

第19课：SVM——直观理解拉格朗日乘子法

上一篇，我们获得了线性可分 SVM 的目标函数：一个带约束条件的求极值问题。

而拉格朗日乘子法，恰恰是一种多元函数在变量受到条件约束时，求极值的方法。正好可以用来解决 SVM 的目标函数最优化。

我们在此不做严格的拉格朗日乘数法正确性的数学证明，而是以最简单的函数形式为例，从直观带大家来领略整个方法的每一个步骤。

换句话说，本文是帮我们积累一些对于“为什么将目标函数转化成拉格朗日函数再最优化是可行的”这件事的感性认识。

### 可视化函数及其约束条件

我们用二元函数——也就是自变量为2维的函数——来做个例子（为了看着更习惯一点，我们直接用 x，yx，y 作为自变量的两个维度）。

#### （被约束的）函数

我们之前有过可视化函数本身的经验。此处我们先要可视化一个二元函数 f(x,y)f(x,y)。

用一个大家熟悉的表达方式：z=f(x,y)z=f(x,y) 。这就涉及到了3个变量：xx，yy 和 zz。

如果在三维直角坐标系中将 f(x,y)f(x,y) 做出图来——把 f(x,y)f(x,y) “画出图来”——会是一个三维空间的曲面——这样一个函数实际上表达了 x，y，zx，y，z 三者之间的关系。

比如下面几幅图，分别对应不同的 f(x,y)f(x,y)：

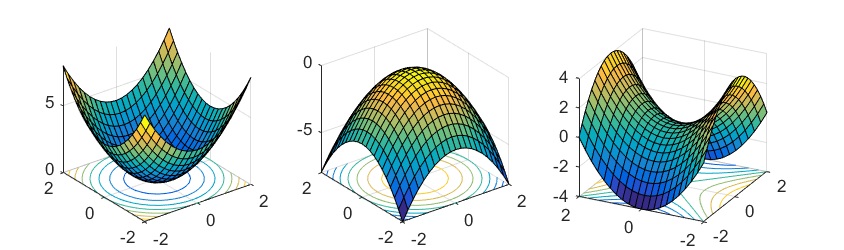


图 1

#### （函数的）约束条件

函数 f(x,y)f(x,y) 的约束条件为：g(x,y)=0g(x,y)=0。

g(x,y)=0g(x,y)=0 又可以写成：y=h(x)y=h(x) 的形式——它所表达的是 xx 与 yy 两者之间的相互关系！

y=h(x)y=h(x)，如果在二维直角坐标中作图，形状应该是一条曲线。

注意：直线也属于广义的曲线；平面也属于曲面。我们说的曲线是包括直线的，曲面则包括平面。

#### 约束条件对函数的约束

那么，一个二维图形对一个三维图形的约束从何体现呢？让我们这样做：

1. 在自己的头脑中建立一个三维直角坐标系，有 xx 轴、yy 轴、zz 轴， 它们互相垂直。
2. f(x,y)f(x,y) 对应图形是三维坐标系里面的一个曲面——一个（可能是奇形怪状的）“体”的“外皮”。
3. 在 z=0z=0 的平面上，把 y=h(x)y=h(x) 的图形（一条曲线）画出来。
4. 将第 3 步做出的那条曲线沿着平行 z 轴的方向平移，它平移过的轨迹也形成了一个曲面，这个曲面和 z=f(x,y)z=f(x,y) 形成的曲面会相交，交叠的部分是一条三维空间中的曲线。
5. 换个角度考虑，第 4 步形成的曲线，其实就是 g(x,y)=0g(x,y)=0 在 z=0z=0 平面上形成的曲线，在 z=f(x,y)z=f(x,y) 形成的曲面上的投影。

#### 一个例子

图2是一个将上面 1-5 步骤综合起来的实例：

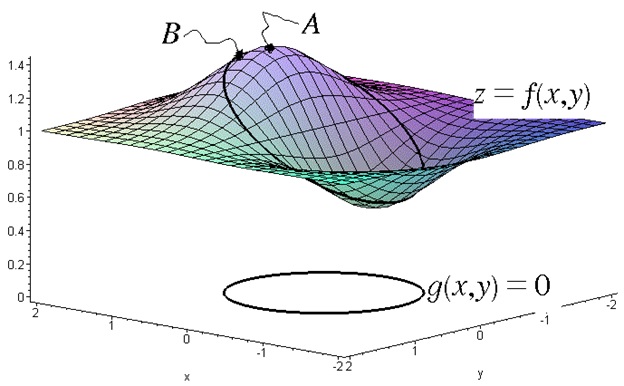


图 2

在这个例子中，我们可以看到 f(x,y)f(x,y) 是存在极大值的，同时因为约束条件是 g(x,y)=0g(x,y)=0，所以，如果我们要取如下目标的话：

maxf(x,y)maxf(x,y)

