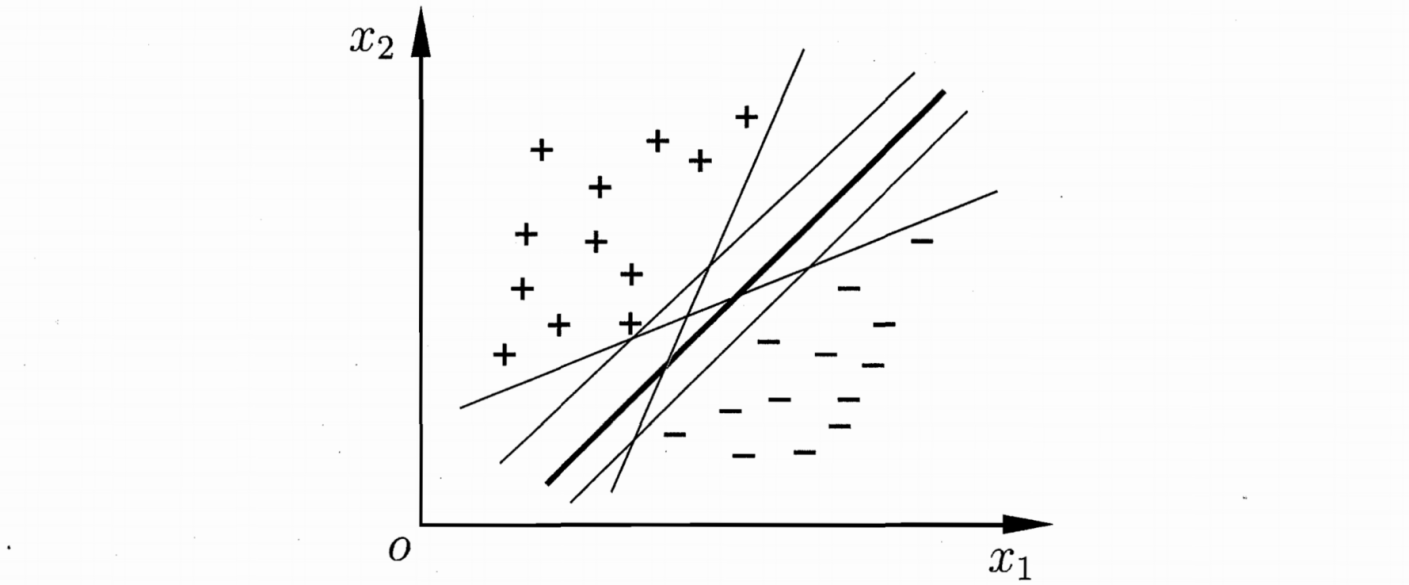
# 6.支持向量机

支持向量机是一种经典的二分类模型，基本模型定义为特征空间中最大间隔的线性分类器，其学习的优化目标便是间隔最大化，因此支持向量机本身可以转化为一个凸二次规划求解的问题。

## 6.1.间隔与支持向量

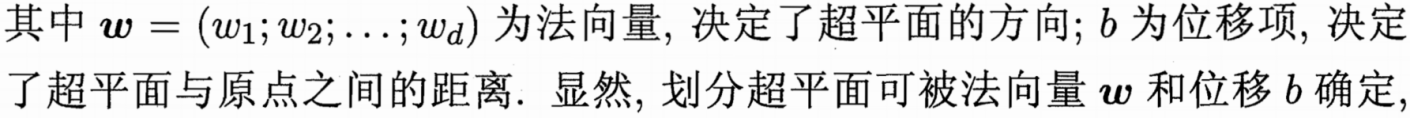
对于二分类学习，假设现在的数据是线性可分的，这时分类学习最基本的想法就是找到一个合适的超平面，该超平面能够将不同类别的样本分开，但能将训练样本分开的划分超平面可能有很多，我们应该努力去找到哪一个呢？

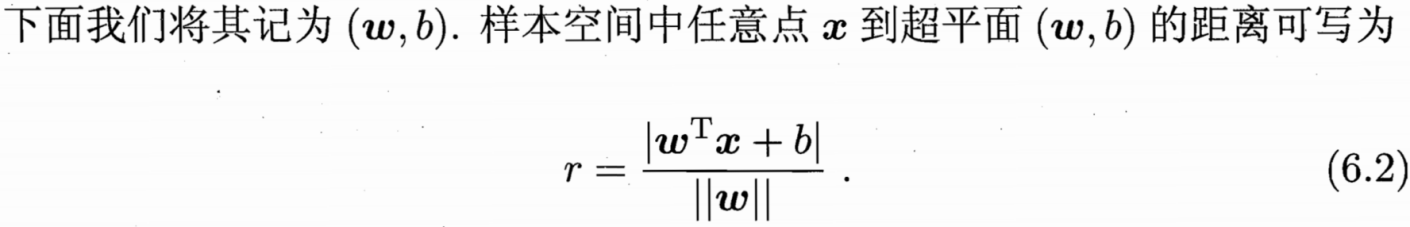


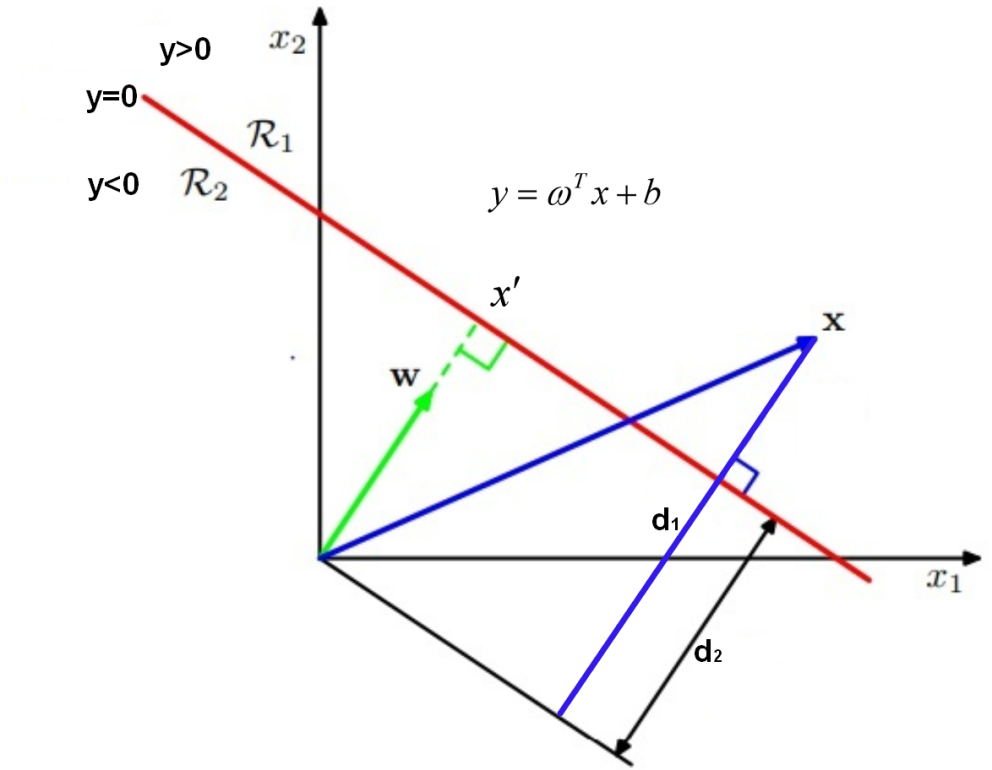
易知：当超平面距离与它最近的数据点的间隔越大，分类的鲁棒性越好，即当新的数据点加入时，超平面对这些点的适应性最强，出错的可能性最小。因此需要让所选择的超平面能够最大化这个间隔Gap

