2.5 Summary

The crucial ingredient of SVMs and other kernel methods is the so-called kernel trick (see (2.7) and Remark 2.8), which permits the computation of dot products in high-dimensional feature spaces, using simple functions defined on pairs of input patterns. This trick allows the formulation of nonlinear variants of any algorithm that can be cast in terms of dot products, SVMs being but the most prominent example. The mathematical result underlying the kernel trick is almost a century old [359]. Nevertheless, it was only much later that it was exploited by the machine learning community for the analysis [4] and construction of algorithms [62], and that it was described as a general method for constructing nonlinear generalizations of dot product algorithms [480].

The present chapter has reviewed the mathematical theory of kernels. We started with the class of polynomial kernels, which can be motivated as computing a combinatorially large number of monomial features rather efficiently. This led to the general question of which kernel can be used, or: which kernel can be represented as a dot product in a linear feature space. We defined this class and discussed some of its properties. We described several ways how, given such a kernel, one can construct a representation in a feature space. The most well-known representation employs Mercer's theorem, and represents the feature space as an ℓ_2 space defined in terms of the eigenfunctions of an integral operator associated with the kernel. An alternative representation uses elements of the theory of reproducing kernel Hilbert spaces, and yields additional insights, representing the linear space as a space of functions written as kernel expansions. We gave an in-depth discussion of the kernel trick in its general form, including the case where we are interested in dissimilarities rather than similarities; that is, when we want to come up with nonlinear generalizations of distance-based algorithms rather than dot-product-based algorithms.

In both cases, the underlying philosophy is the same: we are trying to express a complex nonlinear algorithm in terms of simple geometrical concepts, and we are then dealing with it in a linear space. This linear space may not always be readily available; in some cases, it may even be hard to construct explicitly. Nevertheless, for the sake of design and analysis of the algorithms, it is sufficient to know that the linear space exists, empowering us to use the full potential of geometry, linear algebra and functional analysis.

2.6 Problems

2.1 (Monomial Features in $\mathbb{R}^2$ •) *Verify the second equality in (2.9).*

2.2 (Multiplicity of Monomial Features in $\mathbb{R}^N$ [515] ••) *Consider the monomial kernel $k(x, x') = \langle x, x' \rangle^d$ (where $x, x' \in \mathbb{R}^N$), generating monomial features of order d. Prove*

that a valid feature map for this kernel can be defined coordinate-wise as

$$\Phi_{\mathbf{m}}(\mathbf{x}) = \sqrt{\frac{d!}{\prod_{i=1}^{n}[\mathbf{m}]_i!}} \prod_{i=1}^{n} [\mathbf{x}]_i^{[\mathbf{m}]_i} \tag{2.95}$$

for every $\mathbf{m} \in \mathbb{N}^n$, $\sum_{i=1}^{n}[\mathbf{m}]_i = d$ *(i.e., every such* $\mathbf{m}$ *corresponds to one dimension of* $\mathcal{H}$*).*

2.3 (Inhomogeneous Polynomial Kernel ••) *Prove that the kernel (2.70) induces a feature map into the space of all monomials up to degree d. Discuss the role of c.*

2.4 (Eigenvalue Criterion of Positive Definiteness •) *Prove that a symmetric matrix is positive definite if and only if all its eigenvalues are nonnegative (see Appendix B).*

2.5 (Dot Products are Kernels •) *Prove that dot products (Definition B.7) are positive definite kernels.*

2.6 (Kernels on Finite Domains ••) *Prove that for finite* $\mathcal{X}$*, say* $\mathcal{X} = \{x_1, \ldots, x_m\}$*, k is a kernel if and only if the* $m \times m$ *matrix* $(k(x_i, x_j))_{ij}$ *is positive definite.*

2.7 (Positivity on the Diagonal •) *From Definition 2.5, prove that a kernel satisfies* $k(x, x) \geq 0$ *for all* $x \in \mathcal{X}$*.*

2.8 (Cauchy-Schwarz for Kernels ••) *Give an elementary proof of Proposition 2.7.*

Hint: start with the general form of a symmetric 2×2 *matrix, and derive conditions for its coefficients that ensure that it is positive definite.*

2.9 (PD Kernels Vanishing on the Diagonal •) *Use Proposition 2.7 to prove that a kernel satisfying* $k(x, x) =$ *for all* $x \in \mathcal{X}$ *is identically zero.*

How does the RKHS look in this case? Hint: use (2.31).

2.10 (Two Kinds of Positivity •) *Give an example of a kernel which is positive definite according to Definition 2.5, but not positive in the sense that* $k(x, x') \geq 0$ *for all* x, x'*.*

Give an example of a kernel where the contrary is the case.