s.t.g(x,y)=0s.t.g(x,y)=0

对应点一定位于 g(x,y)=0g(x,y)=0 投影到 f(x,y)f(x,y) 上之后形成的那条曲线上，比如图2中的点 B。

尽管它不一定是 z=f(x,y)z=f(x,y) 的极大值，但却是符合约束条件 g(x,y)=0g(x,y)=0 的极大值！

### 拉格朗日乘子法

#### 目标函数

上面一节介绍了可视化函数及其约束条件的方法。现在我们回到线性可分 SVM 的目标函数：

min||w||22min||w||22

s.t.1−yi(w⋅xi+b)⩽0，i=1，2，…,ms.t.1−yi(w·xi+b)⩽0，i=1，2，…,m

其中的自变量为 ww 和 bb，也是一个二元函数（xixi 和 yiyi 都是样本值，已知数值，||w||22||w||22 虽然只包含 ww，但我们可以想象它也是一个二元函数，只不过变量 bb 的系数为0）。

所以，我们可以再抽象一步，将其概括为：

minf(x,y)minf(x,y)

s.t.g(x,y)⩽0s.t.g(x,y)⩽0

这个约束条件其实可以写作：

s.t.g(x,y)=0ORg(x,y)<0s.t.g(x,y)=0ORg(x,y)<0

因为不等式约束条件难度稍大，我们先从等式约束条件开始。

#### 等式约束条件

假设，我们的目标函数是：

minf(x,y)minf(x,y)

s.t.g(x,y)=0s.t.g(x,y)=0

现在我们在一张图中做出 f(x,y)f(x,y) 和 g(x,y)g(x,y) 的等高线（三维图形投影到二维平面后的结果），形如图3：

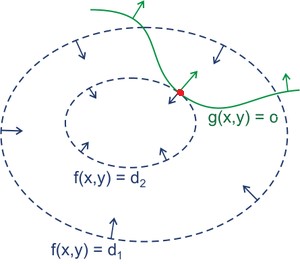


图 3

绿线是 g(x,y)g(x,y) 的等高线，因为约束条件是 g(x,y)=0g(x,y)=0，因此，只有一条 g(x,y)g(x,y)的等高线——g(x,y)=0g(x,y)=0 对我们有意义，因此只画它即可。

蓝线是 f(x,y)f(x,y) 的等高线。图中两条蓝线具体对应的函数分别是 f(x,y)=d1f(x,y)=d1 和 f(x,y)=d2f(x,y)=d2。

d1d1 和 d2d2 是两个常数，对应上图中两个蓝圈对应的 zz 轴坐标。此处，d2<d1d2<d1。

图3中红点，映射到 f(x,y)f(x,y) 上，就是取得 f(x,y)f(x,y) 符合 g(x,y)=0g(x,y)=0 的约束的极小值的点。

我们设红点的自变量值为 (x∗,y∗)(x∗,y∗)。

此处我们需要用到一个概念——****函数的梯度****：表示该函数在某点处的方向导数，方向导数是某个多维函数上的点沿每个维度分别求导后，再组合而成的向量。

记作：

∇x,yf=(∂f∂x,∂f∂y)∇x,yf=(∂f∂x,∂f∂y)， ∇x,yg=(∂g∂x,∂g∂y)∇x,yg=(∂g∂x,∂g∂y)

我们知道，一个****函数的梯度与它的等高线垂直****。

因此，在红点处，f(x∗,y∗)f(x∗,y∗) 的梯度与 f(x,y)=d2f(x,y)=d2 在 (x∗，y∗)(x∗，y∗) 处的切线垂直，g(x∗,y∗)g(x∗,y∗) 的梯度与 g(x,y)=0g(x,y)=0 在 (x∗，y∗)(x∗，y∗) 处的切线垂直。

又因为 f(x,y)=d2f(x,y)=d2 对应的蓝线与 g(x,y)=0g(x,y)=0 对应的绿线在 (x∗，y∗)(x∗，y∗) 处是相切的。

注意：为什么相切？反过来想，如果它们不相切会怎样？

要么不相交，这样的话 g(x,y)=0g(x,y)=0 就没法约束 f(x,y)f(x,y)了；

要么有两个交点，如果这样，我们可以再做另一条 f(x,y)f(x,y) 的等高线，让新等高线和 g(x,y)=0g(x,y)=0 相切，再取那个切点为 (x∗,y∗)(x∗,y∗)。如此一来，它们还是在 (x∗,y∗)(x∗,y∗) 处相切。

