EQ2415 – Machine Learning and Data Science
HT22

Tutorial 2

A. Honoré, A. Ghosh

1 Kernel substitution

Material: Bishop's book Chapter 6.4.1 and 6.4.2

Valid kernels Let $n, d > 0$. A function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to be a valid kernel iif: the matrix $K \in \mathbb{R}^{n\times n}$ associated to k, whose elements are given by $k(\mathbf{x}_i, \mathbf{x}_j)$ with $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$, is positive semi-definite for all possible choices of $\mathbf{x}_i, \mathbf{x}_j$ (Bishop page 295).

By definition, a matrix $K \in \mathbb{R}^{n\times n}$ is said to be positive semi-definite iif

$$\mathbf{a}^T K\mathbf{a} \geq 0, \text{ for } \mathbf{a} \in \mathbb{R}^n, \tag{1}$$

this is not the same thing as a matrix whose elements are non-negative.

1.1 Linear Kernel

Let a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be such that:

$$k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle = \mathbf{x}^T \mathbf{x}', \text{ for } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d. \tag{2}$$

Let $K \in \mathbb{R}^{n\times n}$ denote the matrix with elements $K_{i,j} = k(\mathbf{v}_i, \mathbf{v}_j)$ with $\mathbf{v}_i, \mathbf{v}_j$ in a set of n vectors of $\mathbb{R}^d$.

Question 1. Show that the function k is a valid kernel, by showing that K is positive semi-definite.

Solution: General solution:
Let $\mathbf{a} \in \mathbb{R}^n$. We have

$$\begin{aligned}
\mathbf{a}^T K\mathbf{a} &= \sum_{i,j} a_i a_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle \\
&= \sum_{i,j} \langle a_i \mathbf{v}_i, a_j \mathbf{v}_j \rangle \\
&= \langle \sum_i a_i \mathbf{v}_i, \sum_j a_j \mathbf{v}_j \rangle, \text{ by linearity of scalar products} \\
&= || \sum_i a_i \mathbf{v}_i ||^2 \geq 0,
\end{aligned} \tag{3}$$

Thus K is positive semi-definite and is a Gram matrix associated with k, thus k is a valid kernel.

Note: This is true for any scalar product. Thus, to prove that a kernel is valid, it is sometimes easier to show that the kernel function k can be expressed as the scalar product of some arbitrary functions of $\mathbf{x}$ and $\mathbf{x}'$! ■

1.2 Constructing valid kernels

Bishop exercise 6.7. Suppose that $k_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $k_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are two valid kernels.

Question 1. Show that

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}'), \text{ with } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d, \tag{4}$$

is a valid kernel.

Solution: Let K_1 and K_2 be two Gram matrices associated with k_1 and k_2 respectively. By definition, K_1 and K_2 are positive semi-definite, thus $K = K_1 + K_2$ is positive semi-definite and is a Gram matrix associated with k. Thus k is a valid kernel. ■

Question 2. Show that

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}'), \text{ with } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d, \tag{5}$$

is a valid kernel.

Solution: Let $N, M > 0$. We write $k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}^{(1)}(\mathbf{x})^T\boldsymbol{\phi}^{(1)}(\mathbf{x}')$ with $\boldsymbol{\phi}^{(1)} : \mathbb{R}^d \to \mathbb{R}^M$ and $k_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}^{(2)}(\mathbf{x})^T\boldsymbol{\phi}^{(2)}(\mathbf{x}')$ with $\boldsymbol{\phi}^{(2)} : \mathbb{R}^d \to \mathbb{R}^N$. Then

$$\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}^{(1)}(\mathbf{x})^T\boldsymbol{\phi}^{(1)}(\mathbf{x}')\boldsymbol{\phi}^{(2)}(\mathbf{x})^T\boldsymbol{\phi}^{(2)}(\mathbf{x}') \\
&= \sum_{i=1}^{M} \boldsymbol{\phi}_i^{(1)}(\mathbf{x})\boldsymbol{\phi}_i^{(1)}(\mathbf{x}') \sum_{j=1}^{N} \boldsymbol{\phi}_j^{(2)}(\mathbf{x})\boldsymbol{\phi}_j^{(2)}(\mathbf{x}') = \sum_{i=1}^{M}\sum_{j=1}^{N} [\boldsymbol{\phi}_i^{(1)}(\mathbf{x})\boldsymbol{\phi}_j^{(2)}(\mathbf{x})][\boldsymbol{\phi}_i^{(1)}(\mathbf{x}')\boldsymbol{\phi}_j^{(2)}(\mathbf{x}')] \\
&= \sum_{k=1}^{MN} \boldsymbol{\phi}_k(\mathbf{x})\boldsymbol{\phi}_k(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}'), \text{ where } \boldsymbol{\phi}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^{MN}.
\end{aligned} \tag{6}$$

