凸优化、对偶与拉格朗日函数

信息学院-黄浩

2022年3月14日星期一

概述

- 约束条件分为等式约束与不等式约束，对于等式约束的优化问题，可以直接应用拉格朗日乘子法去求取最优值；
- 对于含有不等式约束的优化问题，可以转化为在满足 KKT 约束条件下应用拉格朗日乘子法求解。
- 拉格朗日求得的并不一定是最优解，只有在凸优化的情况下，才能保证得到的是**最优解**，所以拉格朗日乘子法得到的为**可行解**，其实就是局部极小值。

概述

无约束优化

首先考虑一个不带任何约束的优化问题，对于变量 $x \in \mathbb{R}^N$ 的函数 $f(x)$,无约束优化问题如下:

$$\min_{x} f(x)$$

该问题很好解，直接找到使目标函数得 0 的点即可 即 $\nabla_x f(x) = 0$ ，如果没有解析解的话，可以使用梯度下降或牛顿方法等迭代的手段来使 x 沿负梯度方向逐步逼近极小值点。

概述

等式约束优化

当目标函数加上约束条件之后，问题就变成如下形式：

$$\min_x \ f(x)$$
$$s.t. \quad h_i(x) = 0, i = 1, 2, \ldots, m$$

约束条件会将解的范围限定在一个可行域，此时不一定能找到使得 $\nabla_x f(x)$ 为 0 的点，只需找到在可行域内使得 $f(x)$ 最小的值即可，常用的方法即为拉格朗日乘子法，该方法首先引入 Lagrange Multiplier $\alpha \in \mathbb{R}^m$ ，构建 Lagrangian 如下：

$$L(x, \alpha) = f(x) + \sum_{i=1}^{m} \alpha_i h_i(x)$$

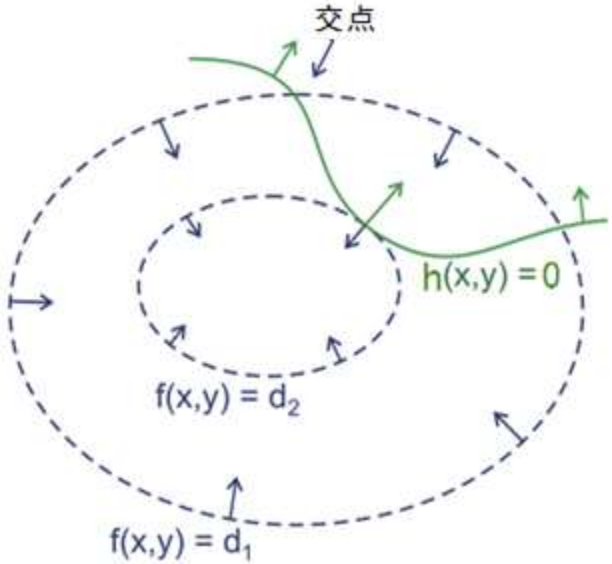
求解方法如下：首先对 Lagrangian 关于 α 与 x 求：

$$\begin{cases} \nabla_x L(x, \alpha) = 0 \\ \nabla_\alpha L(x, \alpha) = 0 \end{cases}$$

概述

令导数为 0，求得 x 、α 的值后，将 x 带入 $f(x)$ 即为在约束条件 $h_i(x)$ 下的可行解。这样做的意义是什么呢? 接下来看一个直观的示例，对于二维情况下的目标函数是 $f(x,y)$，在平面中画出 $f(x,y)$ 的等高线，如下图的虚线所示， 并只给出一个约束等式 $h(x,y)=0$ ，如下图的绿线所示，目标函数 $f(x,y)$ 与约束 $g(x,y)$ 只有三种情况，相交、相切或者没有交集，没交集肯定不是解,只有相交或者相切可能是解，但相交得到的一定不是最优值，因为相交意味着肯定还存在其它的等高线在该条等高线的内部或者外部，使得新的等高线与目标函数的交点的值更大或者更小，这就意味着只有等高线与目标函数的曲线相切的时候，才可能得到可行解.

