Pattern Recognition and Machine Learning
Cristopher Bishop

Exercise Solutions

Stefan Stefanache

April 5, 2022

Chapter 1

Kernel Methods

Exercise 6.1 ⋆⋆

Consider the dual formulation of the least squares linear regression problem given in Section 6.1. Show that the solution for the components a_n of the vector $\mathbf{a}$ can be expressed as a linear combination of the elements of the vector $\boldsymbol{\phi}(\mathbf{x}_n)$. Denoting these coefficients by the vector $\mathbf{w}$, show that the dual of the dual formulation is given by the original representation in terms of the parameter vector $\mathbf{w}$.

Proof. By rewriting (6.4), one has that

$$\begin{aligned} a_n &= -\frac{1}{\lambda}\{\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n) - t_n\} \\ &= -\frac{1}{\lambda}\left\{\sum_{i=1}^{M} w_i\phi_i(\mathbf{x}_n) - \frac{t_n}{\sum_{i=1}^{M}\phi_i(\mathbf{x}_n)}\sum_{i=1}^{M}\phi_i(\mathbf{x}_n)\right\} \\ &= \sum_{i=1}^{M}\left(\frac{t_n}{\lambda\sum_{i=1}^{M}\phi_i(\mathbf{x}_n)} - \frac{w_i}{\lambda}\right)\phi_i(\mathbf{x}_n) \\ &= \sum_{i=1}^{M}\Omega_{ni}\phi_i(\mathbf{x}_n) \\ &= \boldsymbol{\Omega}_n^T\boldsymbol{\phi}(\mathbf{x}_n) \end{aligned}$$

where

$$\Omega_{ni} = \frac{t_n}{\lambda\sum_{i=1}^{M}\phi_i(\mathbf{x}_n)} - \frac{w_i}{\lambda}$$

Therefore, a_n can be written as a linear combination of the elements of $\boldsymbol{\phi}(\mathbf{x}_n)$ and

$$\mathbf{a} = \mathrm{diag}(\boldsymbol{\Omega}\boldsymbol{\Phi})$$

□

Exercise 6.3 ⋆

The nearest-neighbour classifier (Section 2.5.2) assigns a new input vector $\mathbf{x}$ to the same class as that of the nearest input vector $\mathbf{x}_n$ from the training set, where in the simple case, the distance

is defined by the Euclidean metric $\|\mathbf{x} - \mathbf{x}_n\|^2$. By expressing this rule in terms of scalar product and then making use of kernel substitution, formulate the nearest-neighbour classifier for a general nonlinear kernel.

Proof. Since we're dealing with inner products over $\mathbb{R}$, the Euclidian metric can be rewritten as

$$\|\mathbf{x} - \mathbf{x}_n\|^2 = \langle \mathbf{x} - \mathbf{x}_n, \mathbf{x} - \mathbf{x}_n \rangle = \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{x}_n \rangle + \langle \mathbf{x}_n \mathbf{x}_n \rangle$$

Similarly to what happens in Section 6.2, using kernel substitution above to replace $\langle \mathbf{x}, \mathbf{x}' \rangle$ with a nonlinear kernel $\kappa(\mathbf{x}, \mathbf{x}')$ yields the nearest-neighbour classifier for a general nonlinear kernel:

$$k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x}, \mathbf{x}) - 2\kappa(\mathbf{x}, \mathbf{x}_n) + \kappa(\mathbf{x}_n, \mathbf{x}_n)$$

□

Exercise 6.4 ⋆

In Appendix C, we give an example of a matrix that has positive elements but that ahas a negative eigenvalue and hence that is not positive definite. Find an example of the converse property, namely a 2×2 matrix with positive eigenvalues that has at least one negative element.

Proof. Consider the matrix

$$A = \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$$

A contains one negative element and the eigenvalues of A are $\lambda_1 = 1$ and $\lambda_2 = 3$, which proves that a matrix can be positive definite and have negative elements. □

Exercise 6.5 ⋆

Verify the results (6.13) and (6.14) for constructing valid kernels.