Thus, k can be written as a scalar product, thus the associated matrix is a Gram matrix, thus it is positive semi-definite, thus k is a valid kernel. ■

1.3 The exponential kernel

Remember that the Taylor series expansion of the exponential function around 0 is:

$$\exp(x) = \sum_{k=0}^{+\infty} \frac{x^k}{k!}, \text{ for } x \in \mathbb{R}. \tag{7}$$

The radial basis function (RBF) is expressed:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right), \text{ for } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d. \tag{8}$$

Question 1. Show that the RBF is a valid kernel.

Solution: We expand (8):

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2}\right)\exp\left(\frac{\mathbf{x}^T\mathbf{x}'}{\sigma^2}\right)\exp\left(-\frac{(\mathbf{x}')^T\mathbf{x}'}{2\sigma^2}\right). \tag{9}$$

Using 7 we see that the exponential of a kernel is a sum and product of kernels. The linear kernel is valid, the products of valid kernels are valid, and thus the RBF is a valid kernel. ■

Question 2. Show that the RBF can be expressed as the inner product of an infinite-dimensional feature vector. First assume that $d = 1$, and then try to generalize to arbitrary finite d using the multinomial theorem. Bishop 6.11 (p 321)

Solution: We use the expansion:

$$\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2}\right)\exp\left(\frac{\mathbf{x}^T\mathbf{x}'}{\sigma^2}\right)\exp\left(-\frac{(\mathbf{x}')^T\mathbf{x}'}{2\sigma^2}\right) \\
&= \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2}\right) \cdot \left(\sum_{k=0}^{+\infty} \frac{\left(\frac{\mathbf{x}^T\mathbf{x}'}{\sigma^2}\right)^k}{k!}\right) . \exp\left(-\frac{(\mathbf{x}')^T\mathbf{x}'}{2\sigma^2}\right)
\end{aligned} \tag{10}$$

In what follows assume $\sigma = 1$, or replace $\mathbf{x}$ (and $\mathbf{x}'$) with scaled versions $\mathbf{y} = \frac{1}{\sigma}\mathbf{x}$. Let us further expand $(\mathbf{x}^T\mathbf{x}')^k$ for $k \in \mathbb{N}$:

$$\begin{aligned}
(\mathbf{x}^T\mathbf{x}')^k &= (x_1x_1' + \ldots + x_dx_d')^k \text{, using the multinomial theorem:} \\
&= \sum_{\substack{n_1+\ldots+n_d=k \\ n_1,\cdots,n_d>0}} \frac{k!}{n_1!n_2!\ldots n_d!} \prod_{i=1}^{d} (x_ix_i')^{n_i} \\
&= \sum_{\substack{n_1+\ldots+n_d=k \\ n_1,\ldots,n_d>0}} k! \frac{\prod_{i=1}^{d} x_i^{n_i}}{\sqrt{n_1!n_2!\ldots n_d!}} \frac{\prod_{i=1}^{d} (x_i')^{n_i}}{\sqrt{n_1!n_2!\ldots n_d!}}
\end{aligned} \tag{11}$$

This gives:

$$k(\mathbf{x},\mathbf{x}') = \sum_{k=0}^{+\infty} \sum_{\substack{n_1+\ldots+n_d=k \\ n_1,\ldots,n_d>0}} \left(\exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right) \frac{\prod_{i=1}^{d} x_i^{n_i}}{\sqrt{n_1!n_2!\ldots n_d!}} \right) \cdot \left(\exp\left(-\frac{(\mathbf{x}')^T\mathbf{x}'}{2}\right) \frac{\prod_{i=1}^{d} (x_i')^{n_i}}{\sqrt{n_1!n_2!\ldots n_d!}} \right) \tag{12}$$

The number of d-tuples of positive integers which sum to k (the index of the second sum), varies with k. In fact, there are exactly $l_k = \binom{k+d-1}{d-1}$ such d-tuples.

This means that we can define an intermediate vector $v_k(\mathbf{x})$ of length l_k, where the jth element:

$$v_k(\mathbf{x})_j = \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right) \frac{\prod_{i=1}^{d} x_i^{n_i^j}}{\sqrt{n_1^j!n_2^j!\ldots n_d^j!}}, \tag{13}$$

where $n_1^j, \ldots, n_d^j$, is the jth d-tuples of positive integers who sum to k.