在 (x∗,y∗)(x∗,y∗) 点处 f(x,y)f(x,y) 与 g(x,y)g(x,y) 的梯度，要么方向相同，要么方向相反。

所以，一定存在 λ≠0λ≠0， 使得：

∇x,yf(x∗,y∗)+λ∇x,yg(x∗,y∗)=0∇x,yf(x∗,y∗)+λ∇x,yg(x∗,y∗)=0

这时我们将 λλ****称为拉格朗日乘子****！

#### 拉格朗日乘子和拉格朗日函数

****定义拉格朗日函数****：L(x,y,λ)=f(x,y)+λg(x,y)L(x,y,λ)=f(x,y)+λg(x,y) —— 其中 λλ 是拉格朗日乘子。

拉格朗日函数把原本的目标函数和其限制条件整合成了一个函数。

拉格朗日函数对 x，yx，y 求偏导：

L′x,y(x,y,λ)=f′x,y(x,y)+λg′x,y(x,y)Lx,y′(x,y,λ)=fx,y′(x,y)+λgx,y′(x,y)

我们令拉格朗日函数对 x，yx，y 求偏导的结果为 00，则有：

f′x,y(x,y)+λg′x,y(x,y)=0fx,y′(x,y)+λgx,y′(x,y)=0

我们刚才已经知道了，f(x,y)f(x,y) 符合 g(x,y)=0g(x,y)=0 约束的极小值的点 (x∗,y∗)(x∗,y∗) 满足：∇x,yf(x∗,y∗)+λ∇x,yg(x∗,y∗)=0∇x,yf(x∗,y∗)+λ∇x,yg(x∗,y∗)=0，用导函数表示就是：f′x,y(x∗,y∗)+λg′x,y(x∗,y∗)=0fx,y′(x∗,y∗)+λgx,y′(x∗,y∗)=0。也就是说我们要求的极小值点正好满足拉格朗日函数对 x、yx、y 求导后，令其结果为 0 形成的导函数。

既然如此，我们当然也就是可以直接引入一个新的变量：λλ —— 拉格朗日乘子，和目标函数、约束条件一起，先构造出拉格朗日函数 L(x,y,λ)L(x,y,λ)。

然后再令 L′x,y(x,y,λ)=0Lx,y′(x,y,λ)=0，之后通过最优化 L′x,y(x,y,λ)=0Lx,y′(x,y,λ)=0，求取 (x∗,y∗)(x∗,y∗) 的值。

另外，而令拉格朗日函数对 λλ 求偏导的结果为 00：L′λ(x,y,λ)=0Lλ′(x,y,λ)=0，就得到了约束条件：g(x,y)=0g(x,y)=0。

拉格朗日函数也可以通过求导变化成约束条件本身。

于是，****原本有约束的优化问题，就可以转化为对拉格朗日函数的无约束优化问题了****。

#### 不等式约束条件

了解了****约束条件是等式的情况****，我们再来看****约束条件是不等式****的情况。

当条件是：g(x,y)⩽0g(x,y)⩽0 时, 我们可以它拆分成：g(x,y)=0ORg(x,y)<0g(x,y)=0ORg(x,y)<0，两种情况来看。

这两种情况下可以先各取一个极小值，再比较一下，哪个更小就用哪个。

##### ****拆分后的严格不等式约束条件：****g(x,y)<0g(x,y)<0

而对于 g(x,y)<0g(x,y)<0 情况，因为 <0<0 是一个区间，而不再是对应到三维坐标中的某个固定的曲面。

这种情况下，如果 f(x,y)f(x,y) 的极小值就在这个区间里，然后直接对 f(x,y)f(x,y) 求无约束的极小值就好了。

对应到拉格朗日函数就是：****令****λ=0λ=0****，直接通过令****f(x,y)f(x,y)****的梯度为****00****求****f(x,y)f(x,y)****的极小值****。

求出极小值 f(x∗,y∗)f(x∗,y∗) 之后，再判断 (x∗,y∗)(x∗,y∗) 是否符合约束条件。

##### ****拆分后的等式约束条件：****g(x,y)=0g(x,y)=0

g(x,y)=0g(x,y)=0 可以参照上一小节的做法。

有一点要****注意****，在仅有等式约束条件时，我们设 f′x,y(x∗,y∗)+λg′x,y(x∗,y∗)=0fx,y′(x∗,y∗)+λgx,y′(x∗,y∗)=0，只要常数 λ≠0λ≠0 就可以了。