概述



概述

因此给出结论：拉格朗日乘子法取得极值的必要条件是目标函数与约束函数相切，这时两者的法向量是平行的，即

$$\nabla_x f(x) - \alpha \nabla_x h(x) = 0$$

所以只要满足上述等式，且满足之前的约束 $h_i(x) = 0, i = 1, 2, \ldots, m$ ，即可得到解，联立起来，正好得到就是拉格朗日乘子法。

概述

不等式约束优化

当约束加上不等式之后，情况变得更加复杂，首先来看一个简单的情况，给定如下不等式约束问题：

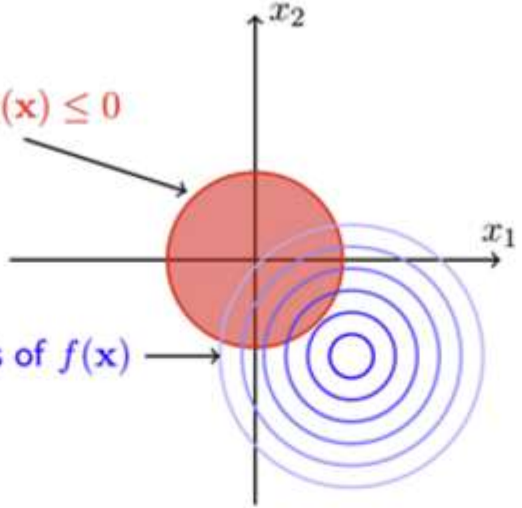
$$\min_{x}\ f(x)$$
$$s.t.\ \ g(x)\leq 0$$

对应的 Lagrangian 与图形分别如下所示：

$$L(x,\lambda)=f(x)+\lambda g(x)$$

这时的可行解必须落在约束区域 $g(x)$ 之内，下图给出了目标函数的等高线与约束：

概述

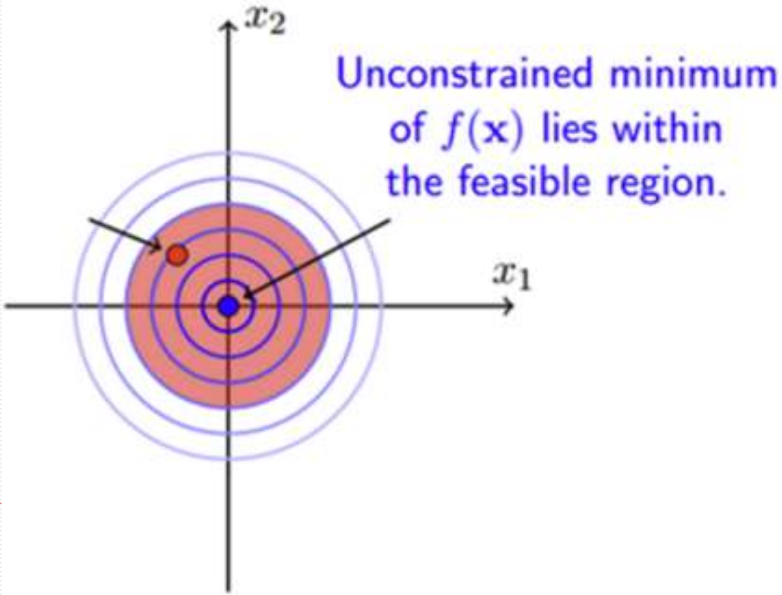


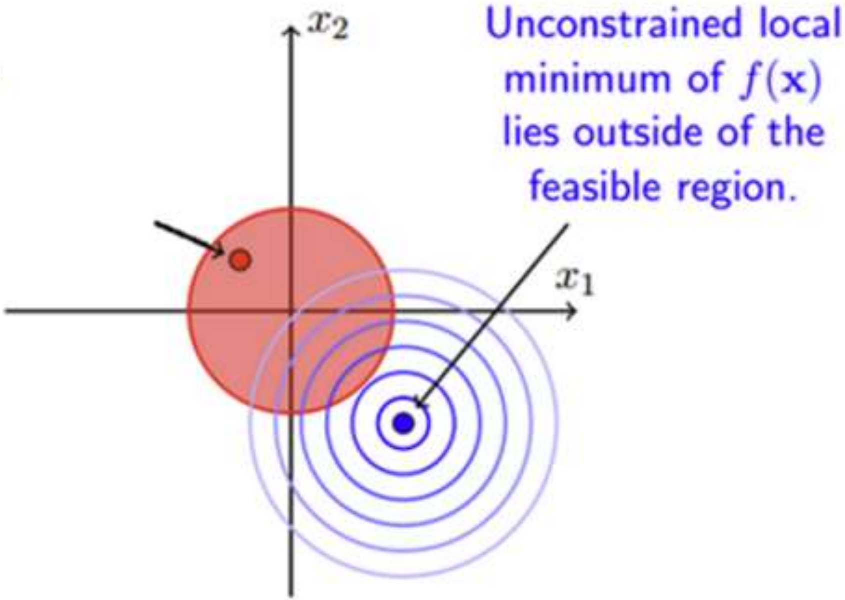
由图可见可行解 x 只能在 $g(x) < 0$ 或者 $g(x) = 0$ 的区域里取得:

- 当可行解 x 落在 $g(x) < 0$ 的区域内，此时直接极小化 $f(x)$ 即可;
- 当可行解 x 落在 $g(x) = 0$ 即边界上，此时等价于等式约束优化问题.

概述

当约束区域包含目标函数原有的的可行解时，此时加上约束可行解扔落在约束区域内部，对应 $g(x) < 0$ 的情况，这时约束条件不起作用；当约束区域不包含目标函数原有的可行解时，此时加上约束后可行解落在边界 $g(x) = 0$ 上。下图分别描述了两种情况，右图表示加上约束可行解会落在约束区域的边界上。





概述

以上两种情况就是说，要么可行解落在约束边界上即得 $g(x)=0$ ，要么可行解落在约束区域内部，此时约束不起作用，另 $\lambda=0$ 消去约束即可，所以无论哪种情况都会得到：

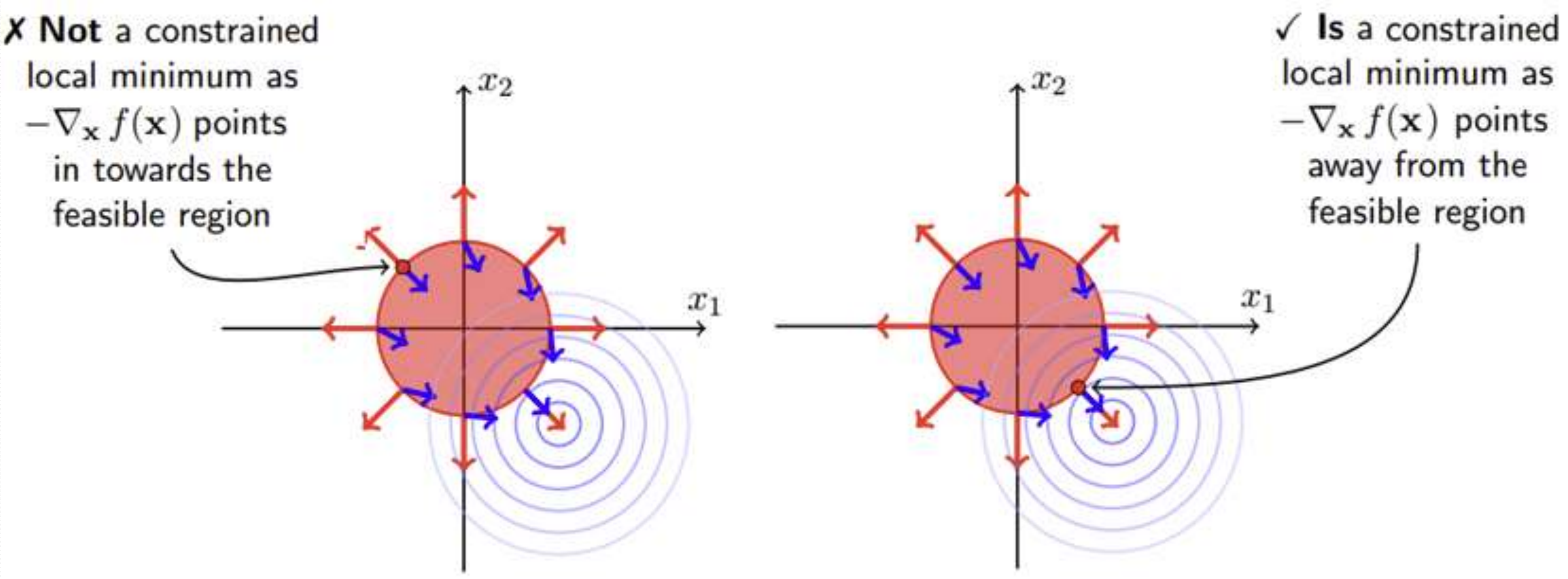
$$\lambda g(x)=0$$

还有一个问题是 λ 的取值，在等式约束优化中，约束函数与目标函数的梯度只要满足平行即可，而在不等式约束中则不然，若 $\lambda\neq 0$，这便说明 可行解 x 是落在约束区域的边界上的，这时可行解应尽量靠近无约束时的解，所以在约束边界上，目标函数的负梯度方向应该远离约束区域朝向无约束时的解，此时正好可得约束函数的梯度方向与目标函数的负梯度方向应相同：

$$-\nabla_x f(x)=\lambda\nabla_x g(x)$$

上式需要满足的要求是拉格朗日乘子 $\lambda>0$ ，这个问题可以举一个形象的例子，假设你去爬山，目标是山顶，但有一个障碍挡住了通向山顶的路，所以只能沿着障碍爬到尽可能靠近山顶的位置，然后望着山顶叹叹气，这里山顶便是目标函数的可行解，障碍便是约束函数的边界，此时的梯度方向一定是指向山顶的，与障碍的梯度同向，下图描述了这种情况：

概述


✗ **Not** a constrained local minimum as $-\nabla_{\mathbf{x}} f(\mathbf{x})$ points in towards the feasible region
✓ **Is** a constrained local minimum as $-\nabla_{\mathbf{x}} f(\mathbf{x})$ points away from the feasible region
