@totoo 2016-08-31 15:42 字数 6886 阅读 0

# 机器学习之拉格朗日乘数法

机器学习, 拉格朗日

拉格朗日乘数法是求解多元函数在一组约束条件下的极值的方法。直接阅读一些机器学习教材对拉格朗日乘数法的介绍可能会觉得难以理解，这篇笔记从高等数学教材中简单的求解二元函数极值问题开始，慢慢引申到拉格朗日乘数法的一般性描述。

## 高数中的拉格朗日乘数法

对于二元函数的极值问题，通过求偏导可以直接解决：

> **定理** 设函数 $z=f(x,y)$ 在点 $(x_0,y_0)$ 的某邻域内连续且有一阶及二阶连续偏导数，又 $f_x(x_0,y_0)=0,\ f_y(x_0,y_0)=0$，令
>
> $$f_{xx}(x_0,y_0)=A,\ f_{xy}(x_0,y_0)=B,\ f_{yy}(x_0,y_0)=C$$
>
> 则 $f(x,y)$ 在 $(x_0,y_0)$ 处是否取得极值的条件如下：
> （1）$AC-B^2>0$ 时具有极值，且当 $A<0$ 时具有极大值，当 $A>0$ 时具有极小值；
> （2）$AC-B^2<0$ 时没有极值；
> （3）$AC-B^2=0$ 时可能有极值，也可能没有极值，还需另作讨论。

但在实际应用中，除了函数的定义域，一般还会有其它的约束条件。例如，求表面积为 $a^2$ 而体积为最大的长方体的体积。

拉格朗日乘数法就是求解这种**条件极值问题**的方法。通过引入拉格朗日乘子，可将有 $d$ 个变量与 $k$ 个约束条件的极值问题（最优化问题）转化为具有 $d+k$ 个变量的无约束优化问题。

> **拉格朗日乘数法** 要找函数 $z=f(x,y)$ 在附加条件 $\varphi(x,y)=0$ 下的可能极值点，可以先作拉格朗日函数
>
> $$L(x,y)=f(x,y)+\lambda\varphi(x,y)$$
>
> 其中 $\lambda$ 为参数。求其对 $x$ 与 $y$ 的一阶偏导数，并使之为零，然后与方程 $\varphi(x,y)=0$ 联立起来：
>
> $$\begin{cases} f_x(x,y)+\lambda\varphi_x(x,y)=0 \\ f_y(x,y)+\lambda\varphi_y(x,y)=0 \\ \varphi(x,y)=0 \end{cases}$$
>
> 由这方程组解出 $x,y$及 $\lambda$，这样得到的 $(x,y)$ 就是函数 $f(x,y)$ 在附加条件 $\varphi(x,y)=0$ 下的可能极值点。
>
> 条件多于一个的情形也是类似：
> 函数：$u=f(x,y,z,t)$
> 条件：$\varphi(x,y,z,t)=0,\ \psi(x,y,z,t)=0$
> 拉格朗日函数：$L(x,y,z,t)=f(x,y,z,t)+\lambda\varphi(x,y,z,t)+\mu\psi(x,y,z,t)$

## 机器学习中的拉格朗日乘数法

# 机器学习中的拉格朗日乘数法

## 等式约束问题

这里的内容与上面类似，只是描述方式不一样。

假定 $\boldsymbol{x}$ 为 $d$ 维向量，欲寻找 $\boldsymbol{x}$ 的某个取值 $\boldsymbol{x}^*$，使目标函数 $f(\boldsymbol{x})$ 最小且同时满足 $g(\boldsymbol{x}) = 0$ 的约束。
拉格朗日函数

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}) \tag{1}$$

$\boldsymbol{x}$ 的偏导 $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda)$ 置零

