零基础学SVM—Support Vector Machine(一)

耳东陈
高校教师，机器学习与计算机视觉

如果你是一名模式识别专业的研究生，又或者你是机器学习爱好者，SVM是一个你避不开的问题。如果你只是有一堆数据需要SVM帮你处理一下，那么无论是Matlab的SVM工具箱，LIBSVM还是python框架下的SciKit Learn都可以提供方便快捷的解决方案。但如果你要追求的不仅仅是会用，还希望挑战一下“理解”这个层次，那么你就需要面对一大堆你可能从来没听过的名词，比如：非线性约束条件下的最优化、KKT条件、拉格朗日对偶、最大间隔、最优下界、核函数等等。这些名词往往会跟随一大堆天书一般的公式。如果你稍微有一点数学基础，那么单个公式你可能看得明白，但是怎么从一个公式跳到另一个公式就让人十分费解了，而最让人糊涂的其实并不是公式推导，而是如果把这些公式和你脑子里空间构想联系起来。

我本人就是上述问题的受害者之一。我翻阅了很多关于SVM的书籍和资料，但没有找到一份材料能够在公式推导、理论介绍，系统分析、变量说明、代数和几何意义的解释等方面完整地对SVM加以分析和说明的。换言之，对于普通的一年级非数学专业的研究生而言，要想看懂SVM需要搜集很多资料，然后对照阅读和深入思考，才可能比较透彻地理解SVM算法。由于我本人也在东北大学教授面向一年级硕士研究生的《模式识别技术与应用》课程，因此希望能总结出一份相对完整、简单和透彻的关于SVM算法的介绍文字，以便学生能够快速准确地理解SVM算法。

以下我会分为四个步骤对最基础的线性SVM问题加以介绍，分别是1）问题原型，2）数学模型，3）最优化求解，4）几何解释。我尽可能用最简单的语言和最基本的数学知识对上述问题进行介绍，希望能对困惑于SVM算法的学生有所帮助。

由于个人时间有限，只能找空闲的时间更新，速度会比较慢，请大家谅解。

一、SVM算法要解决什么问题

SVM的全称是Support Vector Machine，即支持向量机，主要用于解决模式识别领域中的数据分类问题，属于有监督学习算法的一种。SVM要解决的问题可以用一个经典的二分类问题加以描述。如图1所示，红色和蓝色的二维数据点显然是可以被一条直线分开的，在模式识别领域称为线性可分问题。然而将两类数据点分开的直线显然不止一条。图1(b)和(c)分别给出了A、B两种不同的分类方案，其中黑色实线为分界线，术语称为“决策面”。每个决策面对应了一个线性分类器。虽然在目前的数据上看，这两个分类器的分类结果是一样的，但如果考虑潜在的其他数据，则两者的分类性能是有差别的。

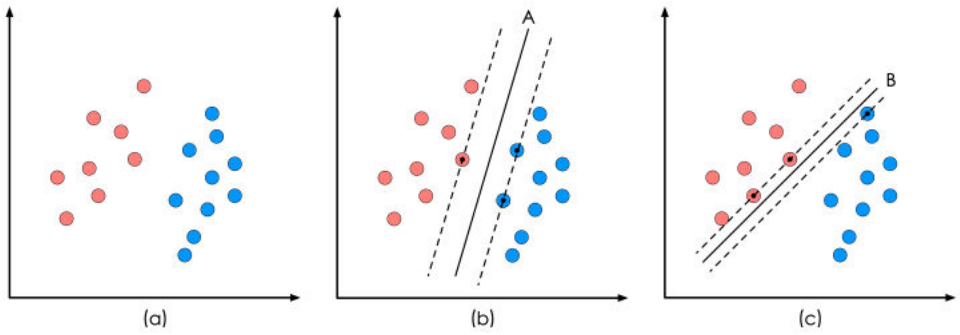


图1 二分类问题描述

SVM算法认为图1中的分类器A在性能上优于分类器B，其依据是A的分类间隔比B要大。这里涉及到第一个SVM独有的概念“分类间隔”。在保证决策面方向不变且不会出现错分样本的情况下移动决策面，会在原来的决策面两侧找到两个极限位置（越过该位置就会产生错分现象），如虚线所示。虚线的位置由决策面的方向和距离原决策面最近的几个样本点决定，而这两条虚线之间的分界线就是在保持当前决策面方向不变的

是这个最优决策面对应的分类间隔。显然每一个可能把数据集正确分开的方向都有一个最优决策面（*有些方向无论如何移动决策面的位置也不可能将两类样本完全分开*），而不同方向的最优决策面的分类间隔通常是不同的，那个具有“最大间隔”的决策面就是SVM要寻找的最优解。而这个真正的最优解对应的两侧虚线所穿过的样本点，就是SVM中的支持样本点，称为“支持向量”。对于图1中的数据，A决策面就是SVM寻找的最优解，而相应的三个位于虚线上的样本点在坐标系中对应的向量就叫做支持向量。

从表面上看，我们优化的对象似乎是这个决策面的方向和位置。但实际上最优决策面的方向和位置完全取决于选择哪些样本作为支持向量。而在经过漫长的公式推导后，你最终会发现，其实与线性决策面的方向和位置直接相关的参数都会被约减掉，最终结果只取决于样本点的选择结果。

到这里，我们明确了SVM算法要解决的是一个最优分类器的设计问题。既然叫作最优分类器，其本质必然是个最优化问题。所以，接下来我们要讨论的就是如何把SVM变成用数学语言描述的最优化问题模型，这就是我们在第二部分要讲的“线性SVM算法的数学建模”。

*关于“决策面”，为什么叫决策面，而不是决策线？好吧，在图1里，样本是二维空间中的点，也就是数据的维度是2，因此1维的直线可以分开它们。但是在更加一般的情况下，样本点的维度是n，则将它们分开的决策面的维度就是n-1维的超平面（*可以想象一下3维空间中的点集被平面分开*），所以叫“决策面”更加具有普适性，或者你可以认为直线是决策面的一个特例。

二、线性SVM算法的数学建模

一个最优化问题通常有两个最基本的因素：1）目标函数，也就是你希望什么东西的什么指标达到最好；2）优化对象，你期望通过改变哪些因素来使你的目标函数达到最优。在线性SVM算法中，目标函数显然就是那个“分类间隔”，而优化对象则是决策面。所以要对SVM问题进行数学建模，首先要对上述两个对象（*“分类间隔”和“决策面”*）进行数学描述。按照一般的思维习惯，我们先描述决策面。

2.1 决策面方程

（*请注意，以下的描述对于线性代数及格的同学可能显得比较啰嗦，但请你们照顾一下用高数课治疗失眠的同学们。*）

请你暂时不要纠结于n维空间中的n-1维超平面这种超出正常人想象力的情景。我们就老老实实地看看二维空间中的一根直线，我们从初中就开始学习的直线方程形式很简单。

$$y = ax + b \quad (2.1)$$

现在我们做个小小的改变，让原来的 x 轴变成 x_1 轴， y 变成 x_2 轴，于是公式(2.1)中的直线方程会变成下面的样子

$$x_2 = ax_1 + b \quad (2.2)$$

$$ax_1 + (-1)x_2 + b = 0 \quad (2.3)$$

公式（2.3）的向量形式可以写成

$$[a, -1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b = 0 \quad (2.4)$$

考虑到我们在等式两边乘上任何实数都不会改变等式的成立，所以我们可以写出一个更加一般的向量表达形式:

$$\boldsymbol{\omega}^T \boldsymbol{x} + \gamma = 0 \quad (2.5)$$

看到变量 $\boldsymbol{\omega}, \boldsymbol{x}$ 略显粗壮的身体了吗？他们是黑体，表示变量是个向量，$\boldsymbol{\omega} = [\omega_1, \omega_2]^T$，$\boldsymbol{x} = [x_1, x_2]^T$。一般我们提到向量的时候，都默认他们是个列向量，所以我在方括号[ ]后面加上了上标T，表示转置（我知道我真的很啰嗦可以帮忙把行向量竖过来变成列向量，所以

回到行向量。这样一个行向量 $\boldsymbol{\omega}^T$ 和一个列向量 $\boldsymbol{x}$ 就可快快乐乐的按照矩阵乘法的方式结合，变成一个标量，然后好跟后面的标量 γ 相加后相互抵消变成0。

