3.2 Characterisation of kernels

Recall that a kernel function computes the inner product of the images under an embedding $\boldsymbol{\phi}$ of two data points

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z}) \rangle .$$

We have seen how forming a matrix of the pairwise evaluations of a kernel function on a set of inputs gives a positive semi-definite matrix. We also saw in Chapter 2 how a kernel function implicitly defines a feature space that in many cases we do not need to construct explicitly. This second observation suggests that we may also want to create kernels without explicitly constructing the feature space. Perhaps the structure of the data and our knowledge of the particular application suggest a way of comparing two inputs. The function that makes this comparison is a candidate for a kernel function.

A general characterisation So far we have only one way of verifying that the function is a kernel, that is to construct a feature space for which the function corresponds to first performing the feature mapping and then computing the inner product between the two images. For example we used this technique to show the polynomial function is a kernel and to show that the exponential of the cardinality of a set intersection is a kernel.

We will now introduce an alternative method of demonstrating that a candidate function is a kernel. This will provide one of the theoretical tools needed to create new kernels, and combine old kernels to form new ones.

One of the key observations is the relation with positive semi-definite matrices. As we saw above the kernel matrix formed by evaluating a kernel on all pairs of any set of inputs is positive semi-definite. This forms the basis of the following definition.

Definition 3.10 [Finitely positive semi-definite functions] A function

$$\kappa : X \times X \longrightarrow \mathbb{R}$$

satisfies the finitely positive semi-definite property if it is a symmetric function for which the matrices formed by restriction to any finite subset of the space X are positive semi-definite. ■

Note that this definition does not require the set X to be a vector space. We will now demonstrate that the finitely positive semi-definite property characterises kernels. We will do this by explicitly constructing the feature space assuming only this property. We first state the result in the form of a theorem.

Theorem 3.11 (Characterisation of kernels) *A function*

$$\kappa : X \times X \longrightarrow \mathbb{R},$$

which is either continuous or has a finite domain, can be decomposed

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z}) \rangle$$

into a feature map ϕ into a Hilbert space F applied to both its arguments followed by the evaluation of the inner product in F if and only if it satisfies the finitely positive semi-definite property.

Proof The 'only if' implication is simply the result of Proposition 3.7. We will now show the reverse implication. We therefore assume that κ satisfies the finitely positive semi-definite property and proceed to construct a feature mapping $\boldsymbol{\phi}$ into a Hilbert space for which κ is the kernel.

There is one slightly unusual aspect of the construction in that the elements of the feature space will in fact be functions. They are, however, points in a vector space and will fulfil all the required properties. Recall our observation in Section 3.1.1 that learning a weight vector is equivalent to identifying an element of the feature space, in our case one of the functions. It is perhaps natural therefore that the feature space is actually the set of functions that we will be using in the learning problem

$$\mathcal{F} = \left\{ \sum_{i=1}^{\ell} \alpha_i \kappa(\mathbf{x}_i, \cdot) : \ \ell \in \mathbb{N}, \ \mathbf{x}_i \in X, \ \alpha_i \in \mathbb{R}, \ i = 1, \ldots, \ell \right\}.$$

We have chosen to use a caligraphic $\mathcal{F}$ reserved for function spaces rather than the normal F of a feature space to emphasise that the elements are functions. We should, however, emphasise that this feature space is a set of points that are in fact functions. Note that we have used a $\cdot$ to indicate the position of the argument of the function. Clearly, the space is closed under multiplication by a scalar and addition of functions, where addition is defined by

$$f, g \in \mathcal{F} \Longrightarrow (f + g)(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}).$$

Hence, $\mathcal{F}$ is a vector space. We now introduce an inner product on $\mathcal{F}$ as follows. Let $f, g \in \mathcal{F}$ be given by