$$\nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0 \tag{2}$$

$\lambda$ 的偏导 $\nabla_{\lambda} L(\boldsymbol{x}, \lambda)$ 置零

$$g(\boldsymbol{x}) = 0 \tag{3}$$

因此，原约束优化问题可转化为对拉格朗日函数（式(1)）的无约束优化问题。

> 从几何角度看，该问题的目标是在由方程 $g(\boldsymbol{x}) = 0$ 确定的 $d-1$ 维曲面上寻找能使目标函数 $f(\boldsymbol{x})$ 最小化的点。因此，有以下结论：
> - 对于约束曲面上任意点 $\boldsymbol{x}$，该点的梯度 $\nabla g(\boldsymbol{x})$ 正交于约束曲面；
> - 在最优点 $\boldsymbol{x}^*$，目标函数在该点的梯度 $\nabla f(\boldsymbol{x}^*)$ 正交于约束曲面。（反证法：若梯度 $\nabla f(\boldsymbol{x}^*)$ 与约束曲面不正交，则仍可在约束曲面上移动该点使函数值进一步下降）
>
> 即在最优点 $\boldsymbol{x}^*$，梯度 $\nabla g(\boldsymbol{x})$ 和 $\nabla f(\boldsymbol{x})$ 的方向必相同或相反，即式(2)成立：
>
> $$\nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}) = 0$$

## 不等式约束问题

如果上面的约束换成不等式约束 $g(\boldsymbol{x}) \leq 0$

可转化为在如下约束下最小化拉格朗日函数（式(1)）：

$$\begin{cases} g(\boldsymbol{x}) \leq 0 \\ \lambda \geq 0 \\ \lambda g(\boldsymbol{x}) = 0 \end{cases} \tag{4}$$

式(4)称为Karush-Kuhn-Tucker（简称KKT）条件。

> **KKT条件的解读：**
> 最优点 $\boldsymbol{x}^*$ 或在 $g(\boldsymbol{x}) < 0$ 的区域中，或在边界 $g(\boldsymbol{x}) = 0$ 上。
> - $g(\boldsymbol{x}) < 0$ 时，约束 $g(\boldsymbol{x}) \leq 0$ 不起作用，可直接通过条件 $\nabla f(\boldsymbol{x}) = 0$来获得最优点；这等价于将式(1)中的 $\lambda$ 置零然后对 $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda)$ 置零得到最优点。
> - $g(\boldsymbol{x}) = 0$ 时，就是上面的等式约束问题，需要注意的是，此时 $\nabla f(\boldsymbol{x}^*)$ 的方向必与 $\nabla g(\boldsymbol{x}^*)$ 相反，即式(2)中的$\lambda$ 有 $\lambda > 0$
>
> 整合上面两种情况，可以得到式(4)的约束。

扩展到 $m$ 个等式约束和 $n$ 个不等式约束，且可行域 $\mathbb{D} \subset \mathbb{R}^d$ 非空的优化问题

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) && (5)\\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0 \ (i = 1, \cdots, m), && (6)\\ & g_j(\boldsymbol{x}) \leq 0 \ (j = 1, \cdots, n). && (7) \end{aligned}$$

引入拉格朗日乘子 $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_m)^{\mathrm{T}}$ 和 $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^{\mathrm{T}}$，相应的拉格朗日函数为

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x}) \tag{8}$$

由不等式约束引入的KKT条件（ $j = 1, 2, \cdots, n$ ）为

$$\begin{cases} g_j(\boldsymbol{x}) \leq 0 \\ \mu_j \geq 0 \\ \mu_j g_j(\boldsymbol{x}) = 0 \end{cases} \tag{9}$$

到目前为止，我们将含有不等式约束的最优化问题转换成了KKT条件下的拉格朗日函数求解，但具体怎么解，我们还不清楚。其实，解的方法就是下面要介绍的**对偶法**。

## 拉格朗日对偶性

利用拉格朗日对偶性（Lagrange duality）可以将原始问题转换为对偶问题，再通过解对偶问题可以得到原始问题的解。

### 原始问题

考虑 $\boldsymbol{x}$的函数：

$$\theta_P(\boldsymbol{x}) = \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}: \mu_j \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{10}$$

这里，下标P表示原始问题。有

$$\theta_P(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}), & \boldsymbol{x}\text{满足原始问题约束} \\ +\infty, & \text{其它} \end{cases} \tag{11}$$

- $\boldsymbol{x}$ 满足原始问题约束时，由式(8)可知，$\theta_P(\boldsymbol{x}) = f(\boldsymbol{x})$
- 不满足约束时，即存在某个 $i$ 使得 $h_i(\boldsymbol{x}) \neq 0$ 或者存在某个 $j$ 使得 $g_j(\boldsymbol{x}) > 0$，若 $h_i(\boldsymbol{x}) \neq 0$，令 $\lambda_i h_i(\boldsymbol{x}) \to +\infty$，若 $g_j(\boldsymbol{x}) > 0$，令 $\mu_j > 0$，且令其它 $\lambda$、$\mu$ 等于0，则 $\theta_P(\boldsymbol{x}) = +\infty$