在样本空间中，划分超平面可通过如下线性方程来描述：









证明为法向量

设为划分超平面上任意两点，则



两式相减得



由于为超平面上的任意向量，根据上式内积为0得为法向量

求点到超平面的距离？

首先求原点到超平面距离

如图，令，则

所以为原点到直线距离

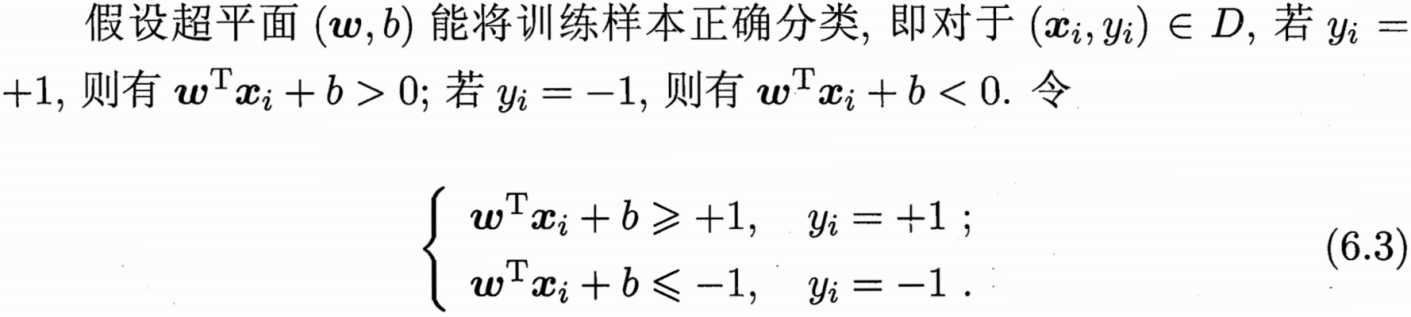
然后，求任意点到超平面距离

如图，到超平面距离为

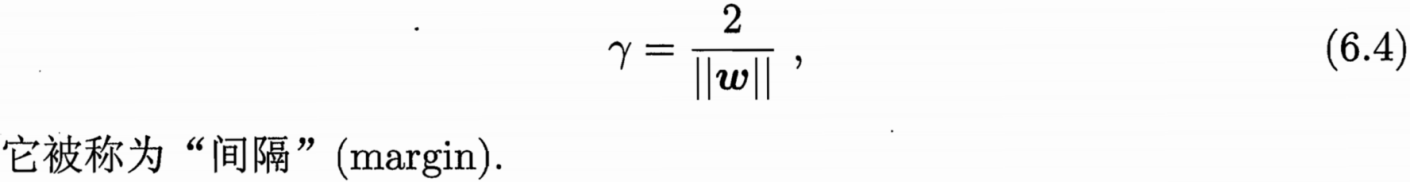
，即与法向量的内积

，即原点到超平面的距离

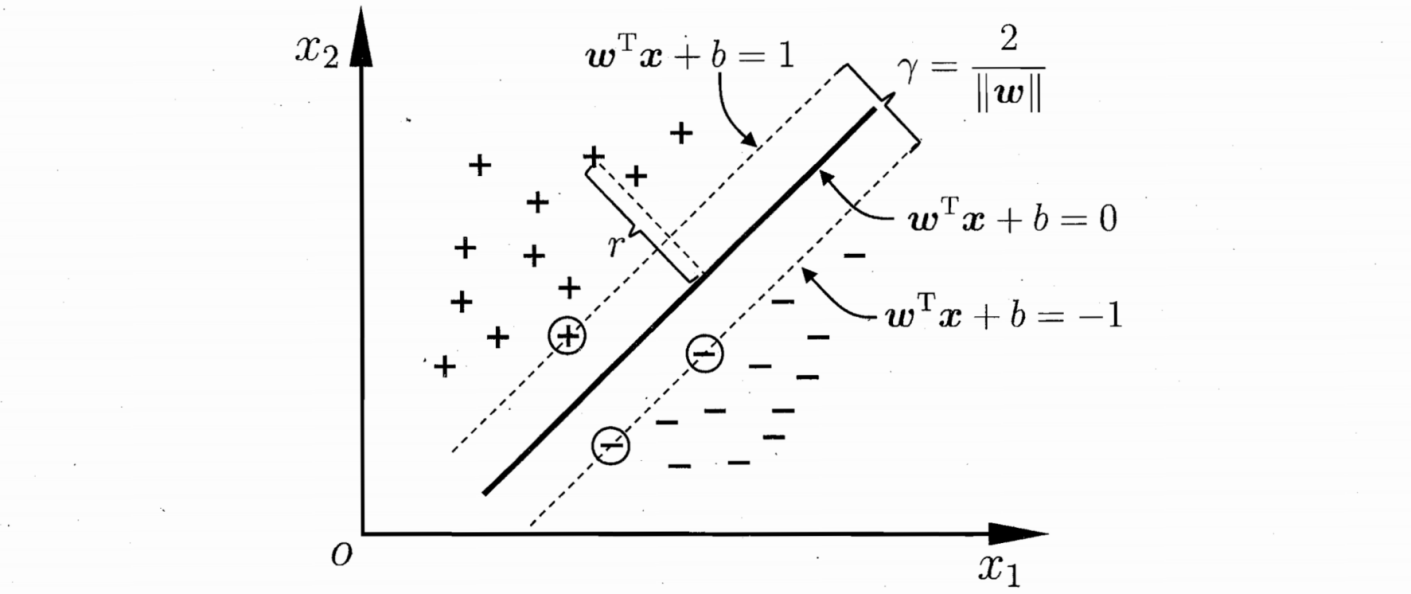
所以，即为上式6.2



如下图所示，距离超平面最近的这几个训练样本点使式（6.3）的等号成立，它们被称为“支持向量”（support vector），两个异类支持向量到超平面的距离之和为



它被称为“间隔”（margin）.