2.11 (General Coordinate Transformations •) *Prove that if* $\sigma : \mathcal{X} \to \mathcal{X}$ *is a bijection, and* $k(x, x')$ *is a kernel, then* $k(\sigma(x), \sigma(x'))$ *is a kernel, too.*

2.12 (Positivity on the Diagonal •) *Prove that positive definite kernels are positive on the diagonal,* $k(x, x) \geq 0$ *for all* $x \in \mathcal{X}$*. Hint: use* $m = 1$ *in (2.15).*

2.13 (Symmetry of Complex Kernels ••) *Prove that complex-valued positive definite kernels are symmetric (2.18).*

2.14 (Real Kernels vs. Complex Kernels •) *Prove that a real matrix satisfies (2.15) for all* $c_i \in \mathbb{C}$ *if and only if it is symmetric and it satisfies (2.15) for* real *coefficients* c_i*.*

Hint: decompose each c_i *in (2.15) into real and imaginary parts.*

2.15 (Rank-One Kernels •) *Prove that if f is a real-valued function on $\mathcal{X}$, then $k(x, x') := f(x)f(x')$ is a positive definite kernel.*

2.16 (Bayes Kernel ••) *Consider a binary pattern recognition problem. Specialize the last problem to the case where $f : \mathcal{X} \to \{\pm 1\}$ equals the Bayes decision function $y(x)$, i.e., the classification with minimal risk subject to an underlying distribution $P(x, y)$ generating the data.*

Argue that this kernel is particularly suitable since it renders the problem linearly separable in a 1D feature space: State a decision function (cf. (1.35)) that solves the problem (hint: you just need one parameter α, and you may set it to 1; moreover, use $b = 0$) [124].

The final part of the problem requires knowledge of Chapter 16: Consider now the situation where some prior $P(f)$ over the target function class is given. What would the optimal kernel be in this case? Discuss the connection to Gaussian processes.

2.17 (Inhomogeneous Polynomials •) *Prove that the inhomogeneous polynomial (2.70) is a positive definite kernel, e.g., by showing that it is a linear combination of homogeneous polynomial kernels with positive coefficients. What kind of features does this kernel compute [561]?*

2.18 (Normalization in Feature Space •) *Given a kernel k, construct a corresponding normalized kernel $\tilde{k}$ by normalizing the feature map $\tilde{\Phi}$ such that for all $x \in \mathcal{X}$, $\|\tilde{\Phi}(x)\| = 1$ (cf. also Definition 12.35). Discuss the relationship between normalization in input space and normalization in feature space for Gaussian kernels and homogeneous polynomial kernels.*

2.19 (Cosine Kernel •) *Suppose $\mathcal{X}$ is a dot product space, and $x, x' \in \mathcal{X}$. Prove that $k(x, x') = \cos(\angle(x, x))$ is a positive definite kernel. Hint: use Problem 2.18.*

2.20 (Alignment Kernel •) *Let $\langle K, K' \rangle_F := \sum_{ij} K_{ij} K'_{ij}$ be the Frobenius dot product of two matrices. Prove that the* empirical alignment *of two Gram matrices [124], $A(K, K') := \langle K, K' \rangle_F / \sqrt{\langle K, K \rangle_F \langle K', K' \rangle_F}$, is a positive definite kernel.*

Note that the alignment can be used for model selection, putting $K'_{ij} := y_i y_j$ (cf. Problem 2.16) and $K_{ij} := \operatorname{sgn}(k(x_i, x_j))$ or $K_{ij} := \operatorname{sgn}(k(x_i, x_j)) - b$ (cf. [124]).

2.21 (Equivalence Relations as Kernels •••) *Consider a similarity measure $k : \mathcal{X} \to \{0, 1\}$ with*

$$k(x, x) = 1 \text{ for all } x \in \mathcal{X}. \tag{2.96}$$

Prove that k is a positive definite kernel if and only if, for all $x, x', x'' \in \mathcal{X}$,

$$k(x, x') = 1 \iff k(x', x) = 1 \text{ and} \tag{2.97}$$

$$k(x, x') = k(x', x'') = 1 \implies k(x, x'') = 1. \tag{2.98}$$

Equations (2.96) to (2.98) amount to saying that $k = I_T$, where $T \subset \mathcal{X} \times \mathcal{X}$ is an equivalence relation.

As a simple example, consider an undirected graph, and let $(x, x') \in T$ whenever x and x' are in the same connected component of the graph. Show that T is an equivalence relation.

