The Kernel Trick

As was previously shown we don't need to know the x_i explicitly. It is enough to be able to compute $K(x_i, x_j)$ for any two vectors x_i, x_j. This means that we can extend each feature-vector x by adding more features, computed as nonlinear functions of the original features, to create a new feature-vector $\phi(x)$, as long as we can compute the kernel function $K(x_i, x_j) = \phi(x_i)'\phi(x_j)$. In fact, **it is not even required that $\phi()$ is known explicitly**, as long as it can be shown that such ϕ exists.

Example: For the two dimensional vectors $x = (x_1, x_2)$ define:

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

In this case direct computation shows:

$$\phi(x_i)'\phi(x_j) = (1 + x_{i1}x_{j1} + x_{i2}x_{j2})^2$$

Therefore, we can use the following kernel:

$$K(x_i, x_j) = (1 + x_{i1}x_{j1} + x_{i2}x_{j2})^2 = (1 + x_i'x_j)^2$$

The kernel trick is the observation that we can use kernels without explicitly using the feature vector $\phi(x)$. Intuitively, $K(x_i, x_j)$ measures the similarity between x_i and x_j.

Commonly used kernels

Polynomial kernels with degree d:

$$K(x_i, x_j) = (x_i'x_j + 1)^d$$

Exponential (radial-basis) kernels with width σ:

$$K(x_i, x_j) = e^{-|x_i - x_j|^2/(2\sigma^2)}$$

What functions are kernels

Not all functions can be used as kernels. A theorem by Mercer characterizes the functions that can be used as kernels. For the cases that we are interested in it can be stated as follows.

A function $K(x, y)$ can be used as a kernel function if and only if the following condition holds. The Gram matrix of any finite subset of vectors $x_1, \ldots, x_m$ is positive definite. The Gram matrix of the above m vectors is an $m \times m$ matrix defined as follows:

$$G = (G_{ij}), \quad G_{ij} = K(x_i, x_j)$$