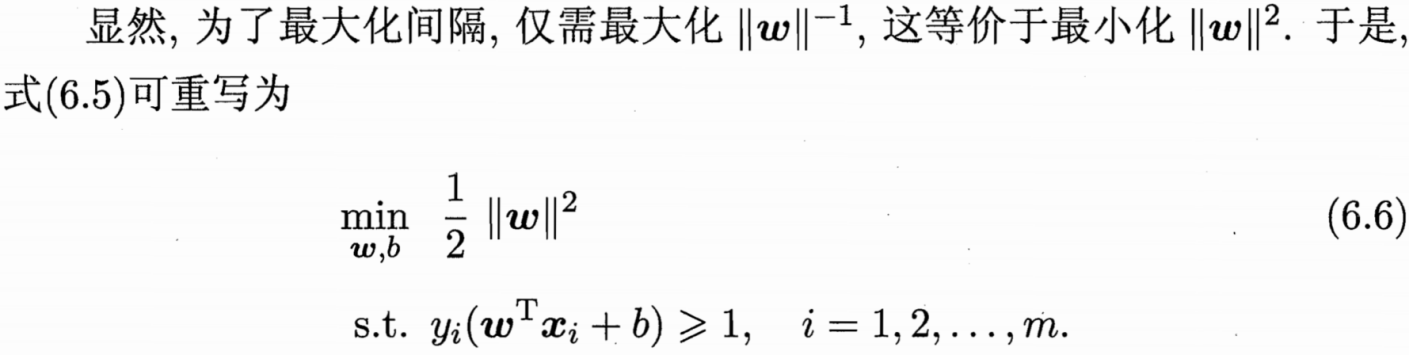
这里，令支持向量所在的直线分别为和

因为支持向量所在直线可以设置为，两边同时乘以则等号不变，并设

所以有，将看做新的，看做新的，则得出上式。

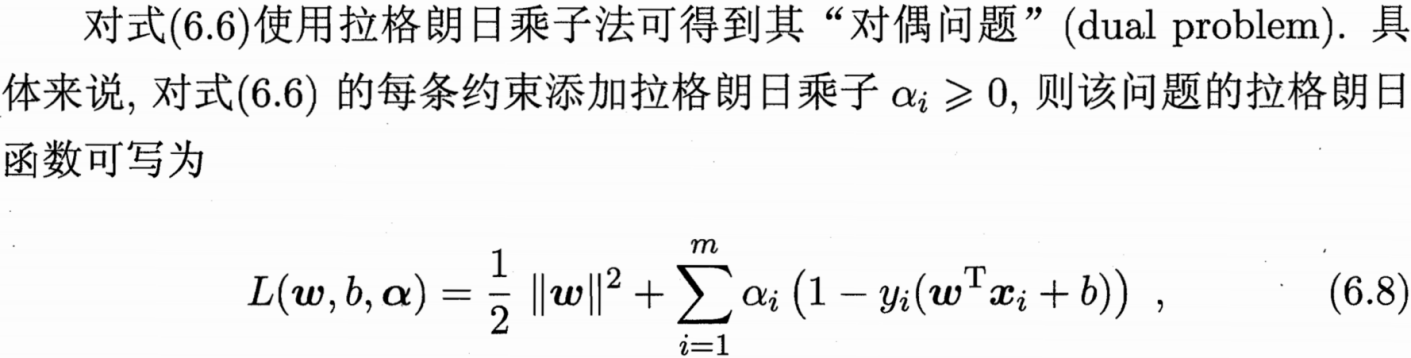
因此，将两条支持向量直线带入式6.2即得，即式6.4

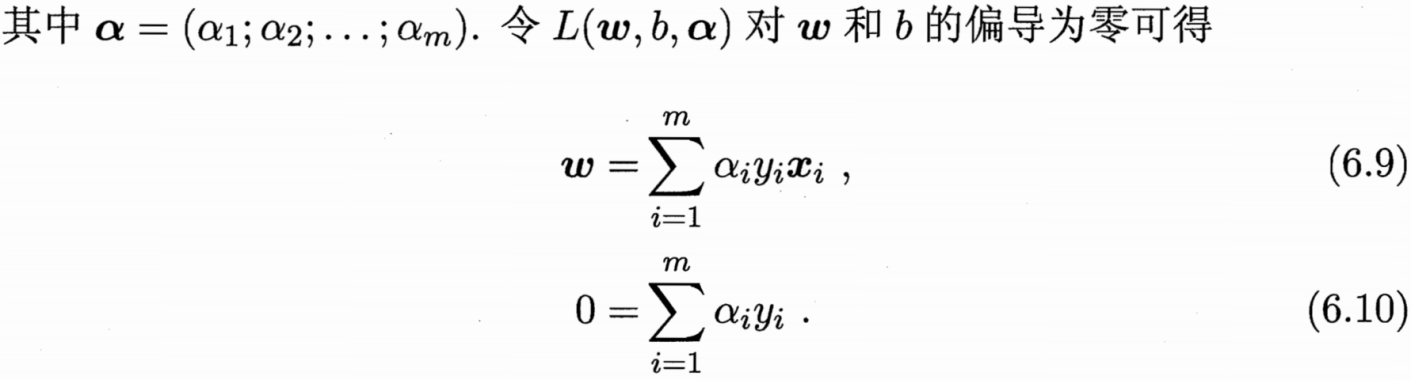




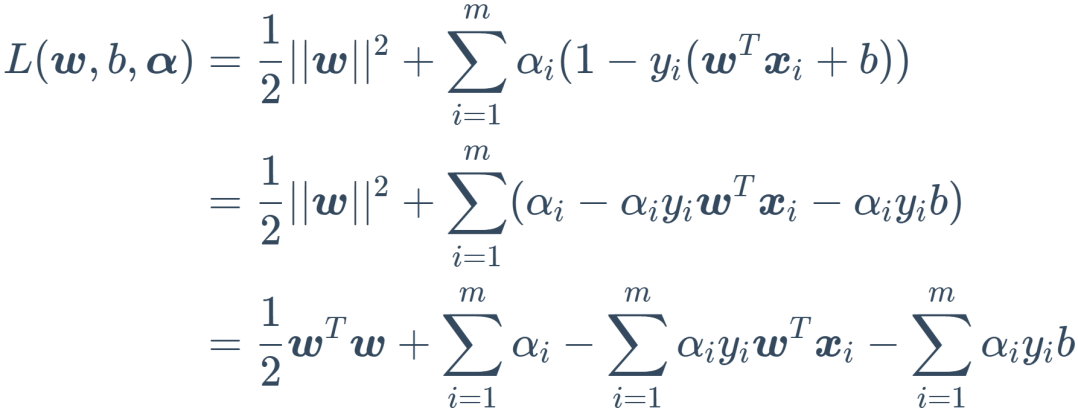
这就是支持向量机（Support Vector Machine，简称SVM）的基本型。