Find examples of equivalence relations that lend themselves to an interpretation as similarity measures. Discuss whether there are other relations that one might want to use as similarity measures.

2.22 (Different Feature Spaces for the Same Kernel •) *Give an example of a kernel with two valid feature maps Φ_1, Φ_2, mapping into spaces $\mathcal{H}_1, \mathcal{H}_2$ of different dimensions.*

2.23 (Converse of Mercer's Theorem •) *Prove that if an integral operator kernel k admits a uniformly convergent dot product representation on some compact set $\mathcal{X} \times \mathcal{X}$,*

$$k(x, x') = \sum_{i=1}^{\infty} \psi_i(x)\psi_i(x'), \tag{2.99}$$

then it is positive definite. Hint: show that

$$\int_{\mathcal{X}\times\mathcal{X}} \left(\sum_{i=1}^{\infty} \psi_i(x)\psi_i(x') \right) f(x)f(x')\,dx\,dx' = \sum_{i=1}^{\infty} \left(\int_{\mathcal{X}} \psi_i(x)f(x)\,dx \right)^2 \geq 0.$$

Argue that in particular, polynomial kernels (2.67) satisfy Mercer's conditions.

2.24 (∞-Norm of Mercer Eigenfunctions ••) *Prove that under the conditions of Theorem 2.10, we have, up to sets of measure zero,*

$$\sup_j \left\| \sqrt{\lambda_j}\psi_j \right\|_\infty \leq \sqrt{\|k\|_\infty} < \infty. \tag{2.100}$$

Hint: note that $\|k\|_\infty \geq k(x, x)$ up to sets of measures zero, and use the series expansion given in Theorem 2.10. Show, moreover, that it is not generally the case that

$$\sup_j \|\psi_j\|_\infty < \infty. \tag{2.101}$$

Hint: consider the case where $\mathcal{X} = \mathbb{N}$, $\mu(\{n\}) := 2^{-n}$, and $k(i, j) := \delta_{ij}$. Show that

1. *$T_k((a_j)) = (a_j 2^{-j})$ for $(a_j) \in L_2(\mathcal{X}, \mu)$,*
2. *T_k satisfies $\langle (a_j), T_k(a_j) \rangle = \sum_j (a_j 2^{-j})^2 \geq 0$ and is thus positive definite,*
3. *$\lambda_j = 2^{-j}$ and $\psi_j = 2^{j/2} e_j$ form an orthonormal eigenvector decomposition of T_k (here, e_j is the jth canonical unit vector in ℓ_2), and*
4. *$\|\psi_j\|_\infty = 2^{j/2} = \lambda_j^{-1/2}$.*

Argue that the last statement shows that (2.101) is wrong and (2.100) is tight.[12]

2.25 (Generalized Feature Maps •••) *Via (2.38), Mercer kernels induce compact (integral) operators. Can you generalize the idea of defining a feature map associated with an*

12. Thanks to S. Smale and I. Steinwart for this exercise.

operator to more general bounded positive definite operators T? Hint: use the multiplication operator representation of T [467].

2.26 (Nyström Approximation (cf. [603]) •) *Consider the integral operator obtained by substituting the distribution* P *underlying the data into (2.38), i.e.,*

$$(T_k f)(x) = \int_{\mathcal{X}} k(x, x') f(x)\, d\mathrm{P}(x). \tag{2.102}$$

If the conditions of Mercer's theorem are satisfied, then k can be diagonalized as

$$k(x, x') = \sum_{j=1}^{N_{\mathcal{H}}} \lambda_j \psi_j(x) \psi_j(x'), \tag{2.103}$$

where λ_j and ψ_j satisfy the eigenvalue equation

$$\int_{\mathcal{X}} k(x, x') \psi_j(x)\, d\mathrm{P}(x) = \lambda_j \psi_j(x') \tag{2.104}$$

and the orthonormality conditions

$$\int_{\mathcal{X}} \psi_i(x) \psi_j(x) d\mathrm{P}(x) = \delta_{ij}. \tag{2.105}$$

Show that by replacing the integral by a summation over an iid sample $X = \{x_1, \ldots, x_m\}$ from P(x)*, one can recover the kernel PCA eigenvalue problem (Section 1.7). Hint: Start by evaluating (2.104) for $x' \in X$, to obtain m equations. Next, approximate the integral by a sum over the points in X, replacing $\int_{\mathcal{X}} k(x, x') \psi_j(x)\, d\mathrm{P}(x)$ by $\frac{1}{m} \sum_{n=1}^{m} k(x_n, x') \psi_j(x_n)$.*

Derive the orthogonality condition for the eigenvectors $(\psi_j(x_n))_{n=1,\ldots,m}$ from (2.105).