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) \quad \text{and} \quad g(\mathbf{x}) = \sum_{i=1}^{n} \beta_i \kappa(\mathbf{z}_i, \mathbf{x})$$

then we define

$$\langle f, g\rangle = \sum_{i=1}^{\ell}\sum_{j=1}^{n} \alpha_i\beta_j\kappa(\mathbf{x}_i, \mathbf{z}_j) = \sum_{i=1}^{\ell} \alpha_i g(\mathbf{x}_i) = \sum_{j=1}^{n} \beta_j f(\mathbf{z}_j), \tag{3.4}$$

where the second and third equalities follow from the definitions of f and g. It is clear from these equalities that $\langle f, g\rangle$ is real-valued, symmetric and bilinear and hence satisfies the properties of an inner product, provided

$$\langle f, f\rangle \geq 0 \text{ for all } f \in \mathcal{F}.$$

But this follows from the assumption that all kernel matrices are positive semi-definite, since

$$\langle f, f\rangle = \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} \alpha_i\alpha_j\kappa(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} \geq 0,$$

where $\boldsymbol{\alpha}$ is the vector with entries α_i, $i = 1, \ldots, \ell$, and $\mathbf{K}$ is the kernel matrix constructed on $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell$.

There is a further property that follows directly from the equations (3.4) if we take $g = \kappa(\mathbf{x}, \cdot)$

$$\langle f, \kappa(\mathbf{x}, \cdot)\rangle = \sum_{i=1}^{\ell} \alpha_i\kappa(\mathbf{x}_i, \mathbf{x}) = f(\mathbf{x}). \tag{3.5}$$

This fact is known as the *reproducing property* of the kernel. It remains to show the two additional properties of completeness and separability. Separability will follow if the input space is countable or the kernel is continuous, but we omit the technical details of the proof of this fact. For completeness consider a fixed input $\mathbf{x}$ and a Cauchy sequence $(f_n)_{n=1}^{\infty}$. We have

$$(f_n(\mathbf{x}) - f_m(\mathbf{x}))^2 = \langle f_n - f_m, \kappa(\mathbf{x}, \cdot)\rangle^2 \leq \|f_n - f_m\|^2\kappa(\mathbf{x}, \mathbf{x})$$

by the Cauchy–Schwarz inequality. Hence, $f_n(\mathbf{x})$ is a bounded Cauchy sequence of real numbers and hence has a limit. If we define the function

$$g(\mathbf{x}) = \lim_{n\to\infty} f_n(\mathbf{x}),$$

and include all such limit functions in $\mathcal{F}$ we obtain the Hilbert space F_κ associated with the kernel κ.

We have constructed the feature space, but must specify the image of an input $\mathbf{x}$ under the mapping $\boldsymbol{\phi}$

$$\boldsymbol{\phi} : \mathbf{x} \in X \longmapsto \boldsymbol{\phi}(\mathbf{x}) = \kappa(\mathbf{x}, \cdot) \in F_\kappa.$$

We can now evaluate the inner product between an element of F_κ and the image of an input $\mathbf{x}$ using equation (3.5)

$$\langle f, \boldsymbol{\phi}(\mathbf{x})\rangle = \langle f, \kappa(\mathbf{x}, \cdot)\rangle = f(\mathbf{x}).$$

This is precisely what we require, namely that the function f can indeed be represented as the linear function defined by an inner product (with itself) in the feature space F_κ. Furthermore the inner product is strict since if $\|f\| = 0$, then for all $\mathbf{x}$ we have that

$$f(\mathbf{x}) = \langle f, \boldsymbol{\phi}(\mathbf{x})\rangle \leq \|f\| \, \|\boldsymbol{\phi}(\mathbf{x})\| = 0.$$

□

Given a function κ that satisfies the finitely positive semi-definite property we will refer to the corresponding space F_κ as its *Reproducing Kernel Hilbert Space (RKHS)*. Similarly, we will use the notation $\langle \cdot, \cdot \rangle_{F_\kappa}$ for the corresponding inner product when we wish to emphasise its genesis.