就着公式(2.5)，我们再稍稍尝试深入一点。那就是探寻一下向量 $\boldsymbol{\omega} = [\omega_1, \omega_2]^T$ 和标量 γ 的几何意义是什么。让我们回到公式(2.4)，对比公式(2.5)，可以发现此时的 $\boldsymbol{\omega} = [a, -1]^T$ 。然后再去看公式(2.2)，还记得那条我们熟悉的直线方程中的a的几何意义吗？对的，那是直线的斜率。如果我们构造一个向量 $\boldsymbol{\phi} = [1, a]^T$ ，它应该跟我们的公式(2.2)描述的直线平行。然后我们求一下两个向量的点积 $\boldsymbol{\omega}^T\boldsymbol{\phi}$ ，你会惊喜地发现结果是0。我们管这种现象叫作“两个向量相互正交”。通俗点说就是两个向量相互垂直。当然，你也可以在草稿纸上自己画出这两个向量，比如让 $a = \sqrt{3}$,你会发现 $\boldsymbol{\phi} = [1, a]^T$ 在第一象限，与横轴夹角为60°，而 $\boldsymbol{\omega} = [a, -1]^T$ 在第四象限与横轴夹角为30°，所以很显然他们两者的夹角为90°。

你现在是不是已经忘了我们讨论正交或者垂直的目的是什么了？那么请把你的思维从坐标系上抽出来，回到决策面方程上来。我是想告诉你向量 $\boldsymbol{\omega} = [\omega_1, \omega_2]^T$ 跟直线 $\boldsymbol{\omega}^T\boldsymbol{x} + \gamma = 0$ 是相互垂直的，也就是说 $\boldsymbol{\omega} = [\omega_1, \omega_2]^T$ 控制了直线的方向。另外，还记得小时候我们学过的那个叫做截距的名词吗？对了， γ 就是截距，它控制了直线的位置。

然后，在本小节的末尾，我冒昧地提示一下，在n维空间中n-1维的超平面的方程形式也是公式(2.5)的样子，只不过向量 $\boldsymbol{\omega}, \boldsymbol{x}$ 的维度从原来的2维变成了n维。如果你还是想不出来超平面的样子，也很正常。那么就请你始终记住平面上它们的样子也足够了。

到这里，我们花了很多篇幅描述一个很简单的超平面方程（其实只是个直线方程），这里真正有价值的是这个控制方向的参数 $\boldsymbol{\omega}$ 。接下来，你会有很长一段时间要思考它到底是个什么东西，对于SVM产生了怎样的影响。

2.2 分类“间隔”的计算模型

我们在第一章里介绍过分类间隔的定义及其直观的几何意义。间隔的大小实际上就是支持向量对应的样本点到决策面的距离的二倍，如图2所示。

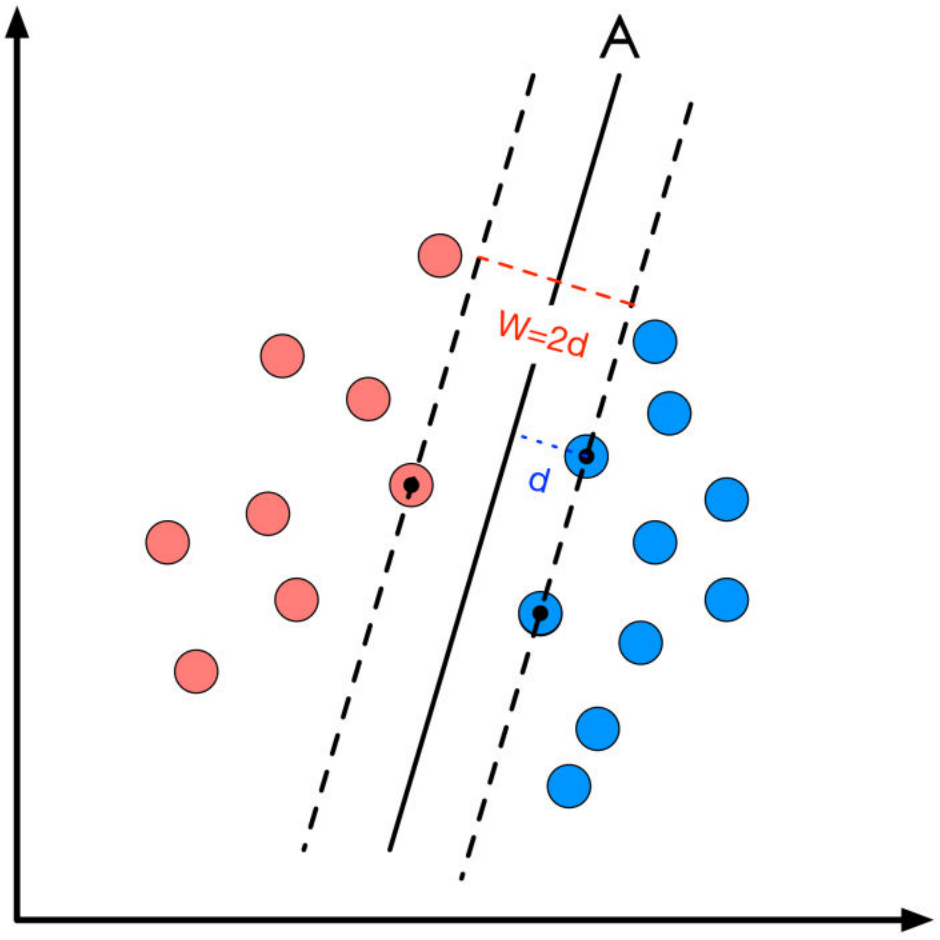


图2 分类间隔计算

所以分类间隔计算似乎相当简单，无非就是点到直线的距离公式。如果你想要回忆高中老师在黑板上推导的过程，可以随便在百度文库里搜索关键词“点到直线距离推导公式”，你会得到至少6、7种推导方法。但这里，请原谅我给出一个简单的公式如下：

$$d = \frac{|\omega^T x + \gamma|}{||\omega||} \quad (2.6)$$

这里$||\omega||$是向量ω的模，表示在空间中向量的长度，$x = [x_1, x_2]^T$就是支持向量样本点的坐标。ω, γ就是决策面方程的参数。而追求W的最大化也就是寻找d的最大化。看起来我们已经找到了目标函数的数学形式。

但问题当然不会这么简单，我们还需要面对一连串令人头疼的麻烦。

2.3 约束条件

接着2.2节的结尾，我们讨论一下究竟还有哪些麻烦没有解决：

1）并不是所有的方向都存在能够实现100%正确分类的决策面，我们如何判断一条直线是否能够将所有的样本点都正确分类？

2）即便找到了正确的决策面方向，还要注意决策面的位置应该在间隔区域的中轴线上，所以用来确定决策面位置的截距γ也不能自由的优化，而是受到决策面方向和样本点分布的约束。

3）即便取到了合适的方向和截距，公式(2.6)里面的x不是随随便便的一个样本点，而是支持向量对应的样本点。对于一个给定的决策面，我们该如何找到对应的支持向量？

以上三条麻烦的本质是“约束条件”，也就是说我们要优化的变量的取值范围受到了限制和约束。事实上约束条件一直是最优化问题里最让人头疼的东西。但既然我们已经论证了这些约束条件确实存在，就不得不用数学语言对他们进行描述。尽管上面看起来是3条约束，但SVM算法通过一些巧妙的小技巧，将这三条约束条件融合在了一个不等式里面。

我们首先考虑一个决策面是否能够将所有的样本都正确分类的约束。图2中的样本点分成两类（*红色和蓝色*），我们为每个样本点x_i加上一个类别标签y_i：

$$y_i = \begin{cases} +1 & \text{for blue points} \\ -1 & \text{for red points} \end{cases} \quad (2.7)$$

如果我们的决策面方程能够完全正确地对图2中的样本点进行分类，就会满足下面的公式

$$\begin{cases} \omega^T x_i + \gamma > 0 & \text{for}\sim\sim y_i = 1 \\ \omega^T x_i + \gamma < 0 & \text{for}\sim\sim y_i = -1 \end{cases} \quad (2.8)$$

如果我们要求再高一点，假设决策面正好处于间隔区域的中轴线上，并且相应的支持向量对应的样本点到决策面的距离为d，那么公式(2.8)就可以进一步写成：

$$\begin{cases} (\omega^T x_i + \gamma)/||\omega|| \geq d & \forall\, y_i = 1 \\ (\omega^T x_i + \gamma)/||\omega|| \leq -d & \forall\, y_i = -1 \end{cases} \quad (2.9)$$

符号$\forall$是“对于所有满足条件的” 的缩写。我们对公式(2.9)中的两个不等式的左右两边除上d，就可得到：

$$\begin{cases} \omega_d^T x_i + \gamma_d \geq 1 & \text{for}\sim\sim y_i = 1 \\ \omega_d^T x_i + \gamma_d \leq -1 & \text{for}\sim\sim y_i = -1 \end{cases} \quad (2.10)$$

