浅谈拉格朗日乘数法及对偶问题在SVM中的应用(二、三)

原创 马旭淼 LOA算法学习笔记 2021-12-12 23:05

最近在算法课堂上学习了拉格朗日乘数法及其对偶问题,因此我在课后结合课堂所学知识和相关文献资料对拉格朗日乘数法及其相关问题进行了归纳总结。(文中的表格与公式可以滑动查看。)

> 这个系列会分为三部分，第一部分讨论Lagrange乘数法，第二部分讨论Lagrange对偶问题，第三部分讨论Lagrange对偶问题在硬间隔支持向量机(SVM)推导过程中的应用。本文为第二、三部分的内容。

PART 2 Lagrange对偶问题

2.1 Lagrange对偶函数

对于PART 1中式(3)所描述问题的拉格朗日函数为：

$$\mathcal{L}(\mathbf{x},\lambda,\boldsymbol{\mu})=f(\mathbf{x})-\sum_{i=1}^{m}\lambda_i g_i(\mathbf{x})-\sum_{j=1}^{n}\mu_j h_j(\mathbf{x})$$

设函数$g(\lambda,\mu)=inf\mathcal{L}_{x\in D}(x,\lambda,\mu)$，其中$inf$表示下确界。则$g(\lambda,\mu)$为拉格朗日对偶函数。很明显有：

$$f(x)\geq\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu})\geq g(\boldsymbol{\lambda},\boldsymbol{\mu}),\boldsymbol{\lambda}\leq 0$$

如果我们取最紧的下确界，即：

$$\begin{array}{ll} max & g(\lambda,\mu) \\ s.t. & \lambda\leq 0 \end{array}\tag{1}$$

则该问题被称为拉格朗日对偶问题。

2.2 标准形式下的Lagrange对偶问题

标准形式下的不等式约束优化问题一般如下所示：

$$\begin{gathered}\min \mathbf{c}^T\mathbf{x}\\ \text{s.t. } \mathbf{A}\mathbf{x}\leq\mathbf{b}\\ x\geq 0\end{gathered}\tag{2}$$

则有：

$$\begin{gathered}\mathcal{L}(\mathbf{x},\lambda,\boldsymbol{\mu})=\mathbf{c}^T\mathbf{x}-\sum_{i=1}^{m}\lambda_i\left(\sum_{j=1}^{n}a_{ij}\mathbf{x_j}-b_i\right)-\sum_{i=1}^{n}\mu_i\mathbf{x_i},\lambda\leq 0,\mu\geq 0\\ \mathbf{c}^T\mathbf{x}\geq L(\mathbf{x},\lambda,\boldsymbol{\mu})\geq\inf_{\mathbf{x}\in\mathcal{D}}L(\mathbf{x},\lambda,\boldsymbol{\mu})\end{gathered}$$

其中，$g(\lambda,\mu)=inf\mathcal{L}_{x\in D}(\mathbf{x},\lambda,\mu)$为拉格朗日对偶函数。继续推导：

$$\begin{aligned}g(\lambda,\mu)=&\ inf_{\mathbf{x}\in D}\mathcal{L}(\mathbf{x},\lambda,\mu)\\ =&\ inf_{\mathbf{x}\in D}(\mathbf{c}^T\mathbf{x}-\sum_{i=1}^{m}\lambda_i(\sum_{j=1}^{n}a_{ij}\mathbf{x}_j-b_i)-\sum_{i=1}^{n}\mu_i\mathbf{x}_i)\end{aligned}$$

可以看到，当$\mathbf{c}^T\geq\lambda^T\mathbf{A}+\boldsymbol{\mu}^T$时，$g(\lambda,\mu)$存在明确的下确界。也就是说，对于相应的拉格朗日对偶问题有：

$$\max g(\lambda,\mu)=\begin{cases}\lambda^T\mathbf{b} & \text{if } c^T\geq\lambda^T\mathbf{A}+\boldsymbol{\mu}^T\\ -\infty & \text{otherwise}\end{cases}$$

其中，$\lambda \leq 0, \mu \geq 0$。因此，原问题有以下对偶形式：

$$\begin{aligned} &\max \lambda^T \mathbf{b} \\ &\text{s.t. } \lambda^T \mathbf{A} + \boldsymbol{\mu}^T \leq \mathbf{c}^T \\ &\lambda \leq 0, \mu \geq 0 \end{aligned} \tag{4}$$