## 6.2.对偶问题

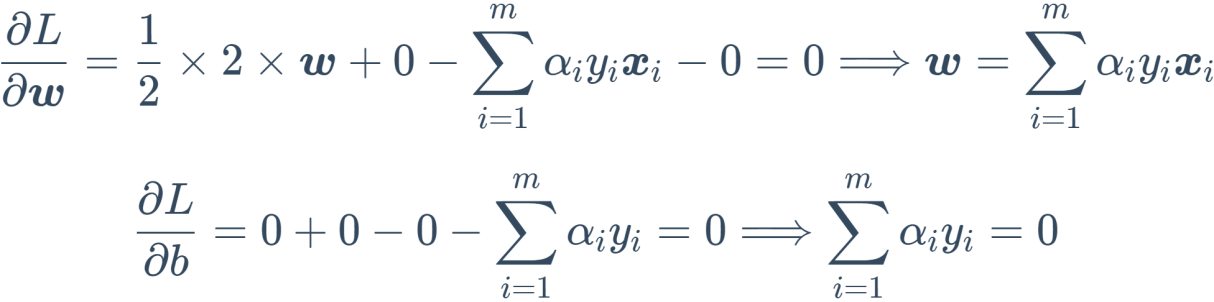


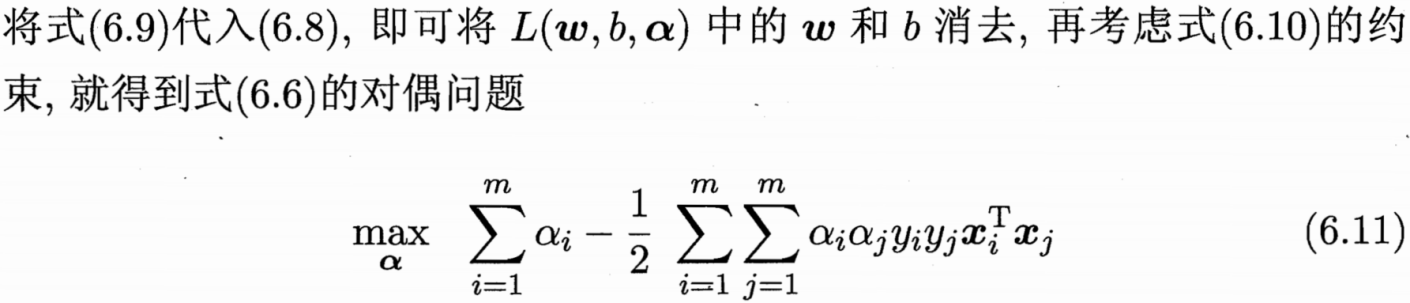


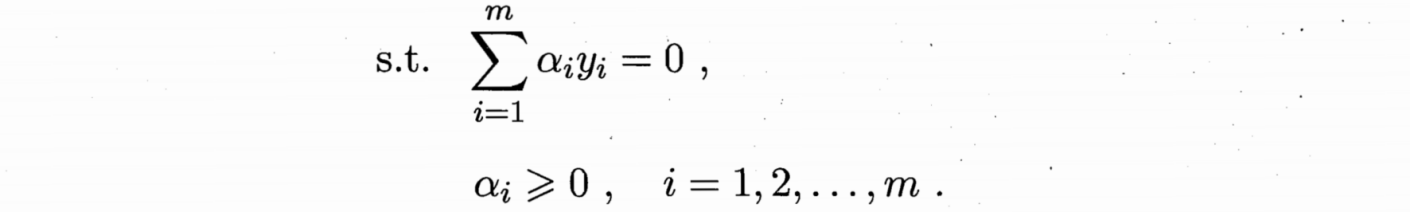
公式(6.8)可作如下展开



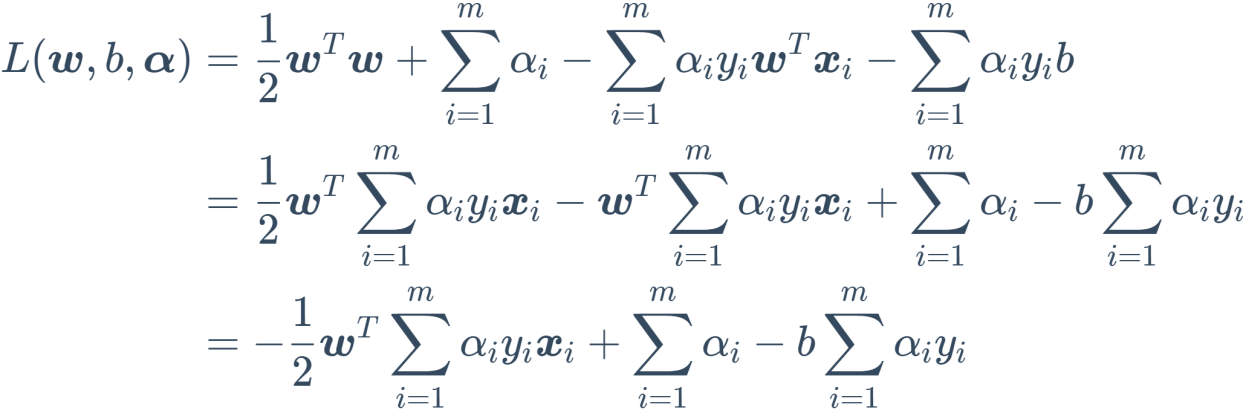
对和分别求偏导数并令其等于0



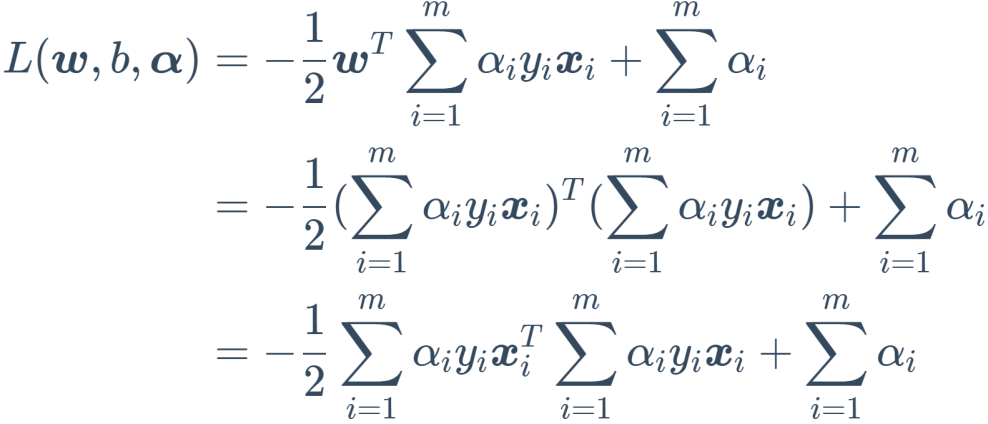


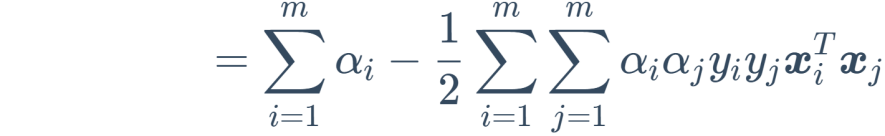


这里，首先将代入得

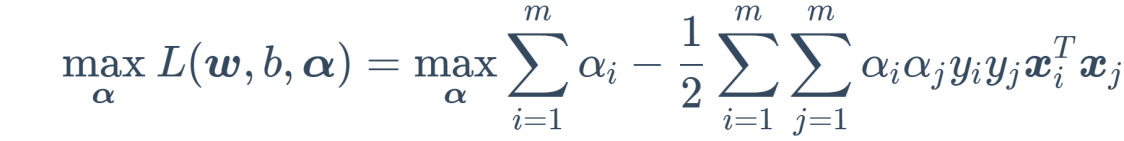


由于，所以上式最后一项可化为0，于是得





所以

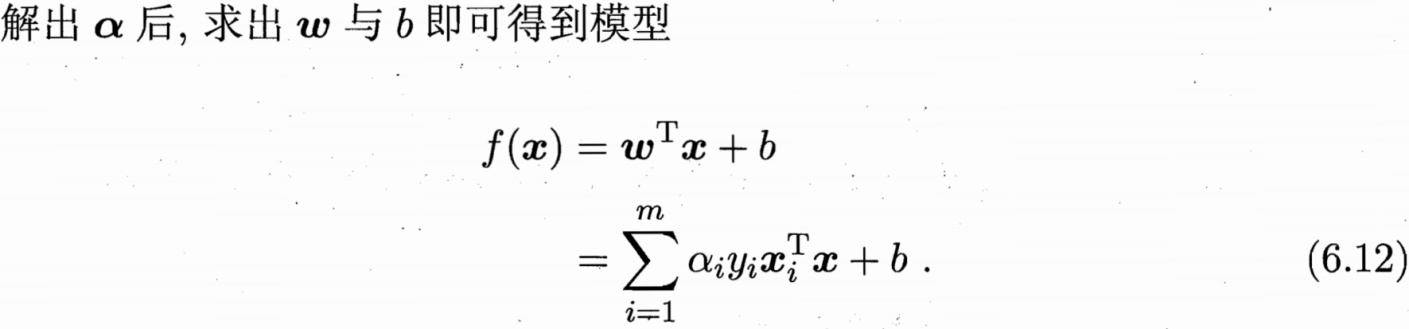


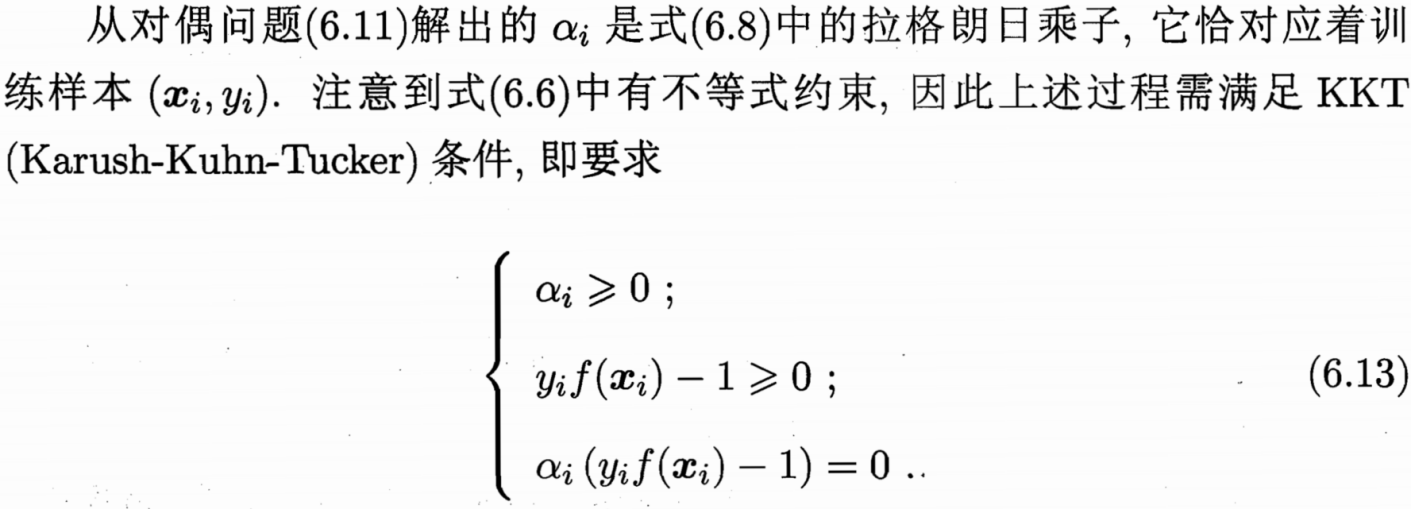
即得式6.11

同时，上式等价于



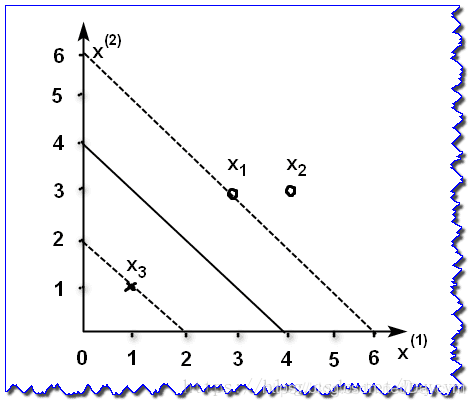
解出后，求出与即可得到模型



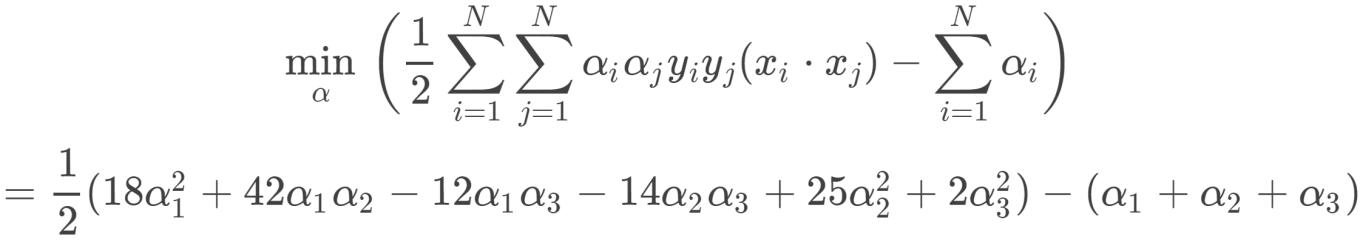


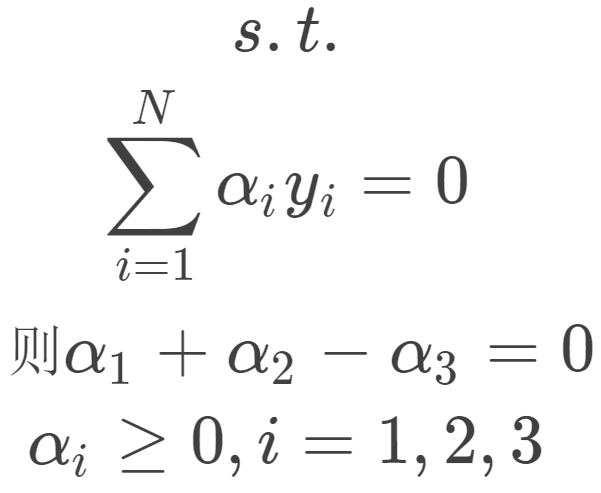
例题

