SVM(Support Vector Machine)

by Xing Chao

首先简要介绍SVM分类器，然后再结合凸优化相关理论作进一步分析。

1 SVM分类器

1.1 最大间隔分离超平面

线性可分的训练样例

$$T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\}$$

其中$\boldsymbol{x}_i\in\mathbb{R}^n, y_i\in\{+1,-1\}, i=1,2,\cdots,N$

超平面

$$\begin{aligned}\boldsymbol{w}\cdot\boldsymbol{x}+b &= 0\\ \frac{\boldsymbol{w}\cdot\boldsymbol{x}+b}{\|\boldsymbol{w}\|} &= 0\end{aligned}$$

$\boldsymbol{x}_i$到超平面距离

$$\gamma_i = \frac{y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)}{\|\boldsymbol{w}\|}$$

当$\boldsymbol{w}\cdot\boldsymbol{x}_i+b=\pm1$时，$\gamma_i=\frac{1}{\|\boldsymbol{w}\|}$。

最大间隔分离超平面问题：

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\| \qquad\qquad s.t.\quad y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geqslant 1$$

有约束的原始目标函数转换为无约束的新构造的拉格朗日目标函数

$$L(\boldsymbol{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{N}\alpha_i(y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1)$$

其中$\alpha_i\geqslant 0$为拉格朗日乘子。原约束问题等价于

$$\min_{\boldsymbol{w},b}\max_{\alpha_i\geqslant 0}L(\boldsymbol{w},b,\boldsymbol{\alpha})$$

可转换为对偶问题：

$$\max_{\alpha_i\geqslant 0}\min_{\boldsymbol{w},b}L(\boldsymbol{w},b,\boldsymbol{\alpha})$$

由

$$\min_{\boldsymbol{w},b}L(\boldsymbol{w},b,\boldsymbol{\alpha})\leqslant L(\boldsymbol{w},b,\boldsymbol{\alpha})\leqslant\max_{\alpha_i\geqslant 0}L(\boldsymbol{w},b,\boldsymbol{\alpha})$$

得：

$$\max_{\alpha_i\geqslant 0}\min_{\boldsymbol{w},b}L(\boldsymbol{w},b,\boldsymbol{\alpha})\leqslant\min_{\boldsymbol{w},b}\max_{\alpha_i\geqslant 0}L(\boldsymbol{w},b,\boldsymbol{\alpha})$$

1.2 拉格朗日对偶

KKT(Karush-Kuhn-Tucker)条件

$$\begin{aligned}\alpha_i &\geqslant 0\\ y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1 &\geqslant 0\\ \alpha_i(y(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1) &= 0\end{aligned}$$

由

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{w}} &= 0\\ \frac{\partial L}{\partial b} &= 0\end{aligned}$$

得：

$$\begin{aligned}\boldsymbol{w} &= \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\\ \sum_{i=1}^{N}\alpha_i y_i &= 0\end{aligned}$$

代入拉格朗日目标函数，得：

$$\begin{aligned}\min_{\boldsymbol{w},b} L(\boldsymbol{w},b,\boldsymbol{\alpha}) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\sum_{i=1}^{N}\alpha_i y_i\left(\left(\sum_{j=1}^{N}\alpha_j y_j \boldsymbol{x}_j\right)\cdot\boldsymbol{x}_j+b\right)+\sum_{i=1}^{N}\alpha_i\\ &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)+\sum_{i=1}^{N}\alpha_i\end{aligned}$$

可得对偶问题：

$$\max_{a_i\geqslant 0} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)+\sum_{i=1}^{N}\alpha_i \qquad\qquad s.t. \sum_{i=1}^{N}\alpha_i y_i=0$$

根据KKT条件，有$\alpha_i=0$或$y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1=0$，当$\alpha_i>0$时，对应的$\boldsymbol{x}_i$为支持向量。可得：

$$\begin{aligned}\boldsymbol{w}^* &= \sum_{i=1}^{N}\alpha_i^* y_i \boldsymbol{x}_i\\ b^* &= y_j-\sum_{i=1}^{N}\alpha_i^* y_i(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)\end{aligned}$$