Remark 3.12 [Reproducing property] We have shown how any kernel can be used to construct a Hilbert space in which the reproducing property holds. It is fairly straightforward to see that if a symmetric function $\kappa(\cdot, \cdot)$ satisfies the reproducing property in a Hilbert space F of functions

$$\langle \kappa(\mathbf{x}, \cdot), f(\cdot)\rangle_{\mathcal{F}} = f(\mathbf{x}), \text{ for } f \in \mathcal{F},$$

then κ satisfies the finitely positive semi-definite property, since

$$\begin{aligned}
\sum_{i,j=1}^{\ell} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i,j=1}^{\ell} \alpha_i \alpha_j \langle \kappa(\mathbf{x}_i, \cdot), \kappa(\mathbf{x}_j, \cdot)\rangle_{\mathcal{F}} \\
&= \left\langle \sum_{i=1}^{\ell} \alpha_i \kappa(\mathbf{x}_i, \cdot), \sum_{j=1}^{\ell} \alpha_j \kappa(\mathbf{x}_j, \cdot) \right\rangle_{\mathcal{F}} \\
&= \left\| \sum_{i=1}^{\ell} \alpha_i \kappa(\mathbf{x}_i, \cdot) \right\|_{\mathcal{F}}^2 \geq 0.
\end{aligned}$$

■

Mercer kernel We are now able to show Mercer's theorem as a consequence of the previous analysis. Mercer's theorem is usually used to construct a feature space for a valid kernel. Since we have already achieved this with the RKHS construction, we do not actually require Mercer's theorem itself. We include it for completeness and because it defines the feature

space in terms of an explicit feature vector rather than using the function space of our RKHS construction. Recall the definition of the function space $L_2(X)$ from Example 3.4.

Theorem 3.13 (Mercer) *Let X be a compact subset of $\mathbb{R}^n$. Suppose κ is a continuous symmetric function such that the integral operator $T_\kappa : L_2(X) \to L_2(X)$*

$$(T_\kappa f)(\cdot) = \int_X \kappa(\cdot, \mathbf{x}) f(\mathbf{x}) d\mathbf{x},$$

is positive, that is

$$\int_{X \times X} \kappa(\mathbf{x}, \mathbf{z}) f(\mathbf{x}) f(\mathbf{z}) d\mathbf{x} d\mathbf{z} \geq 0,$$

for all $f \in L_2(X)$. Then we can expand $\kappa(\mathbf{x}, \mathbf{z})$ in a uniformly convergent series (on $X \times X$) in terms of functions $\boldsymbol{\phi}_j$, satisfying $\langle \boldsymbol{\phi}_j, \boldsymbol{\phi}_i \rangle = \delta_{ij}$

$$\kappa(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{\infty} \boldsymbol{\phi}_j(\mathbf{x}) \boldsymbol{\phi}_j(\mathbf{z}).$$

Furthermore, the series $\sum_{i=1}^{\infty} \|\boldsymbol{\phi}_i\|^2_{L_2(X)}$ is convergent.

Proof The theorem will follow provided the positivity of the integral operator implies our condition that all finite submatrices are positive semi-definite. Suppose that there is a finite submatrix on the points $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ that is not positive semi-definite. Let the vector $\boldsymbol{\alpha}$ be such that

$$\sum_{i,j=1}^{\ell} \kappa(\mathbf{x}_i, \mathbf{x}_j) \alpha_i \alpha_j = \epsilon < 0,$$

and let

$$f_\sigma(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2} \right) \in L_2(X),$$

where d is the dimension of the space X. We have that

$$\lim_{\sigma \to 0} \int_{X \times X} \kappa(\mathbf{x}, \mathbf{z}) f_\sigma(\mathbf{x}) f_\sigma(\mathbf{z}) d\mathbf{x} d\mathbf{z} = \epsilon.$$

But then for some $\sigma > 0$ the integral will be less than 0 contradicting the positivity of the integral operator.

Now consider an orthonormal basis $\boldsymbol{\phi}_i(\cdot)$, $i = 1, \ldots$ of F_κ the RKHS of the kernel κ. Then we have the Fourier series for $\kappa(\mathbf{x}, \cdot)$

$$\kappa(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} \langle \kappa(\mathbf{x}, \cdot), \boldsymbol{\phi}_i(\cdot) \rangle \boldsymbol{\phi}_i(\mathbf{z}) = \sum_{i=1}^{\infty} \boldsymbol{\phi}_i(\mathbf{x}) \boldsymbol{\phi}_i(\mathbf{z}),$$

as required.

