The Kernel Trick

Carlos C. Rodríguez
http://omega.albany.edu:8008/

October 25, 2004

Why don't we do it in higher dimensions?

If SVMs were able to handle only linearly separable data, their usefulness would be quite limited. But someone around 1992 (correct?) had a moment of true enlightenment and realized that if instead of the euclidean inner product $< x_i, x_j >$ one fed the QP solver with a function $K(x_i, x_j)$ the boundary between the two classes would then be,

$$K(x, w) + b = 0$$

and the set of $x \in R^d$ on that boundary becomes a curved surface embedded in R^d when the function $K(x, w)$ is non-linear. I don't really know how the history went (Vladimir?), but it could have perfectly been the case that a courageous soul just fed the QP solver with $K(x, w) = \exp(-|x - w|^2)$ and waited for the picture of the classification boundary to appear on the screen and saw the first non linear boundary computed with just linear methods. Then, depending on that person's mathematical training, s/he either said aha! $K(x, w)$ is a kernel in a reproducing kernel Hilbert space or rushed to the library or to the guy next door to find out, and probably very soon after that said aha!, $K(x, w)$ is the kernel of a RKHS. After knowing about RKHS that conclusion is inescapable but that, I believe, does not diminish the importance of this discovery. No. Mathematicians did not know about the kernel trick then, and most of them don't know about the kernel trick now.

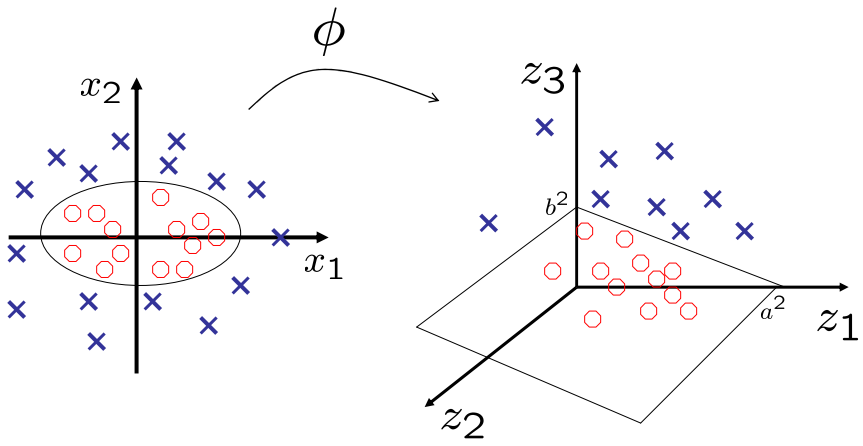
Let's go back to the function $K(x, w)$ and try to understand why, even when it is non linear in its arguments, still makes sense as a proxy for $< x, w >$. The way to think about it is to consider $K(x, w)$ to be the inner product not of the coordinate vectors x and w in R^d but of vectors $\phi(x)$ and $\phi(w)$ in higher dimensions. The map,

$$\phi : \mathcal{X} \longrightarrow \mathcal{H}$$

is called a feature map from the data space $\mathcal{X}$ into the feature space $\mathcal{H}$. The feature space is assumed to be a Hilbert space of real valued functions defined

on $\mathcal{X}$. The data space is often R^d but most of the interesting results hold when $\mathcal{X}$ is a compact Riemannian manifold.

The following picture illustrates a particularly simple example where the feature map $\phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ maps data in R^2 into R^3.



$$\phi : (x_1, x_2) \longrightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 = 1 \longrightarrow \frac{z_1}{a^2} + \frac{z_3}{b^2} = 1$$

There are many ways of mapping points in R^2 to points in R^3 but the above has the extremely useful property of allowing the computation of the inner products of feature vectors $\phi(x)$ and $\phi(w)$ in R^3 by just squaring the inner product of the data vectors x and w in R^2! (to appreciate the exclamation point just replace 3 by 10 or by infinity!) i.e., in this case

$$\begin{aligned} K(x, w) &= <\phi(x), \phi(w)> \\ &= x_1^2w_1^2 + 2x_1x_2w_1w_2 + x_2^2w_2^2 \\ &= (x_1w_1 + x_2w_2)^2 \\ &= (<x, w>)^2 \end{aligned}$$

OK. But are we really entitled to just Humpty Dumpty move up to higher, even infinite dimensions?