对比式(2)与式(4)，结合推导过程，就可以解释第一部分中KKT条件中留下来的问题。首先易得原问题的拉格朗日乘子就是对偶问题中的对偶变量，同时原可行性条件、目标函数以及构造的拉格朗日函数共同决定了对偶可行性条件(因为要求最小化目标函数，并且是以减去拉格朗日乘子的形式构造拉格朗日函数，因此$\lambda_i \leq 0, i = 1, \ldots, m$)。下面用对偶问题的结论证明互补松弛性条件：

$$\begin{aligned} f(\mathbf{x}^*) &= g(\lambda^*, \mu^*) \\ &= \inf\left(f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}) - \sum_{j=1}^{n} \mu_j^* h_j(\mathbf{x})\right) \end{aligned} \tag{5}$$

其中，$\mathbf{x}^*$表示原问题的最优解，λ^* 、 μ^*表示对偶问题的最优解。由式(5)可以得出，最后两个不等式必须要取到等号。因此$\sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) = \sum_{j=1}^{n} \mu_i^* h_j(\mathbf{x}^*) = 0$。由原可行性条件，对$\forall i, j$，有$g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0$。因此$\sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*)$必须为0才能满足等式。而对$\forall i, \lambda_i^* g_i(\mathbf{x}^*)$总是非负的，因此每一项都必须为0。即$\lambda_i^* g_i(\mathbf{x}^*) = 0, i = 1, \ldots, m$，互补松弛性条件得证。

2.3 Lagrange对偶问题实例

（本问题来源于UCAS《081203M04001H-计算机算法设计与分析》2021年秋季学期课程Assignment 4）

现在考虑如下问题。如果工厂需要制造A、B、C三种产品，每种产品都需要镍和铝两种材料。每种产品对材料的需要和利润如下表所示：

Product	Profit ($)	Nickel (kg)	Aluminum (kg)
A	10	3	4
B	8	3	3
C	16	2	7

假设工厂有300kg铝和200kg镍，现在需要合理的安排原材料分配来实现利润的最大化。很明显这是一类带不等式约束的优化问题，可以给出如下求解过程：

- 设变量x_1 、 x_2 、 x_3分别表示产品A、B、C的生产量
- 令目标函数$f(x) = 10x_1 + 8x_2 + 16x_3$表示总利润

因此问题可以表达为以下LP模型：

$$\begin{array}{ll} \max & f(x) \\ \text{s.t.} & 3x_1 + 3x_2 + 2x_3 \leq 200 \\ & 4x_1 + 3x_2 + 7x_3 \leq 300 \\ & x_i \geq 0, i = 1, 2, 3 \end{array}$$

此处目标函数需要最大化，尽管这不是上文提到的标准型问题，但是二者实际上是等价的:$minf(x) \Leftrightarrow max(-f(x))$

构造拉格朗日函数：

$$\begin{aligned} L(x_1, x_2, x_3, \lambda_1, \lambda_2) &= f(x) - \lambda_1(3x_1 + 3x_2 + 2x_3 - 200) - \lambda_2(4x_1 + 3x_2 + 7x_3 - 300) \\ &= (10 - 3\lambda_1 - 4\lambda_2)x_1 + (8 - 3\lambda_1 - 3\lambda_2)x_2 + (16 - 2\lambda_1 - 7\lambda_2)x_3 + 200\lambda_1 + 300\lambda_2 \end{aligned}$$

其中$\lambda_1, \lambda_2 \geq 0$。则相应的拉格朗日对偶函数为：

$$g(\lambda_1, \lambda_2) = \sup \mathrm{L}_{x_1, x_2, x_3}(x_1, x_2, x_3, \lambda_1, \lambda_2)$$

其中sup表示上确界。很明显有：

$$\begin{gathered} \max f(x) \leq L(x_1, x_2, x_3, \lambda_1, \lambda_2) \leq g(\lambda_1, \lambda_2) \\ \lambda_1, \lambda_2 \geq 0 \end{gathered}$$

因此，原问题转化为拉格朗日对偶问题：

$$\begin{gathered}\min\ g(\lambda_1,\lambda_2)\\ \text{s.t. } \lambda_1,\lambda_2\ge 0\end{gathered}$$

相当于：

$$\begin{array}{ll}\min & 200y_1+300y_2\\ \text{s.t.} & 3y_1+4y_2\ge 10\\ & 3y_1+3y_2\ge 8\\ & 2y_1+7y_2\ge 16\\ & y_i\ge 0, i=1,2\end{array}$$

