### 附录A 奇异值分解（Singular Value Decomposition，SVD）

奇异值分解是机器学习算法中广泛应用，如主成分分析和线性回归模型求解均有用到。

#### 1．特征值和特征向量

特征值和特征向量的定义如下：

|  |  |
| --- | --- |
| **，** | （A-1） |

其中是一个的矩阵，是一个维向量，则我们说是矩阵的一个特征值，是矩阵的特征值所对应的特征向量。

求出特征值和特征向量的好处是我们可以将矩阵特征分解。如果我们求出了矩阵的个特征值，以及这个特征值所对应的特征向量，如果这个特征向量线性无关，那么矩阵就可以用下式的特征分解表示：

|  |  |
| --- | --- |
| ， | （A-2） |

其中是这个特征向量所张成的维矩阵（矩阵的每一列为一个特征向量），为这个特征值为主对角线的维矩阵。一般我们将的这个特征向量归一化，即，或者说，，。此时的个列向量为标准正交基，满足，即, 也就是说为酉矩阵。

这样我们的特征分解表达式可以写成

|  |  |
| --- | --- |
| 。 | （A-3） |

注意要进行特征分解，矩阵必须为方阵。那么如果不是方阵，即行数目和列数目不相同时，我们只能对矩阵进行SVD分解。

#### 2．SVD的定义

SVD也是对矩阵进行分解，但并不要求要分解的矩阵为方阵。假设我们的矩阵是一个的矩阵，那么我们定义矩阵的SVD为：

|  |  |
| --- | --- |
| ， | （A-4） |

其中是的矩阵，是 的矩阵，除了主对角线上的元素以外全为0，主对角线上的每个元素称为奇异值， 是 的矩阵。 和矩阵都是酉矩阵，即满足：。

那么我们如何求出SVD分解后的这三个矩阵呢？

如果我们将的转置和做矩阵乘法，那么会得到的一个方阵。既然是方阵，那么我们就可以进行特征分解，得到的特征值和特征向量满足下式：

|  |  |
| --- | --- |
| 。 | （A-5） |

这样我们就可以得到矩阵的个特征值和对应的个特征向量了。将的所有特征向量张成一个的矩阵，就是SVD公式里面的矩阵。中的每个特征向量被称为的右奇异向量。

类似的，如果我们将和的转置做矩阵乘法，那么会得到的一个方阵。既然是方阵，那么我们就可以进行特征分解，得到的特征值和特征向量满足下式：

|  |  |
| --- | --- |
| 。 | （A-6） |

这样我们可以得到矩阵的个特征值和对应的个特征向量。将的所有特征向量张成一个的矩阵，就是SVD公式里面的矩阵。中的每个特征向量被称为的左奇异向量。

和我们求出来后，还剩下奇异值矩阵。由于除了对角线上是奇异值其他位置都是0，那我们只需要求出每个奇异值就可以了。我们注意到：

|  |  |
| --- | --- |
| 。 | （A-7） |

这样我们可以求出每个奇异值，进而求出奇异值矩阵。

上面还有一个问题没有讲，即的特征向量组成的矩阵就是SVD中的矩阵，的特征向量组成的矩阵就是SVD中的矩阵。下面我们以矩阵的证明为例。

|  |  |
| --- | --- |
| ， | （A-8） |

其中。可以看出的特征向量组成的矩阵的确就是SVD中的矩阵。类似的方法可以得到的特征向量组成的就是SVD中的矩阵。进一步还可以看出，特征值矩阵等于奇异值矩阵的平方，也就是说特征值和奇异值满足如下关系：

|  |  |
| --- | --- |
| 。 | （A-9） |

可以通过求出的特征值取平方根来求奇异值。

### 附**B** 拉格朗日乘子法与卡罗需-库恩-塔克（Karush–Kuhn–Tucker，KKT）条件

在求带有约束条件的优化问题时，拉格朗日乘子法和KKT条件是非常重要的两个方法。对有等式约束的优化问题，用拉格朗日乘子法求取其最优值。如果含有不等式约束，可以应用KKT条件计算。当然，这两个方法得到的结果只是必要条件。对于凸规划，KKT条件也是充要条件。只要满足KKT条件，则一定是极值点，且得到的一定还是全局最优解。这里凸规划指的是：目标函数为凸函数，不等式约束函数也为凸函数，等式约束函数是仿射函数。