2.27 (Lorentzian Feature Spaces ••) *If a finite number of eigenvalues is negative, the expansion in Theorem 2.10 is still valid. Show that in this case, k corresponds to a Lorentzian symmetric bilinear form in a space with indefinite signature [467].*

Discuss whether this causes problems for learning algorithms utilizing these kernels. In particular, consider the cases of SV machines (Chapter 7) and Kernel PCA (Chapter 14).

2.28 (Symmetry of Reproducing Kernels •) *Show that reproducing kernels (Definition 2.9) are symmetric. Hint: use (2.35) and exploit the symmetry of the dot product.*

2.29 (Coordinate Representation in the RKHS ••) *Write $\langle \cdot, \cdot \rangle$ as a dot product of coordinate vectors by expressing the functions of the RKHS in the basis $(\sqrt{\lambda_n}\psi_n)_{n=1,\ldots,N_{\mathcal{H}}}$, which is orthonormal with respect to $\langle \cdot, \cdot \rangle$, i.e.,*

$$f(x) = \sum_{n=1}^{N_{\mathcal{H}}} \alpha_n \sqrt{\lambda_n} \psi_n(x). \tag{2.106}$$

Obtain an expression for the coordinates α_n, using (2.47) and $\alpha_n = \left\langle f, \sqrt{\lambda_n}\psi_n \right\rangle$. Show that $\mathcal{H}$ has the structure of a RKHS in the sense that for f and g given by (2.106), and

$$g(x) = \sum_{j=1}^{N_{\mathcal{H}}} \beta_j \sqrt{\lambda_j} \psi_j(x), \tag{2.107}$$

we have $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle = \langle f, g \rangle$ *. Show, moreover, that* $f(x) = \langle \boldsymbol{\alpha}, \Phi(x) \rangle$ *in* $\mathcal{H}$*. In other words,* $\Phi(x)$ *is the coordinate representation of the kernel as a function of one argument.*

2.30 (Equivalence of Regularization Terms •) *Using (2.36) and (2.41), prove that* $\|\mathbf{w}\|^2$*, where* $\mathbf{w} = \sum_{i=1}^{m} \alpha_i \Phi(x_i)$*, is the same no matter whether* Φ *denotes the RKHS feature map (2.21) or the Mercer feature map (2.40).*

2.31 (Approximate Inversion of Gram Matrices ••) *Use the kernel PCA map (2.59) to derive a method for approximately inverting a large Gram matrix.*

2.32 (Effective Dimension of Feature Space •) *Building on Section 2.2.7, argue that for a finite data set, we are always effectively working in a finite-dimensional feature space.*

2.33 (Translation of a Dot Product •) *Prove (2.79).*

2.34 (Example of a CPD Kernel ••) *Argue that the hyperbolic tangent kernel (2.69) is effectively conditionally positive definite, if the input values are suitably restricted, since it can be approximated by* $k + b$*, where* k *is a polynomial kernel (2.67) and* $b \in \mathbb{R}$*. Discuss how this explains that hyperbolic tangent kernels can be used for SVMs although, as pointed out in number of works (e.g., [86], cf. the remark following (2.69)), they are not positive definite.*

2.35 (Polarization Identity ••) *Prove the* polarization identity*, stating that for any symmetric bilinear form* $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*, we have, for all* $x, x' \in \mathcal{X}$*,*

$$\langle x, x' \rangle = \frac{1}{4} \left(\langle x + x', x + x' \rangle - \langle x - x', x - x' \rangle \right). \tag{2.108}$$

Now consider the special case where $\langle \cdot, \cdot \rangle$ *is a Euclidean dot product and* $\langle x - x', x - x' \rangle$ *is the squared Euclidean distance between* x *and* x'*. Discuss why the polarization identity does not imply that the value of the dot product can be recovered from the distances alone. What else does one need?*

2.36 (Vector Space Representation of CPD Kernels •••) *Specialize the vector space representation of symmetric kernels (Proposition 2.25) to the case of cpd kernels. Can you identify a subspace on which a cpd kernel is actually pd?*

2.37 (Parzen Windows Classifiers in Feature Space ••) *Assume that* k *is a positive definite kernel. Compare the algorithm described in Section 1.2 with the one of (2.89). Construct situations where the two algorithms give different results. Hint: consider datasets where the class means coincide.*

2.38 (Canonical Distortion Kernel ○○○) *Can you define a kernel based on Baxter's canonical distortion metric [28]?*