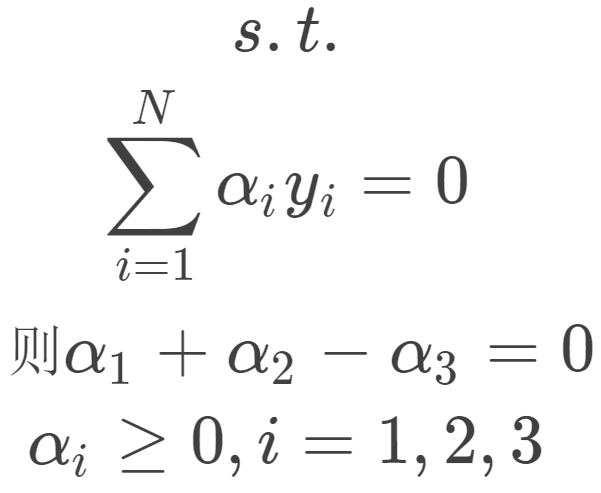
如图，已知，，，，，求最大间隔分离超平面。



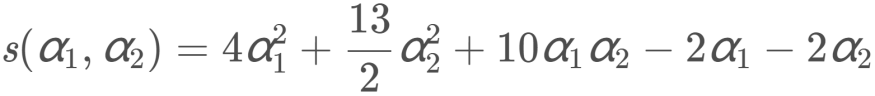
解：首先，写出对偶问题



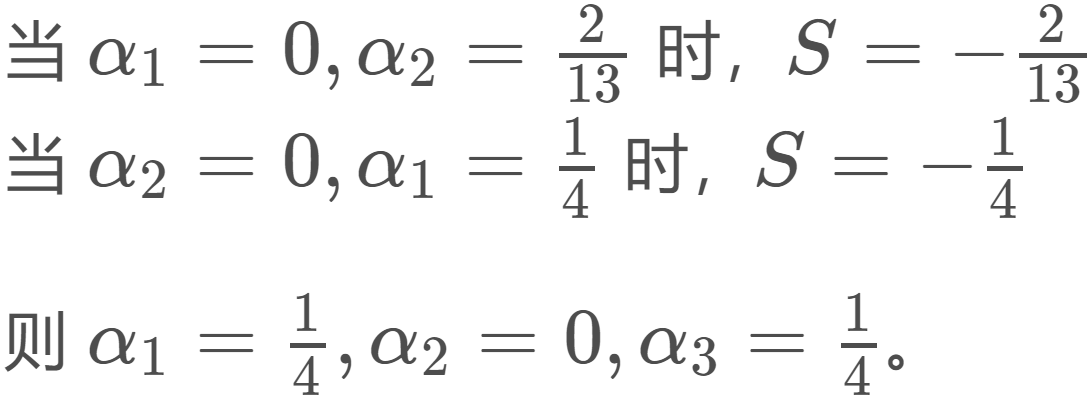


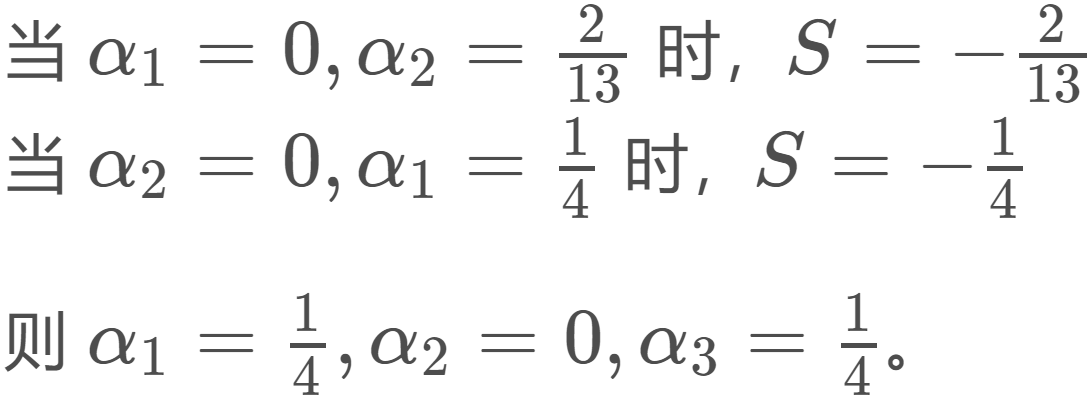


求解该最优化问题。将代入目标函数并记为：

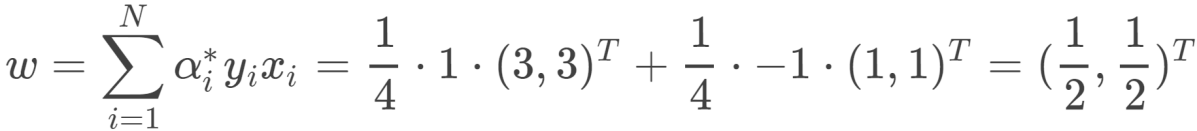


对  求偏导，令其为0，则有，不符合乘子大于0的要求。考虑S为凸函数，在某个范围内，要么有极值；没有极值时，则关注边界值。

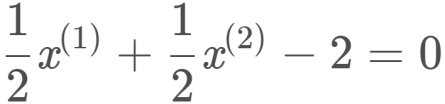




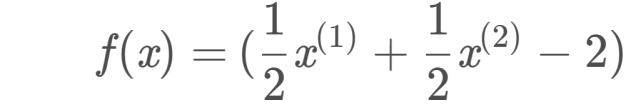
对应的实例点是支持向量，将代入得



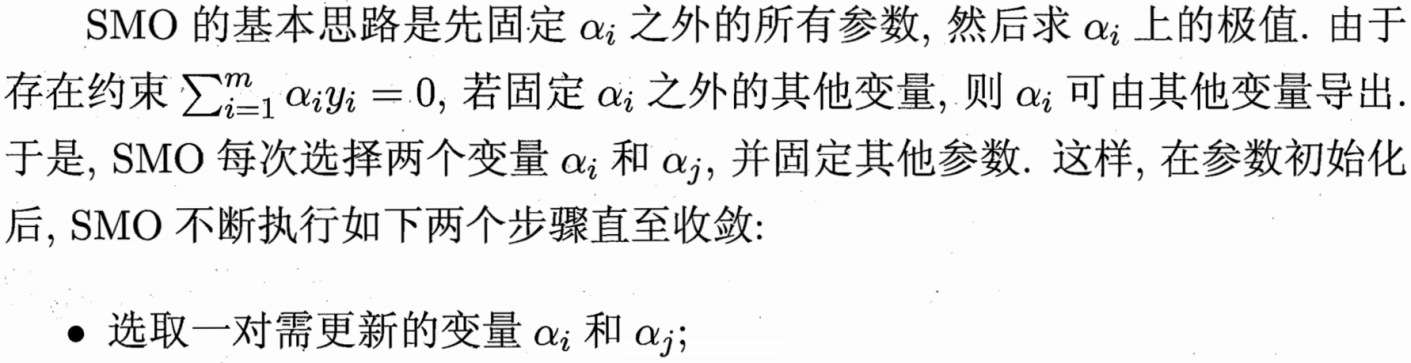
采用两个支撑向量，计算得到的 b 均为-2。所以分离超平面为



超平面所对应的模型为



从以上案例可以看出，求解式（6.11）的规模正比于训练样本数，这会在实际任务中造成很大的开销.为了避开这个障碍，人们通过利用问题本身的特性，提出了很多高效算法，SMO（Sequential Minimal Optimization）是其中一个著名的代表。