其中

$$\omega_d = \frac{\omega}{||\omega|| d}, \quad \gamma_d = \frac{\gamma}{||\omega|| d}$$

把 ω_d 和 γ_d 就当成一条直线的方向矢量和截距。你会发现事情没有发生任何变化，因为直线 $\omega_d^T x + \gamma_d = 0$ 和直线 $\omega^T x + \gamma = 0$ 其实是一条直线。现在，现在让我忘记原来的直线方程参数 ω 和 γ，我们可以把参数 ω_d 和 γ_d 重新起个名字，就叫它们 ω 和 γ。我们可以直接说：“对于存在分类间隔的两类样本点，我们一定可以找到一些决策面，使其对于所有的样本点均满足下面的条件：”

$$\begin{cases} \omega^T x_i + \gamma \geq 1 & \text{for}\sim\sim y_i = 1 \\ \omega^T x_i + \gamma \leq -1 & \text{for}\sim\sim y_i = -1 \end{cases} \quad (2.11)$$

公式(2.11)可以认为是SVM优化问题的约束条件的基本描述。

2.4 线性SVM优化问题基本描述

公式(2.11)里面 $\omega^T x_i + \gamma = 1 \text{ or } -1$ 的情况什么时候会发生呢，参考一下公式(2.9)就会知道，只有当 x_i 是决策面 $\omega^T x + \gamma = 0$ 所对应的支持向量样本点时，等于1或-1的情况才会出现。这一点给了我们另一个简化目标函数的启发。回头看看公式(2.6)，你会发现等式右边分子部分的绝对值符号内部的表达式正好跟公式(2.11)中不等式左边的表达式完全一致，无论原来这些表达式是1或者-1，其绝对值都是1。所以对于这些支持向量样本点有：

$$d = \frac{|\omega^T x_i + \gamma|}{||\omega||} = \frac{1}{||\omega||}, \ \text{if } x_i \text{is a support vector} \quad (2.12)$$

公式(2.12)的几何意义就是，支持向量样本点到决策面方程的距离就是 $1/||\omega||$。我们原来的任务是找到一组参数 ω, γ 使得分类间隔 $W = 2d$ 最大化，根据公式(2.12)就可以转变为 $||\omega||$ 的最小化问题，也等效于 $\frac{1}{2}||\omega||^2$ 的最小化问题。我们之所以要在 $||\omega||$ 上加上平方和1/2的系数，是为了以后进行最优化的过程中对目标函数求导时比较方便，但这绝不影响最优化问题最后的解。

另外我们还可以尝试将公式(2.11)给出的约束条件进一步在形式上精练，把类别标签 y_i 和两个不等式左边相乘，形成统一的表述：

$$y_i(\omega^T x_i + \gamma) \geq 1 \ \forall x_i \quad (2.13)$$

好了，到这里我们可以给出线性SVM最优化问题的数学描述了：

$$\min_{\omega,\gamma} \frac{1}{2}||\omega||^2 \quad (2.14)$$
$$\text{s. t. } \ y_i(\omega^T x_i + \gamma) \geq 1, \ \ i = 1, 2, \ldots, m$$

这里m是样本点的总个数，缩写s. t. 表示“Subject to”，是“服从某某条件”的意思。公式(2.14)描述的是一个典型的不等式约束条件下的二次型函数优化问题，同时也是支持向量机的基本数学模型。（*此时此刻，你也许会回头看2.3节我们提出的三个约束问题，思考它们在公式2.14的约束条件中是否已经得到了充分的体现。但我不建议你现在就这么做，因为2.14采用了一种比较含蓄的方式表示这些约束条件，所以你即便现在不理解也没关系，后面随着推导的深入，这些问题会一点点露出真容。*）

接下来，我们将在第三章讨论大多数同学比较陌生的问题：如何利用最优化技术求解公式(2.14)描述的问题。哪些令人望而生畏的术语，凸二次优化、拉格朗日对偶、KKT条件、鞍点等等，大多出现在这个部分。全面理解和熟练掌握这些概念当然不容易，但如果你的目的主要是了解这些技术如何在SVM问题进行应用的，那么阅读过下面一章后，你有很大的机会可以比较直观地理解这些问题。

come from future_____by Chen

*一点小建议，读到这里，你可以试着在纸上随便画一些点，然后尝试用SVM的思想手动画线将两类不同的点分开。你会发现大多数情况下，1

会在你的脑海中想象去旋转这条线，旋转到某个角度，你就会下意识的停下来，因为如果再旋转下去，就找不到能够成功将两类点分开的直线了。这个过程就是对直线方向的优化过程。对于有些问题，你会发现SVM的最优解往往出现在不能再旋转下去的边界位置，这就是约束条件的边界，对比我们提到的等式约束条件，你会对代数公式与几何想象之间的关系得到一些相对直观的印象。

三、有约束最优化问题的数学模型

（*Hi，好久不见*）就像我们在第二部分结尾时提到的，SVM问题是一个不等式约束条件下的优化问题。绝大多数模式识别教材在讨论这个问题时都会在附录中加上优化算法的简介，虽然有些写得未免太简略，但看总比不看强，所以这时候如果你手头有一本模式识别教材，不妨翻到后面找找看。结合附录看我下面写的内容，也许会有帮助。

我们先解释一下我们下面讲解的思路以及重点关注哪些问题：

1）有约束优化问题的几何意象：闭上眼睛你看到什么？

2）拉格朗日乘子法：约束条件怎么跑到目标函数里面去了？

3）KKT条件：约束条件是不等式该怎么办？

4）拉格朗日对偶：最小化问题怎么变成了最大化问题？

5）实例演示：拉格朗日对偶函数到底啥样子？

6）SVM优化算法的实现：数学讲了辣么多，到底要怎么用啊？

3.1 有约束优化问题的几何意象

约束条件一般分为等式约束和不等式约束两种，前者表示为 $g(\boldsymbol{x})=0$（*注意这里的$\boldsymbol{x}$跟第二章里面的样本x没有任何关系，只是一种通用的表示*）；后者表示为 $g(\boldsymbol{x})\leq 0$（*你可能会问为什么不是$g(\boldsymbol{x})<0$，别着急，到KKT那里你就明白了*）。

假设 $\boldsymbol{x}\in\mathbb{R}^{d}$（*就是这个向量一共有d个标量组成*），则 $g(\boldsymbol{x})=0$ 的几何意象就是d维空间中的d-1维曲面，如果函数 $g(\boldsymbol{x})$ 是线性的，$g(\boldsymbol{x})=0$ 则是个d-1维的超平面。那么有约束优化问题就要求在这个d-1维的曲面或者超平面上找到能使得目标函数最小的点，这个d-1维的曲面就是“可行解区域”。

对于不等式约束条件，$g(\boldsymbol{x})\leq 0$，则可行解区域从d-1维曲面扩展成为d维空间的一个子集。我们可以从d=2的二维空间进行对比理解。等式约束对应的可行解空间就是一条线；不等式约束对应的则是这条线以及线的某一侧对应的区域，就像下面这幅图的样子（图中的目标函数等高线其实就是等值线，在同一条等值线上的点对应的目标函数值相同）。

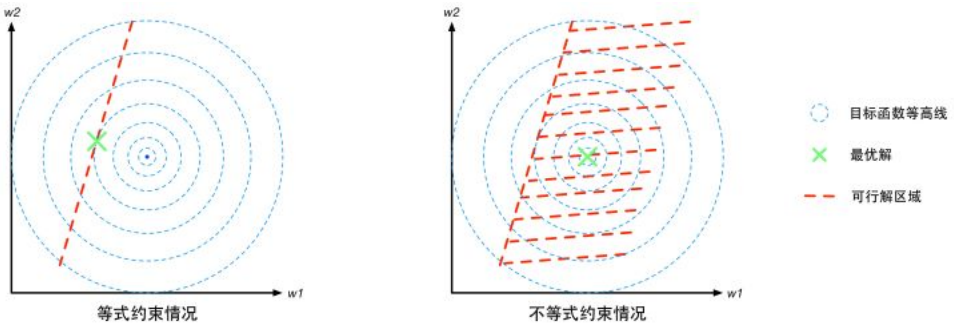


图3 有约束优化问题的几何意象图

3.2 拉格朗日乘子法

尽管在3.1节我们已经想象出有约束优化问题的几何意象。可是如何利用代数方法找到这个被约束了的最优解呢？这就需要用到拉格朗日乘子法。

首先定义原始目标函数 $f(x)$，拉格朗日乘子法的基本思想是把约束条件转化为新的目标函数 $L(x,\lambda)$ 的一部分(*关于 λ 的意义我们一会儿再解释*)，从而使有约束优化问题变成我们习惯的无约束优化问题。那么该如何去改造原来的目标函数 $f(x)$ 使得新的目标函数 $L(x,\lambda)$ 的最优解恰好就在可行解区域中呢？这需要我们去分析可行解区域中最优解的特点。