其中$\alpha_i^*>0$

分类决策函数：

$$f(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}^*\cdot\boldsymbol{x}+b^*)$$

非线性SVM

核函数

$$K(\boldsymbol{x},\boldsymbol{z}) = \phi(\boldsymbol{x})\cdot\phi(\boldsymbol{z})$$

用核函数代替内积

$$\begin{aligned} b^* &= y_j - \sum_{i=1}^{N} \alpha_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ f(\boldsymbol{x}) &= \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^* \right) \end{aligned}$$

2 凸优化，鞍点，拉格朗日对偶

对于优化问题,$f(x)$, $g_i(x)$为凸函数

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leqslant 0 \end{array}$$

设其解为x^*,即

$$\begin{aligned} f(x^*) &\leqslant f(x) \\ g_i(x) &\leqslant 0 \\ g_i(x^*) &\leqslant 0 \end{aligned}$$

考虑

$$L(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x) \qquad \lambda_i \geqslant 0$$

由

$$\min_x L(x, \lambda) \leqslant L(x, \lambda) \leqslant \max_{\lambda \geqslant 0} L(x, \lambda)$$

得：

$$\max_\lambda \min_x L(x, \lambda) \leqslant \min_x \max_\lambda L(x, \lambda) \tag{1}$$

若

$$\min_x \max_\lambda L(x, \lambda) \leqslant L(\bar{x}, \bar{\lambda}) \leqslant \max_\lambda \min_x L(x, \lambda) \tag{2}$$

则$(\bar{x}, \bar{\lambda})$称为$L(x, \lambda)$的鞍点。且有

$$\begin{aligned} \min_x \max_\lambda L(x, \lambda) &= \max_\lambda L(\bar{x}, \lambda) \qquad (3) \\ \max_\lambda L(\bar{x}, \lambda) &= L(\bar{x}, \bar{\lambda}) \\ \max_\lambda \min_x L(x, \lambda) &= \min_x L(x, \bar{\lambda}) \qquad (4) \\ \min_x L(x, \bar{\lambda}) &= L(\bar{x}, \bar{\lambda}) \end{aligned}$$

下面分析鞍点和原优化问题解$f(x^*)$之间的关系。

2.1 从鞍点到极值点

先看式(2)左边与式(3)，当$g_i(\bar{x}) > 0$时，有

$$\max_\lambda L(\bar{x}, \lambda) = \infty$$

可知$g_i(\overline{x}) \leqslant 0$,且有

$$
\begin{aligned}
\max_{\lambda} f(\overline{x}) + \lambda_i g_i(\overline{x}) &= f(\overline{x}) \\
&= L(\overline{x}, \lambda)|_{\lambda_i g_i(\overline{x})=0} \\
L(\overline{x}, \overline{\lambda}) &= f(\overline{x})
\end{aligned}
$$

再看式(2)式右边与式(4)，由 $\min_x L(x, \overline{\lambda}) = L(\overline{x}, \overline{\lambda})$ 得

$$
\begin{aligned}
f(\overline{x}) &= \min_x L(x, \overline{\lambda}) \\
&\leqslant \min_{\substack{x \\ g_i(x) \leqslant 0}} L(x, \overline{\lambda}) \\
&\leqslant \min_{\substack{x \\ g_i(x) \leqslant 0}} f(x) + \sum_i 0 \cdot g_i(x)
\end{aligned}
$$

可知

$$
f(\overline{x}) = \min_{\substack{x \\ g_i(x) \leqslant 0}} f(x)
$$

因此，可由鞍点得原优化问题的极小值点。

2.2 从极值点到鞍点

当已知原优化问题的极小值点时，即

$$
f(x^*) = \min_{\substack{x \\ g_i(x) \leqslant 0}} f(x)
$$

根据凸优化相关理论，可知存在$\lambda_i^* \geqslant 0$，满足：

$$
\begin{aligned}
\nabla f(x^*) + \sum_i \lambda_i^* \nabla g_i(x) &= 0 \qquad (5) \\
\lambda_i^* g_i(x) &= 0 \\
L(x^*, \lambda^*) &= f(x^*)
\end{aligned}
$$