This gives:

$$k(\mathbf{x},\mathbf{x}') = \sum_{k=0}^{+\infty} \sum_{j=1}^{l_k} v_k(\mathbf{x})_j v_k(\mathbf{x}')_j \tag{14}$$

Now, we can write the final vector of infinite dimension:

$$\boldsymbol{\phi}(\mathbf{x}) = [v_0(\mathbf{x}), v_1(\mathbf{x})_1, \ldots, v_1(\mathbf{x})_{l_1}, \ldots, v_n(\mathbf{x})_1, \ldots, v_n(\mathbf{x})_{l_n}, \ldots] \tag{15}$$

Finally,

$$k(\mathbf{x},\mathbf{x}') = \sum_{m=0}^{+\infty} \boldsymbol{\phi}_m(\mathbf{x})\boldsymbol{\phi}_m(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}'), \tag{16}$$

where $\boldsymbol{\phi}$ maps vectors in $\mathbb{R}^d$ to vectors in a space of infinite dimension. ∎

1.4 Gaussian Process for regression

Suppose that you are given N training data points for a regression problem in the form of two matrices: $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$ and $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{q\times N}$. In a Gaussian process model, the joint distribution of the target training data is assumed Gaussian with zero mean and with covariance determined by a Gram matrix K, i.e. :

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_N | \mathbf{x}_1, \ldots, \mathbf{x}_N) = \mathcal{N}(\mathbf{0}, K_N), \tag{17}$$

where the elements of $K_N \in \mathbb{R}^{n\times n}$ are determined from a kernel k on the set of training data points X.

Suppose that you want to predict the target value $\mathbf{y}_{N+1} \in \mathbb{R}^q$ for a new target $\mathbf{x}_{N+1} \in \mathbb{R}^d$ using the Gaussian process model in (17). This consists in finding the posterior distribution of the target value, given the training data and the new input data point:

$$p(\mathbf{y}_{N+1} | \mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}). \tag{18}$$

Question 1.. Bishop 6.20 p322

Find the family and parameters of the joint distribution of the training and new *target* points conditioned on the training and new *data* points:

$$p(\mathbf{y}_{N+1}, \mathbf{y}_1, \ldots, \mathbf{y}_N | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}). \tag{19}$$

Solution: Using the definition of the model in (17), the joint distribution can be written in terms of a Gram matrix K_{N+1}:

$$p(\mathbf{y}_{N+1}, \mathbf{y}_1, \ldots, \mathbf{y}_N | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}) = \mathcal{N}\left(\mathbf{0}, K_{N+1}\right) \tag{20}$$

where $K_{N+1} = \begin{bmatrix} c & \mathbf{k}^T \\ \mathbf{k} & K_N \end{bmatrix}$, with $\mathbf{k} = [k(\mathbf{x}_1, \mathbf{x}_{N+1}), \ldots, k(\mathbf{x}_N, \mathbf{x}_{N+1})]^T$ and $c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1})$. ■

Question 2. Using standard results on Gaussian, we can say that

$$p(\mathbf{y}_{N+1} | \mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}) = \mathcal{N}\left(m(\mathbf{x}_{N+1}), \sigma_2(\mathbf{x}_{N+1})\right), \tag{21}$$

i.e. that the distribution we are looking for is Gaussian. Use the equations on partitioned Gaussian: (2.81)-(2.82) page 87, to determine $m(\mathbf{x}_{N+1})$ and $\sigma_2(\mathbf{x}_{N+1})$.

Solution:

Suppose $\mathbf{x} \in \mathbb{R}^d$ is distributed according to a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose we write $\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$, since $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$ we have $\boldsymbol{\Sigma}_{aa}$ and $\boldsymbol{\Sigma}_{bb}$ symetric and $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^T$.

Then we have that:

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \end{aligned} \tag{22}$$

We consider $\mathbf{y}_{N+1}$ as $\mathbf{x}_a$, $\mathbf{y}_N$ as $\mathbf{x}_b$, c as Σ_{aa}, $\mathbf{k}$ as Σ_{ba} and $\mathbf{k}^T$ as Σ_{ab}.

We find

$$\begin{aligned} m(\mathbf{x}_{N+1}) &= \mathbf{k}^T K_N^{-1} Y \\ \sigma_2(\mathbf{x}_{N+1}) &= c - \mathbf{k}^T K_N^{-1}\mathbf{k} \end{aligned} \tag{23}$$

■

Question 3. Implement the Gaussian Process model in Python.