. We obtain $\mathbf{w} = \sum_{i=1}^{m}\alpha_i\Phi(\mathbf{x}_i)$ which implies that

$$f(\mathbf{x}) = \left\langle \sum_{i=1}^{m}\alpha_i\Phi(\mathbf{x}_i), \Phi(\mathbf{x})\right\rangle = \sum_{i=1}^{m}\alpha_i\langle\Phi(\mathbf{x}_i), \Phi(\mathbf{x})\rangle = \sum_{i=1}^{m}\alpha_i k(\mathbf{x}_i, \mathbf{x}).$$

This means that f is given by a kernel expansion at the patterns $\mathbf{x}_i$.

Linear Regression in Feature Space, III

THE AUSTRALIAN NATIONAL UNIVERSITY

Solving the Expansion

It follows that α_i is given by

$$\alpha_i = \frac{1}{m\lambda}(y_i - f(\mathbf{x}_i)) = \frac{1}{m\lambda}\left(y_i - \sum_{j=1}^{m} \alpha_j k(\mathbf{x}_j, \mathbf{x}_i)\right)$$

In vector notation this reads as

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

Interpretation

This equation resembles the one obtained in the linear case, only that now we replaced $XX^\top$, 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.