x_2
x_1
x_2
x_1

可见**对于不等式约束，只要满足一定的条件，依然可以使用拉格朗日乘子法解决，这里的条件便是 KKT 条件**。接下来给出形式化的 KKT 条件 首先给出形式化的不等式约束优化问题：

$$\begin{aligned} &\min_x \ f(x) \\ &s.t. \ \ h_i(x) = 0, \ i = 1,2,\ldots,m \\ &\qquad g_j(x) \le 0, \ j = 1,2,\ldots,n \end{aligned}$$

列出 Lagrangian 得到无约束优化问题：

$$L(x,\alpha,\beta) = f(x) + \sum_{i=1}^{m} \alpha_i h_i(x) + \sum_{j=1}^{n} \beta_i g_i(x)$$

经过之前的分析，便得知加上不等式约束后可行解 x 需要满足的就是以下的 KKT 条件：

$$\begin{aligned} \nabla_x L(x,\alpha,\beta) &= 0 &(1) \\ \beta_j g_j(x) &= 0, \ j = 1,2,\ldots,n &(2) \\ h_i(x) &= 0, \ i = 1,2,\ldots,m &(3) \\ g_j(x) &\le 0, \ j = 1,2,\ldots,n &(4) \\ \beta_j &\ge 0, \ j = 1,2,\ldots,n &(5) \end{aligned}$$

概述

满足 KKT 条件后极小化 Lagrangian 即可得到在不等式约束条件下的可行解。 KKT 条件看起来很多，其实很好理解：

(1)：拉格朗日取得可行解的必要条件；

(2)：这就是以上分析的一个比较有意思的约束，称作松弛互补条件；

(3) ～ (4)：初始的约束条件；

(5)：不等式约束的 Lagrange Multiplier 需满足的条件。

主要的KKT条件便是 (3) 和 (5)，只要满足这俩个条件便可直接用拉格朗日乘子法， SVM 中的支持向量便是来自于此，需要注意的是 KKT 条件与对偶问题也有很大的联系。

