Statistical Machine Learning 2018

Exercises and answers, week 11

30 November 2018

TUTORIAL

Exercise 1

See Bishop page 294-295 and equation (6.10), (6.11) and (6.12). Consider the kernel

$$k(x, z) = (xz + c)^2 \tag{1}$$

with $c > 0$ a constant. Show that the correponding feature mapping contains constant, linear, and quadratic terms.

ANSWER:

$$\begin{aligned} (xz + c)^2 &= x^2 z^2 + 2cxz + c^2 \\ &= (x^2, \sqrt{2c}x, c)(z^2, \sqrt{2c}z, c)^T \end{aligned}$$

So

$$k(x, z) = \phi(x)^T \phi(z)$$

with the feature mapping $\phi(x) = (x^2, \sqrt{2c}x, c)^T$

Exercise 2

In this exercise we will compute the posterior of a Gaussian process in a simple case and compare it to linear regression.

- Since we want to compare it to linear regression we want to use a kernel which is equivalent to the basis functions $\phi(x) = (x, c)^T$. Compute the corresponding kernel.

  ANSWER:

  $$k(x, z) = \phi(x)^T \phi(z)$$

  So

  $$k(x, z) = \phi(x)^T \phi(z) = (x, c)^T (z, c) = xz + c^2$$

- Given now one observation $x = t = 0$ compute the mean and the variance as a function of x of the noise-free posterior distribution assuming the above constructed kernel and a zero mean $m(x) = 0$ for the prior distribution.

ANSWER:

$$\begin{aligned}\mu &= \frac{k(0,x)}{k(0,0)}t = 0 \\ \Sigma &= k(x,x) - k(0,x)(k(0,0))^{-1}k(0,x) = x^2 + c^2 - \frac{c^4}{c^2} = x^2\end{aligned}$$

- Finally compute the mean and the variance given an observation noise with variance σ^2.

ANSWER:

$$\begin{aligned}\mu &= \frac{k(0,x)}{k(0,0)+\sigma^2}t = 0 \\ \Sigma &= k(x,x) - k(0,x)(k(0,0)+\sigma^2)^{-1}k(0,x) = x^2 + c^2 - \frac{c^4}{c^2+\sigma^2} = x^2 + \frac{c^2\sigma^2}{c^2+\sigma^2}\end{aligned}$$

Exercise 3

In the Bayesian linear regression chapter we learned how to compute the predictive distribution $p(t|x, \mathbf{t}, \mathbf{x}, \alpha, \beta)$ from a data set of two points: $\{x_1, t_1\} = (0.4, 0.05)$ and $\{x_2, t_2\} = (0.6, -0.35)$. Using feature vector $\phi(x) = (1, x)^{\mathrm{T}}$, we assumed an isotropic Gaussian weights prior $p(\mathbf{w}|\alpha = 2) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbf{I})$, and random Gaussian observation noise with variance $\beta^{-1} = 0.1$. We found

$$\begin{aligned}p(t|x, \mathbf{x}, \mathbf{t}) &= \mathcal{N}(t|m(x), s^2(x)), \quad \text{with} \\ m(x) &= \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{m}_N \\ s^2(x) &= \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(x) \\ \mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \\ \mathbf{S}_N &= (\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\end{aligned}$$

with $\boldsymbol{\Phi}$ the 2×2 design matrix (see Bishop 3.16), and

$$\mathbf{S}_N = \begin{pmatrix} 22 & 10 \\ 10 & 7.2 \end{pmatrix}^{-1} \simeq \begin{pmatrix} 0.1233 & -0.1712 \\ -0.1712 & 0.3767 \end{pmatrix}$$

1. Show that the predictive mean as function of x can also be expressed directly as a weighted combination of the output values $\{t_1, t_2\}$ as

$$t(x) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}')t_n$$

by computing the 'equivalent kernel' given as B(3.62):

$$k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\phi(\mathbf{x}').$$

Verify that the two indeed give equivalent predictions, e.g. by comparing some values.