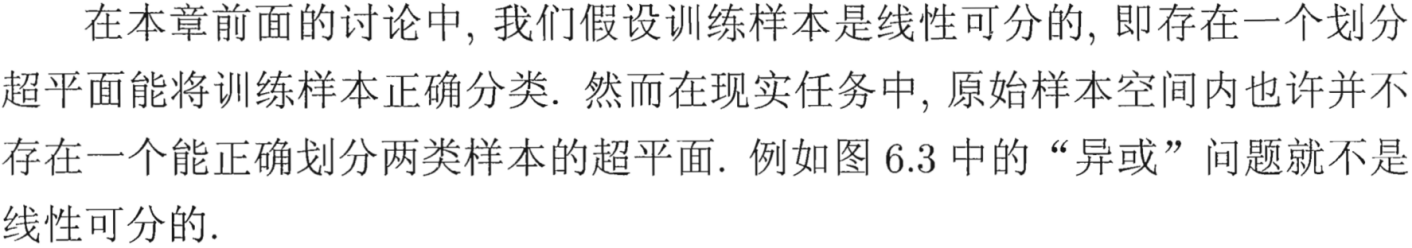


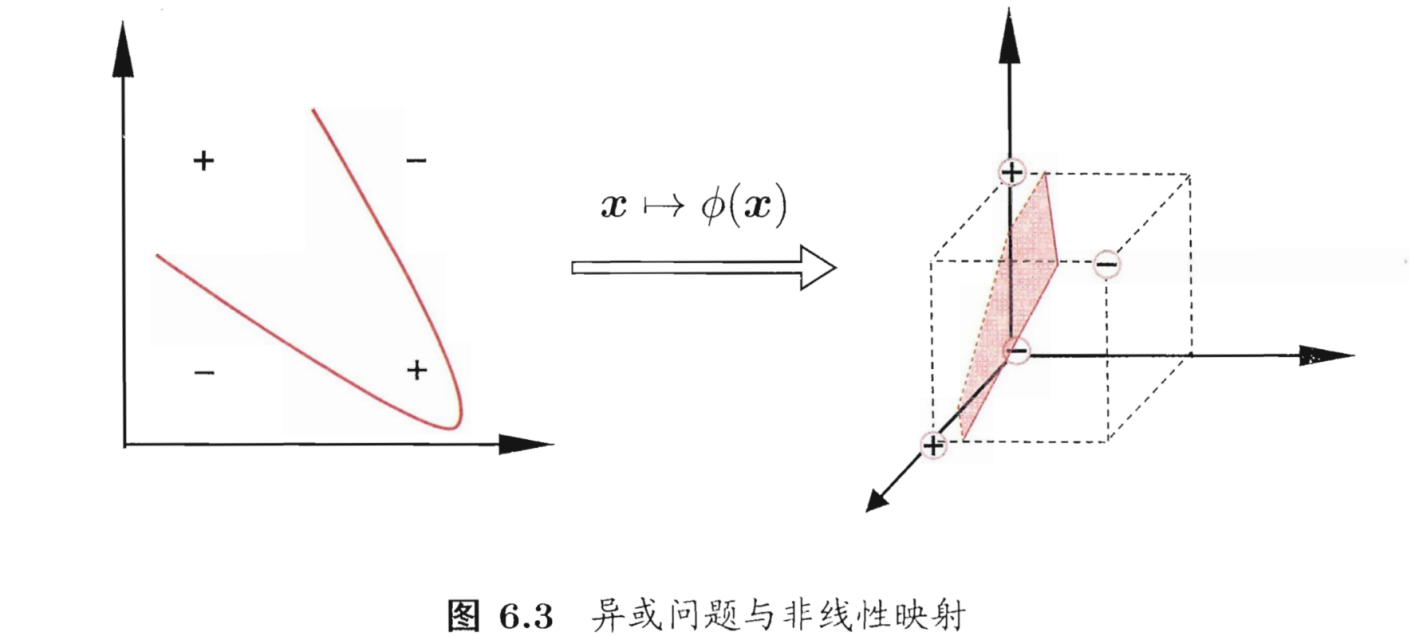
## 6.3.核函数

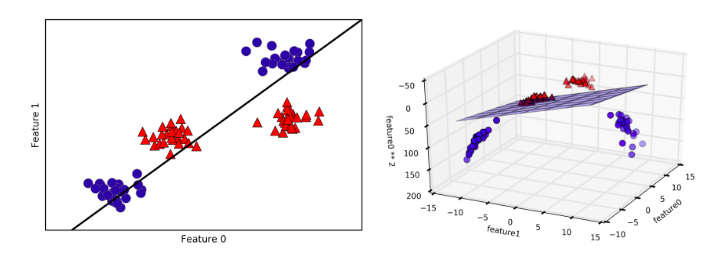
核函数思想：

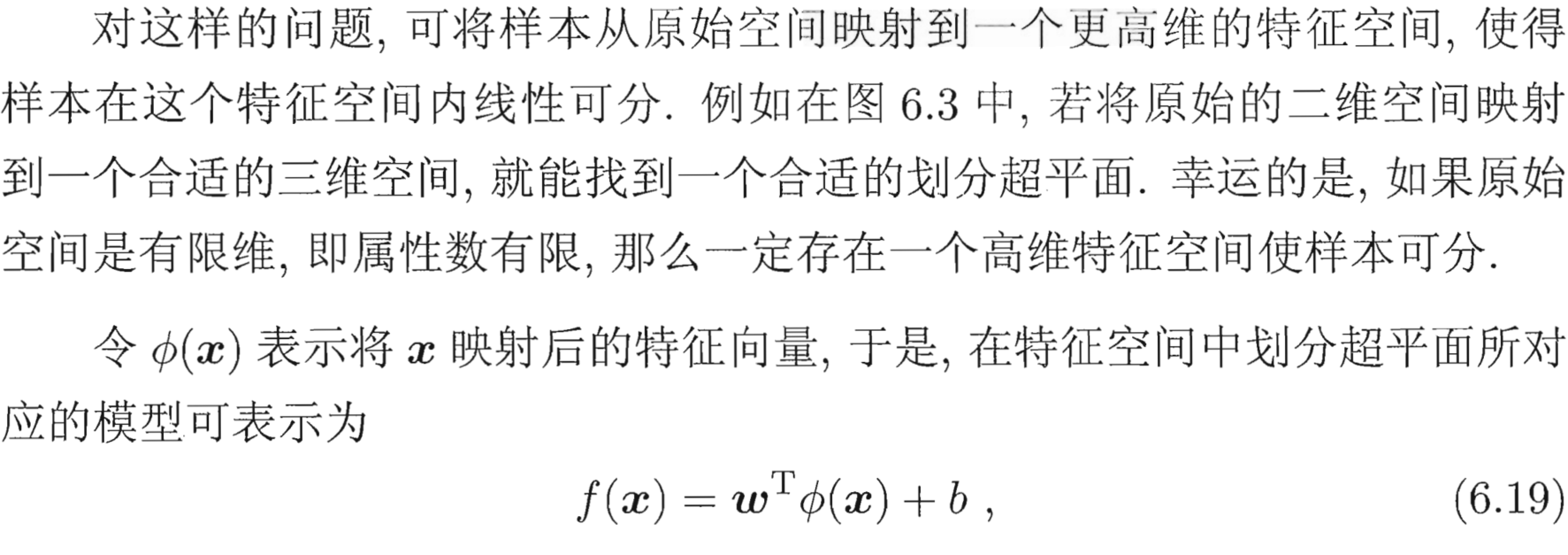
把非线性转换为线性，即低维高维

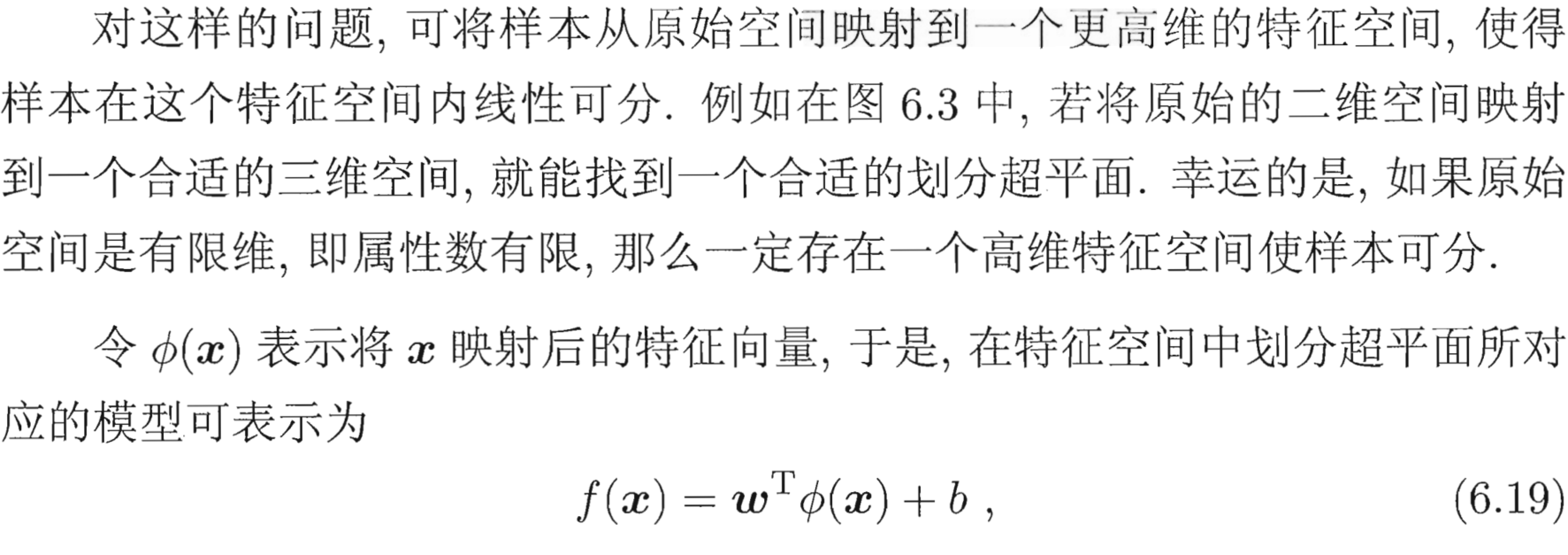
找出对应核函数，即高维低维

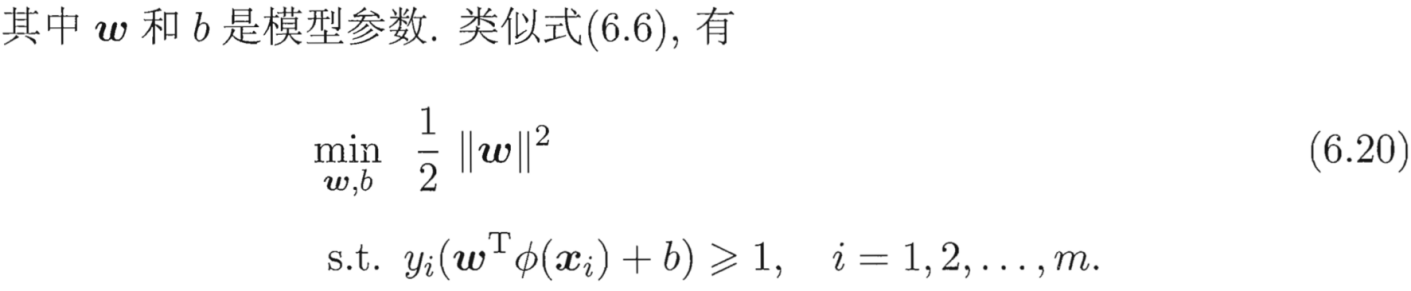


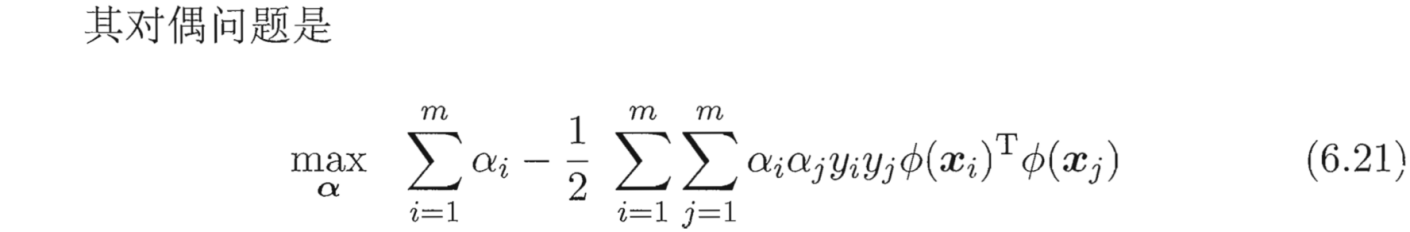


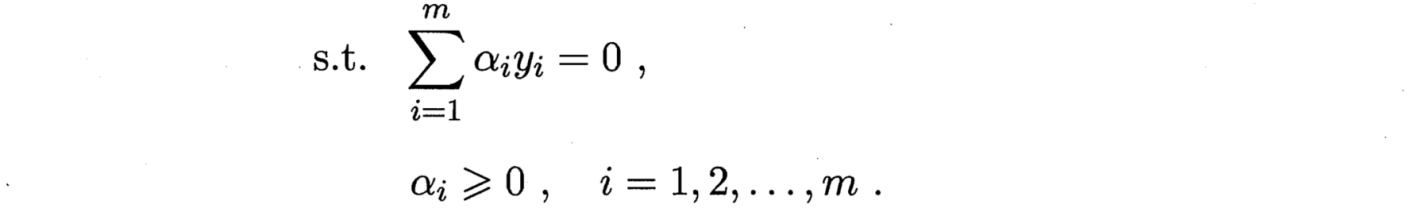