概述

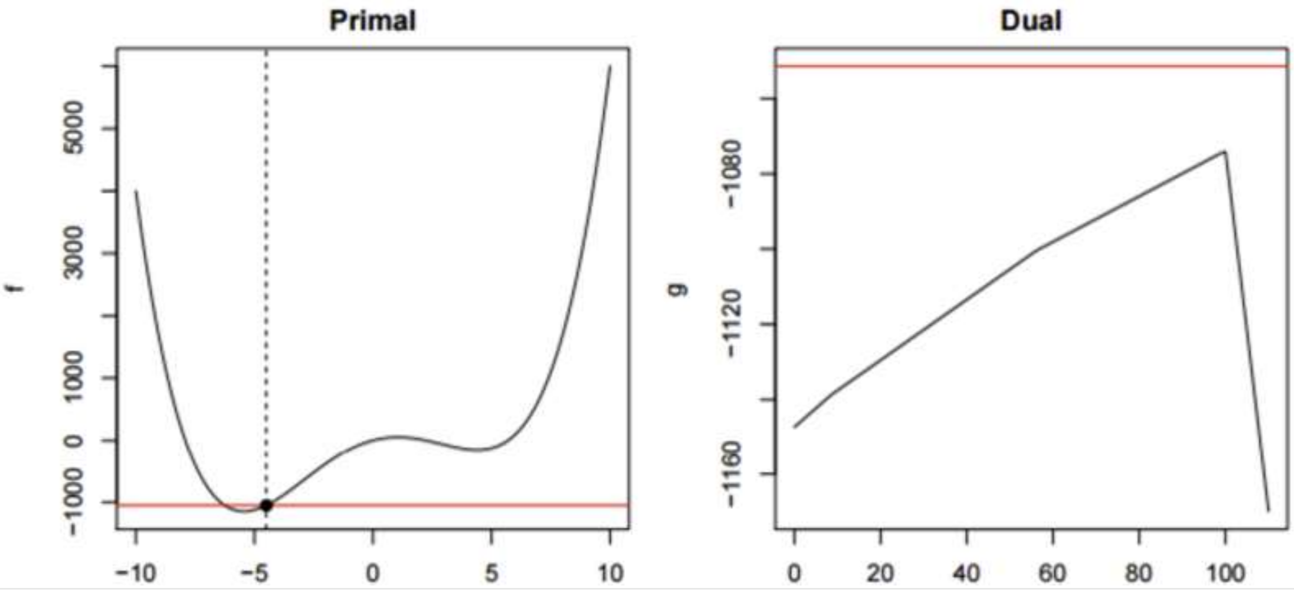
在优化理论中，目标函数 $f(x)$ 会有多种形式：如果目标函数和约束条件都为变量 x 的线性函数，称该问题为**线性规划**；如果目标函数为二次函数，约束条件为线性函数，称该最优化问题为**二次规划**；如果目标函数或者约束条件均为非线性函数，称该最优化问题为**非线性规划**。每个线性规划问题都有一个与之对应的**对偶问题**，对偶问题有非常良好的性质，以下列举几个：

- 对偶问题的对偶是原问题；
- 无论原始问题是否是凸的，对偶问题都是凸优化问题；
- 对偶问题可以给出原始问题一个下界；
- 当满足一定条件时，原始问题与对偶问题的解是完全等价的；

概述

比如下边这个例子，虽然原始问题非凸，但是对偶问题是凸的：

$$\min_x \quad (x^4 - 50x^2 + 100x)$$
$$s.t. \quad x \geq 4.5$$



概述

原始问题

开始步入正题，首先给出不等式约束优化问题：

$$\begin{aligned} &\min_{x} \ f(x) \\ &s.t. \quad h_i(x) = 0, \ i = 1, 2, \ldots, m \\ &\qquad\quad g_j(x) \leq 0, \ j = 1, 2, \ldots, n \end{aligned}$$

定义 Lagrangian 如下：

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m} \alpha_i h_i(x) + \sum_{j=1}^{n} \beta_j g_j(x)$$

根据以上 Lagrangian 便可以得到一个重要结论：

$$f(x) = \max_{\alpha\beta;\beta_i \geq 0} L(x, \alpha, \beta) > L(x, \alpha, \beta) \qquad (*)$$

概述

(*) 式很容易验证，因为满足约束条件的 x 会使得 $h_i(x)=0$ ，因此第二项消掉了；而 $g_j(x)\le 0$ ，并且使得 $\beta_j\ge 0$，因此会有 $\beta_j g_j(x)\le 0$ ，所以最大值只能在它们都取零的时候得到，这个时候就只剩下 $f(x)$ 了。反之如果有任意一个约束条件不满足，则只需令其相应的乘子 $\to +\infty$ ，则会得到 $L(x,\alpha,\beta)\to +\infty$，这样将导致问题无解，因此必须满足约束条件。经过这样一转变，约束都融合到了一起而得到如下的无约束的优化目标：

