Chapter 7 Sparse Kernel Machines

Yue Yu

7.1 Maximum Margin Classifiers

(7.8)

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

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

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.

(7.29)

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

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

Rearranging the equation, we obtain

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

(7.30)

Setting the derivative of L with respect to b to 0, we have

$$\frac{\partial}{\partial b} L = -\sum_{n=1}^{N} a_n t_n = 0,$$

that is

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

(7.31)

Setting the derivative of L with respect to ξ_n to 0, we have

$$\frac{\partial}{\partial \xi_n} L = C - a_n - \mu_n = 0.$$

Rearranging the equation, we obtain

$$a_n = C - \mu_n.$$

(7.32)

Substituting (7.29), (7.30) and (7.31) back into (7.22), we obtain

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

(7.53) – (7.54)

We loose the condition for lying inside the ϵ-tube by ξ_n to the left and $\hat{\xi}_n$ to the right in Figure 7.6, that is

$$-\epsilon - \xi_n \leq y(\mathbf{x}_n) - t_n \leq \epsilon + \hat{\xi}_n,$$

which is equivalent to (7.53) and (7.54).

(7.57)

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

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{n=1}^{N} a_n \boldsymbol{\phi}(\mathbf{x}_n) + \sum_{n=1}^{N} \hat{a}_n \boldsymbol{\phi}(\mathbf{x}_n) = \mathbf{0}.$$

Rearranging the equation, we obtain

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

(7.58)

Setting the derivative of L with respect to b to 0, we have

$$\frac{\partial}{\partial b} L = -\sum_{n=1}^{N} a_n + \sum_{n=1}^{N} \hat{a}_n = 0,$$

that is

$$\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0.$$

(7.59)

Setting the derivative of L with respect to ξ_n to 0, we have

$$\frac{\partial}{\partial \xi_n} L = C - \mu_n - a_n = 0,$$

that is

$$a_n + \mu_n = C.$$

(7.60)

Similar to (7.59), we have

$$\hat{a}_n + \hat{\mu}_n = C.$$

(7.61)

Substituting (7.57) – (7.60) back into (7.56), we obtain

$$\begin{aligned}
\hat{L}(\mathbf{a}, \hat{\mathbf{a}}) = C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n) + \frac{1}{2} \Bigg(\sum_{n=1}^{N} (a_n - \hat{a}_n) \boldsymbol{\phi}(\mathbf{x}_n) \Bigg)^2 - \sum_{n=1}^{N} (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) \\
- \sum_{n=1}^{N} a_n \Bigg(\epsilon + \xi_n + \sum_{m=1}^{N} (a_m - \hat{a}_m) \boldsymbol{\phi}(\mathbf{x}_m)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) + b - t_n \Bigg) \\
- \sum_{n=1}^{N} \hat{a}_n \Bigg(\epsilon + \hat{\xi}_n - \sum_{m=1}^{N} (a_m - \hat{a}_m) \boldsymbol{\phi}(\mathbf{x}_m)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) - b + t_n \Bigg) \\
= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (a_n - \hat{a}_n)(a_m - \hat{a}_m) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_m) - \sum_{n=1}^{N} \sum_{m=1}^{N} (a_n - \hat{a}_n)(a_m - \hat{a}_m) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_m) \\
- \epsilon \sum_{n=1}^{N} (a_n + \hat{a}_n) + \sum_{n=1}^{N} (a_n - \hat{a}_n) t_n + \sum_{n=1}^{N} (C - \mu_n - a_n) \xi_n + \sum_{n=1}^{N} (C - \hat{\mu}_n - \hat{a}_n) \hat{\xi}_n \\
= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (a_n - \hat{a}_n)(a_m - \hat{a}_m) k(\mathbf{x}_n, \mathbf{x}_m) - \epsilon \sum_{n=1}^{N} (a_n + \hat{a}_n) + \sum_{n=1}^{N} (a_n - \hat{a}_n) t_n.
\end{aligned}$$

7.2 Relevance Vector Machines

Skip reading for now.