KKT条件最初以Harold W. Kuhn和Albert W. Tucker的名字命名，他们在1951年首次发表了这些条件。后来学者发现1939年William Karush已经在[硕士论文](https://www.zhihu.com/search?q=%E7%A1%95%E5%A3%AB%E8%AE%BA%E6%96%87&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2897351790%7D" \t "_blank)中声明了必要条件。

#### 1 拉格朗日乘子法

拉格朗日乘子法是一种寻找多元函数在其变量受到一个或多个条件的相等约束时的求局部极值的方法。这种方法可以将一个有个变量和个约束条件的最优化问题转换为一个有个变量的不带约束的优化问题。

考虑最优化问题

|  |  |
| --- | --- |
| ， | （B-1） |

为了求，引入一个新的变量，称为拉格朗日乘子，问题转化为求拉格朗日函数

|  |  |
| --- | --- |
|  | （B-2） |

的极值。

分别让拉格朗日函数对的每维和每个求偏导并置零，得到上述优化问题的解的必要条件为：

|  |  |
| --- | --- |
|  | （B-3） |

或者写成向量形式，用梯度表示：

|  |  |
| --- | --- |
|  | （B-4） |

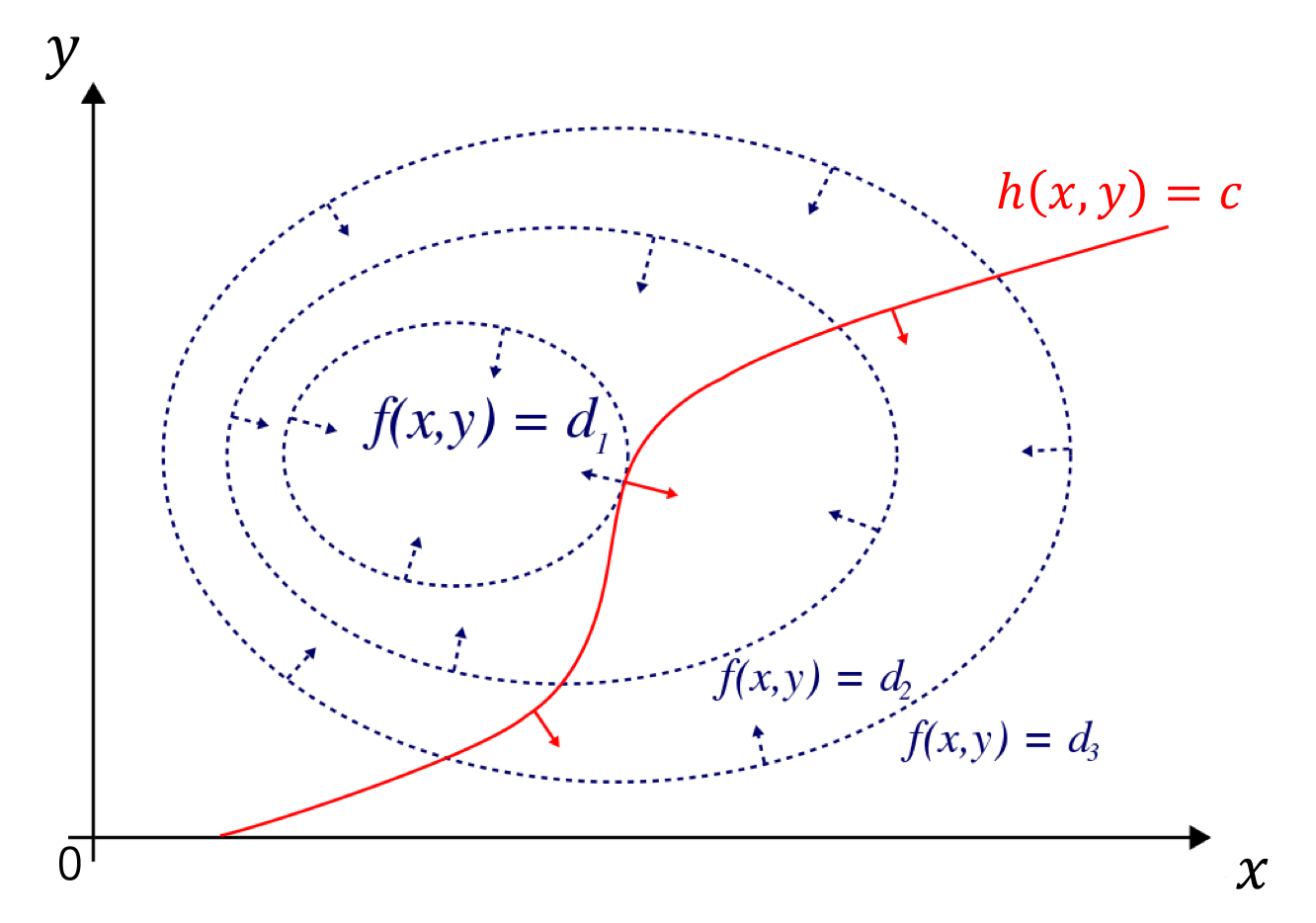
当只有一个等式约束（）时，根据第1行的条件，在最优点 处，梯度 和 的方向必相同或相反。

