

The Representer Theorem, Kernels, and Hilbert Spaces

We will now work with infinite dimensional feature vectors and parameter vectors. The space ℓ_2 is defined to be the set of sequences f_1, f_2, f_3, \dots which have finite norm, i.e., where we have the following.

$$\|f\|^2 = \sum_{i=1}^{\infty} f_i^2 < \infty \quad (1)$$

We are now interested in regression and classification with infinite dimensional feature vectors and weight parameters. In other words we have $\Phi(x) \in \ell_2$ and $\beta \in \ell_2$. In practice there is essentially no difference between the infinite dimensional case and the finite dimensional case with $\Phi(x), \beta \in R^d$ but where $d \gg T$, i.e., where the dimension is large compared to the size of the training data.

It is possible to prove an infinite dimensional version of the Cauchy-Swartz inequality:

$$\sum_{i=1}^{\infty} f_i g_i \leq \|f\| \|g\|$$

This inequality also implies that we have the following for any two vectors f and g in ℓ_2 .

$$\sum_{i=1}^{\infty} |f_i g_i| < \infty$$

In other words sums of products of features are absolutely convergent. Absolutely convergent sums have the property that the terms in the sum can be rearranged in any order while preserving the value of the sum.

Now consider the general regularized regression equation (see the notes on regularized regression).

$$\begin{aligned} \beta^* &= \operatorname{argmin}_{\beta} \sum_{t=1}^T L(m_t(\beta)) + \lambda \|\beta\|^2 \quad (2) \\ m_t(\beta) &= y_t(\beta \cdot \Phi(x_t)) \end{aligned}$$

This regression equation is well defined in the infinite dimensional case except that β^* may have infinite norm. We will assume in these notes that β^* has finite norm.

We can minimize $\|\beta\|$ while holding all $m_t(\beta)$ constant (for all t) by removing the component of β orthogonal to all vectors $\Phi(x_t)$. Without loss of generality we can therefore assume that β^* is in the span of the vectors $\Phi(x_t)$.

$$\beta^* = \sum_{t=1}^T \alpha_t \Phi(x_t) \quad (3)$$

Equation (3) is called the representer theorem. The representer theorem yields the following.

$$\beta^* \cdot \Phi(x) = \sum_{t=1}^T \alpha_t \Phi(x_t) \cdot \Phi(x) \quad (4)$$

$$= \sum_{t=1}^T \alpha_t K(x_t, x) \quad (5)$$

$$K(x_1, x_2) = \Phi(x_1) \cdot \Phi(x_2) \quad (6)$$

Equation (6) introduces $K(x_1, x_2)$ as an abbreviation for $\Phi(x_1) \cdot \Phi(x_2)$. However, it is often possible to compute $K(x_1, x_2)$ efficiently without computing the (infinite) feature vectors $\Phi(x_1)$ or $\Phi(x_2)$. We will consider a variety of easily computed functions K . We now have the following definition.

Definition: A function K on $\mathcal{X} \times \mathcal{X}$ is called a *kernel function* if there exists a function Φ mapping \mathcal{X} into ℓ_2 such that for any $x_1, x_2 \in \mathcal{X}$ we have that $K(x_1, x_2) = \Phi(x_1) \cdot \Phi(x_2)$.

We will show below that for $x_1, x_2 \in \mathbb{R}^q$ the functions $(x_1 \cdot x_2 + 1)^p$ and $\exp(-\frac{1}{2}(x_1 - x_2)^T \Sigma^{-1}(x_1 - x_2))$ are both kernels. The first is called a polynomial kernel and the second is called a Gaussian kernel. The Gaussian kernel is particularly widely used. For the Gaussian kernel we have that $K(x_1, x_2) \leq 1$ where the equality is achieved when $x_1 = x_2$. In this case $K(x_1, x_2)$ expresses a nearness of x_1 to x_2 . When K is a Gaussian kernel equation (5) can be viewed as a classifying x using a weighted nearest neighbor rule where $K(x_t, x)$ gives the “nearness” of x_t to x .

In order to use equation (2) with a Gaussian kernel we need to find an expression for $\lambda \|\beta^*\|^2$ in terms of the parameters α_t . This can be done as follows.

$$\begin{aligned}
\|\beta\|^2 &= \beta \cdot \beta \\
&= \left(\sum_{t=1}^T \alpha_t \Phi(x_t) \right) \cdot \left(\sum_{s=1}^T \alpha_s \Phi(x_s) \right) \\
&= \sum_{t,s} \alpha_t \Phi(x_t) \cdot \Phi(x_s) \alpha_s \\
&= \sum_{t,s} \alpha_t K(x_t, x_s) \alpha_s \\
&= \alpha^T K \alpha
\end{aligned}$$

$$K_{t,s} = K(x_t, x_s) \tag{7}$$

The matrix K defined by (7) is called the kernel matrix or sometimes the Gram matrix. Equation (2) can now be rewritten in terms of α as follows.

$$\alpha^* = \operatorname{argmin}_{\alpha} \left(\sum_{t=1}^T L(m_t(\alpha)) \right) + \lambda \alpha^T K \alpha \tag{8}$$

In (2) the margin $m_t(\alpha)$ is computed using (5). The significance of (8) is the feature vectors $\Phi(x)$ and the parameter vector β are never computed. Instead, the learning is specified by a kernel function K (such as a Gaussian kernel) and a loss function L with no explicit reference to features. When L is taken to be hinge loss the resulting optimization problem in α is a convex quadratic program. This is the kernel form of a support vector machine (SVM). Equation (8) can also be viewed as a way of setting the weights α in a nearest neighbor rule. Empirically (8) works better than other weight setting heuristics.

