Chapter 6 Support Vector Machine

Jinghao.Zhao

Abstract

This is the notes for support vector machine, section 6.1 to 6.3, including formula and key points.

1 Margin and SVM

In classification, we want to find a hyperplane $\omega^T x + b$ to divided different samples, the distance between $\forall \mathbf{x}$ in sample and the hyperplane is

$$\gamma = \frac{| \boldsymbol{\omega^T x} + b |}{||\boldsymbol{\omega}||}, ||\boldsymbol{\omega}|| = \sqrt{\omega_1^2 + \omega_2^2 + ... + \omega_n^2} \tag{1}$$

Then we will have (6.2). When

$$\boldsymbol{\omega^T x_i} + b = 1 \ or \ \boldsymbol{\omega^T x_i} + b = -1 \tag{2}$$

We called such point **Support Vector**(They are actually samples). And the sum of distance between two support vector of two different class and hyperplane is $\frac{2}{||\boldsymbol{\omega}||}$ which is called **margin**. SVM is try to find the maximum margin to plot the hyperplane. Based on (6.3) and (6.5)

$$\max \frac{2}{||\boldsymbol{\omega}||} \Rightarrow \min ||\boldsymbol{\omega}|| \Rightarrow \min ||\boldsymbol{\omega}||^2, \tag{3}$$

since $||\boldsymbol{\omega}||$ always larger than 0.

2 Dual Problem

We derivative (6.8),

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\omega}} &= \boldsymbol{\omega} - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Rightarrow (6.9) \\ \frac{\partial L}{\partial b} &= -\sum_{i=1}^{m} \alpha_i y_i = 0 \Rightarrow (6.10) \end{aligned} \tag{4}$$

Then put (6.9) (6.10) into (6.8)[1],

$$
\begin{aligned}
L(\boldsymbol{\omega}, b, a) = \frac{1}{2}||\boldsymbol{\omega}|| + \sum_{i=1}^{m} \alpha_i (1 - y_i(\boldsymbol{\omega^T x_i} + b)) \\
= \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{\omega^T x_i} - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i, \sum_{i=1}^{m} \alpha_i y_i = 0 \\
= \sum_{i=1}^{m} \alpha_i + \frac{\boldsymbol{\omega}^T\boldsymbol{\omega}}{2} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{\omega^T x_i} \\
= \sum_{i=1}^{m} \alpha_i + \frac{\boldsymbol{\omega}^T\boldsymbol{\omega}}{2} - \boldsymbol{\omega}^T\boldsymbol{\omega} \\
= \sum_{i=1}^{m} \alpha_i - \frac{\boldsymbol{\omega}^T\boldsymbol{\omega}}{2} \\
= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i})^T \sum_{j=1}^{m} \alpha_j y_j \boldsymbol{x_j} \\
= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j}
\end{aligned} \tag{5}
$$

Then we can get (6.11).

2.1 KKT条件

The construction of KKT has mainly three part:

1. The Lagrange multiplier should ≥ 0;
2. The constraints for original function should be meet;
3. The production of Lagrange multiplier and constraints should be equal to 0.

2.2 SMO Optimization

对于不满足KKT条件的α, 采取SMO序列最小优化进行更新，具体步骤如下：

1. 选取两个变量α_i, α_j, 固定其他参数的值;
2. 重写约束$\alpha_i + \alpha_j = c \ and \ \alpha_i, \alpha_j \geq 0$;
3. 由上式消去 α_j, 从而转化为α_i 的单变量二次规划且有$\alpha_i \geq 0$;
4. 将α_i, α_j更新至收敛，从而根据式(6.9)计算$\boldsymbol{\omega}$;
5. 根据式(6.18)计算b, 从而获得超平面.

[1] https://blog.csdn.net/BIT_666/article/details/79865225

3 Kernel Function

When current sample space is not linear separable(线性可分), we can reflect sample to a higher dimensional space and make it linear separable in the new space. Kernel function is the reflection of original $\boldsymbol{x}$.

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j), \tag{6}$$

where $\phi(\boldsymbol{x})$ is the feature vector reflected by $\boldsymbol{x}$. The common types and features of kernel function are shown in table(6.1) and equation(6.25)-(6.27) in book.