先看式(2)左边

$$
\begin{aligned}
\min_x \max_\lambda L(x, \lambda) &\leqslant \min_{\substack{x \\ g_i(x) \leqslant 0}} \max_\lambda f(x) + \sum_i \lambda_i g_i(x) \\
&\leqslant \min_{\substack{x \\ g_i(x) \leqslant 0}} \max_\lambda f(x) + \sum_i \lambda_i \cdot 0 \\
&= \min_{\substack{x \\ g_i(x) \leqslant 0}} f(x) \\
&= f(x^*) + \lambda_i^* g_i(x^*) \\
\min_x \max_\lambda L(x, \lambda) &\leqslant L(x^*, \lambda^*) \qquad (6)
\end{aligned}
$$

再看式(2)右边

$$
\max_\lambda \min_x L(x,\lambda) \geqslant \min_x f(x) + \sum_i \lambda_i^* g_i(x)
$$

由于$L(x,\lambda^*)=f(x)+\sum_i\lambda_i^*g_i(x)$为凸函数，因此有惟一极小值满足

$$\nabla f(x)+\sum_i\lambda_i^*g_i(x) \;=\; 0$$

与式(5)结合可知

$$\begin{aligned} f(x^*) &= \min_x f(x)+\sum_i\lambda_i^*g_i(x) \\ L(x^*,\lambda^*) &\leqslant \max_\lambda \min_x L(x,\lambda) \end{aligned} \tag{7}$$

结合式(6)、式(7)得

$$\min_x\max_\lambda L(x,\lambda)\leqslant L(x^*,\lambda^*)\leqslant \max_\lambda\min_x L(x,\lambda)$$

由式(1),得

$$\min_x\max_\lambda L(x,\lambda)= L(x^*,\lambda^*)= \max_\lambda\min_x L(x,\lambda)$$

即：由原问题极值解可得鞍点。

2.3 $\min_x\max_\lambda L(x,\lambda)$与原问题的关系

下面说明$\min_x\max_\lambda L(x,\lambda)$与原问题的关系，当$g_i(x)>0$时有

$$\max_\lambda L(x,\lambda) \;=\; \infty$$

可得

$$\begin{aligned} \min_{\substack{x\\ g_i(x)>0}}\max_\lambda L(x,\lambda) &= \infty \\ \min_x\max_\lambda L(x,\lambda) &= \min_{\substack{x\\ g_i(x)\leqslant 0}}\max_\lambda L(x,\lambda) \end{aligned}$$

因此

$$\begin{aligned} \min_{\substack{x\\ g_i(x)\leqslant 0}}\max_\lambda L(x,\lambda) &= \min_{\substack{x\\ g_i(x)\leqslant 0}}\max_\lambda f(x)+\sum_i\lambda_i g_i(x) \\ &\leqslant \min_{\substack{x\\ g_i(x)\leqslant 0}}\max_\lambda f(x)+\sum_i\lambda_i\cdot 0 \\ &= \min_{\substack{x\\ g_i(x)\leqslant 0}} f(x) \\ \min_{\substack{x\\ g_i(x)\leqslant 0}}\max_\lambda L(x,\lambda) &\geqslant \min_{\substack{x\\ g_i(x)\leqslant 0}} f(x)+\sum_i 0\cdot g_i(x) \\ &= \min_{\substack{x\\ g_i(x)\leqslant 0}} f(x) \end{aligned}$$

得

$$\min_x\max_\lambda L(x,\lambda)= \min_{\substack{x\\ g_i(x)\leqslant 0}} f(x)$$

即与原问题等价。