#### 2 为什么拉格朗日乘子法能够得到最优值

可以利用反证法来证明。令 ，取不同的值，相当于目标投影在构成的平面上的等高线。如图B-1所示的2维平面上，蓝色虚线是的等高线，红色实线是约束。假设优化问题有解，则会与等高线相交，交点就是同时满足等式约束条件和目标函数的可行域的值。想像此时我们移动上的点，因为是连续的，因此能走到更高或更低的等高线上，也就是说，可以变大或变小。只有当等高线与约束项的曲线相切的时候，会出现极值点或鞍点。

等高线在切点处的切线与等高线和曲线在该点的法线都垂直，因此等高线和约束曲线的法线是共线的。法线可以用该点的梯度表示，所以最优值必须满足：梯度 *，*是常数，表示左右两边同向或反向，得到。于是我们构造拉格朗日函数：，求解它的最优解就等价于求原带约束的优化问题的最优解，将原来带约束的优化问题转化为无约束的优化问题。

关于拉格朗日乘子法更详细的解释请见：<https://www.matongxue.com/madocs/939.html>。



图B-1 拉格朗日图解，其中红色实线是约束的点的轨迹。蓝色虚线是的等高线。箭头表示梯度，和等高线的法向量平行。

#### 3．KKT条件：

对含有不等式约束的优化问题

|  |  |
| --- | --- |
|  | （B-5） |

我们将不等式约束与写为一个式子，也称为拉格朗日函数

|  |  |
| --- | --- |
|  | （B-6） |

系数也称拉格朗日乘子，注意不等式约束时要求。

为了将带不等式约束的优化问题转化为已知的等式约束优化问题，我们引入松弛变量，得到。注意，这里加上平方项而非，是因为不等式的左边必须加上一个正数才能使不等式变为等式。若只加上，又会引入新的约束。

