

Now consider a data set of inputs $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with corresponding target values t_1, \dots, t_N . We group the target variables $\{t_n\}$ into a column vector that we denote by \mathbf{t} where the typeface is chosen to distinguish it from a single observation of a multivariate target, which would be denoted \mathbf{t} . Making the assumption that these data points are drawn independently from the distribution (4.8), we obtain an expression for the likelihood function, which is a function of the adjustable parameters \mathbf{w} and σ^2 :

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \sigma^2) \quad (4.9)$$

where we have used (4.3). Taking the logarithm of the likelihood function and making use of the standard form (2.49) for the univariate Gaussian, we have

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) &= \sum_{n=1}^N \ln \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \sigma^2) \\ &= -\frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) - \frac{1}{\sigma^2} E_D(\mathbf{w}) \end{aligned} \quad (4.10)$$

where the sum-of-squares error function is defined by

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2. \quad (4.11)$$

The first two terms in (4.10) can be treated as constants when determining \mathbf{w} because they are independent of \mathbf{w} . Therefore, as we saw previously, maximizing the likelihood function under a Gaussian noise distribution is equivalent to minimizing the sum-of-squares error function (4.11).

Section 2.3.4

4.1.3 Maximum likelihood

Having written down the likelihood function, we can use maximum likelihood to determine \mathbf{w} and σ^2 . Consider first the maximization with respect to \mathbf{w} . The gradient of the log likelihood function (4.10) with respect to \mathbf{w} takes the form

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\} \phi(\mathbf{x}_n)^T. \quad (4.12)$$

Setting this gradient to zero gives

$$0 = \sum_{n=1}^N t_n \phi(\mathbf{x}_n)^T - \mathbf{w}^T \left(\sum_{n=1}^N \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right). \quad (4.13)$$

Solving for \mathbf{w} we obtain

$$\mathbf{w}_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}, \quad (4.14)$$

which are known as the *normal equations* for the least-squares problem. Here Φ is an $N \times M$ matrix, called the *design matrix*, whose elements are given by $\Phi_{nj} = \phi_j(\mathbf{x}_n)$, so that

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}. \quad (4.15)$$

The quantity

$$\Phi^\dagger \equiv (\Phi^T \Phi)^{-1} \Phi^T \quad (4.16)$$

is known as the *Moore–Penrose pseudo-inverse* of the matrix Φ (Rao and Mitra, 1971; Golub and Van Loan, 1996). It can be regarded as a generalization of the notion of a matrix inverse to non-square matrices. Indeed, if Φ is square and invertible, then using the property $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ we see that $\Phi^\dagger \equiv \Phi^{-1}$.

At this point, we can gain some insight into the role of the bias parameter w_0 . If we make the bias parameter explicit, then the error function (4.11) becomes

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \left\{ t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}_n) \right\}^2. \quad (4.17)$$

Setting the derivative with respect to w_0 equal to zero and solving for w_0 , we obtain

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j \bar{\phi}_j \quad (4.18)$$

where we have defined

$$\bar{t} = \frac{1}{N} \sum_{n=1}^N t_n, \quad \bar{\phi}_j = \frac{1}{N} \sum_{n=1}^N \phi_j(\mathbf{x}_n). \quad (4.19)$$

Thus, the bias w_0 compensates for the difference between the averages (over the training set) of the target values and the weighted sum of the averages of the basis function values.

We can also maximize the log likelihood function (4.10) with respect to the variance σ^2 , giving

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N \{ t_n - \mathbf{w}_{\text{ML}}^T \boldsymbol{\phi}(\mathbf{x}_n) \}^2, \quad (4.20)$$

and so we see that the maximum likelihood value of the variance parameter is given by the residual variance of the target values around the regression function.