1 Some Closure Properties on Kernels

Note that any kernel function K must be symmetric, i.e., $K(x_1, x_2) = K(x_2, x_1)$. It must also be positive semidefinite, i.e., $K(x, x) \geq 0$.

If K is a kernel and $\alpha > 0$ then αK is also a kernel. To see this let Φ be a feature map for K . Define Φ_2 so that $\Phi_2(x) = \sqrt{\alpha} \Phi(x)$. We then have that $\Phi_2(x_1) \cdot \Phi_2(x_2) = \alpha K(x_1, x_2)$. Note that for $\alpha < 0$ we have that αK is not positive semidefinite and hence cannot be a kernel.

If K_1 and K_2 are kernels then $K_1 + K_2$ is a kernel. To see this let Φ_1 be a feature map for K_1 and let Φ_2 be a feature map for K_2 . Let Φ_3 be the feature

map defined as follows.

$$\Phi_3(x) = f_1(x), g_1(x), f_2(x), g_2(x), f_3(x), g_3(x), \dots$$

$$\Phi_1(x) = f_1(x), f_2(x), f_3(x), \dots$$

$$\Phi_2(x) = g_1(x), g_2(x), g_3(x), \dots$$

We then have that $\Phi_3(x_1) \cdot \Phi_3(x_2)$ equals $\Phi_1(x_1) \cdot \Phi_1(x_2) + \Phi_2(x_1) \cdot \Phi_2(x_2)$ and hence Φ_3 is the desired feature map for $K_1 + K_2$.

If K_1 and K_2 are kernels then so is the product $K_1 K_2$. To see this let Φ_1 be a feature map for K_1 and let Φ_2 be the feature map for K_2 . Let $f_i(x)$ be the i th feature value under feature map Φ_1 and let $g_i(x)$ be the i th feature value under the feature map Φ_2 . We now have the following.

$$\begin{aligned} K_1(x_1, x_2) K_2(x_1, x_2) &= (\Phi_1(x_1) \cdot \Phi_1(x_2)) (\Phi_2(x_1) \cdot \Phi_2(x_2)) \\ &= \left(\sum_{i=1}^{\infty} f_i(x_1) f_i(x_2) \right) \left(\sum_{j=1}^{\infty} g_j(x_1) g_j(x_2) \right) \\ &= \sum_{i,j} f_i(x_1) f_i(x_2) g_j(x_1) g_j(x_2) \\ &= \sum_{i,j} (f_i(x_1) g_j(x_1)) (f_i(x_2) g_j(x_2)) \end{aligned}$$

We can now define a feature map Φ_3 with a feature $h_{i,j}(x)$ for each pair $\langle i, j \rangle$ defined as follows.

$$h_{i,j}(x) = f_i(x) g_j(x)$$

. We then have that $K_1(x_1, x_2) K_2(x_1, x_2)$ is $\Phi_3(x_1) \cdot \Phi_3(x_2)$ where the inner product sums over all pairs $\langle i, j \rangle$. Since the number of such pairs is countable, we can enumerate the pairs in a linear sequence to get $\Phi_3(x) \in \ell_2$.

It follows from these closure properties that if p is a polynomial with positive coefficients, and K is a kernel, then $p(K(x_1, x_2))$ is also a kernel. This proves that polynomial kernels are kernels. One can also give a direct proof that if K is a kernel and p is a convergent infinite power series with positive coefficients (an convergent infinite polynomial) then $p(K(x_1, x_2))$ is a kernel. The proof is similar to the proof that a product of kernels is a kernel but uses a countable set of higher order moments as features. The result for infinite power series can then be used to prove that a Gaussian kernel is a kernel. These proofs are homework problems for these notes. Unlike most proofs in the literature, we do not require compactness of the set X on which the Gaussian kernel is defined.

2 Hilbert Space

The set ℓ_2 is an infinite dimensional Hilbert space. In fact, all Hilbert spaces with a countable basis are isomorphic to ℓ_2 . So ℓ_2 is really the only Hilbert space we need to consider. But different feature maps yield different interpretations of the space ℓ_2 as functions on \mathcal{X} . A particularly interesting feature map is the following.

$$\Phi(x) = 1, x, \frac{x^2}{\sqrt{2}}, \frac{x^3}{\sqrt{3!}}, \dots, \frac{x^n}{\sqrt{n!}}, \dots$$

Now consider any function f all of whose derivatives exist at 0. Define $\beta(f)$ to be the following infinite sequence.

$$\beta(f) = f(0), f'(0), \frac{f''(0)}{\sqrt{2}}, \dots, \frac{f^k(0)}{\sqrt{k!}}, \dots$$

For any f with $\beta(f) \in \ell_2$ (which is many familiar functions) we have the following.

$$f(x) = \beta(f) \cdot \Phi(x) \tag{9}$$

So under this feature map, the parameter vectors β in ℓ_2 represent essentially all functions whose Taylor series converges. For any given feature map Φ on \mathcal{X} define $\mathcal{H}(\Phi)$ to be the set of functions f from \mathcal{X} to R such that there exists a parameter vector $\beta(f) \in \ell_2$ satisfying (9). Equation (2) can then be written as follows where $\|f\|^2$ abbreviates $\|\beta(f)\|^2$.

$$f^* = \operatorname{argmin}_{f \in \mathcal{H}(\Phi)} \left(\sum_{t=1}^T L(y_t f(x_t)) \right) + \lambda \|f\|^2$$

This way of writing the equation emphasizes that with a rich feature map selecting β is equivalent to selecting a function from a rich space of functions.