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.

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

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}'), \text{ with } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d, \tag{3}$$

is a valid kernel.

Question 2. Show that

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

is a valid kernel.

1.3 The exponential kernel

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

$$\exp(x) = \sum_{k=0}^{+\infty} \frac{x^k}{k!}, \text{ for } x \in \mathbb{R}. \tag{5}$$

The radial basis function (RBF) is expressed:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right), \text{ for } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d. \tag{6}$$

Question 1. Show that 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)

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. :

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_N|\mathbf{x}_1, \ldots, \mathbf{x}_N) = \mathcal{N}(\mathbf{0}, K_N), \tag{7}$$

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 (7). This consists in finding the posterior distribution of the target value, given the training data and the new input data point:

$$p(\mathbf{y}_{N+1}|\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}). \tag{8}$$

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:

$$p(\mathbf{y}_{N+1}, \mathbf{y}_1, \ldots, \mathbf{y}_N|\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}). \tag{9}$$

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

$$p(\mathbf{y}_{N+1}|\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}) = \mathcal{N}\left(m(\mathbf{x}_{N+1}), \sigma_2(\mathbf{x}_{N+1})\right), \tag{10}$$

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 $\sigma_2(\mathbf{x}_{N+1})$.

Question 3. Implement the Gaussian Process model in Python.