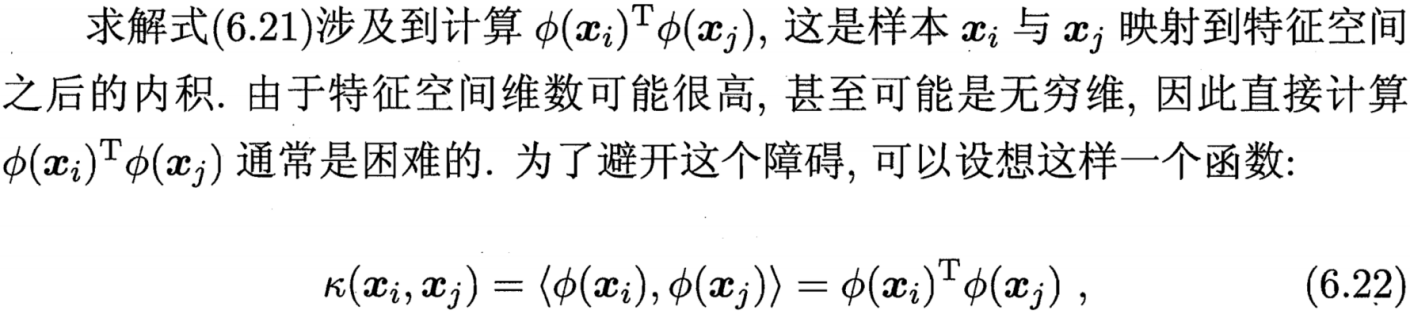


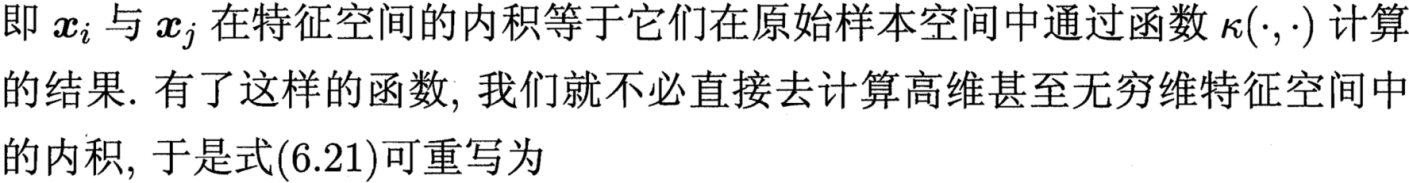


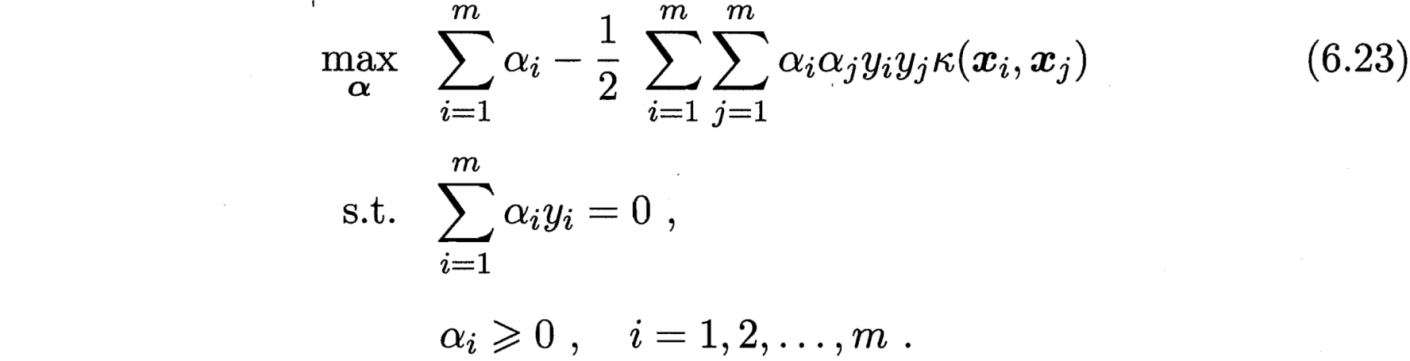


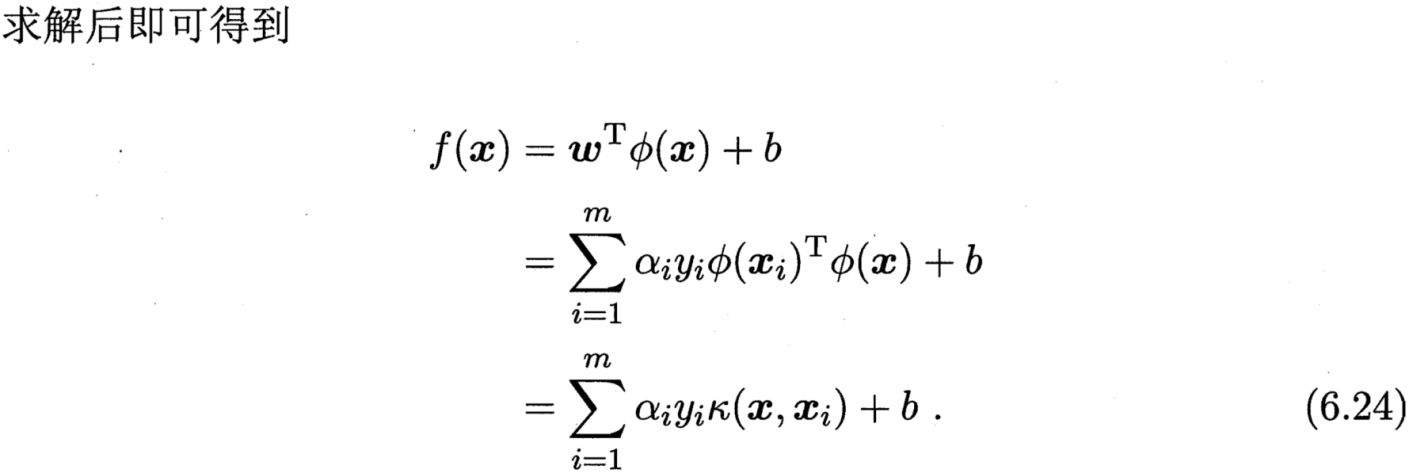


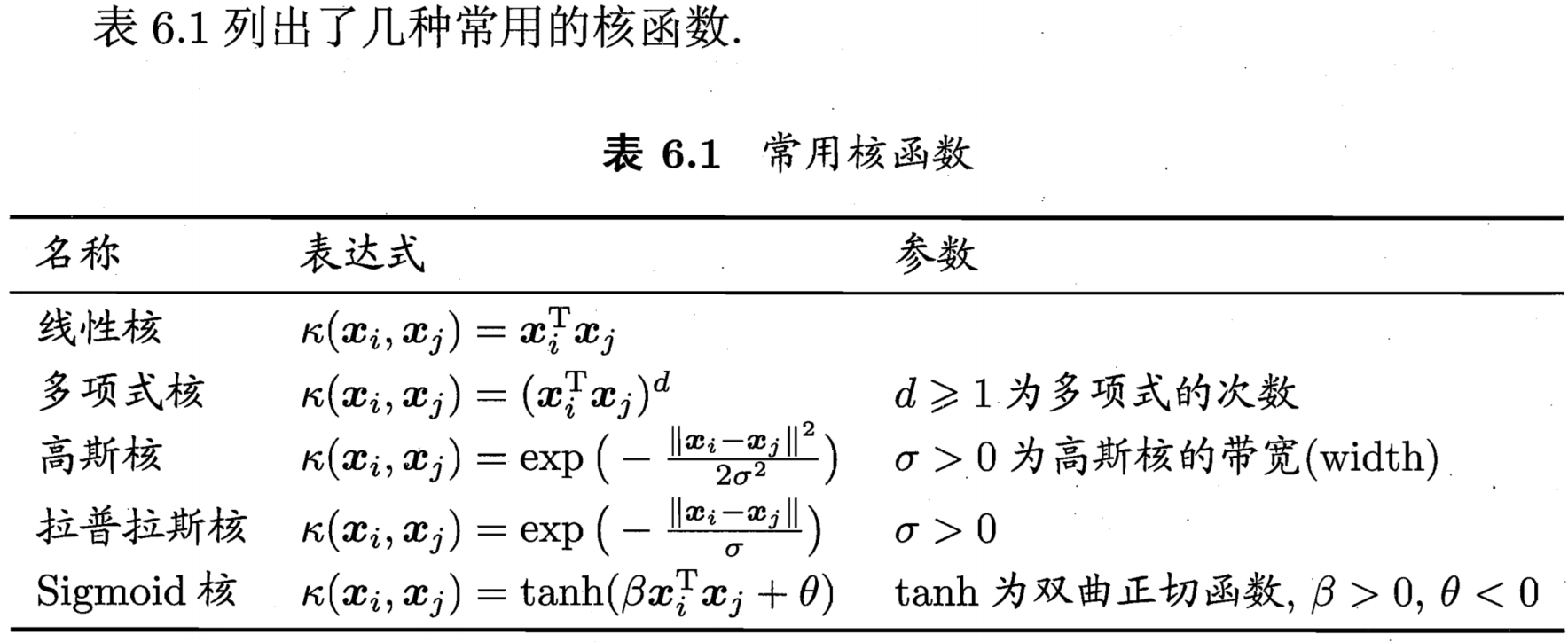


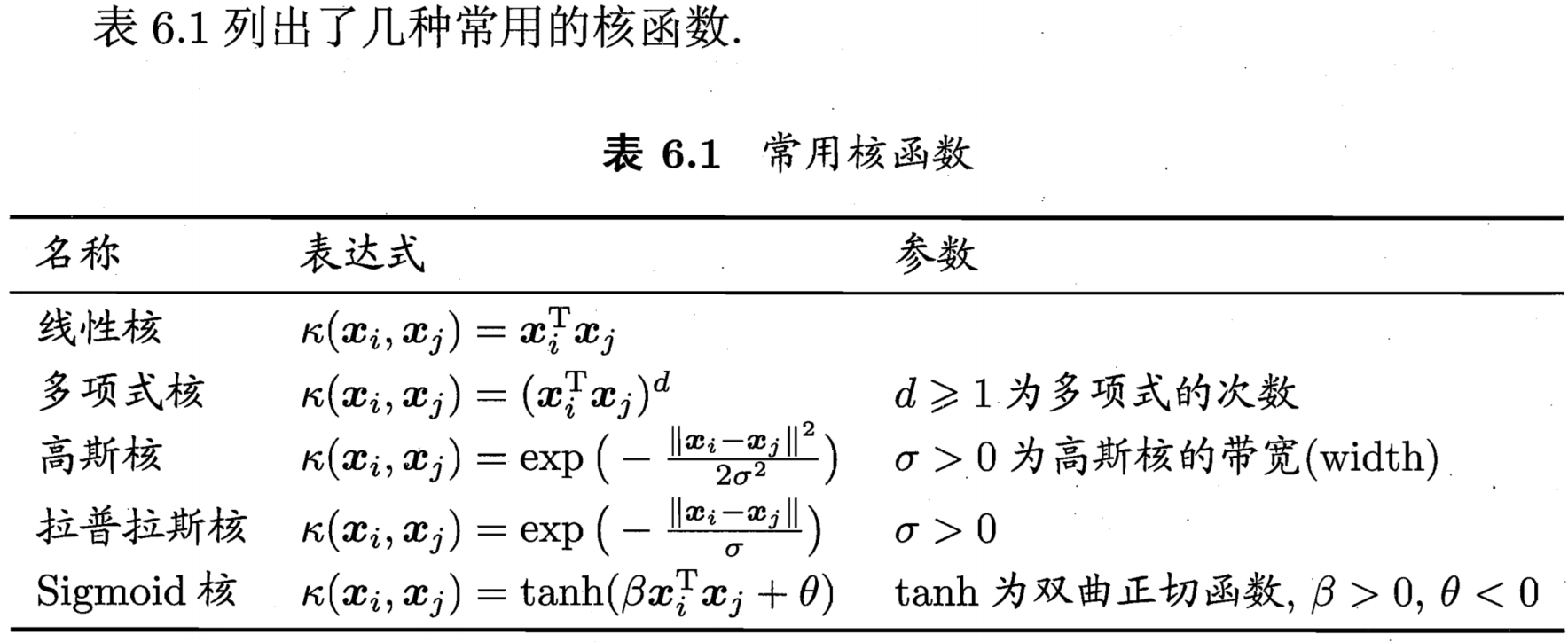












多项式核中以d=2为例

设二维向量映射为

则对于和映射后为和

其中





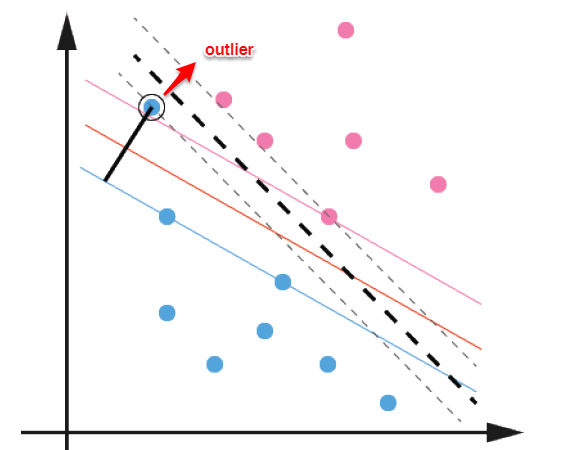