1）最优解的特点分析

这里比较有代表性的是等式约束条件（*不等式约束条件的情况我们在KKT条件里再讲*）。我们观察一下图3中的红色虚线（*可行解空间*）和蓝色虚线（*目标函数的等值线*），发现这个被约束的最优解恰好在二者相切的位置。这是个偶然吗？**我可以负责任地说：“NO！它们温柔的相遇，是三生的宿命。”** 为了解释这个相遇，我们先介绍**梯度的概念。**梯度可以直观的认为是函数的变化量，可以描述为包含变化方向和变化幅度的一个向量。然后我们给出一个推论：

推论1：“在那个宿命的相遇点 x^* （*也就是等式约束条件下的优化问题的最优解*），原始目标函数 $f(x)$ 的梯度向量 $\nabla f(x^*)$ 必然与约束条件 $g(x)=0$ 的切线方向垂直。”

关于推论1的粗浅的论证如下：

*如果梯度矢量 $\nabla f(x^*)$ 不垂直于 $g(x)=0$ 在 x^* 点的切线方向，就会在 $g(x)=0$ 的切线方向上存在不等于0的分量，也就是说在相遇点 x^* 附近，$f(x)$ 还在沿着 $g(x)=0$ 变化。这意味在 $g(x)=0$ 上 x^* 这一点的附近一定有一个点的函数值比 $f(x^*)$ 更小，那么 x^* 就不会是那个约束条件下的最优解了。**所以，梯度向量 $\nabla f(x^*)$ 必然与约束条件 $g(x)=0$ 的切线方向垂直。***

推论2：“函数 $f(x)$ 的梯度方向也必然与函数自身等值线切线方向垂直。”

推论2的粗浅论证：与推论1 的论证基本相同，如果 $f(x)$ 的梯度方向不垂直于该点等值线的切线方向，$f(x)$ 就会在等值线上有变化，这条线也就不能称之为等值线了。

根据推论1和推论2，函数 $f(x)$ 的梯度方向在 x^* 点同时垂直于约束条件 $g(x)=0$ 和自身的等值线的切线方向，也就是说**函数 $f(x)$ 的等值线与约束条件曲线 $g(x)=0$ 在 x^* 点具有相同（或相反）的法线方向，所以它们在该点也必然相切。**

让我们再进一步，约束条件 $g(x)=0$ 也可以被视为函数 $g(x)$ 的一条等值线。按照**推论2中“函数的梯度方向必然与自身的等值线切线方向垂直”**的说法，函数 $g(x)$ 在 x^* 点的梯度矢量 $\nabla g(x^*)$ 也与 $g(x)=0$ 的切线方向垂直。

到此我们可以将目标函数和约束条件视为两个具有平等地位的函数，并得到推论3：

推论3：“函数 $f(x)$ 与函数 $g(x)$ 的等值线在最优解点 x^* 处相切，即两者在 x^* 点的梯度方向相同或相反”，

于是我们可以写出公式(3.1)，用来描述最优解 x^* 的一个特性：

$$\nabla f(x^*)+\lambda\nabla g(x^*)=0 \quad (3.1)$$

这里增加了一个新变量 λ,用来描述两个梯度矢量的长度比例。那么是不是有了公式（3.1）就能确定 x^* 的具体数值了呢？显然不行！从代数解方程的角度看，公式（3.1）相当于d个方程（*假设 x^* 是d维向量，函数 $f(x)$ 的梯度就是d个偏导数组成的向量，所以公式(2.15)实际上是1个d维矢量方程，等价于d个标量方程*），而未知数除了 x^* 的d个分量以外，还有1个 λ。所以相当于用d个方程求解d+1个未知量，应有无穷多组解；

围内的任意实数）上都能至少找到一个满足公式(3.1)的点，也就是可以找到无穷多个这样的相切点。所以我们还需要增加一点限制，使得无穷多个解变成一个解。好在这个限制是现成的，那就是：

$$g(\boldsymbol{x}^*) = 0 \quad (3.2)$$

把公式(3.1)和(3.2)放在一起，我们有d+1个方程，解d+1个未知数，方程有唯一解，这样就能找到这个最优点 $\boldsymbol{x}^*$ 了。

2）构造拉格朗日函数

虽然根据公式(3.1)和(3.2),已经可以求出等式约束条件下的最优解了，但为了在数学上更加便捷和优雅一点，我们按照本节初提到的思想，构造一个拉格朗日函数，将有约束优化问题转为无约束优化问题。拉格朗日函数具体形式如下：

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}) \quad (3.3)$$

新的拉格朗日目标函数有两个自变量 $\boldsymbol{x}, \lambda$，根据我们熟悉的求解无约束优化问题的思路，将公式(3.3)分别对 $\boldsymbol{x}, \lambda$ 求导，令结果等于零，就可以建立两个方程。同学们可以自己试一下，很容易就能发现这两个由导数等于0构造出来的方程正好就是公式(3.1)和(3.2)。说明新构造的拉格朗日目标函数的优化问题完全等价于原来的等式约束条件下的优化问题。

至此，我们说明白了“为什么构造拉格朗日目标函数可以实现等式约束条件下的目标优化问题的求解”。可是，我们回头看一下公式(2.14)，也就是我们的SVM优化问题的数学表达。囧，约束条件是不等式啊！怎么办呢?

3.3 KKT条件

对于不等式约束条件 $g(\boldsymbol{x}) \leq 0$ 的情况，如图4所示，最优解所在的位置 $\boldsymbol{x}^*$ 有两种可能，或者在边界曲线 $g(\boldsymbol{x}) = 0$ 上或者在可行解区域内部满足不等式 $g(\boldsymbol{x}) < 0$ 的地方。

第一种情况：最优解在边界上，就相当于约束条件就是 $g(\boldsymbol{x}) = 0$。参考图4，注意此时目标函数 $f(\boldsymbol{x})$ 的最优解在可行解区域外面，所以函数 $f(\boldsymbol{x})$ 在最优解 $\boldsymbol{x}^*$ 附近的变化趋势是“在可行解区域内侧较大而在区域外侧较小”，与之对应的是函数 $g(\boldsymbol{x})$ 在可行解区域内小于0，在区域外大于零，所以在最优解 $\boldsymbol{x}^*$ 附近的变化趋势是内部较小而外部较大。这意味着目标函数 $f(\boldsymbol{x})$ 的梯度方向与约束条件函数 $g(\boldsymbol{x})$ 的梯度方向相反。因此根据公式(3.1)，可以推断出参数 $\lambda > 0$.

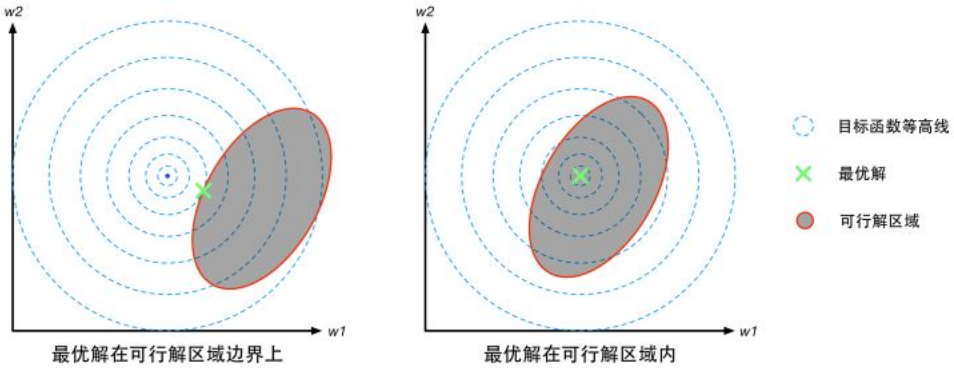


图4：不等式约束条件下最优解位置分布的两种情况

第二种情况：如果在区域内，则相当于约束条件没有起作用，因此公式(3.3)的拉格朗日函数中的参数 $\lambda = 0$。整合这两种情况，可以写出一个约束条件的统一表达，如公式(3.4)所示。

$$\begin{aligned} &g(\boldsymbol{x}) \leq 0 \quad (1) \\ &\boldsymbol{\lambda} \geq 0 \quad (2) \\ &\boldsymbol{\lambda} g(\boldsymbol{x}) = 0 \quad (3) \end{aligned} \tag{3.4}$$