上式中的y_i就是λ_i。可以使用GLPK线性规划求解软件(For windows,version=4.6.5)求解原问题和其对偶问题。求解原问题得到如下结果：

```
Status:     OPTIMAL
Objective:  z = 746.6666667 (MAXimum)

   No.   Row name   St   Activity     Lower bound   Upper bound    Marginal
------ ------------ -- ------------- ------------- ------------- -------------
     1 z            B        746.667
     2 con1         NU           200                         200      0.533333
     3 con2         NU           300                         300       2.13333

   No. Column name  St   Activity     Lower bound   Upper bound    Marginal
------ ------------ -- ------------- ------------- ------------- -------------
     1 x1           NL             0             0                   -0.133333
     2 x2           B        53.3333             0
     3 x3           B             20             0
```

求解对偶问题得到的结果如下：

```
Status:     OPTIMAL
Objective:  z = 746.6666667 (MINimum)

   No.   Row name   St   Activity     Lower bound   Upper bound    Marginal
------ ------------ -- ------------- ------------- ------------- -------------
     1 z            B        746.667
     2 con1         B        10.1333            10
     3 con2         NL             8             8                     53.3333
     4 con3         NL            16            16                          20

   No. Column name  St   Activity     Lower bound   Upper bound    Marginal
------ ------------ -- ------------- ------------- ------------- -------------
     1 y1           B       0.533333             0
     2 y2           B        2.13333             0
```

可以看到两个问题求解的结果是一样的，符合之前的理论推导。实际上这个问题也很好地解释了课堂上老师提到的"对偶变量实际上就是边际成本"。这个问题中，线性规划对偶问题的最优解，就是原材料的"影子价格"，它对于线性规划模型的经济分析起着极为重要的指导作用。

PART 3 拉格朗日对偶问题与SVM

支持向量机(SVM)是一种非常经典的线性二分类模型，其理论基础就是拉格朗日对偶问题。由于篇幅所限，本文只讨论硬间隔支持向量机。对于非线性可分的情况，则还需要引入核函数等概念，此处不予赘述。

3.1 SVM间隔(Margin)定义

如果我们给定一组数据$S=\{(\mathbf{x_1},y_1),\ldots,(\mathbf{x_m},y_m)\}$,其中，对$\forall i, y_i\in\{-1,1\}$。我们希望对这些样本点进行线性二分类。换而言之，就是希望找到一组合适的参数$(\mathbf{w},\mathbf{b})$,可以产生$\mathbf{w}^T\mathbf{x}+\mathbf{b}=0$这样一个超平面，对$\forall \boldsymbol{x_i}$，都有：

$$\begin{cases}\mathbf{w}^T\boldsymbol{x_i}+\mathbf{b}>0 & \text{if}\quad y_i=1\\ \mathbf{w}^T\boldsymbol{x_i}+\mathbf{b}<0 & \text{if}\quad y_i=-1\end{cases}$$

等价于对$\forall i$,都有$y_i(\mathbf{w}^T\mathbf{x}+\mathbf{b})>0$。

由于能划分样本的超平面不止一个，由于随机扰动的存在，通常希望超平面能以较大的置信度将样本分开。对于空间中的某点$\mathbf{p}$，其到超平面$\mathbf{w^T x}+\mathbf{b}=0$的距离为$\mathbf{p}$到超平面上某点$\mathbf{x}$连线在超平面法向量(即$\mathbf{w}$)上的投影：

$$\begin{aligned}\operatorname{proj}_{\mathbf{w}}(\mathbf{p}-\mathbf{x}) &= \|\mathbf{p}-\mathbf{x}\|\cdot|\cos\langle\mathbf{w},\mathbf{p}-\mathbf{x}\rangle| \\ &= \|\mathbf{p}-\mathbf{x}\|\cdot\frac{\left|\mathbf{w}^T(\mathbf{p}-\mathbf{x})\right|}{\|\mathbf{w}\|\|\mathbf{p}-\mathbf{x}\|} \\ & 1 \quad | \quad T \qquad T \quad | \end{aligned} \qquad (6)$$

在SVM中，通常用"间隔"来刻画划分超平面与样本之间的距离。间隔γ的定义为距离划分超平面最近的样本点到划分超平面距离的两倍，即：

$$\gamma = 2\min_i \frac{1}{\|\boldsymbol{w}\|}\left|\boldsymbol{w}^T\boldsymbol{x}_i+\mathbf{b}\right|$$

SVM的目的就是寻找一组合适的参数$(\boldsymbol{w},\mathbf{b})$，使得：

$$\begin{aligned} &\max_{\boldsymbol{w},\boldsymbol{b}}\ \min_i \frac{2}{\|\boldsymbol{w}\|}\left|\boldsymbol{w}^T\boldsymbol{x}_i+\mathbf{b}\right| \\ &\text{s.t.}\quad y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i+\mathbf{b}\right)>0 \\ &\qquad i=1,2,\ldots,m \end{aligned} \qquad (7)$$

