Chapter 3 Linear Models For Regression

Yue Yu

3.1 Linear Basis Function Models

(3.8)

Recall $t = y(\mathbf{x}, \mathbf{w}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \beta^{-1})$. This is equivalent to

$$\begin{aligned} p(\epsilon; \beta) &= \frac{\beta}{\sqrt{2\pi}} \exp\left\{ -\frac{\beta}{2}\epsilon^2 \right\} \\ &= \frac{\beta}{\sqrt{2\pi}} \exp\left\{ -\frac{\beta}{2}(t - y(\mathbf{x}, \mathbf{w}))^2 \right\}, \end{aligned}$$

which implies that

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

(3.13)

This equation should be

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n), \tag{*}$$

because

$$\nabla_{\mathbf{w}}(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)) = \boldsymbol{\phi}(\mathbf{x}_n).$$

(3.14)

According to (*), this equation should be

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

(3.15)

By defining a design matrix $\boldsymbol{\Phi}$ in the form of (3.16), (**) can be reduced to

$$\mathbf{0} = \boldsymbol{\Phi}\mathbf{t} - \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{w}.$$

Solving for $\mathbf{w}$, we obtain

$$\mathbf{w}_{\mathrm{ML}} = (\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi} \mathbf{t}.$$

(3.19)

$$\begin{aligned} \frac{\partial}{\partial w_0} E_D(\mathbf{w}) &= -\sum_{n=1}^{N} \left\{ t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}_n) \right\} \\ &= 0. \end{aligned}$$

Solving for w_0, we obtain

$$
\begin{aligned}
w_0 &= \frac{1}{N}\sum_{n=1}^{N} t_n - \frac{1}{N}\sum_{j=1}^{M-1} w_j \sum_{n=1}^{N} \phi_j(\mathbf{x}_n) \\
&= \bar{t} - \sum_{j=1}^{M-1} w_j \overline{\phi_j},
\end{aligned}
$$

where

$$
\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n, \qquad \overline{\phi_j} = \frac{1}{N}\sum_{n=1}^{N} \phi_j(\mathbf{x}_n).
$$

(3.23)

$$
\begin{aligned}
\nabla_{\mathbf{w}} E_n(\mathbf{w}) &= -(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n) \\
&= -(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)\boldsymbol{\phi}_n,
\end{aligned}
$$

where $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$. Plug it into (3.22), we obtain (3.23).

(3.28)

Recall the regularized error function

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\
&= \frac{1}{2}\big(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\big)^{\mathrm{T}}\big(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\big) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}.
\end{aligned}
$$

Setting the gradient with respect to $\mathbf{w}$ to $\mathbf{0}$, we have

$$
\begin{aligned}
\nabla_{\mathbf{w}} E(\mathbf{w}) &= \frac{1}{2}\nabla_{\mathbf{w}}\big(\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w}\big) \\
&= \big(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda\mathbf{I}\big)\mathbf{w} - \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \\
&= \mathbf{0}.
\end{aligned}
$$

Solving for $\mathbf{w}$, we obtain

$$
\mathbf{w} = \big(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\big)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.
$$

(3.40)

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\big[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\big] \\
&= \mathbb{E}_{\mathcal{D}}\big[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2\big] \\
&= \mathbb{E}_{\mathcal{D}}\big[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 + 2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}\big] \\
&= \mathbb{E}_{\mathcal{D}}\big[\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2\big] + \mathbb{E}_{\mathcal{D}}\big[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\big] + 2\mathbb{E}_{\mathcal{D}}\big[y(\mathbf{x};\mathcal{D})\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - y(\mathbf{x};\mathcal{D})h(\mathbf{x}) - \\
&\quad \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]^2 + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]h(\mathbf{x})\big] \\
&= \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 + \mathbb{E}_{\mathcal{D}}\big[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\big] + 2\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]^2 - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]h(\mathbf{x}) - \\
&\quad \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]^2 + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]h(\mathbf{x})\} \\
&= \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 + \mathbb{E}_{\mathcal{D}}\big[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\big].
\end{aligned}
$$