其中第一个式子是约束条件本身。第二个式子是对拉格朗日乘子 $\boldsymbol{\lambda}$ 的描述。第三个式子是第一种情况和第二种情况的整合：在第一种情况里，$\boldsymbol{\lambda} > 0,\ g(\boldsymbol{x}) = 0$；在第二种情况下，$\boldsymbol{\lambda} = 0,\ g(\boldsymbol{x}) < 0$。所以无论哪一种情况都有 $\boldsymbol{\lambda} g(\boldsymbol{x}) = 0$。公式(3.4)就称为Karush-Kuhn-Tucker条件，简称KKT条件。

推导除了KKT条件，感觉有点奇怪。因为本来问题的约束条件就是一个 $g(\boldsymbol{x}) \leq 0$，怎么这个KKT条件又多弄出来两条，这不是让问题变得更复杂了吗？这里我们要适当的解释一下：

1）KKT条件是对最优解的约束，而原始问题中的约束条件是对可行解的约束。

2）KKT条件的推导对于后面马上要介绍的拉格朗日对偶问题的推导很重要。

3.4 拉格朗日对偶

接下来让我们进入重头戏——拉格朗日对偶。很多教材到这里自然而然的就开始介绍“对偶问题”的概念，这实际上是一种“先知式”的教学方式，对于学生研究问题的思路开拓有害无益。所以，在介绍这个知识点之前，我们先要从宏观的视野上了解一下**拉格朗日对偶问题出现的原因和背景**。

按照前面等式约束条件下的优化问题的求解思路，构造拉格朗日方程的目的是将约束条件放到目标函数中，**从而将有约束优化问题转换为无约束优化问题**。我们仍然秉承这一思路去解决不等式约束条件下的优化问题，那么如何**针对不等式约束条件下的优化问题构建拉格朗日函数呢?**

因为我们要求解的是最小化问题，所以一个直观的想法是如果我能够构造一个函数，使得该函数在可行解区域内与原目标函数完全一致，而在可行解区域外的数值非常大，甚至是无穷大，那么这个**没有约束条件的新目标函数的优化问题**就与原来**有约束条件的原始目标函数的优化**是等价的问题。

拉格朗日对偶问题其实就是沿着这一思路往下走的过程中，为了方便求解而使用的一种技巧。于是在这里出现了三个问题：**1）有约束的原始目标函数优化问题；2）新构造的拉格朗日目标函数优化问题；3）拉格朗日对偶函数的优化问题**。我们希望的是这三个问题具有完全相同的最优解，而在数学技巧上通常第三个问题——拉格朗日对偶优化问题——最好解决。所以**拉格朗日对偶不是必须的，只是一条捷径。**

1）原始目标函数（有约束条件）

为了接下来的讨论，更具有一般性，我们把等式约束条件也放进来，进而有约束的原始目标函数优化问题重新给出统一的描述：

$$\begin{aligned} &\min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) \\ &\text{s. t.} \quad h_i(\boldsymbol{x}) = 0 \quad i = 1, 2, \ldots, m \\ &\qquad\ \ g_j(\boldsymbol{x}) \leq 0 \quad j = 1, 2, \ldots, n \end{aligned} \tag{3.5}$$

公式(3.5)表示m个等式约束条件和n个不等式约束条件下的目标函数 $f(\boldsymbol{x})$ 的最小化问题。

2）新构造的目标函数（没有约束条件）

接下来我们构造一个基于广义拉格朗日函数的新目标函数，记为：

$$\theta_P(\boldsymbol{x}) = \max_{\boldsymbol{\alpha},\, \boldsymbol{\beta};\, \beta_j \geq 0} L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{3.6}$$

其中 $L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ 为广义拉格朗日函数，定义为：

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

这里，$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\ldots,\alpha_m]^T$，$\boldsymbol{\beta}=[\beta_1,\beta_2,\ldots,\beta_n]^T$，是我们在构造新目标函数时加入的系数变量，同时也是公式(3.6)中最大化问题的自变量。将公式(3.7)带入公式(3.6)有：

$$\theta_P(\boldsymbol{x})=\max_{\boldsymbol{\alpha},\,\boldsymbol{\beta};\,\beta_j\geq 0}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})=f(\boldsymbol{x})+\max_{\boldsymbol{\alpha},\,\boldsymbol{\beta};\,\beta_j\geq 0}\left[\sum_{i=1}^{m}\alpha_i h_i(\boldsymbol{x})+\sum_{j=1}^{n}\beta_j g_j(\boldsymbol{x})\right] \tag{3.8}$$

我们对比公式(3.5)中的约束条件，将论域范围分为可行解区域和可行解区域外两个部分对公式（3.8）的取值进行分析，将可行解区域记为 $\boldsymbol{\Phi}$，当 $\beta_j\geq 0, j=1,2,\ldots,n$ 时有：

可行解区域内：由于 $h_i(\boldsymbol{x})=0\ \forall i$，$g_j(\boldsymbol{x})\leq 0$ 且系数 $\beta_j\geq 0,\ \forall j$，所以有：

$$\max_{\boldsymbol{\alpha},\,\boldsymbol{\beta};\,\beta_j\geq 0}\left[\sum_{i=1}^{m}\alpha_i h_i(\boldsymbol{x})+\sum_{j=1}^{n}\beta_j g_j(\boldsymbol{x})\right]=0,\ \text{for}\ \boldsymbol{x}\in\boldsymbol{\Phi} \tag{3.9}$$

可行解区域外：代表公式(3.5)中至少有一组约束条件没有得到满足。如果 $h_i(\boldsymbol{x})\neq 0$，则调整系数 α_i 就可以使 $\alpha_i h_i(\boldsymbol{x})\to+\infty$；如果 $g_j(\boldsymbol{x})>0$，调整系数 β_j 就可以使 $\beta_j g_j(\boldsymbol{x})\to+\infty$。这意味着，此时有

$$\max_{\boldsymbol{\alpha},\,\boldsymbol{\beta};\,\beta_j\geq 0}\left[\sum_{i=1}^{m}\alpha_i h_i(\boldsymbol{x})+\sum_{j=1}^{n}\beta_j g_j(\boldsymbol{x})\right]=+\infty,\ \text{for}\ \boldsymbol{x}\notin\boldsymbol{\Phi} \tag{3.10}$$

把公式(3.8),(3.9)和(3.10)结合在一起就得到我们新构造的目标函数 $\theta_P(\boldsymbol{x})$ 的取值分布情况：

$$\theta_P(\boldsymbol{x})=\begin{cases}f(\boldsymbol{x}) & \boldsymbol{x}\in\boldsymbol{\Phi}\\ +\infty & \text{otherwise}\end{cases} \tag{3.11}$$

此时我们回想最初构造新目标函数的初衷，就是为了建立一个在可行解区域内与原目标函数相同，在可行解区域外函数值趋近于无穷大的新函数。看看公式（3.11）,yeah,我们做到了。

现在约束条件已经没了，接下来我们就可以求解公式(3.12)的问题

$$\min_{\boldsymbol{x}}\left[\theta_P(\boldsymbol{x})\right] \tag{3.12}$$

这个问题的解就等价于有约束条件下的原始目标函数 $f(\boldsymbol{x})$ 最小化问题（公式3.5）的解。

3）对偶问题

尽管公式(3.12)描述的无约束优化问题看起来很美好，但一旦你尝试着手处理这个问题，就会发现一个麻烦。什么麻烦呢？那就是我们很难建立 $\theta_P(\boldsymbol{x})$ 的显示表达式。如果再直白一点，我们很难直接从公式(3.8)里面把 $\boldsymbol{\alpha},\boldsymbol{\beta}$ 这两组参数拿掉，这样我们就没法通过令 $\partial\theta_P(\boldsymbol{x})/\partial\boldsymbol{x}=\boldsymbol{0}$ 的方法求解出最优解 $\boldsymbol{x}^*$。

要解决这个问题，就得用一点数学技巧了，这个技巧就是对偶问题。我们先把公式(3.6)和公式(3.12)放在一起，得到关于新构造的目标函数的无约束优化的一种表达：

$$\min_{\boldsymbol{x}}\left[\theta_P(\boldsymbol{x})\right]=\min_{\boldsymbol{x}}\left[\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right] \tag{3.13}$$

