SVM (Support Vector Machine)

SVM is abbreviated from Support vector machine. It has both implemation on supervise-learning and un-supervised learning. It is a really powerful machine learning techniques.

Lagrange and KKT method explaination

Lagrange equation is most important part in SVM algorithm. It can easily solve the problem on "subject to" problem.

$$\begin{gathered}\min f = 2x_1^2 + 3x_2^2 + 7x_3^2 \\ \text{s.t.} \quad 2x_1 + x_2 = 1 \\ 2x_2 + 3x_3 = 2\end{gathered} \tag{1}$$

If we want slove the problem on first equation, we can easily take the sub-differential on x_1 to x_3, but for here, we have restriction on two equations, we can not take derivative directly.

Instead, we can change the restriction equations and multiply some parameters α_1 and α_2, then combine all equations.

$$\min f = 2x_1^2 + 3x_2^2 + 7x_3^2 + \alpha_1(2x_1 + x_2 - 1) + \alpha_2(2x_2 + 3x_3 - 2) \tag{2}$$

then take derivative:

$$\begin{gathered}\frac{\partial f}{\partial x_1} = 4x_1 + 2\alpha_1 = 0 \Rightarrow x_1 = -0.5\alpha_1 \\ \frac{\partial f}{\partial x_2} = 6x_2 + \alpha_1 + 2\alpha_2 = 0 \Rightarrow x_2 = -\frac{\alpha_1 + 2\alpha_2}{6} \\ \frac{\partial f}{\partial x_3} = 14x_3 + 3\alpha_2 = 0 \Rightarrow x_3 = -\frac{3\alpha_2}{14}\end{gathered} \tag{3}$$

It will the same as if the subjective has unequivalent equation. Like the example here,

$$\begin{gathered}\min f = x_1^2 - 2x_1 + 1 + x_2^2 + 4x_2 + 4 \\ \text{s.t.} \quad x_1 + 10x_2 > 10 \\ 10x_1 - 10x_2 < 10\end{gathered} \tag{4}$$

simply chage all un-equivalent to less than 0 on right side.

$$\begin{gathered}\text{s.t. } 10 - x_1 - 10x_2 < 0 \\ 10x_1 - x_2 - 10 < 0\end{gathered}$$

Then use the same way to combine the equation:

$$\begin{gathered}L(x, \alpha) = f(x) + \alpha_1 g1(x) + \alpha_2 g2(x) \\ = x_1^2 - 2x_1 + 1 + x_2^2 + 4x_2 + 4 + \alpha_1\left(10 - x_1 - 10x_2\right) + \\ \alpha_2\left(10x_1 - x_2 - 10\right)\end{gathered} \tag{5}$$

Then the problem becomes

$$L(x, \alpha, \beta) = f(x) + \sum \alpha_i g_i(x) + \sum \beta_i h_i(x) \tag{6}$$

Here, we convert the problem to solve the equation above, and meet all the requirements below:

1. $L(x, \alpha, \beta)$ has sub-differential to all $x_i, i \in (1, N)$
2. $h(x) = 0$
3. $\sum \alpha_i g_i(x) = 0, \alpha_i \geq 0$

Follow up the previous complicated equation, we take the derivative on both x_1 and x_2

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= 2x_1 - 2 - \alpha_1 + 10\alpha_2 = 0 \Rightarrow x_1 = 0.5\left(\alpha_1 - 10\alpha_2 + 2\right) \\ \frac{\partial L}{\partial x_2} &= 2x_2 + 4 - 10\alpha_1 - \alpha_2 = 0 \Rightarrow x_2 = 0.5\left(10\alpha_1 + \alpha_2 - 4\right) \end{aligned} \tag{7}$$

The principle of SVM

In order to get a perfect line to divide all the data to two parts, in the every beginning, we can roughly draw a line. The function of line is $y = kx + b$. So, the directly distance between any node and line that we can represent to the equation:

$$d = \frac{|c_2 - c_1|}{\sqrt{w_1^2 + w_2^2}} = \frac{1}{||W||} \tag{8}$$

For the equation above, we assume all the data has two features, so we have w_1 and w_2 for genal weight. Also, it is important to get two distance d_1 and d_2. d_i is the distance between the closest nodes to linear boundary.