Support we have input space χ, $\kappa(\cdot, \cdot)$ is symmetric function on $\chi \times \chi$, then κ is kernel function $\Leftrightarrow$(if and only if) for $D = \boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_m}$, the kernel matrix $\boldsymbol{K}$ is always semidefinite(半正定，即有n阶方阵A, 若对任意非零向量X, 均有$A^T \cdot X \cdot A \geq 0$).

4 Soft Margin and Regularization

We tried to find a hyperplane which makes the dataset linear separable, but it may caused over-fitting. Soft margin can solve it by allowing SVM has classification error on some samples, those samples are not meet the constraint $y_i(\boldsymbol{\omega^T x} + b) \geq 1$. The optimization function will be (6.29), we use surrogate loss(替代损失) to replace loss function.

$$\begin{aligned} hinge\ loss &: l_{hinge}(z) = \max(0, 1 - z) \\ exponential\ loss &: l_{exp}(z) = \exp(-z) \\ logistic\ loss &: l_{log}(z) = \log(1 + \exp(-z)) \end{aligned} \tag{7}$$

And the figure of three loss:

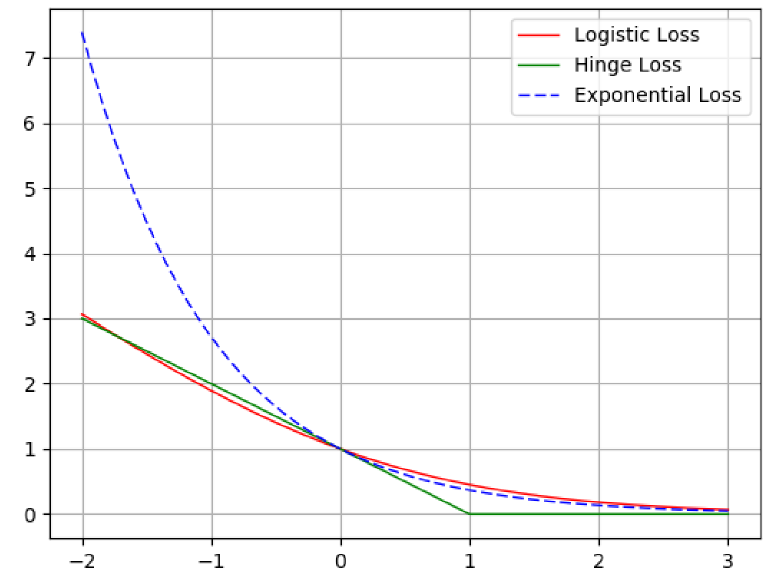

Figure 1: Loss Function

(6.29) can be transfer to (6.35) by adding slack variables ξ_i ($\xi_i \geq 0$). Based on constraints in (6.35), the Lagrange function is (6.36) (即用原式加上拉格朗日乘子 * 约束， 其中拉格朗日乘子的个数等于约束个数). We derivative (6.36) for $(\boldsymbol{\omega}, b, \xi_i)$ and get (6.37) - (6.39)[2].

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \mu_i = 0 \Rightarrow (6.39) \tag{8}$$

Put (6.37) - (6.39) into (6.36) then the dual problem of (6.35) is:

$$\begin{aligned}
L = (6.36) &= \underline{\frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\boldsymbol{\omega}^T \boldsymbol{x} + b))} + \underline{(\mu_i + \alpha_i) \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \xi_i - \sum_{i=1}^{m} \mu_i \xi_i}, \\
where\ \frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\boldsymbol{\omega}^T \boldsymbol{x} + b)) &= (6.11) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T} \boldsymbol{x_j}, \\
&- \sum_{i=1}^{m} \alpha_i \xi_i + (\mu_i + \alpha_i) \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \mu_i \xi_i = 0, \\
&C = \alpha_i + \mu_i\ and\ \alpha_i, \mu_i \geq 0 \Rightarrow (6.40)(for constraint of \alpha_i)
\end{aligned} \tag{9}$$

[2] 仅对第i项求偏导(式(8))

5 Support Vector Regression

SVR is a regression that calculate loss only if $|f(x_i) - y_i| > \epsilon$, that means their is a 2ϵ width field beside $f(\boldsymbol{x})$ which is considered to be correctly classified. By using previous Lagrange multiplier method.