换而言之，SVM希望在特征空间找到一个划分超平面，能有效分离正负样本，并且距离所有样本点越远越好。

3.2 SVM基本型

由于式(7)所描述的优化问题比较难以求解，我们需要对其进行一定的简化。由于划分超平面对样本点分类的判断依据就是$\mathbf{w}^T\mathbf{x_i}+\mathbf{b}$的正负性。因此，对$(\mathbf{w}^T,\mathbf{b})$进行正向放缩为$(\eta\mathbf{w}^T,\eta\mathbf{b})$,其中$\eta>0$。这样的放缩并不会影响分类结果。我们可以找到一个合适的η，使得:

$$\min_i|\boldsymbol{w}^T\boldsymbol{x}_i+\mathbf{b}|=1 \qquad (8)$$

因此，式(7)可以表达为

$$\begin{aligned} &\max_{\boldsymbol{w},\boldsymbol{b}}\quad \frac{2}{\|\boldsymbol{w}\|} \\ &\text{s.t. } y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i+\mathbf{b}\right)>0 \\ &\quad i=1,2,\ldots,m \end{aligned} \qquad (9)$$

而：

$$\max_{\boldsymbol{w},\boldsymbol{b}}\frac{2}{\|\boldsymbol{w}\|}\Leftrightarrow\min_{\boldsymbol{w},\boldsymbol{b}}\frac{\|\boldsymbol{w}\|}{2}\Leftrightarrow\min_{\boldsymbol{w},\boldsymbol{b}}\frac{\boldsymbol{w}^T\boldsymbol{w}}{2}$$

并且由于$\min_i|\boldsymbol{w}^T\boldsymbol{x}_i+\mathbf{b}|=1$（离划分超平面最近的样本点）,那么对于其他样本点，一定有$|\boldsymbol{w}^T\boldsymbol{x}_j+\mathbf{b}|\ge 1$,又因为对$\forall i, y_i\in(-1,1)$，因此约束条件$y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i+\mathbf{b}\right)>0$等价于$y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i+\mathbf{b}\right)\ge 1$。根据上述分析，式(9)可以转化为：

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{b}}\quad \frac{\boldsymbol{w}^T\boldsymbol{w}}{2} \\ &\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i+\boldsymbol{b})\ge 1 \\ &\quad i=1,2,\ldots,m \end{aligned} \qquad (10)$$

上述表达称为**SVM**基本型。

3.3 SVM对偶型

根据第二部分的分析，SVM标准型描述的问题为带不等式约束的凸二次规划问题，并且满足Slater条件(目标函数与约束函数均可微并且约束函数为仿射函数)，由此可推出问题具有强对偶性。写出SVM的拉格朗日函数如下：

$$\mathcal{L}(\mathbf{w},\mathbf{b},\boldsymbol{\lambda})=\frac{\mathbf{w}^T\mathbf{w}}{2}+\sum_{i=1}^{m}\lambda_i\left(1-y_i(\mathbf{w}^T\mathbf{x_i}+\mathbf{b})\right) \qquad (11)$$

其对偶问题为：

$$\begin{aligned}\max_{\boldsymbol{\lambda}}\inf_{\boldsymbol{w},\boldsymbol{b}}\ \frac{\boldsymbol{w}^T\boldsymbol{w}}{2}+\sum_{i=1}^{m}\lambda_i\big(1-y_i\big(\boldsymbol{w}^T\boldsymbol{x_i}+\boldsymbol{b}\big)\big)\\ \text{s.t. } \lambda_i\geq 0\\ i=1,2,\ldots,m\end{aligned} \tag{12}$$

式(11)中，对$\boldsymbol{b}$求偏导得：

$$\begin{aligned}\frac{\partial\mathscr{L}}{\partial\boldsymbol{b}}&=\frac{\partial}{\partial\boldsymbol{b}}\left[\sum_{i=1}^{m}\lambda_i-\sum_{i=1}^{m}\lambda_iy_i\big(\boldsymbol{w}^T\boldsymbol{x_i}+\boldsymbol{b}\big)\right]\\&=\frac{\partial}{\partial\boldsymbol{b}}\left[-\sum_{i=1}^{m}\lambda_iy_i\boldsymbol{b}\right]\\&=-\sum_{i=1}^{m}\lambda_iy_i=0\\&\Rightarrow\sum_{i=1}^{m}\lambda_iy_i=0\end{aligned} \tag{13}$$