Finally, to show that the series $\sum_{i=1}^{\infty} \|\boldsymbol{\phi}_i\|^2_{L_2(X)}$ is convergent, using the compactness of X we obtain

$$\begin{aligned} \infty &> \int_X \kappa(\mathbf{x}, \mathbf{x}) d\mathbf{x} = \lim_{n \to \infty} \int_X \sum_{i=1}^{n} \boldsymbol{\phi}_i(\mathbf{x}) \boldsymbol{\phi}_i(\mathbf{x}) d\mathbf{x} \\ &= \lim_{n \to \infty} \sum_{i=1}^{n} \int_X \boldsymbol{\phi}_i(\mathbf{x}) \boldsymbol{\phi}_i(\mathbf{x}) d\mathbf{x} = \lim_{n \to \infty} \sum_{i=1}^{n} \|\boldsymbol{\phi}_i\|^2_{L_2(X)} \end{aligned}$$

□

Example 3.14 Consider the kernel function $\kappa(\mathbf{x}, \mathbf{z}) = \kappa(\mathbf{x} - \mathbf{z})$. Such a kernel is said to be *translation invariant*, since the inner product of two inputs is unchanged if both are translated by the same vector. Consider the one-dimensional case in which κ is defined on the interval $[0, 2\pi]$ in such a way that $\kappa(u)$ can be extended to a continuous, symmetric, periodic function on $\mathbb{R}$. Such a function can be expanded in a uniformly convergent Fourier series

$$\kappa(u) = \sum_{n=0}^{\infty} a_n \cos(nu).$$

In this case we can expand $\kappa(x - z)$ as follows

$$\kappa(x - z) = a_0 + \sum_{n=1}^{\infty} a_n \sin(nx) \sin(nz) + \sum_{n=1}^{\infty} a_n \cos(nx) \cos(nz).$$

Provided the a_n are all positive this shows $\kappa(x, z)$ is the inner product in the feature space defined by the orthogonal features

$$\{\boldsymbol{\phi}_i(x)\}_{i=0}^{\infty} = (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots, \sin(nx), \cos(nx), \ldots),$$

since the functions, 1, $\cos(nu)$ and $\sin(nu)$ form a set of orthogonal functions on the interval $[0, 2\pi]$. Hence, normalising them will provide a set of Mercer features. Note that the embedding is defined independently of the parameters a_n, which subsequently control the geometry of the feature space. ■

Example 3.14 provides some useful insight into the role that the choice of kernel can play. The parameters a_n in the expansion of $\kappa(u)$ are its Fourier coefficients. If, for some n, we have $a_n = 0$, the corresponding features are removed from the feature space. Similarly, small values of a_n mean that the feature is given low weighting and so will have less influence on the choice of hyperplane. Hence, the choice of kernel can be seen as choosing a filter with a particular spectral characteristic, the effect of which is to control the influence of the different frequencies in determining the optimal separation.

Covariance kernels Mercer's theorem enables us to express a kernel as a sum over a set of functions of the product of their values on the two inputs

$$\kappa(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{\infty} \boldsymbol{\phi}_j(\mathbf{x})\boldsymbol{\phi}_j(\mathbf{z}).$$

This suggests a different view of kernels as a covariance function determined by a probability distribution over a function class. In general, given a distribution $q(f)$ over a function class $\mathcal{F}$, the covariance function is given by

$$\kappa_q(\mathbf{x}, \mathbf{z}) = \int_{\mathcal{F}} f(\mathbf{x})f(\mathbf{z})q(f)df.$$

