SVM

Shumin An

June 2019

1 Constraint Optimization

In order to solving SVM problems, we need the Lagrange Multiplier Method, so here's a small introduction of LMM. In general, we could write the problem in the way as follows:

$$\begin{aligned} &\min_{x\in\mathbb{R}^{'}} f(x) \\ &s.t.\ m_i(x) \leq 0, i = 1, 2, \cdots, M \\ &\quad\quad n_j(x) = 0, j = 1, 2, \cdots, N \end{aligned} \tag{1}$$

Define Lagrange function:

$$L(x, \lambda, \eta) = f(x) + \sum_{i=1}^{M} \lambda_i m_i(x) + \sum_{i=1}^{N} \eta_i n_i(x)$$

So the original one equals:

$$\min_{x\in\mathbb{R}^p} \max_{\lambda,\eta} L(x, \lambda, \eta)\ s.t.\ \lambda_i \geq 0$$

That's because when subjected to the constrains in (1), $\lambda_i = 0$ has the max value. If not, the maximum value becomes infinity, which obviously not the minimum value we want to get.

2 Hard-Margin SVM

The purpose is to find the largest distance, and the distance here means the distance between hyperplane and data. So the function could be represented as:

$$\underset{w,b}{argmax}[\min_i \frac{|w^T x_i + b|}{||w||}]\ s.t.\ y_i(w^T x_i + b) > 0$$

$$\Longrightarrow \underset{w,b}{argmax}[\min_i \frac{y_i(w^T x_i + b)}{||w||}]\ s.t.\ y_i(w^T x_i + b) > 0$$

Since if we changed the parameter of hyperplane to scale the hyperplane, it won't changed. So we define: $\min y_i(w^T x_i + b) = 1 > 0$. Then we get:

$$\underset{w,b}{argmin}\, \frac{1}{2} w^T w \ s.t. \ \min_i y_i(w^T x_i + b) = 1$$

$$\Rightarrow \underset{w,b}{argmin}\, \frac{1}{2} w^T w \ s.t. \ y_i(w^T x_i + b) \geq 1, i = 1, 2, \cdots, N$$

Here comes to a convex optimization problem which contains N constrains. And we could solve it by computer. However, if considering high-dimension situation, the problem is hard to solve. Therefore, we use the Lagrange function:

$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_{i=1}^{N} \lambda_i (1 - y_i(w^T x_i + b))$$

The primal problem becomes:

$$\underset{w,b}{argmin}\, \max_{\lambda} L(w, b, \lambda_i) \ s.t. \ \lambda_i \geq 0$$

We then exchange the max and min:

$$\max_{\lambda_i} \min_{w,b} L(w, b, \lambda_i) \ s.t. \ \lambda_i \geq 0$$

Since the constraint is a affine function, the dual problem equals the original one. So:

$$b : \frac{\partial}{\partial b} L = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0$$

For w:

$$L(w, b, \lambda_i) = \frac{1}{2} w^T w + \sum_{i=1}^{N} \lambda_i (1 - y_i w^T x_i - y_i b) = \frac{1}{2} w^T w + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i y_i w^T x_i$$

$$\frac{\partial}{\partial w} L = 0 \Rightarrow w = \sum_{i=1}^{N} \lambda_i y_i x_i$$

So we have L:

$$L(w, b, \lambda_i) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \lambda_i$$

The dual problem is:

$$\max_{\lambda} -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \lambda_i, \ s.t. \ \lambda_i \geq 0$$

According to KKT:

$$\begin{aligned}
&\frac{\partial L}{\partial w} = 0, \frac{\partial L}{\partial b} = 0 \\
&\lambda_k(1 - y_k(w^T x_k + b)) = 0 \\
&\lambda_i \geq 0 \\
&1 - y_i(w^T x_i + b) \leq 0
\end{aligned} \tag{2}$$

Finally we get the parameter:

$$\hat{w} = \sum_{i=1}^{N} \lambda_i y_i x_i$$

$$\hat{b} = y_k - w^T x_k = y_k - \sum_{i=1}^{N} \lambda_i y_i x_i^T x_k, \exists k, 1 - y_k(w^T x_k + b) = 0$$

3 Soft-Margin SVM

The Soft-Margin is a kind fault-tolerance I think. The purpose is to consider the possibility of mis-classification. The number of mis-classification is:

$$error = \sum_{i=1}^{N} \mathbb{I}\{y_i(w^T x_i + b)1\}$$

But this function is not continuous. So we changed it into another form:

$$error = \sum_{i=1}^{N} \max\{0, 1 - y_i(w^T x_i + b)\}$$

So the Hard-Margin SVM becomes:

$$\underset{w,b}{argmin} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \max\{0, 1 - y_i(w^T x_i + b)\} \; s.t. \; y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, \cdots, N$$

In this expression, constant C could be seen as the level of "fault-tolerance". And we could simplify the expression further:

$$\underset{w,b}{argmin} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \; s.t. \; y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \cdots, N$$

$\xi_i = 1 - y_i(w^T x_i + b)$