但是在以从 g(x,y)⩽0g(x,y)⩽0 中拆分出来的等式为约束条件时，我们要求：存在常数 λ>0λ>0, 使得 f′x,y(x∗,y∗)+λg′x,y(x∗,y∗)=0fx,y′(x∗,y∗)+λgx,y′(x∗,y∗)=0。

这是为什么呢？我们先来可视化地看看下面不等式约束的两种情况：

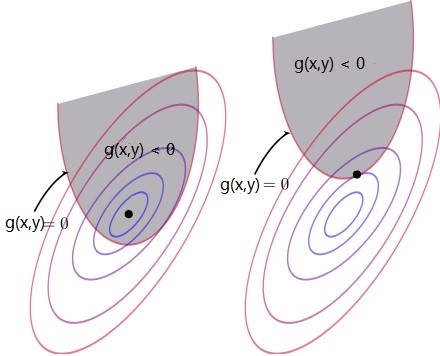


图 4

图4中黑点表示最终满足约束条件的函数极小值。

如果是图4左侧的状况，限制条件就又变成了 g(x,y)<0g(x,y)<0，且直接求 f(x,y)f(x,y) 极小值就行。

只有是右侧情况时，不等式约束条件才又变成了等式。这种情况，才是我们在拆分了不等式条件后有意义的情况。

在这种情况下，最终找到的符合约束条件的极小值 f(x∗,y∗)f(x∗,y∗) 肯定不是 f(x,y)f(x,y) 的极小值。

怎么用数学方式来表现这一点呢？

大家回想一下，****函数梯度的物理意义****：实际上，一个函数在某一点的梯度方向（向量方向）指明了该函数的函数值相对于当前函数值增量的方向。

也就是说：我们在 g(x∗,y∗)g(x∗,y∗) 的梯度方向上前进一个长度：ϵϵ（ϵϵ 很小很小），到达点 (x∗∗,y∗∗)(x∗∗,y∗∗)，则 g(x∗∗，y∗∗)>g(x∗,y∗)g(x∗∗，y∗∗)>g(x∗,y∗)。

注意：此处可以用 z=x+yz=x+y 和 z=−x−yz=−x−y 来直观感受一下。二维平面（比如 z=0z=0 平面）上的 y=xy=x 是这两个函数共同的等高线投影，不过 z=x+yz=x+y 的导数方向为 (1,1)(1,1)，而 z=−x−yz=−x−y 的导数方向为 (−1,−1)(−1,−1)，两个导向量共线，但方向相反，正好分别指向各自上升的区域。

因为 g(x∗,y∗)=0g(x∗,y∗)=0，又根据梯度的物理意义，则 (x∗,y∗)(x∗,y∗) 点处 g(⋅)g(⋅) 函数的梯度方向指向的是 g(x,y)>0g(x,y)>0 的区域。

偏偏我们的约束条件是 g(x，y)<0g(x，y)<0，那说明不等式约束条件所对应的区域与 (x∗,y∗)(x∗,y∗) 点处 g(⋅)g(⋅) 函数的梯度方向是相反的。

我们现在考虑的是图4右侧的情况，此时 f(⋅)f(⋅) 函数带约束条件的极小值点 (x∗,y∗)(x∗,y∗) 就位于 g(x,y)=0g(x,y)=0 对应的曲线上。

而 (x∗,y∗)(x∗,y∗) 点处 f(⋅)f(⋅) 函数的梯度方向正是 f(⋅)f(⋅) 函数本身的增量方向，因此在 g(x,y)<0g(x,y)<0 所在的区域对应的应该是 f(x,y)f(x,y) 的递增方向，如若不然，就应该是图4左侧的情况了，不是吗？

只有当 f(x,y)f(x,y) 梯度方向和 g(x,y)<0g(x,y)<0 区域所在方向相同，也就是和 (x∗,y∗)(x∗,y∗) 点处 g(⋅)g(⋅) 函数梯度方向相反，那么 f(x,y)f(x,y) 的约束条件极小值才会出现在 g(x,y)=0g(x,y)=0 的曲线上。