然后我们再构造另一个函数，叫做 $\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$，然后给出另外一个优化问题的描述：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}\left[\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})\right]=\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}\left[\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right] \tag{3.14}$$

对比公式(3.13)和(3.14)，发现两者之间存在一种对称的美感。所以我们就把(3.14)称作是(3.13)的对偶问题。现在我们可以解释一下 $\theta_P(\boldsymbol{x})$ 中的P是原始问题Primary的缩写，$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})$ 中的D是对偶问题Dual的缩写。**如果我们能够想办法证明(3.14)和(3.13)存在相同的解 $\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$，那我们就可以在对偶问题中选择比较简单的一个来求解。**

4）对偶问题同解的证明

对偶问题和原始问题到底有没有相同的最优解呢？关于这个问题的根本性证明其实没有在这里给出，而且在几乎我看到的所有有关SVM的资料里都没有给出。但我比较厚道的地方是我至少可以告诉你哪里能找到这个证明。在给出证明的链接地址之前，我们先给一个定理，帮助大家做一点准备，同时也减少一点看那些更简略的资料时的困惑。

定理一：对于任意 $\boldsymbol{\alpha},\boldsymbol{\beta}$ 和 $\boldsymbol{x}$ 有：

$$d^*=\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}\left[\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right]\leq\min_{\boldsymbol{x}}\left[\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right]=q^*$$

定理一的证明：

$$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\min_{\boldsymbol{x}}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\leq L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\leq\max_{\boldsymbol{\alpha},\,\boldsymbol{\beta};\,\beta_j\geq 0}L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})=\theta_P(\boldsymbol{x})$$

即

$$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})\leq\theta_P(\boldsymbol{x})\ \forall\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{x}$$

所以

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\beta_j\geq 0}\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})\leq\min_{\boldsymbol{x}}\theta_P(\boldsymbol{x})$$

即：

$$d^*\leq q^*$$

这里的 d^*,q^* 分别是对偶问题和原始问题的最优值。

定理一既引入了 d^*,q^* 的概念，同时也描述了两者之间的关系。我们可以在这个基础上再给一个**推论**：如果能够找到一组 $\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$ 使得 $\theta_D(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)=\theta_P(\boldsymbol{x}^*)$，那么就应该有：

$$\theta_D(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)=d^*,\ \theta_P(\boldsymbol{x}^*)=p^*,\ d^*=p^*$$

这个推论实际上已经涉及了原始问题与对偶问题的“强对偶性”。当 $d^*\leq q^*$ 时，我们称原始问题与对偶问题之间“弱对偶性”成立；若 $d^*=q^*$，则称“强对偶性”成立。

如果我们希望能够使用拉格朗日对偶问题替换原始问题进行求解，则需要“强对偶性”作为前提条件。于是我们的问题变成了什么情况下，强对偶性才能够在SVM问题中成立。关于这个问题我们给出定理二：

定理二：对于原始问题和对偶问题，假设函数 $f(\boldsymbol{x})$ 和不等式约束条件 $g_j(\boldsymbol{x}),\ \forall j$ 为凸函数，等式约束条件中的 $h_i(\boldsymbol{x})$ 为仿射函数（即由一阶多项式构成的函数，$h_i(\boldsymbol{x})=\boldsymbol{a}_i^T\boldsymbol{x}+b_i$，$\boldsymbol{a}_i,\boldsymbol{x}$ 均为列向量，b_i 为标量）；并且至少存在一个 $\boldsymbol{x}$ 使所有不等式约束条件严格成立，即 $g_j(\boldsymbol{x})<0,\ \forall j$，则存在 $\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$ 使得 $\boldsymbol{x}^*$ 是原始问题的最优解，$\boldsymbol{\alpha}^*,\ \boldsymbol{\beta}^*$ 是对偶问题的最优解且有：$d^*=p^*=L(\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$，并其充分必要条件如下：

$$\nabla_{\boldsymbol{x}}(\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)=0 \quad (1)$$
$$\nabla_{\boldsymbol{\alpha}}(\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)=0 \quad (2)$$
$$\nabla_{\boldsymbol{\beta}}(\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)=0 \quad (3)$$
$$g_j(\boldsymbol{x}^*)\le 0,\ j=1,2,\dots,n \quad (4) \quad (3.15)$$
$$\beta_j^*\ge 0,\ j=1,2,\dots,n \quad (5)$$
$$\beta_j^* g_j(\boldsymbol{x}^*)=0,\ j=1,2,\dots,n \quad (6)$$
$$h_i(\boldsymbol{x}^*)=0,\ i=1,2,\dots,m \quad (7)$$

再次强调一下，公式(3.15)是使 $\boldsymbol{x}^*$ 为原始问题的最优解，$\boldsymbol{\alpha}^*$，$\boldsymbol{\beta}^*$ 为对偶问题的最优解，且 $d^*=p^*=L(\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$ 的充分必要条件。公式(3.15)中的(1)~(3)，是为了求解最优化要求目标函数相对于三个变量 $\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}$ 的梯度为0；(4)~(6)为KKT条件（见公式3.4(3)），这也是我们为什么要在3.3节先介绍KKT条件的原因；(7)为等式约束条件。

定理二的证明详见《Convex Optimization》， by Boyd and Vandenberghe. Page-234, 5.3.2节。stanford.edu/~boyd/cvxb...

关于拉格朗日对偶的一些参考资料：

1. 简易解说拉格朗日对偶（Lagrange duality），这一篇对对偶问题的来龙去脉说的比较清楚，但是在强对偶性证明方面只给出了定理，没有给出证明过程，同时也缺少几何解释。

2.优化问题中的对偶性理论，这一篇比较专业，关于对偶理论的概念，条件和证明都比较完整，在数学专业文献里属于容易懂的，但在我们这种科普文中属于不太好懂的，另外就是论述过程基本跟SVM没啥关系。

3.5 拉格朗日对偶函数示例

尽管上述介绍在代数层面已经比较浅显了，但是没有可视化案例仍然不容易建立起直观的印象。所以我尽可能的在这里给出一个相对简单但是有代表性的可视化案例。

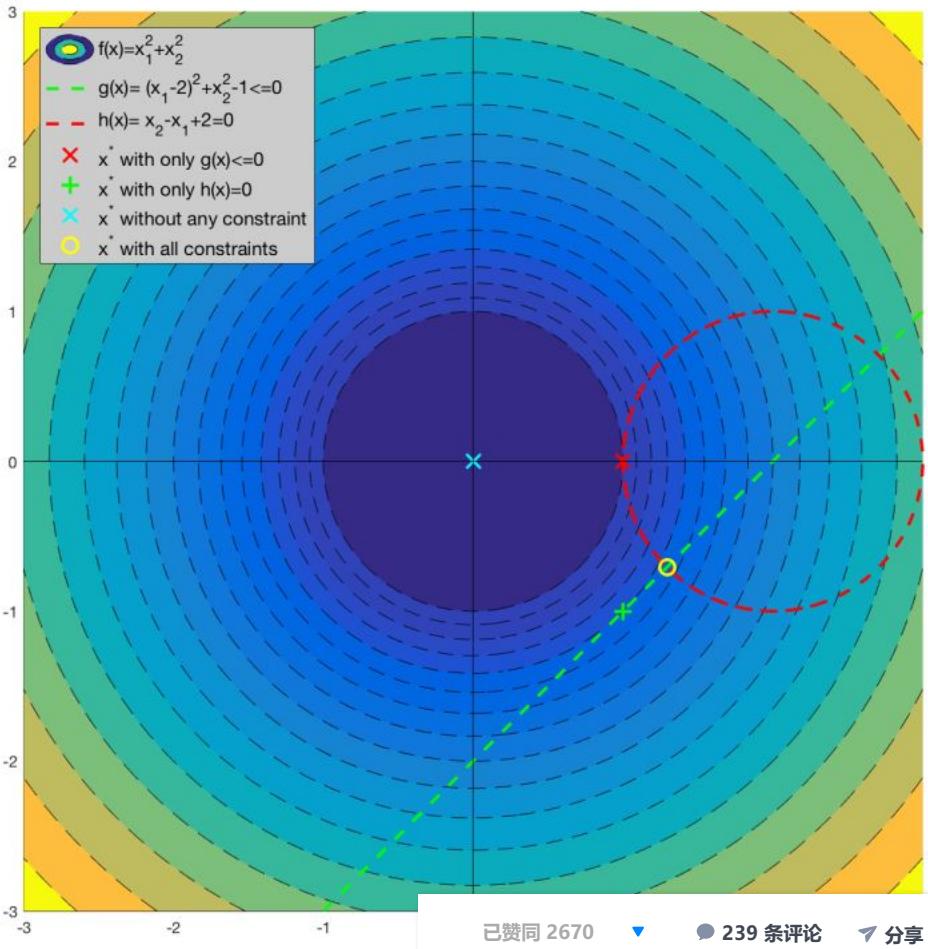


图5：有约束条件下的最优化问题可视化案例。

图5中的优化问题可以写作：

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = x_1^2 + x_2^2 \\ \text{s. t.} \quad & h(\boldsymbol{x}) = x_1 - x_2 - 2 = 0 \\ & g(\boldsymbol{x}) = (x_1 - 2)^2 + x_2^2 - 1 \leq 0 \end{aligned} \tag{3.16}$$

