ECE1513
Tutorial 5:
Selected Exercises from Chapter 6

TA: Faye Pourghasem, fateme.pourghasem@mail.utoronto.ca

October 16, 2023

1 Example 6.1

($\star\star$) WWW 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 $\mathbf{w}$.

Solution

We first of all 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 $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$ will be rank deficient. There will then be M eigenvectors of $\mathbf{K}$ having non-zero eigenvalues, and $N - M$ eigenvectors with eigenvalue zero. We can then decompose $\mathbf{a} = \mathbf{a}_{\parallel} + \mathbf{a}_{\perp}$ where $\mathbf{a}_{\parallel}^{\mathrm{T}}\mathbf{a}_{\perp} = 0$ 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

$$\mathbf{a} = \mathbf{\Phi}\mathbf{u}. \tag{199}$$

Substituting (199) into (6.7) we obtain

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2}\left(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}\mathbf{u} \\ &= \frac{1}{2}\left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} \end{aligned} \tag{200}$$

Since the matrix $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ has full rank we can define an equivalent parametrization given by

$$\mathbf{w} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u}$$

and substituting this into (200) we recover the original regularized error function (6.2).

2 Example 6.12

($\star\star$) **www** Consider the space of all possible subsets A of a given fixed set D. Show that the kernel function (6.27) corresponds to an inner product in a feature space of dimensionality $2^{|D|}$ defined by the mapping $\phi(A)$ where A is a subset of D and the element $\phi_U(A)$, indexed by the subset U, is given by

$$\phi_U(A) = \begin{cases} 1, & \text{if } U \subseteq A; \\ 0, & \text{otherwise.} \end{cases} \tag{6.95}$$

Here $U \subseteq A$ denotes that U is either a subset of A or is equal to A.

Solution

NOTE: In the 1^{st} printing of PRML, there is an error in the text relating to this exercise. Immediately following (6.27), it says: $|A|$ denotes the number of *subsets* **in A; it should have said: $|A|$ denotes the number of** *elements* **in A.**

Since A may be equal to D (the subset relation was not defined to be strict), $\phi(D)$ must be defined. This will map to a vector of $2^{|D|}$ 1s, one for each possible subset of D, including D itself as well as the empty set. For $A \subset D$, $\phi(A)$ will have 1s in all positions that correspond to subsets of A and 0s in all other positions. Therefore, $\phi(A_1)^{\mathrm{T}}\phi(A_2)$ will count the number of subsets shared by A_1 and A_2. However, this can just as well be obtained by counting the number of elements in the intersection of A_1 and A_2, and then raising 2 to this number, which is exactly what (6.27) does.

3 Example 6.14

($\star$) **www** Write down the form of the Fisher kernel, defined by (6.33), for the case of a distribution $p(\mathbf{x}|\boldsymbol{\mu}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S})$ that is Gaussian with mean $\boldsymbol{\mu}$ and fixed covariance $\mathbf{S}$.

Solution

In order to evaluate the Fisher kernel for the Gaussian we first note that the covariance is assumed to be fixed, and hence the parameters comprise only the elements of the mean μ. The first step is to evaluate the Fisher score defined by (6.32). From the definition (2.43) of the Gaussian we have

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) = \nabla_{\boldsymbol{\mu}} \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) = \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

Next we evaluate the Fisher information matrix using the definition (6.34), giving

$$\mathbf{F} = \mathbb{E}_{\mathbf{x}}\left[\mathbf{g}(\boldsymbol{\mu}, \mathbf{x})\mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^{\mathrm{T}}\right] = \mathbf{S}^{-1}\mathbb{E}_{\mathbf{x}}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\right]\mathbf{S}^{-1}.$$

Here the expectation is with respect to the original Gaussian distribution, and so we can use the standard result

$$\mathbb{E}_{\mathbf{x}}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\right] = \mathbf{S}$$

from which we obtain

$$\mathbf{F} = \mathbf{S}^{-1}.$$

Thus the Fisher kernel is given by

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}' - \boldsymbol{\mu}),$$

which we note is just the squared Mahalanobis distance.

4 Example 6.16

($\star\star$) Consider a parametric model governed by the parameter vector $\mathbf{w}$ together with a data set of input values $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and a nonlinear feature mapping $\boldsymbol{\phi}(\mathbf{x})$. Suppose that the dependence of the error function on $\mathbf{w}$ takes the form

$$J(\mathbf{w}) = f(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_N)) + g(\mathbf{w}^{\mathrm{T}}\mathbf{w}) \tag{6.97}$$

where $g(\cdot)$ is a monotonically increasing function. By writing $\mathbf{w}$ in the form

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(\mathbf{x}_n) + \mathbf{w}_{\perp} \tag{6.98}$$

show that the value of $\mathbf{w}$ that minimizes $J(\mathbf{w})$ takes the form of a linear combination of the basis functions $\boldsymbol{\phi}(\mathbf{x}_n)$ for $n = 1, \ldots, N$.