We will refer to such a kernel as a *covariance kernel*. We can see that this is a kernel by considering the mapping

$$\boldsymbol{\phi} : \mathbf{x} \longmapsto (f(\mathbf{x}))_{f\in\mathcal{F}}$$

into the space of functions on $\mathcal{F}$ with inner product given by

$$\langle a(\cdot), b(\cdot)\rangle = \int_{\mathcal{F}} a(f)\, b(f)\, q(f)\, df.$$

This definition is quite natural if we consider that the ideal kernel for learning a function f is given by

$$\kappa_f(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z}), \tag{3.6}$$

since the space $\mathcal{F} = \mathcal{F}_{\kappa_f}$ in this case contains functions of the form

$$\sum_{i=1}^{\ell} \alpha_i \kappa_f(\mathbf{x}_i, \cdot) = \sum_{i=1}^{\ell} \alpha_i f(\mathbf{x}_i) f(\cdot) = Cf(\cdot).$$

So for the kernel κ_f, the corresponding $\mathcal{F}$ is one-dimensional, containing only multiples of f. We can therefore view κ_q as taking a combination of these

simple kernels for all possible f weighted according to the prior distribution q. Any kernel derived in this way is a valid kernel, since it is easily verified that it satisfies the finitely positive semi-definite property

$$\begin{aligned}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j\kappa_q(\mathbf{x}_i,\mathbf{x}_j) &= \sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j\int_{\mathcal{F}} f(\mathbf{x}_i)f(\mathbf{x}_j)q(f)df \\ &= \int_{\mathcal{F}}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j f(\mathbf{x}_i)f(\mathbf{x}_j)q(f)df \\ &= \int_{\mathcal{F}}\left(\sum_{i=1}^{\ell}\alpha_i f(\mathbf{x}_i)\right)^2 q(f)df \geq 0.\end{aligned}$$

Furthermore, if the underlying class $\mathcal{F}$ of functions are $\{-1,+1\}$-valued, the kernel κ_q will be normalised since

$$\kappa_q(\mathbf{x},\mathbf{x}) = \int_{\mathcal{F}} f(\mathbf{x})f(\mathbf{x})q(f)df = \int_{\mathcal{F}} q(f)df = 1.$$

We will now show that every kernel can be obtained as a covariance kernel in which the distribution has a particular form. Given a valid kernel κ, consider the Gaussian prior q that generates functions f according to

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} u_i\boldsymbol{\phi}_i(\mathbf{x}),$$

where $\boldsymbol{\phi}_i$ are the orthonormal functions of Theorem 3.13 for the kernel κ, and u_i are generated according to the Gaussian distribution $\mathcal{N}(0,1)$ with mean 0 and standard deviation 1. Notice that this function will be in $L_2(X)$ with probability 1, since using the orthonormality of the $\boldsymbol{\phi}_i$ we can bound its expected norm by

$$\begin{aligned}\mathbb{E}\left[\|f\|^2_{L_2(X)}\right] &= \mathbb{E}\left[\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}u_iu_j\left\langle\boldsymbol{\phi}_i,\boldsymbol{\phi}_j\right\rangle_{L_2(X)}\right] \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\mathbb{E}\left[u_iu_j\right]\left\langle\boldsymbol{\phi}_i,\boldsymbol{\phi}_j\right\rangle_{L_2(X)} \\ &= \sum_{i=1}^{\infty}\mathbb{E}[u_i^2]\|\boldsymbol{\phi}_i\|^2_{L_2(X)} = \sum_{i=1}^{\infty}\|\boldsymbol{\phi}_i\|^2_{L_2(X)} < \infty,\end{aligned}$$

where the final inequality follows from Theorem 3.13. Since the norm is a positive function it follows that the measure of functions not in $L_2(X)$ is 0,

as otherwise the expectation would not be finite. But curiously the function will almost certainly not be in $\mathcal{F}_\kappa$ for infinite-dimensional feature spaces. We therefore take the distribution q to be defined over the space $L_2(X)$.