对$\boldsymbol{w}$求偏导得：

$$\begin{aligned}\frac{\partial\mathscr{L}}{\partial\boldsymbol{w}}&=\boldsymbol{w}-\sum_{i=1}^{m}\lambda_iy_i\boldsymbol{x_i}=0\\&\Rightarrow\boldsymbol{w}=\sum_{i=1}^{m}\lambda_iy_i\boldsymbol{x_i}\end{aligned} \tag{14}$$

将式(13)、式(14)代入式(12),就可以得到**SVM**对偶型：

$$\begin{aligned}\min_{\boldsymbol{\lambda}}\ \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_jy_iy_j\boldsymbol{x_i}^T\boldsymbol{x_j}-\sum_{i=1}^{m}\lambda_i\\ \text{s.t.}\quad\sum_{i=1}^{m}\lambda_iy_i=0,\\ \lambda_i\geq 0,i=1,2,\ldots,m\end{aligned} \tag{15}$$

对比式(10)和式(15)，我们可以看到相对于SVM标准型复杂的约束条件，SVM对偶型摆脱了$\boldsymbol{w}$和$\boldsymbol{b}$的限制。用计算机来实现SVM时，可以在原问题上直接对其进行优化，比如Pegasos算法等；也可以将其转化为对偶问题来求解，如SMO算法等等。一般来说，后者算法的时间复杂度往往较低，在样本数据的维度较高时，这种差异尤为明显。

3.4 SVM的KKT条件与支持向量

SVM的KKT条件如下所示：

- $\nabla_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\lambda}}\mathscr{L}(\boldsymbol{w}^*,\boldsymbol{b}^*,\boldsymbol{\lambda}^*)=0$
- $1-y_i\big(\boldsymbol{w}^T\boldsymbol{x_i}+\boldsymbol{b}\big)\leq 0$
- $\lambda_i\geq 0$
- $\boldsymbol{\lambda}^*\big(1-y_i\big(\boldsymbol{w}^T\boldsymbol{x}_i+\boldsymbol{b}\big)\big)=0$

由互补松弛性条件可知，当$\lambda_i>0$时，$1-y_i(\boldsymbol{w}^T\boldsymbol{x_i}+\boldsymbol{b})=0$，即$y_i(\boldsymbol{w}^T\boldsymbol{x_i}+\boldsymbol{b})=1$，满足这个条件的样本点就是距离划分超平面最近的样本点，也被称为支持向量，其对应的对偶变量$\lambda_i>0$。由式(14)，结合$\lambda_i>0$，可得：

$$\boldsymbol{w}^* = \sum_{i=1}^{m} \lambda_i y_i \boldsymbol{x}_i = \sum^{m} 0 \cdot y_i \boldsymbol{x}_i + \sum^{m} \lambda_i y_i \boldsymbol{x}_i \quad (16)$$

其中，SV代表支持向量数据样本下标的集合。对应的$\boldsymbol{b}^*$可以由互补松弛性条件得出。对某支持向量$\boldsymbol{x_s}$,由于$y_s(\boldsymbol{w}^{T^*}\boldsymbol{x_s}+\boldsymbol{b}^*)=1$,则有：

$$\mathbf{b}^* = y_s - \mathbf{w}^{T^*}\mathbf{x_s} = y_s - \sum_{i \in SV} \lambda_i\, y_i \mathbf{x}_i^T \mathbf{x_s} \qquad (17)$$

实际求解时，通常对所有支持向量求解$\boldsymbol{b}^*$的平均值，以期提高准确性。

由以上分析可知，SVM的参数$(\boldsymbol{w}^*, \boldsymbol{b}^*)$完全由其支持向量所决定，与其他样本点没有关系，这也是该模型名称的由来。

拉格朗日对偶问题在硬间隔SVM中的使用使得问题的求解变的更加简单。以上讨论的都是数据样本线性可分的情况，如果遇到数据集中存在异常点或者不能线性可分时，就要引入核函数与软间隔等概念来求解问题。尽管问题变的复杂了，但是将问题转化为对偶问题的基本策略是不变的。因此，拉格朗日对偶问题对于SVM来说，是基础性的理论支撑。有了前者，才在这其上诞生了SVM的各种拓展版本。

参考文献

[1]李航. 统计学习方法[M]. 清华大学出版社, 2012.

[2]周志华. 机器学习[M]. 清华大学出版社, 2015.

[3]卜东波.《计算机算法设计与分析》课程讲义[EB/OL].中国科学院大学，2021.

[4]Christopher M. Bishop. Pattern Recognition and Machine Learning[M]. Springer, 2006.

[5]Cortes C ,Vapnik V , et al. Support-vector networks[J]. 1995.

[6]张皓.从零推导支持向量机[EB/OL].https://zhuanlan.zhihu.com/p/31652569