Solution

Based on the total derivative of function f, we have:

$$f\left((\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_1, (\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_2, \ldots, (\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_N\right) = \sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T\boldsymbol{\phi}_n)} \cdot \Delta\mathbf{w}^T\boldsymbol{\phi}_n$$

Which can be further written as:

$$f\left((\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_1, (\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_2, \ldots, (\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_N\right) = \left[\sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T\boldsymbol{\phi}_n)} \cdot \boldsymbol{\phi}_n^T\right] \Delta\mathbf{w}$$

Note that here $\boldsymbol{\phi}_n$ is short for $\boldsymbol{\phi}(\mathbf{x}_n)$. Based on the equation above, we can obtain:

$$\nabla_{\mathbf{w}} f = \sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T\boldsymbol{\phi}_n)} \cdot \boldsymbol{\phi}_n^T$$

Now we focus on the derivative of function g with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}} g = \frac{\partial g}{\partial(\mathbf{w}^T \mathbf{w})} \cdot 2\mathbf{w}^T$$

In order to find the optimal $\mathbf{w}$, we set the derivative of J with respect to $\mathbf{w}$ equal to $\mathbf{0}$, yielding:

$$\nabla_{\mathbf{w}} J = \nabla_{\mathbf{w}} f + \nabla_{\mathbf{w}} g = \sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T \boldsymbol{\phi}_n)} \cdot \boldsymbol{\phi}_n^T + \frac{\partial g}{\partial(\mathbf{w}^T \mathbf{w})} \cdot 2\mathbf{w}^T = \mathbf{0}$$

Rearranging the equation above, we can obtain:

$$\mathbf{w} = \frac{1}{2a} \sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T \boldsymbol{\phi}_n)} \cdot \boldsymbol{\phi}_n$$

Where we have defined: $a = 1 \div \frac{\partial g}{\partial(\mathbf{w}^T \mathbf{w})}$, and since g is a monotonically increasing function, we have $a > 0$.

5 Example 6.17

($\star\star$) **www** Consider the sum-of-squares error function (6.39) for data having noisy inputs, where $\nu(\boldsymbol{\xi})$ is the distribution of the noise. Use the calculus of variations to minimize this error function with respect to the function $y(\mathbf{x})$, and hence show that the optimal solution is given by an expansion of the form (6.40) in which the basis functions are given by (6.41).

Solution

NOTE: In the 1^{st} printing of PRML, there are typographical errors in the text relating to this exercise. In the sentence following immediately after (6.39), $f(\mathbf{x})$ should be replaced by $y(\mathbf{x})$. Also, on the l.h.s. of (6.40), $y(\mathbf{x}_n)$ should be replaced by $y(\mathbf{x})$. There were also errors in Appendix D, which might cause confusion; please consult the errata on the PRML website.

Following the discussion in Appendix D we give a first-principles derivation of the solution. First consider a variation in the function $y(\mathbf{x})$ of the form

$$y(\mathbf{x}) \to y(\mathbf{x}) + \epsilon\eta(\mathbf{x}).$$

Substituting into (6.39) we obtain

$$E[y + \epsilon\eta] = \frac{1}{2}\sum_{n=1}^{N}\int \left\{y(\mathbf{x}_n + \boldsymbol{\xi}) + \epsilon\eta(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\right\}^2 \nu(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}.$$

Now we expand in powers of ϵ and set the coefficient of ϵ, which corresponds to the functional first derivative, equal to zero, giving

$$\sum_{n=1}^{N}\int \left\{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\right\}\eta(\mathbf{x}_n + \boldsymbol{\xi})\nu(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = 0. \qquad \textbf{(207)}$$

This must hold for every choice of the variation function $\eta(\mathbf{x})$. Thus we can choose

$$\eta(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{z})$$

where $\delta(\,\cdot\,)$ is the Dirac delta function. This allows us to evaluate the integral over $\boldsymbol{\xi}$ giving

$$\sum_{n=1}^{N} \int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\} \, \delta(\mathbf{x}_n + \boldsymbol{\xi} - \mathbf{z}) \nu(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} = \sum_{n=1}^{N} \{y(\mathbf{z}) - t_n\} \, \nu(\mathbf{z} - \mathbf{x}_n).$$

Substituting this back into (207) and rearranging we then obtain the required result (6.40).

6 Example 6.23

($\star\star$) **www** Consider a Gaussian process regression model in which the target variable $\mathbf{t}$ has dimensionality D. Write down the conditional distribution of $\mathbf{t}_{N+1}$ for a test input vector $\mathbf{x}_{N+1}$, given a training set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{N+1}$ and corresponding target observations $\mathbf{t}_1, \ldots, \mathbf{t}_N$.

Solution

NOTE: In the 1^{st} printing of PRML, a typographical mistake appears in the text of the exercise at line three, where it should say ". a training set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$".

If we assume that the target variables, $t_1, \ldots, t_D$, are independent given the input vector, $\mathbf{x}$, this extension is straightforward.

Using analogous notation to the univariate case,

$$p(\mathbf{t}_{N+1}|\mathbf{T}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{m}(\mathbf{x}_{N+1}), \sigma(\mathbf{x}_{N+1})\mathbf{I}),$$

where $\mathbf{T}$ is a $N \times D$ matrix with the vectors $\mathbf{t}_1^{\mathrm{T}}, \ldots, \mathbf{t}_N^{\mathrm{T}}$ as its rows,

$$\mathbf{m}(\mathbf{x}_{N+1})^{\mathrm{T}} = \mathbf{k}^{\mathrm{T}}\mathbf{C}_N\mathbf{T}$$

and $\sigma(\mathbf{x}_{N+1})$ is given by (6.67). Note that $\mathbf{C}_N$, which only depend on the input vectors, is the same in the uni- and multivariate models.

7 Example 6.25

(⋆) **www** Using the Newton-Raphson formula (4.92), derive the iterative update formula (6.83) for finding the mode $\mathbf{a}_N^\star$ of the posterior distribution in the Gaussian process classification model.

Solution

Substituting the gradient and the Hessian into the Newton-Raphson formula we obtain

$$\begin{aligned}
\mathbf{a}_N^{\text{new}} &= \mathbf{a}_N + (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1}\mathbf{a}_N\right] \\
&= (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\right] \\
&= \mathbf{C}_N(\mathbf{I} + \mathbf{W}_N\mathbf{C}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\right]
\end{aligned}$$