Proof. Since k_1 is a valid kernel, let $\boldsymbol{\alpha}$ be a feature mapping such that

$$k_1(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\alpha}(\mathbf{x}), \boldsymbol{\alpha}(\mathbf{x}') \rangle$$

Using the fact that an inner product on a real vector space is a positive-definite symmetric bilinear form, we have that

$$ck_1(\mathbf{x}, \mathbf{x}') = c\langle \boldsymbol{\alpha}(\mathbf{x}), \boldsymbol{\alpha}(\mathbf{x}') \rangle = \langle \sqrt{c}\boldsymbol{\alpha}(\mathbf{x}), \sqrt{c}\boldsymbol{\alpha}(\mathbf{x}') \rangle = \langle \boldsymbol{\beta}(\mathbf{x}), \boldsymbol{\beta}(\mathbf{x}') \rangle$$

where $c > 0$ is a constant and $\boldsymbol{\beta}(\mathbf{x}) = \sqrt{c}\boldsymbol{\alpha}(\mathbf{x})$. Therefore, the new kernel

$$k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}') \tag{6.13}$$

is valid. Analogously, since $f(\cdot)$ is a real-valued function,

$$f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') = f(\mathbf{x})\langle \boldsymbol{\alpha}(\mathbf{x}), \boldsymbol{\alpha}(\mathbf{x}') \rangle f(\mathbf{x}') = \langle f(\mathbf{x})\boldsymbol{\alpha}(\mathbf{x}), f(\mathbf{x}')\boldsymbol{\alpha}(\mathbf{x}') \rangle = \langle \boldsymbol{\gamma}(\mathbf{x}), \boldsymbol{\gamma}(\mathbf{x}') \rangle$$

where $\boldsymbol{\gamma}(\mathbf{x}) = f(\mathbf{x})\boldsymbol{\alpha}(\mathbf{x})$. As a result, the kernel

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') \tag{6.14}$$

will also be valid. □

Exercise 6.7 ⋆

Verify the results (6.17) and (6.18) for constructing valid kernels.

Proof. Let K_1 and K_2 be the Gram matrices corresponding to the kernels k_1 and k_2. Therefore, they are positive semidefinite matrices, so for any $\mathbf{a} \in \mathbb{R}^n$, one has that

$$\mathbf{a}^T\mathbf{H}\mathbf{a} = \mathbf{a}^T\big(\mathbf{H}_1 + \mathbf{H}_2\big)\mathbf{a} = \mathbf{a}^T\mathbf{H}_1\mathbf{a} + \mathbf{a}^T\mathbf{H}_2\mathbf{a} > 0$$

Since $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2$ is positive semidefinite and corresponds to the Gram matrix of the kernel $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$, one has that the kernel

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \tag{6.17}$$

is valid. Now, let $\boldsymbol{\alpha}, \boldsymbol{\beta}$ be feature mappings such that

$$k_1(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\alpha}(\mathbf{x}), \boldsymbol{\alpha}(\mathbf{x}') \rangle$$

$$k_2(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\beta}(\mathbf{x}), \boldsymbol{\beta}(\mathbf{x}') \rangle$$

As a result,

$$\begin{aligned}
k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') &= \langle \boldsymbol{\alpha}(\mathbf{x}), \boldsymbol{\alpha}(\mathbf{x}') \rangle \langle \boldsymbol{\beta}(\mathbf{x}), \boldsymbol{\beta}(\mathbf{x}') \rangle \\
&= \boldsymbol{\alpha}(\mathbf{x})^T \boldsymbol{\alpha}(\mathbf{x}') \boldsymbol{\beta}^T(\mathbf{x}) \boldsymbol{\beta}^T(\mathbf{x}') \\
&= \left[\sum_{i=1}^{N} \alpha_i(\mathbf{x}) \alpha_i(\mathbf{x}') \right] \left[\sum_{j=1}^{N} \beta_i(\mathbf{x}) \beta_i(\mathbf{x}') \right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(\mathbf{x}) \beta_j(\mathbf{x}) \alpha_i(\mathbf{x}') \beta_j(\mathbf{x}') \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}(\mathbf{x}) A_{ij}(\mathbf{x}') \\
&= \langle \mathbf{A}(\mathbf{x}), \mathbf{A}(\mathbf{x}') \rangle_{\mathbf{F}}
\end{aligned}$$

