Chapter 6 Kernel Methods

Yue Yu

6.1 Dual Representations

(6.3)

By setting the gradient of $J(\mathbf{w})$ with respect to $\mathbf{w}$ to $\mathbf{0}$, it is easy to see that

$$\begin{aligned}
\mathbf{w} &= \arg_{\mathbf{w}} \nabla_{\mathbf{w}} J(\mathbf{w}) = \mathbf{0} \\
&= \arg_{\mathbf{w}} \left(\sum_{n=1}^{N} \left\{ \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) - t_n \right\} + \lambda \mathbf{w} = \mathbf{0} \right) \\
&= -\frac{1}{\lambda} \sum_{n=1}^{N} \left\{ \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) - t_n \right\} \boldsymbol{\phi}(\mathbf{x}_n) \\
&= \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{a}
\end{aligned}$$

where

$$\mathbf{a} = -\frac{1}{\lambda}(\boldsymbol{\Phi}\mathbf{w} - \mathbf{t}).$$

(6.5)

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

(6.8)

From the derivation of (6.3), we have

$$\begin{aligned}
\mathbf{a} &= -\frac{1}{\lambda}(\boldsymbol{\Phi}\mathbf{w} - \mathbf{t}) \\
&= -\frac{1}{\lambda}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{a} - \mathbf{t}) \\
&= -\frac{1}{\lambda}(\mathbf{K}\mathbf{a} - \mathbf{t}).
\end{aligned}$$

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

$$\mathbf{a} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \mathbf{t}.$$

(6.9)

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}) = (\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}))^{\mathrm{T}}\mathbf{a} = \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} + \lambda\mathbf{I}_N)^{-1}\mathbf{t}.$$