Chapter 7 Sparse Kernel Machines

Yue Yu

7.1 Maximum Margin Classifiers

(7.8)

Setting the derivative of $L(\mathbf{w}, b, \mathbf{a})$ with respect to $\mathbf{w}$ to $\mathbf{0}$, we have

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \mathbf{a}) = \mathbf{w} - \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\mathbf{x}_n) = \mathbf{0}.$$

Rearranging the equation, we obtain

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\mathbf{x}_n).$$

(7.9)

Setting the derivative of $L(\mathbf{w}, b, \mathbf{a})$ with respect to b to 0, we have

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \mathbf{a}) = -\sum_{n=1}^{N} a_n t_n = 0,$$

that is

$$0 = \sum_{n=1}^{N} a_n t_n.$$

(7.10)

Substituting (7.8) and (7.9) in (7.7), we obtain

$$\begin{aligned}
\tilde{L}(\mathbf{a}) &= \frac{1}{2}\left(\sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\mathbf{x}_n)\right)^2 - \sum_{n=1}^{N} a_n \left\{ t_n \left(\sum_{m=1}^{N} a_m t_m \boldsymbol{\phi}(\mathbf{x}_m)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) + b \right) - 1 \right\} \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_m) - \sum_{n=1}^{N} a_n t_n \sum_{m=1}^{N} a_m t_m \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_m) - b \sum_{n=1}^{N} a_n t_n + \sum_{n=1}^{N} a_n \\
&= \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_m) \\
&= \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m).
\end{aligned}$$

(7.17)

For any support vector $\mathbf{x}_n$, we have $t_n y(\mathbf{x}_n) = 1$, which can be expanded in the form of

$$\begin{aligned} t_n y(\mathbf{x}_n) &= t_n(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) + b) \\ &= t_n\left(\sum_{m=1}^{N} a_m t_m \boldsymbol{\phi}(\mathbf{x}_m)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) + b\right) \\ &= t_n\left(\sum_{m=1}^{N} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b\right) \\ &= 1. \end{aligned}$$

According to the KKT conditions, either $a_m = 0$ or $t_m y(\mathbf{x}_m) = 1$ must hold. In other words, the data point $\mathbf{x}_m$ either vanishes or is a support vector. Hence, we have

$$t_n\left(\sum_{m\in\mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b\right) = 1$$

where $\mathcal{S}$ denotes the set of indices of the support vectors.