where $\mathbf{A}$ is a square matrix with

$$A_{ij}(\mathbf{x}) = \alpha_i(\mathbf{x}) \beta_j(\mathbf{x})$$

and $\langle \cdot, \cdot \rangle_{\mathbf{F}}$ is the Frobenius inner product. Since the product kernel can be rewritten as a valid inner product in the feature space defined by the feature mapping $\mathbf{A}(\mathbf{x})$, the new kernel

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \tag{6.18}$$

is valid. □

Exercise 6.6 ⋆

Verify the results (6.15) and (6.16) for constructing valid kernels.

Proof. Let $q(\cdot)$ be a polynomial with nonnegative coefficients. Since in the polynomial kernels are summed and multiplied by nonnegative constants or other kernels, combining (6.13), (6.17) and (6.18) proves that the kernel

$$k(\mathbf{x}, \mathbf{x}') = q\big(k_1(\mathbf{x}, \mathbf{x}')\big) \tag{6.15}$$

is valid. Now, the exponential function is defined as

$$\exp(x) \coloneqq \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

, so

$$\exp\big(k_1(\mathbf{x}, \mathbf{x}')\big) = \sum_{i=0}^{\infty} \frac{k_1(\mathbf{x}, \mathbf{x}')^i}{i!}$$

Note that the exponential of a kernel is an infinite sequence of kernel sums and products (with itself or nonnegative constants), so by using (6.13), (6.17), (6.18) again, one has that the new kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp\big(k_1(\mathbf{x}, \mathbf{x}')\big) \tag{6.16}$$

is valid. □

Exercise 6.7 ⋆

Verify the results (6.19) and (6.20) for constructing valid kernels.

Proof. Let $\boldsymbol{\psi}$ be a feature mapping such that

$$k_3(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle$$

Then,

$$\begin{aligned} k_3\big(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\big) &= \langle \boldsymbol{\psi}\big(\boldsymbol{\phi}(\mathbf{x})\big), \boldsymbol{\psi}\big(\boldsymbol{\phi}(\mathbf{x}')\big) \rangle \\ &= \langle \big(\boldsymbol{\psi} \circ \boldsymbol{\phi}\big)(\mathbf{x}), \big(\boldsymbol{\psi} \circ \boldsymbol{\phi}\big)(\mathbf{x}') \rangle \\ &= \langle \boldsymbol{\gamma}(\mathbf{x}), \boldsymbol{\gamma}(\mathbf{x}') \rangle \end{aligned}$$

where $\boldsymbol{\phi}$ is a function from $\mathbf{x}$ to $\mathbb{R}^M$ and $\boldsymbol{\gamma} = \boldsymbol{\psi} \circ \boldsymbol{\phi}$. Therefore, the kernel

$$k(\mathbf{x}, \mathbf{x}') = k_3\big(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\big) \tag{6.19}$$

is valid. For the second part, since $\mathbf{A}$ is a symmetric, positive semidefinite matrix, one can use the Cholesky decomposition to obtain a matrix $\mathbf{L}$ such that

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

As a result, one can show that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{L}\mathbf{L}^T\mathbf{x} = \big(\mathbf{L}^T\mathbf{x}\big)^T\big(\mathbf{L}^T\mathbf{x}\big) = \langle \boldsymbol{\zeta}(\mathbf{x}), \boldsymbol{\zeta}(\mathbf{x}') \rangle$$

where $\boldsymbol{\zeta}(\mathbf{x}) = \mathbf{L}^T\mathbf{x}$. Hence, the kernel

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{A}\mathbf{x} \tag{6.20}$$

is valid. □