所以，在这种情况下，存在常数 λ>0λ>0，使得 f′x,y(x∗,y∗)=−λg′x,y(x∗,y∗)fx,y′(x∗,y∗)=−λgx,y′(x∗,y∗)，进一步导出：

f′x,y(x∗,y∗)+λg′x,y(x∗,y∗)=0，λ>0fx,y′(x∗,y∗)+λgx,y′(x∗,y∗)=0，λ>0

#### KKT 约束条件

我们将上面拆分开的严格不等式和等式两种情况再整合起来:

1. 如果严格不等式成立时得到 f(⋅)f(⋅) 函数的极小值，则有 λ=0λ=0，这样才能将拉格朗日函数直接转换为原始函数。则有 λg(x,y)=0λg(x,y)=0。
2. 如果等式成立时得到 f(⋅)f(⋅) 函数的约束条件极小值，则必然存在 λ>0λ>0，且同时 g(x,y)=0g(x,y)=0, 因此也有 λg(x,y)=0λg(x,y)=0。

于是，对于不等式约束条件 g(x,y)⩽0g(x,y)⩽0，最终的约束条件变成了：

g(x,y)⩽0g(x,y)⩽0;

λ⩾0λ⩾0;

λg(x,y)=0λg(x,y)=0

这样由1变3的约束条件，叫做 ****KKT 约束条件****。

#### 目标函数转化为拉格朗日函数

当然，KKT 条件不是单独成立的，它是拉格朗日乘子法的一部分！

当我们在约束条件下求解函数最优化问题，且约束条件为不等式时（问题描述如下）：

Optimizef(x,y)Optimizef(x,y)

s.t.g(x,y)⩽0s.t.g(x,y)⩽0

我们针对目标函数生成拉格朗日函数为：

L(x,y,λ)=f(x,y)+λg(x,y)L(x,y,λ)=f(x,y)+λg(x,y)

同时有 KKT 条件：

g(x,y)⩽0g(x,y)⩽0

λ⩾0λ⩾0

λg(x,y)=0λg(x,y)=0

假设最优化结果对应点为 (x∗,y∗)(x∗,y∗)，则当目标函数为：

minf(x,y)minf(x,y)

s.t.g(x,y)⩽0s.t.g(x,y)⩽0

时，若存在 λ≠0λ≠0，使得 f′x,y(x∗,y∗)+λg′x,y(x∗,y∗)=0fx,y′(x∗,y∗)+λgx,y′(x∗,y∗)=0，则 λ>0λ>0。

而当目标函数为：

maxf(x,y)maxf(x,y)

s.t.g(x,y)⩽0s.t.g(x,y)⩽0

时，若存在 λ≠0λ≠0，使得 f′x,y(x∗,y∗)+λg′x,y(x∗,y∗)=0fx,y′(x∗,y∗)+λgx,y′(x∗,y∗)=0，则 λ<0λ<0。

或者写作：此时，若存在 λ≠0λ≠0，使得 f′x,y(x∗,y∗)−λg′x,y(x∗,y∗)=0fx,y′(x∗,y∗)−λgx,y′(x∗,y∗)=0，则λ>0λ>0。

注意：上面几个式子要注意：

L(x,y,λ)=f(x,y)+λg(x,y)L(x,y,λ)=f(x,y)+λg(x,y) 是拉格朗日函数的定义；

λ⩾0λ⩾0 是之前我们推导出来的 KKT 条件；

存在常数 λ>0λ>0, 使得 f′x,y(x∗,y∗)+λg′x,y(x∗,y∗)=0fx,y′(x∗,y∗)+λgx,y′(x∗,y∗)=0 或者 f′x,y(x∗,y∗)−λg′x,y(x∗,y∗)=0fx,y′(x∗,y∗)−λgx,y′(x∗,y∗)=0 是推导 KKT 条件过程中经历的一个步骤——这个步骤中涉及到的是在 (x∗,y∗)(x∗,y∗) 这个确定点处 ff 函数和 gg 函数的梯度。

前面我们已经推导了求极小值时有 f′x,y(x∗,y∗)+λg′x,y(x∗,y∗)=0，λ>0fx,y′(x∗,y∗)+λgx,y′(x∗,y∗)=0，λ>0，至于求极大值时为什么 λλ 符号不同，就需要同学们自己再去推导一遍了。[邀请好友一起学，获得 25% 返现奖励](https://gitbook.cn/m/mazi/columns/5bc6ac7442d7d32f50f19a98/topics/5bf28289fd72950cafdcd0a3)

[IMG_261](https://gitbook.cn/m/mazi/comp/column?columnId=5bc6ac7442d7d32f50f19a98%26tag=2#catalog)