ANSWER: Filling in $k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\phi(\mathbf{x}')$ gives

$$
\begin{aligned}
k(x,x') &= \beta \begin{pmatrix} 1 & x \end{pmatrix} \mathbf{S}_N \begin{pmatrix} 1 \\ x' \end{pmatrix} \\
&= 10 \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} 0.1233 & -0.1712 \\ -0.1712 & 0.3767 \end{pmatrix} \begin{pmatrix} 1 \\ x' \end{pmatrix} \\
&= 10 * (0.1233 - 0.1712x - 0.1712x' + 0.3767xx') \\
&= 1.233 - 1.712(x + x') + 3.767xx'
\end{aligned}
$$

Filling in for the predictive mean gives $m(x) \approx -0.0446 - 0.2022\,x$.

Using the equivalent kernel we obtain:

$$
\begin{aligned}
t(x) &= \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}')t_n \\
&= k(x, x_1) * t_1 + k(x, x_2) * t_2 \\
&\approx (1.233 - 1.712(x + x_1) + 3.767 * x * x_1) * t_1 + (1.233 - 1.712(x + x_2) + 3.767 * x * x_2) * t_2 \\
&= 0.0274 - 0.0103\,x - 0.0720 - 0.1919\,x \\
&= -0.0446 - 0.2022\,x
\end{aligned}
$$

(or compare for a few specific values of x).

2. Make a sketch of the equivalent kernel as function of $x' = 0.2$ over the interval $[0, 1]$. Compare with Bishop, Figure 3.10 (left), and explain the difference.

   ANSWER: For $x' = 0.2$ the equivalent kernel reduces to a straight line:

$$
\begin{aligned}
k(x, 0.2) &= 1.233 - 1.712(x + 0.2) + 3.767 * x * 0.2 \\
&= 0.8906 - 0.9586\,x
\end{aligned}
$$

   The equivalent kernel in Fig.3.10 is based on the 11 Gaussian basis functions depicted in Figure 3.1 (mid), which introduces the 'wiggly' effect. Here we have just one linear feature, so no fancy graphs.

BONUS PRACTICE

Exercise 4

(Bishop 6.1) Consider the dual formulation of the least squares linear regression problem, which is given in Section 6.1:

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{t} - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a}, \tag{2}$$

where $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ is the Gram matrix.

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 $\mathbf{w}$.

ANSWER:

We first note that $J(\mathbf{a})$ depends on $\mathbf{a}$ only through the form $\mathbf{K}\mathbf{a}$. Since typically the number N of data points is greater than the number M of basis functions, the matrix $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ will be rank deficient. There will be then M eigenvectors of $\mathbf{K}$ having non-zero eigenvalues, and $N - M$ eigenvectors with eigenvalue zero. We can then decompose $\mathbf{a} = \mathbf{a}_{\|} + \mathbf{a}_{\perp}$, where $\mathbf{a}_{\|}^{\mathrm{T}}\mathbf{a}_{\perp} = 0$ and $\mathbf{K}\mathbf{a}_{\perp} = \mathbf{0}$. Thus the value of $\mathbf{a}_{\perp}$ is not determined by $J(\mathbf{a})$. We can remove the ambiguity by setting $\mathbf{a}_{\perp} = \mathbf{0}$, or equivalently by adding a regularizer term $\frac{\epsilon}{2}\mathbf{a}_{\perp}^{\mathrm{T}}\mathbf{a}_{\perp}$ to $J(\mathbf{a})$, where ϵ is a small positive constant. Then $\mathbf{a} = \mathbf{a}_{\|}$, where $\mathbf{a}_{\|}$ lies in the span of $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ and hence can be written as a linear combination of the columns of $\mathbf{\Phi}$, so that in component notation

$$a_n = \sum_{i=1}^{M} u_i \phi_i(\mathbf{x}_n)$$

or equivalently in vector notation

$$\mathbf{a} = \mathbf{\Phi}\mathbf{u}.$$

Substituting the last result into (2), we obtain

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{t} - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} \\ &= \frac{1}{2}(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t})^{\mathrm{T}}(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}\mathbf{u} \\ &= \frac{1}{2}(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t})^{\mathrm{T}}(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u}. \end{aligned}$$

Since the matrix $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ has full rank, we can define an equivalent parametrization given by $\mathbf{w} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u}$. Making this substitution into (2), we recover the original regularized error function:

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{\Phi}\mathbf{w} - \mathbf{t})^{\mathrm{T}}(\mathbf{\Phi}\mathbf{w} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}.$$

Exercise 5

Gaussian Processes versus Bayesian Linear Regression (Exercise 6.21 in Bishop)

We consider the standard linear regression model with Gaussian noise, where the regression function is defined as a linear combination of (potentially non-linear) fixed basis functions given by the elements of the vector $\boldsymbol{\phi}(\mathbf{x})$. In other words:

$$\mathbf{y}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),$$

where $\mathbf{x}$ is the input vector and $\mathbf{w}$ is the weight vector.