$$D = d1 + d2 = \frac{2}{||W||} = \frac{2}{\sqrt{W^T W}} \tag{9}$$

We want the margin D get as large as possible, so we want $max(D)$, on the other hand, the problem becomes $min(\frac{1}{2}W^T W)$

$$\text{s.t.} \quad y_i\left(W x_i + b\right) \geq 1$$

then, we transform the equation to

$$\text{s.t.} \quad 1 - y_i\left(W x_i + b\right) \leq 0 \tag{10}$$

Using the lagrange equation, we can get

$$\begin{gathered} L(w, b, \alpha) = \frac{1}{2} w^T w + \alpha_1 h_1(x) + \ldots + \alpha_n h_n(x) \\ = \frac{1}{2} w^T w - \alpha_1\left[y_1\left(w x_1 + b\right) - 1\right] - \ldots - \alpha_n\left[y_n\left(w x_n + b\right) - 1\right] \\ = \frac{1}{2} w^T w - \sum_{i=1}^N \alpha_i y_i\left(w x_i + b\right) + \sum_{i=1}^N \alpha_i \end{gathered} \tag{11}$$

The best solution on this equation is that:

$$\begin{aligned}\frac{\partial L}{\partial w}&=w-\sum_{i=1}^{N}\alpha_iy_ix_i=0\Rightarrow w=\sum_{i=1}^{N}\alpha_iy_ix_i\\ \frac{\partial L}{\partial b}&=-\sum_{i=1}^{N}\alpha_iy_i=0\Rightarrow\sum_{i=1}^{N}\alpha_iy_i=0\end{aligned}\tag{12}$$

Next, re-plugin the results to orginal equation, we can cancel out a lot of thing and get:

$$\begin{aligned}W(\alpha)=L(w,b,\alpha)&=\tfrac{1}{2}\left(\sum_{i=1}^{N}\alpha_iy_ix_i\right)^T\left(\sum_{j=1}^{N}\alpha_jy_jx_j\right)-\\ &\sum_{i=1}^{N}\alpha_iy_i\left(\left(\sum_{i=1}^{N}\alpha_iy_ix_i\right)x_i+b\right)+\sum_{i=1}^{N}\alpha_i\\ &=\tfrac{1}{2}\left(\sum_{i,j=1}^{N}\alpha_iy_i\alpha_jy_jx_i*x_j\right)-\sum_{i,j=1}^{N}\alpha_iy_i\alpha_jy_jx_i*x_j+b\sum_{i=1}^{N}\alpha_iy_i+\sum_{i=1}^{N}\alpha_i\\ &=-\tfrac{1}{2}\left(\sum_{i,j=1}^{N}\alpha_iy_i\alpha_jy_jx_i*x_j\right)+\sum_{i=1}^{N}\alpha_i\end{aligned}\tag{13}$$

and we can make the problem to be:

$$\begin{aligned}\max\quad W(\alpha)=-\tfrac{1}{2}\left(\sum_{i,j=1}^{N}\alpha_iy_i\alpha_jy_jx_i*x_j\right)+\sum_{i=1}^{N}\alpha_i\\ \text{s.t.}\quad\alpha_i\geq 0\\ \sum_{i=1}^{N}\alpha_iy_i=0\end{aligned}\tag{14}$$

For more advanced situation we can define a slack variable ϵ_i, regard the data on the other sied of line and if the distance less than ϵ, that is considered to be okay.

$$\begin{aligned}\min\tfrac{1}{2}W^TW+C\sum_{i=1}^{N}\epsilon_i\\ \text{s.t.}\quad 1+\epsilon_i-y_i\left(Wx_i+b\right)\leq 0\\ \epsilon_i\geq 0\end{aligned}\tag{15}$$

$$\begin{aligned}L(x,\alpha,\beta)=\tfrac{1}{2}W^TW-\sum_{i=1}^{N}\alpha_i\left(y_i\left(Wx_i+b\right)+\epsilon_i-1\right)+\\ C\sum_{i=1}^{N}\epsilon_i-\sum_{i=1}^{N}r_i\epsilon_i\end{aligned}\tag{16}$$

$$\begin{aligned}\frac{\partial L}{\partial w}&=w-\sum_{i=1}^{N}\alpha_iy_ix_i=0\Rightarrow w=\sum_{i=1}^{N}\alpha_iy_ix_i\\ \frac{\partial L}{\partial b}&=-\sum_{i=1}^{N}\alpha_iy_i=0\Rightarrow\sum_{i=1}^{N}\alpha_iy_i=0\\ \frac{\partial L}{\partial\epsilon_i}&=0\Rightarrow C-\alpha_i-r_i=0\end{aligned}\tag{17}$$

Then re-plug into the equation again, then we get:

$$\begin{aligned}W(\alpha)=-\frac{1}{2}\left(\sum_{i,j=1}^{N}\alpha_iy_i\alpha_jy_jx_i*x_j\right)+\sum_{i=1}^{N}\alpha_i\\ \text{s.t. }0\leq\alpha_i\leq C\\ \sum_{i=1}^{N}\alpha_iy_i=0\end{aligned}\tag{18}$$

So, for here, compare (18) to (14), the only change is that we have one more limitation C