$$\min_x f(x)=\min_x \max_{\alpha,\beta;\beta_i\ge 0} L(x,\alpha,\beta)$$

对偶问题

上式与原优化目标等价，将之称作原始问题，将原始问题的解记做 p^*，如此便把带约束问题转化为了无约束的原始问题，其实只是一个形式上的重写，方便找到其对应的对偶问题，首先为对偶问题定义一个对偶函数（**dual function**）：

$$D(\alpha,\beta)=\min_x L(x,\alpha,\beta)$$

有了对偶函数就可给出对偶问题了，与原始问题的形式非常类似，只是把 min 和 max 交换了一下：

$$\max_{\alpha,\beta;\beta_i\geq 0}\ \min_x L(x,\alpha,\beta)$$

然后定义对偶问题的最优解即关于 α β 的函数：

$$d^*=\max_{\alpha,\beta;\beta_i\geq 0} D(\alpha,\beta)$$

对偶问题和原始问题的最优解并不相等，而是满足的如下关系：

$$d^* \leq p^*$$

直观地，可以理解为最小的里最大的那个要比最大的中最小的那个要大。具体的证明过程如下：

证明在这里，首先这里的约束要全部满足，对偶问题与原始问题的关系如下：

$$D(\alpha, \beta) = \min_x L(x, \alpha, \beta) \leq L(x, \alpha, \beta) \leq \max_{\alpha,\beta,\beta_i \geq 0} L(x, \alpha, \beta) = f(x)$$

即 $D(\alpha, \beta) \leq f(x)$，所以自然而然可得：

$$d^* = \max_{\alpha,\beta;\beta_i \geq 0} D(\alpha, \beta) \leq \min_x f(x) = p^*$$

即现在通过对偶性，为**原始问题引入一个下界**，$d^* \leq p^*$.

概述

这个性质便叫做**弱对偶性（weak duality）**，对于所有优化问题都成立，即使原始问题非凸。这里还有两个概念：$f(x)-D(\alpha,\beta)$ 叫做对偶间隔（duality gap），p^*-d^* 叫做最优对偶间隔（optimal duality gap）。

之前提过无论原始问题是什么形式，对**偶问题总是一个凸优化的问题**，这样对于那些难以求解的原始问题（甚至是 NP 问题），均可以通过转化为偶问题，通过优化这个对偶问题来得到原始问题的一个下界，与弱对偶性相对应的有一个**强对偶性（strong duality）**，强对偶即满足：

$$d^* = p^*$$

概述

强对偶是一个非常好的性质，**因为在强对偶成立的情况下，可以通过求解对偶问题来得到原始问题的解**，在 SVM 中就是这样做的。当然并不是所有的对偶问题都满足强对偶性，在 SVM 中是直接假定了强对偶性的成立，其实只要满足一些条件，强对偶性是成立的，比如说 **Slater 条件与KKT条件。**

Slater 条件

若原始问题为凸优化问题，且存在严格满足约束条件的点 x ，这里的"严格"是指 $g_i(x) \leq 0$ 中的"$\leq$"严格取到"$<$"，即存在 x 满足 $g_i(x) < 0$,$i = 1, 2, \ldots, n$ ，则存在 x^*, α^*, β^* 使得 x^* 是原始问题的解， α^*, β^* 是对偶问题的解，且满足：

$$p^* = d^* = L(x^*, \alpha^*, \beta^*)$$

概述

也就是说如果原始问题是凸优化问题并且满足 Slater 条件的话，那么强对偶性成立。需要注意的是，这里只是指出了强对偶成立的一种情况，并不是唯一的情况。例如，对于某些非凸优化的问题，强对偶也成立。SVM 中的原始问题 是一个凸优化问题（二次规划也属于凸优化问题），Slater 条件在 SVM 中指的是存在一个超平面可将数据分隔开，即数据是线性可分的。当数据不可分时，强对偶是不成立的，这个时候寻找分隔平面这个问题本身也就是没有意义了，所以对于不可分的情况预先加个 kernel 就可以了。

概述

KKT条件

假设 x^* 与 α^*, β^* 分别是原始问题（并不一定是凸的）和对偶问题的最优解，且满足强对偶性，则相应的极值的关系满足：

$$\begin{aligned}
f(x^*) &= d^* = p^* = D(\alpha^*, \beta^*) \\
&= \min_x f(x) + \sum_{i=1}^{m} \alpha_i^* h_i(x) + \sum_{j=1}^{n} \beta_j^* g_j(x) \\
&\le f(x^*) + \sum_{i=1}^{m} \alpha_i^* h_i(x^*) + \sum_{j=1}^{n} \beta_j^* g_j(x^*) \\
&\le f(x^*)
\end{aligned}$$