考虑如下极小化问题

$$\min_{\boldsymbol{x}} \theta_P(\boldsymbol{x}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}: \mu_j \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{12}$$

它是与原始最优化问题是等价的，称为拉格朗日函数的**极小极大问题。**

定义原始问题的最优值

$$p^* = \min_{x} \theta_P(\boldsymbol{x}) \tag{13}$$

称为原始问题的值。

对偶问题

定义

$$\theta_D(\boldsymbol{\lambda},\boldsymbol{\mu})=\min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{14}$$

有拉格朗日函数的**极大极小**问题：

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\mu_j\geq 0}\theta_D(\boldsymbol{\lambda},\boldsymbol{\mu})=\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\mu_j\geq 0}\min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{15}$$

表示为约束最优化问题：

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}}\theta_D(\boldsymbol{\lambda},\boldsymbol{\mu})=\max_{\boldsymbol{\lambda},\boldsymbol{\mu}}\min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{16}$$

$$\text{s.t.}\quad \mu_i\geq 0,\ i=1,2,\cdots,j \tag{17}$$

称为原始问题的**对偶问题**。

定义对偶问题的最优值

$$d^*=\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\mu_j\geq 0}\theta_D(\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{18}$$

称为对偶问题的值。

原始问题与对偶问题的关系

**定理1** 若原始问题和对偶问题都有最优解，则

$$d^*=\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\mu_j\geq 0}\min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})\leq\min_{\boldsymbol{x}}\max_{\boldsymbol{\lambda},\boldsymbol{\mu}:\mu_j\geq 0} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})=p^*$$

**推论1** 设 $\boldsymbol{x}^*$ 和 $\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*$ 分别是原始问题(5)~(7)和对偶问题(16)~(17)的可行解，并且 $d^*=p^*$，则 $\boldsymbol{x}^*$ 和 $\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*$ 分别是原始问题和对偶问题的最优解。

**定理2** 考虑原始问题(5)~(7)和对偶问题(16)~(17)。假设函数 $f(\boldsymbol{x})$ 和 $g_j(\boldsymbol{x})$ 是凸函数，$h_i(\boldsymbol{x})$ 是仿射函数；并且假设不等式 $g_j(\boldsymbol{x})$ 是严格可行的，即存在 $\boldsymbol{x}$ 对所有 $j$ 有$g_j(\boldsymbol{x})<0$，则存在 $\boldsymbol{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*$，使 $\boldsymbol{x}^*$ 是原始问题的解，$\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*$ 是对偶问题的解，并且

$$p^*=d^*=L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)$$

**定理3** 对原始问题(5)~(7)和对偶问题(16)~(17)，假设函数 $f(\boldsymbol{x})$ 和 $g_j(\boldsymbol{x})$ 是凸函数，$h_i(\boldsymbol{x})$ 是仿射函数，并且不等式 $g_j(\boldsymbol{x})$ 是严格可行的，则$\boldsymbol{x}^*$和$\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*$分别是原始问题和对偶问题的解的充分必要条件是$\boldsymbol{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*$满足下面的KKT条件：

$$\begin{aligned}
\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)&=0\\
\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)&=0\\
\nabla_{\boldsymbol{\mu}} L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)&=0\\
\mu_j^* g_j(\boldsymbol{x}^*)&=0,\ j=1,2,\cdots,n\\
g_j(\boldsymbol{x}^*)&\leq 0,\ j=1,2,\cdots,n\\
\mu_j^*&\geq 0,\ j=1,2,\cdots,n\\
h_i(\boldsymbol{x}^*)&=0,\ i=1,2,\cdots,m
\end{aligned}$$

其中，$\mu_j^* g_j(\boldsymbol{x}^*) = 0$ 称为KKT对偶互补条件，由此条件可知，若 $\mu_i^* > 0$，则 $g_j(\boldsymbol{x}^*) = 0$。

# 参考资料


1. 高等数学 同济大学数学系
2. 机器学习 周志华
3. 统计学习方法 李航