由此我们将不等式约束转化为了等式约束，并得到拉格朗日函数

|  |  |
| --- | --- |
| 。 | （B-7） |

我们再按照等式约束优化问题对其求解，得到

|  |  |
| --- | --- |
|  | （B-8） |

对第3行的约束，我们分两种情况：

（1），

由于乘子，因此与其相乘为零，可以理解为约束不起作用，且有；

（2）*，*

根据第2行的约束，此时，且，可以理解为约束起作用，且有。

合并情形1和情形2得：，且在约束起作用，；约束不起作用，。

由此，上述约束条件（极值必要条件）转化为

|  |  |
| --- | --- |
|  | （B-9） |

这些约束条件称为KKT条件，其中第2行的约束条件称为**互补松弛条件**。第4行为原始约束。第3行要求，原因我们稍后讨论。

或用向量的形式用梯度表示：

|  |  |
| --- | --- |
|  | （B-10） |

SVM的目标函数就是带不等式的约束问题。因为，如果要满足这个等式，必须或者。这是SVM的很多重要性质的来源，如支持向量的概念。

当只有一个约束条件（）时，省略约束条件的下标，最优点 或在 的区域内，或在边界上。

（1）如果最优解恰好在边界上，那么可以通过拉格朗日乘子法来求解。 时，就是等式约束问题。需要注意的是，此时梯度 和 的方向必须相反，即KKT条件中。原因入下。

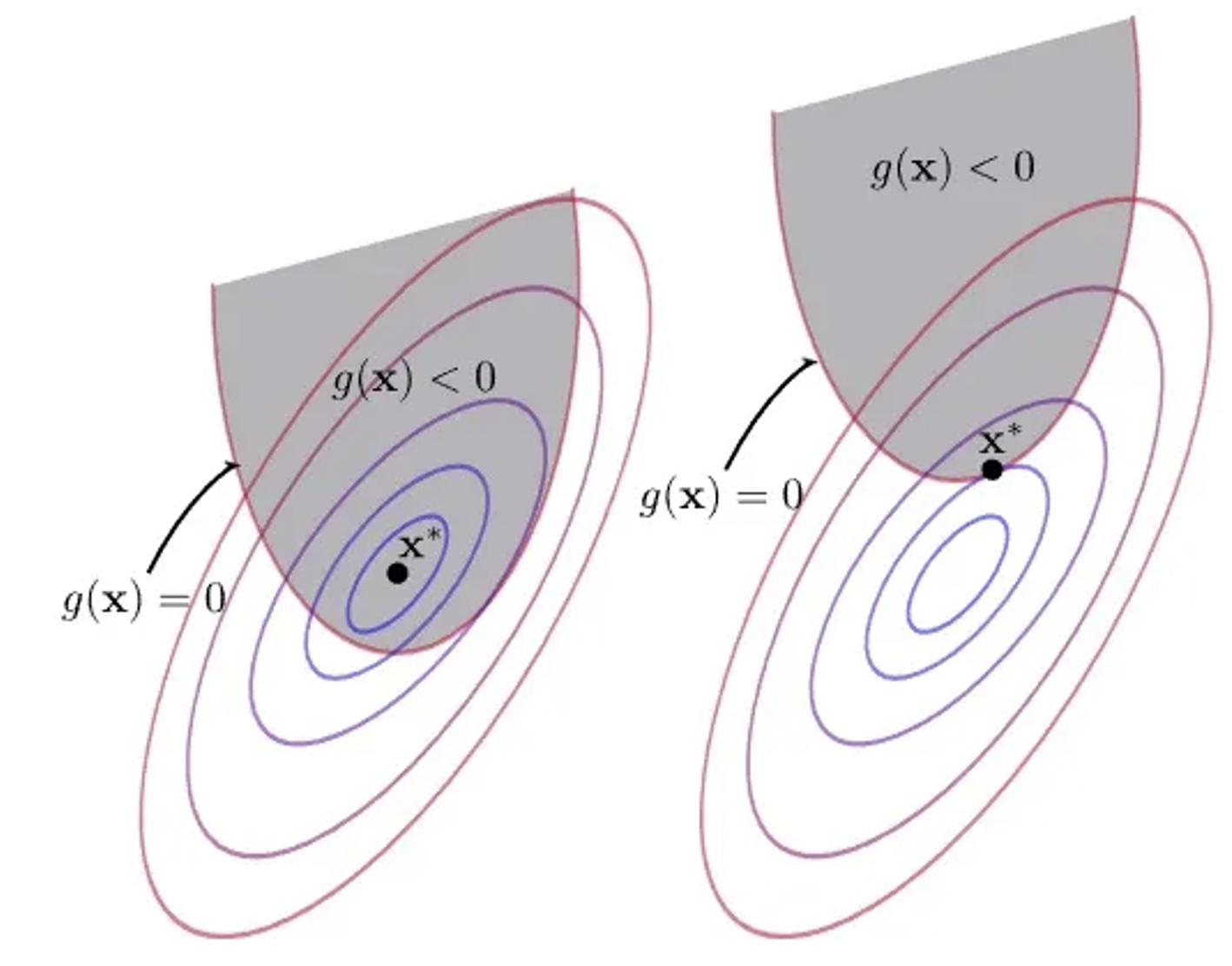
如图B-2所示，函数值下降方向为下方。因为目标是最小化，若下降方向为上方，则最优解一定不是在上，而是在可行域内部（阴影区域，）。由于曲线是，上方阴影区域是，值在下降，因此函数值增长的方向（梯度方向）是下方。而函数值下降方向为下方，增长方向（梯度方向）为上方，与就是梯度方向相反，所以。

