Chapter *9* Support Vector Machine

Siheng Zhang
zhangsiheng@cvte.com

2022 年 8 月 16 日

This part corresponds to **Chapter ? of PRML, Chapter of UML**. It mainly introduces support vector machines (SVMs), a class of model built from linear model with **margin**, and can be extended to non-linear case.

目录

1 Hard SVM 2

2 Kernel trick 2

2.1 Representor Theorem 2

3 Soft SVM 3

1 Hard SVM

$$\min \frac{1}{2}\mathbf{w}^\top \mathbf{w}, \quad s.t.\ y_i \mathbf{w}^\top \mathbf{x}_i \geq 1, \forall i \in \{1, ..., m\} \tag{1}$$

Using Lagrange multipliers $\lambda_i, i \in \{1, ..., m\}$, the above objective can be written as:

$$\min \mathcal{L} = \frac{1}{2}\mathbf{w}^\top \mathbf{w} + \sum_{i=1}^{m} \lambda_i (1 - y_i \mathbf{x}_i \mathbf{w}_i)$$

Taking derivatives *w.r.t* $\mathbf{w}$ and letting it to be zero leads to

$$\mathbf{w} = \sum_{i=1}^{m} \lambda_i y_i \mathbf{x}_i$$

Bring it back to $\mathcal{L}$, we have

$$\mathcal{L} = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \mathbf{x}_i \mathbf{x}_j$$

Take derivatives *w.r.t* λ_i and letting it to be zero leads to

$$y_i \mathbf{x}_i \sum_{j=1} \lambda_i y_j \mathbf{x}_j = 1$$

Hence,

$$\mathbf{w}^\top \mathbf{w} = \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \mathbf{x}_i \mathbf{x}_j = \sum_{i=1}^{m} \lambda_i$$

2 Kernel trick

2.1 Representor Theorem

Theorem 1 *(Nonparametric Representor Theorem) Given a nonempty set $\mathcal{X}$, a positive definite real-valued kernel k on $\mathcal{X} \times \mathcal{X}$, a training set $S = (\vec{x}_i, y_i), i = 1, ..., m \in \mathcal{X} \times \mathcal{R}$, a strictly monotonically increasing real-valued function g on $[0, \infty]$, and an arbitrary cost function $c : (\mathcal{R} \times \mathcal{R})^m \to \mathcal{R}$, and a class of hypothesis*

$$\mathcal{F} = \left\{ f \in \mathcal{R}^{\mathcal{X}} | f(\cdot) = \sum_{i=1}^{\infty} \beta_i k(\cdot, z_i), \beta_i \in \mathcal{R}, z_i \in \mathcal{X}, \|f\| < \infty \right\}$$

Here, $\|\cdot\|$ is the norm in the RKHS. Then, any $f \in \mathcal{F}$ minimizing the regularized risk function

$$c((y_1, f(\vec{x}_1)), ..., (y_m, f(\vec{x}_m))) + g(\|f\|)$$

admits a representation of the form:

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \tag{2}$$

remark1. If we discarded the strictness of the monotonicity of g, it would no longer follow that each minimizer (there might be multiple minimizers) admits such an expansion. However, it would still follow that there is always one minimizer that DOES admit the expansion.

Proof *Any $f \in \mathcal{F}$ can be decomposed into a part that lives in the span of the $\phi(x_i), i = 1, ..., m$ and another part that*

lives in its ortho-complement space (thus the two parts are orthogonal), i.e.,

$$f = \sum_{i=1}^{m} \alpha_i \phi(\vec{x}_i) + \vec{v}$$

in which $\vec{v} \cdot \phi(\vec{x}) = 0, \forall \vec{x}$. *Then for arbitrary point* $\vec{x}$, *the first term of loss yields*

$$f(\vec{x}_j) = \left(\sum_{i=1}^{m} \alpha_i \phi(\vec{x}_i) + \vec{v} \right) \cdot \phi(\vec{x}) = \left(\sum_{i=1}^{m} \alpha_i \phi(\vec{x}_i) \right) \cdot \phi(\vec{x})$$

The second term of loss yields

$$g(\|f\|) = g\left(\left(\sum_{i=1}^{m} \alpha_i \phi(\vec{x}_i) + \vec{v} \right) \cdot \left(\sum_{i=1}^{m} \alpha_i \phi(\vec{x}_i) + \vec{v} \right) \right) = g\left(\| \sum_{i=1}^{m} \alpha_i \phi(\vec{x}_i) \| + \|\vec{v}\| \right) \geq g\left(\| \sum_{i=1}^{m} \alpha_i \phi(\vec{x}_i) \| \right)$$

Hence, any minimizer must have $\vec{v} = 0$, *which conclude the proof.*

3 Soft SVM

For non-linearly separable case, build new data $\mathbf{z}_i = (\mathbf{x}_i, \rho e_i) \in \mathbb{R}^{d+m}$, where $\rho > 0$ and e_i is the $m-$dimensional vector all of whose components are zero except for the $i-$th component which is equal to 1. Obviously, the dataset $\{(\mathbf{z}_i, y_i)\}_{i=1}^{m}$ is linearly separable. Again, apply hard SVM on it.

$$\min \frac{1}{2} \mathbf{v}^\top \mathbf{v}, \quad s.t.\ y_i \mathbf{v}^\top \mathbf{z}_i \geq 1, \forall i \in \{1, ..., m\}$$

Denote $\mathbf{v} = (\mathbf{w}, \xi_1/(\rho * y_1), \cdots, \xi_m/(\rho * y_m))$, in which $\mathbf{w} \in \mathbb{R}^d$, using Lagrange multipliers the objective can be written as

$$\min \frac{1}{2} \mathbf{w}^\top \mathbf{w} + \frac{1}{2\rho^2} \sum_{i=1}^{m} \xi_i^2 + \sum_{i=1}^{m} \lambda_i (1 - \xi_i - y_i \mathbf{w}^\top \mathbf{x}_i)$$

which is equivalent to

$$\min \mathbf{w}^\top \mathbf{w} + \frac{1}{2\rho^2} \sum_{i=1}^{m} \xi_i^2, \quad s.t.\ y_i \mathbf{w}^\top \mathbf{x}_i \geq 1 - \xi_i, \forall i \in \{1, ..., m\}$$