Previously, we analyzed this model in the *weight space view* by deriving a posterior distribution for the weights, which represent the parameters of the model. In this exercise, we instead take the *function space view* by defining a distribution directly over the functions to be modeled. We consider the Gaussian process defined by linear regression model, in which the kernel function is expressed in terms of the basis functions, i.e. $k(\mathbf{x}, \mathbf{x}') = \alpha^{-1}\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')$.

Show that the predictive distribution is identical to the result obtained for the Bayesian linear regression model:

$$p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{N+1}), \sigma^2(\mathbf{x}_{N+1})),$$

where $\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t})$, $\sigma^2(\mathbf{x}_{N+1}) = \beta^{-1} + \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})$, $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$, $\mathbf{m}_0 = \mathbf{0}$, and $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}_M$.

Hints: Note that both the Gaussian process and the linear regression model give rise to Gaussian predictive distributions, so it is only necessary to show that the conditional mean and variance are the same. For the mean, make use of the matrix identity $\mathbf{A}(\mathbf{I}+\mathbf{BA})^{-1} = (\mathbf{I}+\mathbf{AB})^{-1}\mathbf{A}$ (Check its validity!). For the variance, make use of the Woodbury identity from a previous exercise.

ANSWER: Both the Gaussian process and the linear regression model give rise to Gaussian predictive distributions $p(t_{N+1}|\mathbf{x}_{N+1})$, so we simply need to show that these have the same mean and variance.

We start from the definitions for the mean (Bishop 6.66) and variance (Bishop 6.67) of the Gaussian process predictive distribution and show that they are equivalent to those of the linear regression model predictive distribution. We use the definition of the covariance matrix for the joint distribution of $(\mathbf{t}, t_{N+1})$ (see Bishop 6.53, 6.54, 6.62, and 6.65):

$$C(\mathbf{x}_n, \mathbf{x}_m) = (\mathbf{C}_{N+1})_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm},$$

$$\mathbf{C}_{N+1} = \begin{bmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^{\mathrm{T}} & c \end{bmatrix}.$$

Mean

$$\begin{aligned}
m(\mathbf{x}_{N+1}) &= \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{t} \\
&= \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N)^{-1}\mathbf{t} \\
&= \alpha^{-1}\beta\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\beta\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \mathbf{I}_N)^{-1}\mathbf{t} \\
&= \alpha^{-1}\beta\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}(\alpha^{-1}\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \mathbf{I}_M)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \\
&= \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}(\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \alpha\mathbf{I}_M)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \\
&= \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\mathbf{S}_N(\beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}) \\
&= \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\mathbf{m}_N \quad \square.
\end{aligned}$$

Variance

$$
\begin{aligned}
\sigma^2(\mathbf{x}_{N+1}) &= c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k} \\
&= k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1} - \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N)^{-1}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}_{N+1})\alpha^{-1} \\
&= \beta^{-1} + \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{N+1}) - \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N)^{-1}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}_{N+1})\alpha^{-1} \\
&= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\left[(\alpha^{-1}\mathbf{I}_M) - (\alpha^{-1}\mathbf{I}_M)\boldsymbol{\Phi}^{\mathrm{T}}(\beta^{-1}\mathbf{I}_N + \boldsymbol{\Phi}(\alpha^{-1}\mathbf{I}_M)\boldsymbol{\Phi}^{\mathrm{T}})^{-1}(\alpha^{-1}\mathbf{I}_M)\boldsymbol{\Phi}\right]\boldsymbol{\phi}(\mathbf{x}_{N+1}) \\
&= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}(\alpha^{-1}\mathbf{I}_M + \boldsymbol{\Phi}^{\mathrm{T}}(\beta^{-1}\mathbf{I}_N)\boldsymbol{\Phi})^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1}) \\
&= \beta^{-1} + \phi(\mathbf{x}_{N+1})^{\mathrm{T}}\mathbf{S}_N\phi(\mathbf{x}_{N+1}) \quad \square.
\end{aligned}
$$