图示, 维恩图

描述已自动生成

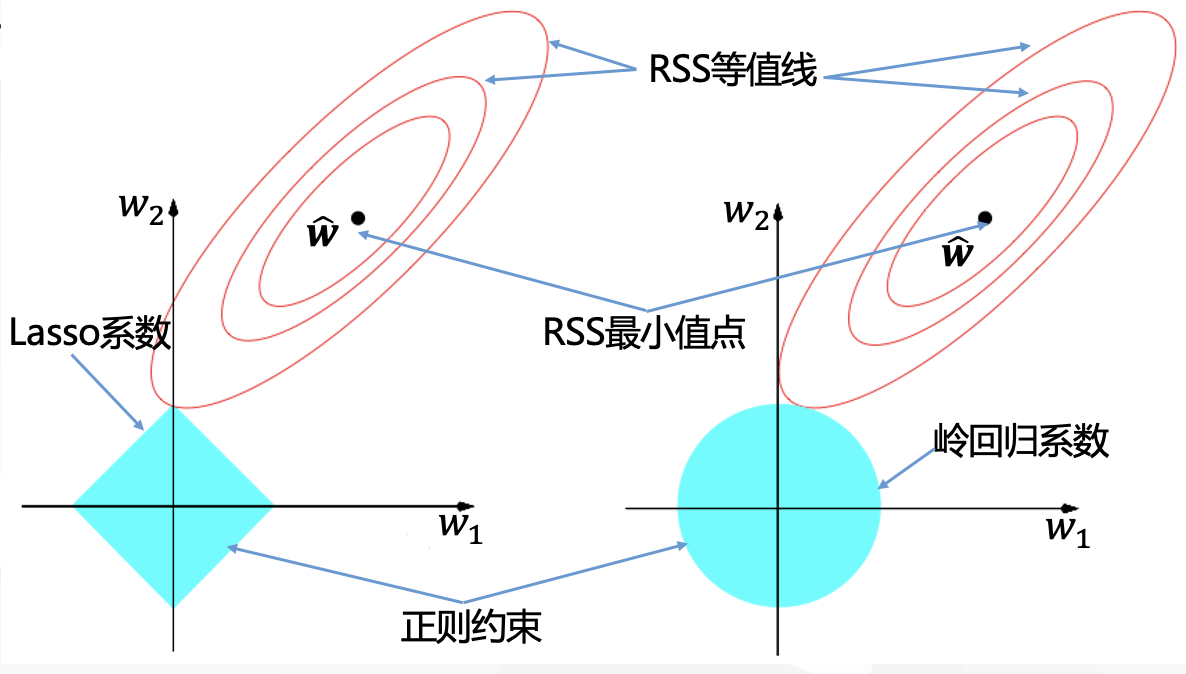
图B-2 最优解恰好在边界上，拉格朗日系数以，与的方向相反。

（2）如果最优解在阴影区内部，这个解一定是单纯考虑的最优解，也就是直接求的解。时，约束不起作用，可直接通过条件 来获得最优点；这等价于将拉格朗日函数中的置零，然后对 置零得到最优点。这种情况如图B-3左边所示，右边为第（1）种情况。



图B-3 不等式约束的两种情况。左边为最优解在可行域内部，此时也是的最小值；右边最优解在边界上，也在与等值线的切点处。

例如，在Lasso和岭回归中，，约束条件分别为和。一般情况最优的，所以最优解为或与等高线相切的切点。如果，则约束项/正则项不起作用，最优解为的最小值（最小二乘）。



图B-4 Lasso和岭回归可视为带正则不等式约束的优化问题。最优解通常在约束边界和目标函数等值线的切点处。

### 附C对偶问题

对偶是求解最优化问题的一种手段，它将一个最优化问题转化为另一个更容易求解的问题。转化为对偶问题的好处有：（1）约束减少了，对偶问题只剩个简单的不等式约束；（2）即使原主问题不是凸优化问题（局部最优不一定是全局最优），对偶问题一定是凸优化问题。