之所以说这个案例比较典型是因为它与线性SVM的数学模型非常相似，且包含了等式和不等式两种不同的约束条件。更重要的是，这两个约束条件在优化问题中都起到了作用。如图5所示，如果没有任何约束条件，最优解在坐标原点(0, 0)处（青色X）；如果只有不等式约束条件 $g(\boldsymbol{x}) \leq 0$ ，最优解在坐标(1,0)处（红色X）；如果只有等式约束条件 $h(\boldsymbol{x}) = 0$ ，最优解在坐标(1,-1)处（绿色+）；如果两个约束条件都有，最优解在 $(2 - \sqrt{2}/2, -\sqrt{2}/2)$ 处(黄色O)。

针对这一问题，我们可以设计拉格朗日函数如下：

$$L(\boldsymbol{x}, \alpha, \beta) = (x_1^2 + x_2^2) + \alpha(x_1 - x_2 - 2) + \beta\left[(x_1 - 2)^2 + x_2^2 - 1\right] \tag{3.17}$$

根据公式（3.11），函数 $\theta_p(\boldsymbol{x})$ 只在绿色直线在红色圆圈内的部分——也就是直线 $h(\boldsymbol{x}) = 0$ 在圆 $g(\boldsymbol{x}) = 0$ 上的弦——与原目标函数 $f(\boldsymbol{x})$ 取相同的值，而在其他地方均有 $\theta_P(\boldsymbol{x}) = +\infty$ ，如图6所示。

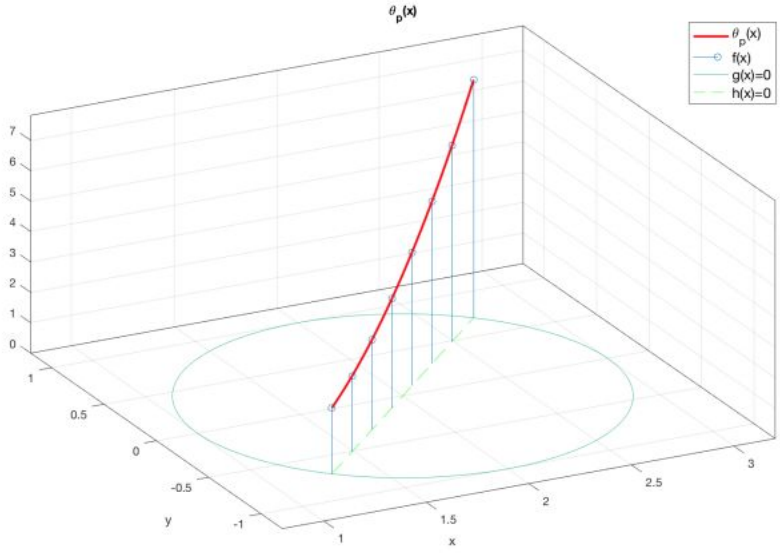


图6： $\theta_P(\boldsymbol{x})$ （除了图中绿色虚线部分，其他部分的函数值均为无穷大）。

（需要注意的是，此处不能使用对 α, β 求导等于0的方式消掉 α, β ，因为函数 $L(\boldsymbol{x}, \alpha, \beta)$ 在 $\boldsymbol{x}$ 为确定值时，是 α, β 的线性函数，其极大值并不在梯度为0的地方）。由于函数 $f(\boldsymbol{x})$ 在没有约束条件下的最优解并不在这条弦上，所以显然对 $\theta_P(\boldsymbol{x})$ 求导等于零的方法是找不到最优解 $\boldsymbol{x}^*$ 的。但是对于这个简单的问题，还是能够从图中看到最优解应该在：

$$[x_1^*, x_2^*] = [2 - \sqrt{2}/2, -\sqrt{2}/2]$$

由于该最优解是在 $h(\boldsymbol{x}) = 0$ 和 $g(\boldsymbol{x}) = 0$ 的交点处，所以可以很容易地理解：当 $\boldsymbol{x} = [x_1^*, x_2^*]^T$ 时，无论 α, β 取什么值都可以使函数 $L(\boldsymbol{x}, \alpha, \beta)$ 达到最小值。

然而这个最优解是依靠几何推理的方式找到的，对于复杂的问题，这种方法似乎不具有可推广性。

那么，我们不妨尝试一下，用拉格朗日对偶的方式看看这个问题。我们将 α, β 视为常数，这时 $L(\boldsymbol{x}, \alpha, \beta)$ 就只是 $\boldsymbol{x}$ 的函数。我们可以通过求导等于零的方式寻找其最小值，即 $\theta_D(\alpha, \beta) = \min_{\boldsymbol{x}}\left[L(\boldsymbol{x}, \alpha, \beta)\right]$ 。我们对公式(3.17)对 x_1, x_2 分别求偏导，令其等于0，有：

$$\begin{cases} \alpha + 2x_1 + \beta(2x_1 - 4) = 0 \\ 2x_2 - \alpha + 2\beta x_2 = 0 \end{cases} \quad (3.18)$$

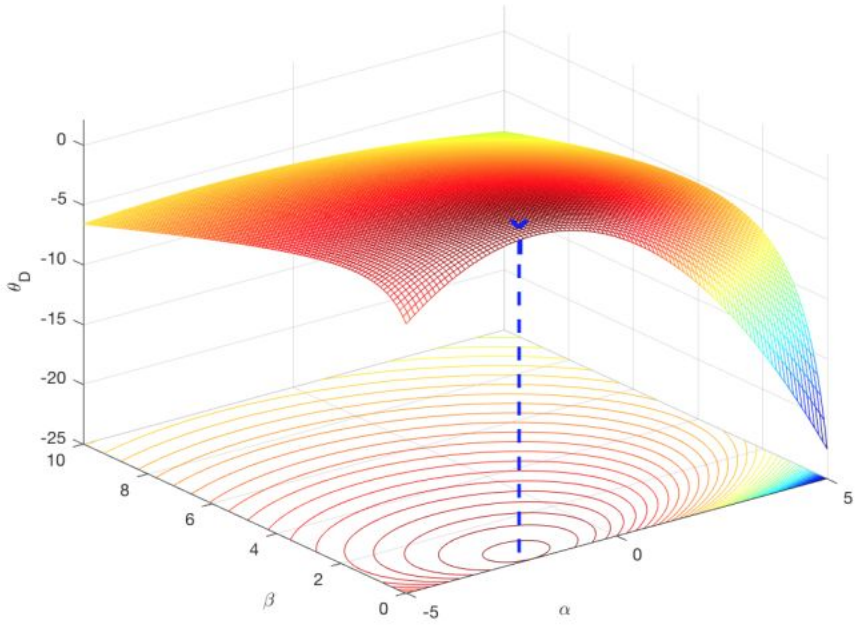
可以解得:

$$\begin{cases} x_1 = \frac{4\beta - a}{2\beta + 2} \\ x_2 = \frac{\alpha}{2\beta + 2} \end{cases} \quad (3.19)$$

将(3.19)带入(3.17)可以得到:

$$\theta_D(\alpha, \beta) = -\frac{\alpha^2 + 4\alpha + 2\beta^2 - 6\beta}{2(\beta + 1)} \quad (3.20)$$

考虑到（3.15）中的条件（5），我们将函数(3.20)在 $\beta \geq 0$ 的 $\alpha \times \beta$ 论域画出来，如图7所示。可以通过 $\theta_D(\alpha, \beta)$ 对 α, β 求导等于0的方式解出最优解 $\alpha^* = -2,\ \beta^* = \sqrt{2} - 1$，将其带入公式（3.19）可以得到 $x^* = [x_1^*, x_2^*] = [2 - \sqrt{2}/2, -\sqrt{2}/2]$