Sure, why not? Just remember that even the coordinate vectors x and w are just that, coordinates, arbitrary labels, that stand as proxy for the objects that they label, perhaps images or speech. What is important is the inter relation of these objects as measured, for example by the Grammian matrix of abstract inner products ($< x_i, x_j >$) and not the labels themselves with the euclidean dot product. If we think of the x_i as abstract labels (coordinate free), as pointers to the objects that they represent, then we are free to choose values for $< x_i, x_j >$ representing the similarity between x_i and x_j and that can be provided with a non linear function $K(x, w)$ producing the n^2 numbers $k_{ij} = K(x_i, x_j)$ when the abstract labels x_i are replaced with the values of the observed data. In fact, the QP solver sees the data only through the n^2 numbers k_{ij}. These could, in principle be chosen without even a kernel function and the QP solver will deliver (w, b). The kernel function becomes useful for choosing the classification boundary but even that could be empirically approximated. Now, of course, arbitrary n^2 numbers k_{ij} that disregard the observed data completely will not be of much help, unless they happen to contain cogent prior information about the problem. So, what's the point that I am trying to make? The point is that it is obvious that a choice of kernel function is an ad-hoc way of sweeping under the rug prior information into the problem, indutransductibly (!) ducking the holy Bayes Theorem. There is a kosher path that flies under the banner of *gaussian processes* and RVMs but they are not cheap. We'll look at the way of the Laplacians (formerly a.k.a. bayesians) later but let's stick with the SVMs for now.

The question is: What are the constraints on the function $K(x, w)$ so that there exists a Hilbert space $\mathcal{H}$ of abstract labels $\phi(x) \in \mathcal{H}$ such that $< \phi(x_i), \phi(x_j) >= k_{ij}$?.

The answer: It is sufficient for $K(x, w)$ to be continuous, symmetric and positive definite. i.e., for $\mathcal{X}$ either R^d or a d dimensional compact Riemannian manifold,

$$K : \mathcal{X} \times \mathcal{X} \longrightarrow R$$

satisfies,

1. $K(x, w)$ is continuous.
2. $K(x, w) = K(w, x)$. Symmetric
3. $K(x, w)$ is p.d. (positive definite). i.e., for any set $\{z_1, \ldots, z_m\} \subset \mathcal{X}$ the m by m matrix $K[z] = (K(z_i, z_j))_{ij}$ is positive definite.

Such function is said to be a *Mercer kernel*. We have,

Theorem If $K(x, w)$ is a Mercer kernel then there exists a Hilbert space $\mathcal{H}_K$ of real valued functions defined on $\mathcal{X}$ and a feature map $\phi : \mathcal{X} \longrightarrow \mathcal{H}_K$ such that,

$$< \phi(x), \phi(w) >_K = K(x, w)$$

where $<,>_K$ is the innerproduct on $\mathcal{H}_K$.

Proof: Let L be the vector space containing all the real valued functions f defined on $\mathcal{X}$ of the form,

$$f(x) = \sum_{j=1}^{m} \alpha_j K(x_j, x)$$

where m is a positive integer, the α_j's are real numbers and $\{x_1, \ldots, x_m\} \subset \mathcal{X}$. Define in L the inner product,

$$< \sum_{i=1}^{m} \alpha_i K(x_i, \cdot), \sum_{j=1}^{n} \beta_j K(w_j, \cdot) >= \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j K(x_i, w_j)$$

Since K is a Mercer kernel the above definition makes L a well defined inner product space. The Hilbert space $\mathcal{H}_K$ is obtained when we add to L the limits of all the Cauchy sequences (w.r.t. the $<,>$) in L.

Notice that the inner product in H_K was defined so that

$$< K(x, \cdot), K(w, \cdot) >_K = K(x, w).$$

We can then take the feature map to be,

$$\phi(x) = K(x, \cdot) \bullet$$

RKHS

The space H_K is said to be a Reproducing Kernel Hilbert Space (RKHS). Moreover, for all $f \in \mathcal{H}_K$

$$f(x) =< K(x, \cdot), f >_K .$$

This follows from the reproducing property when $f \in L$ and by continuity for all $f \in \mathcal{H}_K$. It is also easy to show that the reproducing property is equivalent to the continuity of the evaluation functionals $\delta_x(f) = f(x)$.

The Mercer kernel $K(x, w)$ naturally defines an integral linear operator that (abusing notation a bit) we also denote by $K : \mathcal{H}_K \longrightarrow \mathcal{H}_K$, where,

$$(Kf)(x) = \int K(x, y) f(y) dy$$

and since K is symmetric and positive definite, it is orthogonally diagonalizable (just as in the finite dimensional case). Thus, there is an ordered orthonormal basis $\{\phi_1, \phi_2, \ldots\}$ of eigen vectors of K, i.e., for $j = 1, 2, \ldots$

$$K\phi_j = \lambda_j \phi_j$$

with $< \phi_i, \phi_j >_K = \delta_{ij}$, $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ and such that,

$$K(x, w) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x) \phi_j(w)$$

from where it follows that the feature map is,

$$\phi(x) = \sum_{j=1}^{\infty} \lambda_j^{1/2} \phi_j(x) \phi_j$$

producing,

$$K(x, w) =< \phi(x), \phi(w) >_K$$

Notice also that any continuous map $\phi : \mathcal{X} \longrightarrow \mathcal{H}$ where $\mathcal{H}$ is a Hilbert space, defines $K(x, w) =< \phi(x), \phi(w) >$ which is a Mercer Kernel.