#### 1．原问题

对含有等式约束和不等式约束的优化问题

|  |  |
| --- | --- |
| ， | （C-1） |

这里我们无需假定原函数的凹凸性（可以是非凸/非凹函数）。

根据拉格朗日乘子法，构造广义拉格朗日函数：

|  |  |
| --- | --- |
| ， | （C-2） |

其中系数为拉格朗日乘子。

现在我们已经有了拉格朗日函数，要想办法把拉格朗日函数与原问题对等起来，为此构建一个新的关于的函数：

|  |  |
| --- | --- |
|  | （C-3） |
|  |

这个函数是关于变量的函数，对于每一个给定的，去寻找合适的参数，使得找到的使拉格朗日函数的值最大。针对的不同取值，我们观察一下这个函数的结果。

（1）假设给定某个满足原问题约束条件的，即且，则当时，，，所以

，

所以

|  |  |
| --- | --- |
|  |  |

（2）假设给定某个违反原问题约束条件的，即存在某个使得 或。若，可令，使得；若，可令，使，使得。将其余均取值为0，则

|  |  |
| --- | --- |
|  |  |

综合上述两种情况，得

|  |  |
| --- | --- |
|  | （C-4） |

这样一来，我们成功地用拉格朗日函数表示了原始问题，并将原问题的约束条件去掉了。所以拉格朗日函数就是为了表述“符合约束条件的本身”。

原问题求满足约束条件的的最小值，所以最终原问题的表达形式如下：

|  |  |
| --- | --- |
|  | （C-5） |

称为广义拉格朗日函数的极小极大问题。

定义原问题的最优值对应的目标函数值为

|  |  |
| --- | --- |
|  | （C-6） |

但有时候，这个问题的求解仍然比较复杂，比如在约束多变量少的时候，直接求解会很麻烦，这个时候需要考虑对偶问题。

#### 2．对偶问题

对偶问题就是将 转化成 。

我们构造关于的函数：

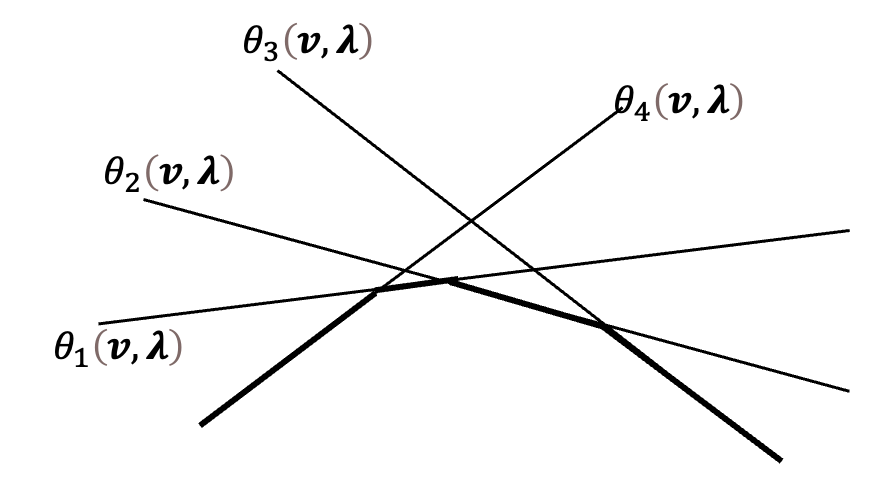
|  |  |
| --- | --- |
|  | （C-7） |

（1）对偶函数一定是一个凹函数。由于对偶函数是凹函数，故拉格朗日对偶问题一定是凸优化问题。

证明：

 可以看作是一个无限的函数集，这个函数集中每个元素是，取遍其在定义域上的所有不同的  。针对不同的 ， 的表达式不一样，由于这个表达式是只关于和 的，故用来表示。所以有