最后通过对比，我们看到拉格朗日原始问题和对偶问题得到了相同的最优解（原始问题的最优解中 α^*, β^* 可以是任何值）。

最后，我来解释一下鞍点的问题。鞍点的概念大家可以去网上找，形态上顾名思义，就是马鞍的中心点，在一个方向上局部极大值，在另一个方向上局部极小值。这件事跟我们的拉格朗日函数有什么关系呢？由于这个例子中的拉格朗日函数包含 x_1, x_2, α, β 四个自变量，无法直接显示。为了更好的可视化，我们固定住其中两个变量，令 $x_2 = -\sqrt{2}/2, \beta = \sqrt{2} - 1$。此时拉格朗日函数就变成一个可以可视化的二元函数 $L(x_1, \alpha)$，我们把它的曲面画出来。

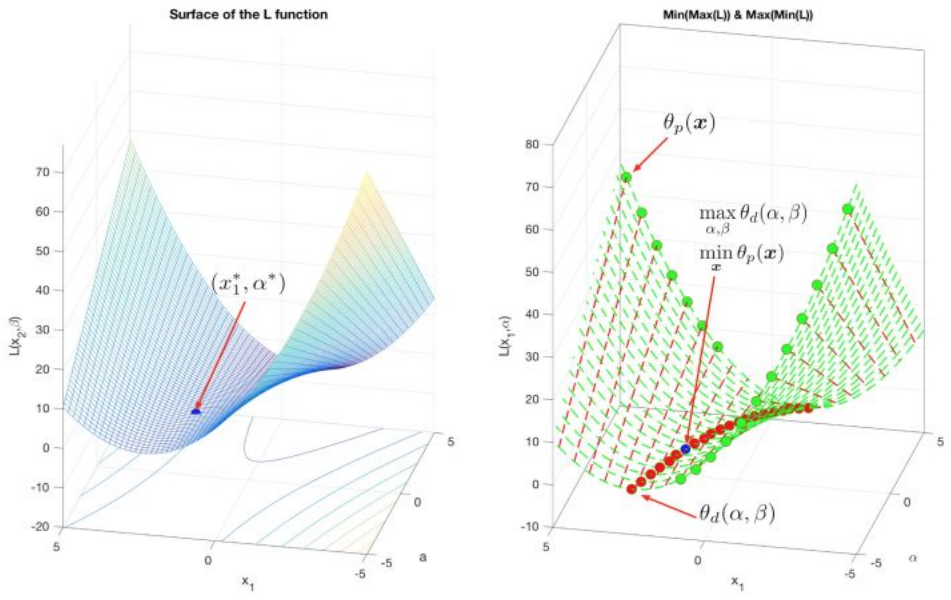


图8： $L(x_1, \alpha)$ 可视化效果

图8(a)中的最优点 (x_1^*, α^*) 可以能够两个角度去定义，如图8(b)所示。(为加以区别二维和四维的情况，我们将四维情况对应的 θ_P, θ_D 大写的下角标P和D改写为小写的p和d)。

第一种定义：沿着与 α 轴平行的方向将曲面切成无数条曲线（红色虚线），在每条红色虚线上找到最大值（绿色圆点），即 $\theta_p(x_1) = \max_{\alpha}(L(x_1, \alpha))$ ，然后在所有的 $\theta_p(x_1)$ 找到最小的那个（蓝色圆点），即 $\min_{x_1} \theta_p(x_1)$ 。

第二种定义：沿着与 x_1 轴平行的方向将曲面切成无数条曲线（绿色虚线），在每条绿色虚线上找到最小值（红色圆点），即 $\theta_d(\alpha) = \min_{x_1}(L(x_1, \alpha))$ ，然后在所有的 $\theta_d(\alpha)$ 中找到最大的那个（蓝色圆点），即 $\max_{\alpha} \theta_d(\alpha)$ 。

从图8的二维情况思考神秘的四维空间中的拉格朗日函数， $\theta_p(x_1)$ 就变成了 $\theta_p(x)$ ，$\theta_d(\alpha) = \theta_d(\alpha, \beta)$ ，如图8(b)所示。其实四元函数 $L(x_1, x_2, \alpha, \beta)$ 就是一个定义在4维空间上的鞍形函数，这个从两种思路共同得到的蓝色点就是函数 $L(x_1, x_2, \alpha, \beta)$ 的鞍点，也就是我们要找的最优解。在这个二元化的图中，拉格朗日对偶问题和拉格朗日原始问题的差别就是：原始问题采用第一种定义去求解鞍点，对偶问题采用第二种方法去求解鞍点。

至此，我们比较形象地描述了一个有约束条件下的函数优化问题的拉格朗日对偶问题求解过程以及相应的几何解释。

（未完待续~~）

编辑于 2018-05-08

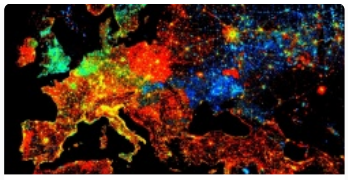
模式识别　机器学习　SVM

推荐阅读

SVM解决多分类问题

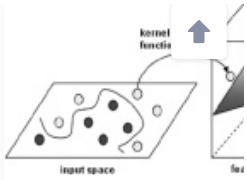
SVM解决多分类问题的方法 SVM算法最初是为二值分类问题设计的，当处理多类问题时，就需要构造合适的多类分类器。目前，构造SVM多类分类器的方法主要有两类：一类是直接法，直接在目标函数上...

Turin... 发表于图灵Ai



GEE 学习笔记 14：数字化训练样本利用SVM进行土地利用...

猿人充电站 发表于Googl...



《统计学习方法》系列

Pikachu5808

239 条评论 ⇌ 切换为时间排序

写下你的评论...

朱师傅 2016-12-30
坐等更新
4

五角 2016-12-30
坐等更新
4

OrNot 2016-12-30
功德无量
8

wood hawk 2016-12-30
等更，SVM真心有点难懂，核函数那里到现在还是有点不太明白
8

Zheng Qin 2016-12-30
坐等更新
1

付德智 2016-12-31
收藏，等更
2

木头 2016-12-31
等更。。
赞

一步 2016-12-31
陈老师你的立意是好的，持续关注，希望能交流一些阅读材料
赞

耳东陈 (作者) 回复 一步 2016-12-31
看了前半句，感觉后半句应该说：“但是水平确实一般，还是放弃吧！”
21

sbit 回复 耳东陈 (作者) 2017-06-30
谢谢陈老师分享啦~收获很多~
赞

展开其他 2 条回复

已赞同 2670 ▼ 239 条评论 分享 收藏 ···

cuixs 2016-12-31

👏👏

赞

维修工 2017-01-02

😃

赞

Clemens 2017-01-02

科普专业知识功德无量，支持陈老师=W=

赞

黄源Bjergsen 2017-01-02

小板凳围观更新

赞

OrNot 2017-01-03

公式推导到2.9， 我个人感觉应该基本能说明大部分情况了，似乎无必要非要推导2.10和2.11 去强求公式右边非要为1. 约定几何间隔大于d/||w||, 函数间隔大于d, 就可以作为约束条件。 不知道是不是这样?

2

南楼 回复 OrNot 2017-01-08

归一化吧?

赞

兔翻鹿 回复 OrNot 2017-11-10

当然不是的。等于1是边界条件

赞

展开其他 1 条回复

刘锐 2017-01-08

SVM是机器学习必须理解的算法，我记得我上机器学习的时候，期末project就是手动编写SVM内核，并实现一个应用。 当时大量参考哦了bishop的Pattern Recognition and Machine Learning这本书里的推导过程，想学习的同学可以看这本经典书

6

mini 回复 刘锐 2018-05-05

老师讲的太好了👍

赞

小鱼骆驼 2017-01-09

内容很好，很易于理解，坐等更新

赞

王鹏越 2017-01-10

台大林轩田老师的机器学习技法这部分讲的很清楚

3

王司徒 回复 王鹏越 2018-08-02

嗯嗯 感觉比ng和李宏毅的好一些

赞

杨大卫 2017-01-10

很好，期待更新。

赞

雕栏玉砌 2017-01-10

老师好，没想到在知乎遇到东大的老师，倍感亲切。讲的真是非常具体，没有比这在详细的了，(连st都注释成原单词了，0_-0）。

赞

Stinor 回复 雕栏玉砌 2017-08-27

莫非你也东大的?

赞

vagrant 2017-01-14

老师等着跟新了!

赞

的了哈 2017-01-16

lowB的知乎不给提供文章收藏功能

赞

克里斯573 回复 的了哈 04-26

可以收藏的

赞

1 2 3 4 ... 11 下一页