The covariance function κ_q is now equal to

$$\begin{aligned}
\kappa_q(\mathbf{x}, \mathbf{z}) &= \int_{L_2(X)} f(\mathbf{x}) f(\mathbf{z}) q(f) df \\
&= \lim_{n\to\infty} \sum_{i,j=1}^{n} \boldsymbol{\phi}_i(\mathbf{x}) \boldsymbol{\phi}_j(\mathbf{z}) \int_{\mathbb{R}^n} u_i u_j \prod_{k=1}^{n} \left(\frac{1}{\sqrt{2\pi}} \exp(-u_k^2/2) du_k \right) \\
&= \lim_{n\to\infty} \sum_{i,j=1}^{n} \boldsymbol{\phi}_i(\mathbf{x}) \boldsymbol{\phi}_j(\mathbf{z}) \delta_{ij} = \sum_{i=1}^{\infty} \boldsymbol{\phi}_i(\mathbf{x}) \boldsymbol{\phi}_i(\mathbf{z}) \\
&= \kappa(\mathbf{x}, \mathbf{z}).
\end{aligned}$$

3.3 The kernel matrix

Given a training set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ and kernel function $\kappa(\cdot, \cdot)$, we introduced earlier the kernel or Gram matrix $\mathbf{K} = (\mathbf{K}_{ij})_{i,j=1}^{\ell}$ with entries

$$\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j), \text{ for } i, j = 1, \ldots, \ell.$$

The last subsection was devoted to showing that the function κ is a valid kernel provided its kernel matrices are positive semi-definite for all training sets S, the so-called finitely positive semi-definite property. This fact enables us to manipulate kernels without necessarily considering the corresponding feature space. Provided we maintain the finitely positive semi-definite property we are guaranteed that we have a valid kernel, that is, that there exists a feature space for which it is the corresponding kernel function. Reasoning about the similarity measure implied by the kernel function may be more natural than performing an explicit construction of its feature space.

The intrinsic modularity of kernel machines also means that any kernel function can be used provided it produces symmetric, positive semi-definite kernel matrices, and any kernel algorithm can be applied, as long as it can accept as input such a matrix together with any necessary labelling information. In other words, the kernel matrix acts as an interface between the data input and learning modules.

Kernel matrix as information bottleneck In view of our characterisation of kernels in terms of the finitely positive semi-definite property, it becomes clear why the kernel matrix is perhaps the core ingredient in the theory of kernel methods. It contains all the information available in order

to perform the learning step, with the sole exception of the output labels in the case of supervised learning. It is worth bearing in mind that it is only through the kernel matrix that the learning algorithm obtains information about the choice of feature space or model, and indeed the training data itself.

The finitely positive semi-definite property can also be used to justify intermediate processing steps designed to improve the representation of the data, and hence the overall performance of the system through manipulating the kernel matrix before it is passed to the learning machine. One simple example is the addition of a constant to the diagonal of the matrix. This has the effect of introducing a soft margin in classification or equivalently regularisation in regression, something that we have already seen in the ridge regression example. We will, however, describe more complex manipulations of the kernel matrix that correspond to more subtle tunings of the feature space.

In view of the fact that it is only through the kernel matrix that the learning algorithm receives information about the feature space and input data, it is perhaps not surprising that some properties of this matrix can be used to assess the generalization performance of a learning system. The properties vary according to the type of learning task and the subtlety of the analysis, but once again the kernel matrix plays a central role both in the derivation of generalisation bounds and in their evaluation in practical applications.

The kernel matrix is not only the central concept in the design and analysis of kernel machines, it can also be regarded as the central data structure in their implementation. As we have seen, the kernel matrix acts as an interface between the data input module and the learning algorithms. Furthermore, many model adaptation and selection methods are implemented by manipulating the kernel matrix as it is passed between these two modules. Its properties affect every part of the learning system from the computation, through the generalisation analysis, to the implementation details.

Remark 3.15 [Implementation issues] One small word of caution is perhaps worth mentioning on the implementation side. Memory constraints mean that it may not be possible to store the full kernel matrix in memory for very large datasets. In such cases it may be necessary to recompute the kernel function as entries are needed. This may have implications for both the choice of algorithm and the details of the implementation. ■

Another important aspect of our characterisation of valid kernels in terms