这里第一个不等式成立是因为 x^* 为 $L(x, \alpha^*, \beta^*)$ 的一个极大值点，最后一个不等式成立是因为 $h_i(x^*) = 0$,且 $g_j(x^*) \le 0, \beta_j \ge 0$，（$\beta_j \ge 0$ 是之前 (*) 式的约束条件）因此这一系列的式子里的不等号全部都可以换成等号。根据公式还可以得到两个结论：

概述

1）第一个不等式成立是因为 x^* 为 $L(x,\alpha^*,\beta^*)$ 的一个极大值点，由此可得:

$$\nabla_{x^*} L(x,\alpha^*,\beta^*) = 0$$

2）第二个不等式其实就是之前的（*）式，$\beta_j^* g_j(x^*)$ 都是非正的，所以这里有:

$$\beta_j^* g_j(x^*) = 0,\ i = 1,2,\ldots,m$$

也就是说如果 $\beta_j^* > 0$，那么必定有 $g_j(x^*) = 0$；反过来，如果 $g_j(x^*) < 0$ 那么可以得到 $\beta_j^* = 0$，即:

$$\begin{cases} \beta_j^* > 0 \Rightarrow g_j^*(x) = 0 \\ g_j^*(x) < 0 \Rightarrow \beta_j^* = 0 \end{cases}$$

这些条件都似曾相识，把它们写到一起，就是传说中的 KKT (Karush-Kuhn-Tucker) 条件

概述

$$\nabla_x L(x, \alpha, \beta) = 0 \tag{1}$$

$$\beta_j g_j(x) = 0,\ j = 1, 2, \ldots, n \tag{2}$$

$$h_i(x) = 0,\ i = 1, 2, \ldots, m \tag{3}$$

$$g_j(x) \leq 0,\ j = 1, 2, \ldots, n \tag{4}$$

$$\beta_j \geq 0,\ j = 1, 2, \ldots, n \tag{5}$$

总结来说就是说**任何满足强对偶性的优化问题，只要其目标函数与约束函数可微，任一对原始问题与对偶问题的解都是满足 KKT 条件的**。即满足强对偶性的优化问题中，若 x^* 为原始问题的最优解，α^*, β^* 为对偶问题的最优解，则可得 x^*, α^*, β^* 满足 KKT 条件。

概述

上面只是说明了必要性，当满足原始问题为凸优化问题时，必要性也是满足的，也就是说当原始问题是凸优化问题,且存在 x^*, α^*, β^* 满足 KKT 条件，那么它们分别是原始问题和对偶问题的极值点并且强对偶性成立，证明如下：

首先原始问题是凸优化问题，固定 α^*, β^* 之后对偶问题 $D(\alpha^*, \beta^*)$ 也是一个凸优化问题，x^* 是 $L(x, \alpha^*, \beta^*)$ 的极值点：

$$
\begin{aligned}
D(\alpha^*, \beta^*) &= \min_x L(x, \alpha^*, \beta^*) \\
&= L(x^*, \alpha^*, \beta^*) \\
&= f(x^*) + \sum_{i=1}^{m} \alpha_i^* h_i(x^*) + \sum_{j=1}^{n} \beta_j^* g_j(x^*) \\
&= f(x^*)
\end{aligned}
$$

最后一个式子是根据 KKT 条件中的 $h_i(x) = 0$ 与 $\beta_j g_j(x) = 0$ 得到的。这样一来，就证明了对偶间隔为零，也就是说，强对偶成立。

概述

总结一下对偶的基本概念，对于一个约束优化问题，找到其对偶问题，当弱对偶成立时，可以得到原始问题的一个下界。而如果强对偶成立，则可以直接求解对偶问题来解决原始问题。 SVM 就是这样的。对偶问题由于性质良好一般比原始问题更容易求解，在 SVM 中通过引入对偶问题可以将问题表示成数据的内积形式从而使得 kernel trick 的应用更加自然。此外，还有一些情况会同时求解对偶问题与原始问题 ，比如在迭代求解的过程中，通过判断对偶间隔的大小，可以得出一个有效的迭代停止条件。

谢谢