当看成是关于或的函数时，是一个仿射函数（的线性组合），亦即是仿射函数。对仿射函数集取min（粗线条），得到的函数是凹函数，如图C-1所示。



图C-1 仿射函数集的min运算为凹函数

（2）对 ，，如果原问题最优解对应的目标函数值为 ,则。

证明：设原问题的最优解为，则且，有：

假设这个最优解 在处取得，结合性质（1）中的证明部分，有 ：

很显然，由图C-1可以看到，的图像总在下方，即以下不等式总成立：

|  |  |
| --- | --- |
|  | （C-8） |

（3）极大化，就是对偶问题 ，即

|  |  |
| --- | --- |
|  |  |

称为广义拉格朗日函数的极大极小问题。

将广义拉格朗日函数的极大极小问题表示为带约束的最优化问题

|  |  |
| --- | --- |
|  | （C-9） |

称为原问题的对偶问题。注意对偶问题中的约束条件只有，形式比原问题的约束条件简单。

定义对偶问题的最优值

|  |  |
| --- | --- |
|  | （C-10） |

对比原问题，对偶问题是先固定，求最优化的解，再确定参数；原问题是先固定，求最优化的解，再确定。

#### 3．原问题和对偶问题的关系

若原问题和对偶问题都有最优值，那么

|  |  |
| --- | --- |
|  | （C-10） |

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。

证明：根据式（C-8），，由于原问题和对偶问题均有最优值，所以

即

与弱对偶性相对应的有一个强对偶性（Strong Duality） ，强对偶即满足：

|  |  |
| --- | --- |
|  | （C-11） |

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。

**Slater条件**：对于原问题及其对偶问题，如果：

（1）假设函数和是凸函数， 是仿射函数；

（2）不等式约束是严格可行的，即存在，对所有有；

则存在，使 是原始问题的解，是对偶问题的解，并且

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。

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，任意原问题与对偶问题的解都满足 KKT 条件：

|  |  |
| --- | --- |
|  | （C-12） |

拉格朗日对偶法的求解分为两个步骤：

（1）把原始的约束问题通过拉格朗日函数转化为无约束问题；

（2）在满足KKT的条件下用求解对偶问题来代替求解原问题，使得问题求解更加容易。

强对偶性成立时，将拉格朗日函数分别对原变量和对偶变量求偏导，并令偏导为0，得到原变量和对偶变量的数值关系，即可解决主问题变量求解问题。SVM的对偶问题就是如是推导得到。

例：有3个样本点分别为 ，，，其中、为正例，为负例，求最大间隔分离超平面。作图如下

图表, 散点图

描述已自动生成

**求解原问题：**

变成标准形式：

基于该标准形式，构造的拉格朗日函数为

上面等式给出了3个方程，还需要加上松弛条件：，。

1. 负例只有一个，肯定是支持向量，，正例有两个，先假设只有为支持向量，，

根据，得到。

，最优解在边界上：

得到，，最大间隔超平面为，。

还需要验证，满足约束条件，所以该解是最优解。

1. 负例只有一个，肯定是支持向量，，正例有两个，假设只有为支持向量，，

根据，得到。

，最优解在边界上：

，，，最大间隔超平面为，。

但还需要验证，不满足约束条件，所以该解是不是最优解。

1. 假设3个点都是支持向量，

，最优解在边界上：

其中第1行和第2行不能同时满足，所以无解。

综上3种情况，得到，，最大间隔超平面为，。

**求解对偶问题：**

先求解，、都与相关，得到了 也就求得了超平面参数。

，，，

，，，

由，得到将，将其带入目标函数

由，得到，代入。易知时取极大值。但是注意到这里的 不满足约束条件 ，所以此时极大值就不在边界内而是在边界上，需分别考虑 为 0 的情况：

1. ，，，；
2. ，，。

综上两种情况，的极大值为，对应的，，。

通过拉格朗日乘子的特点就可以知道 、为支持向量（因为，） ，所以约束条件  ，，所以 、为是在间隔边界上的。

然后再代入 、的计算公式，求得：

结果同原问题求解相同。

参考文献

[1] KKT条件，原来如此简单 (上) | 理论部：https://zhuanlan.zhihu.com/p/556931657

[2] Karush-Kuhn-Tucker (KKT)条件：<https://zhuanlan.zhihu.com/p/38163970>

[3]如何通俗地讲解对偶问题，尤其是拉格朗日对偶 lagrangian duality？

https://www.zhihu.com/question/58584814