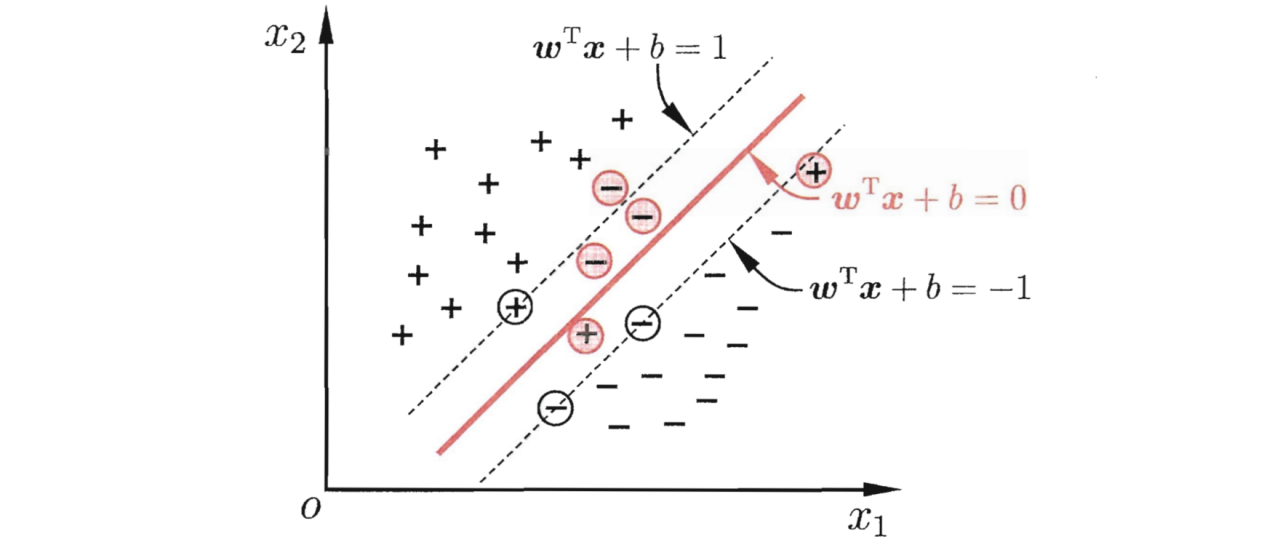
因此，在求解时只需求解原数据的内积即可，同理当次数更高时，可以映射到更高维空间，其中高斯核函数利用泰勒展开，可以映射到无限维中。

## 6.4.软间隔

在前面的讨论中，我们一直假定训练样本在样本空间或特征空间中是线性可分的，也对线性不可分数据集问题通过核函数对原来的线性SVN进行了推广，使得非线性的情况也能处理。

但是如果数据并不是因为本身是非线性结构的，而是因为数据有噪声，或者貌似线性可分的结果是由于过拟合所造成的.

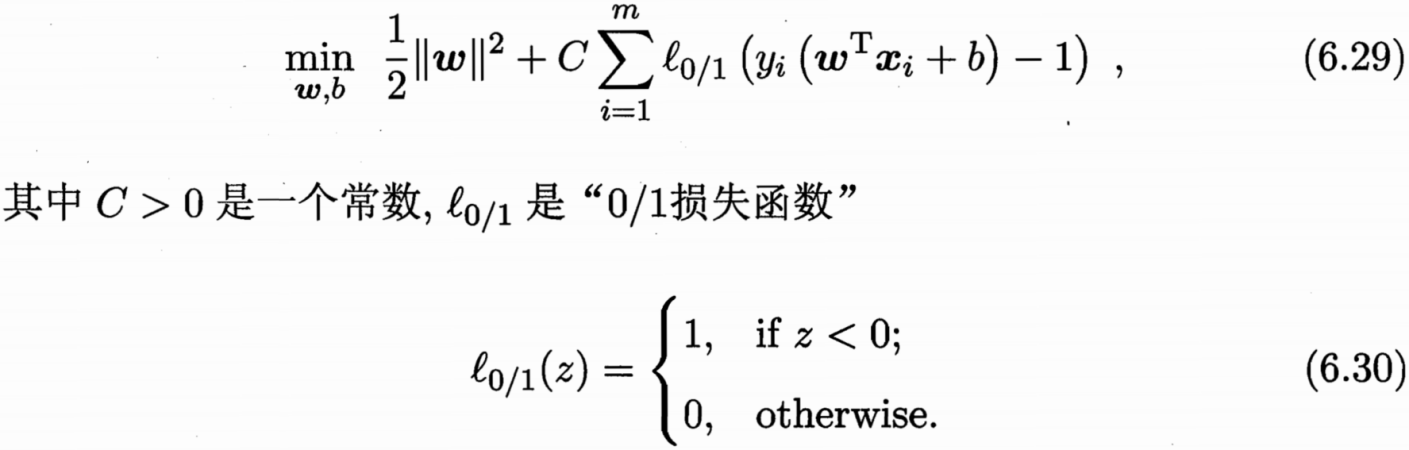
缓解该问题的一个办法是允许支持向量机在一些样本上出错.为此，要引入“软间隔”（soft margin）的概念，如图所示.



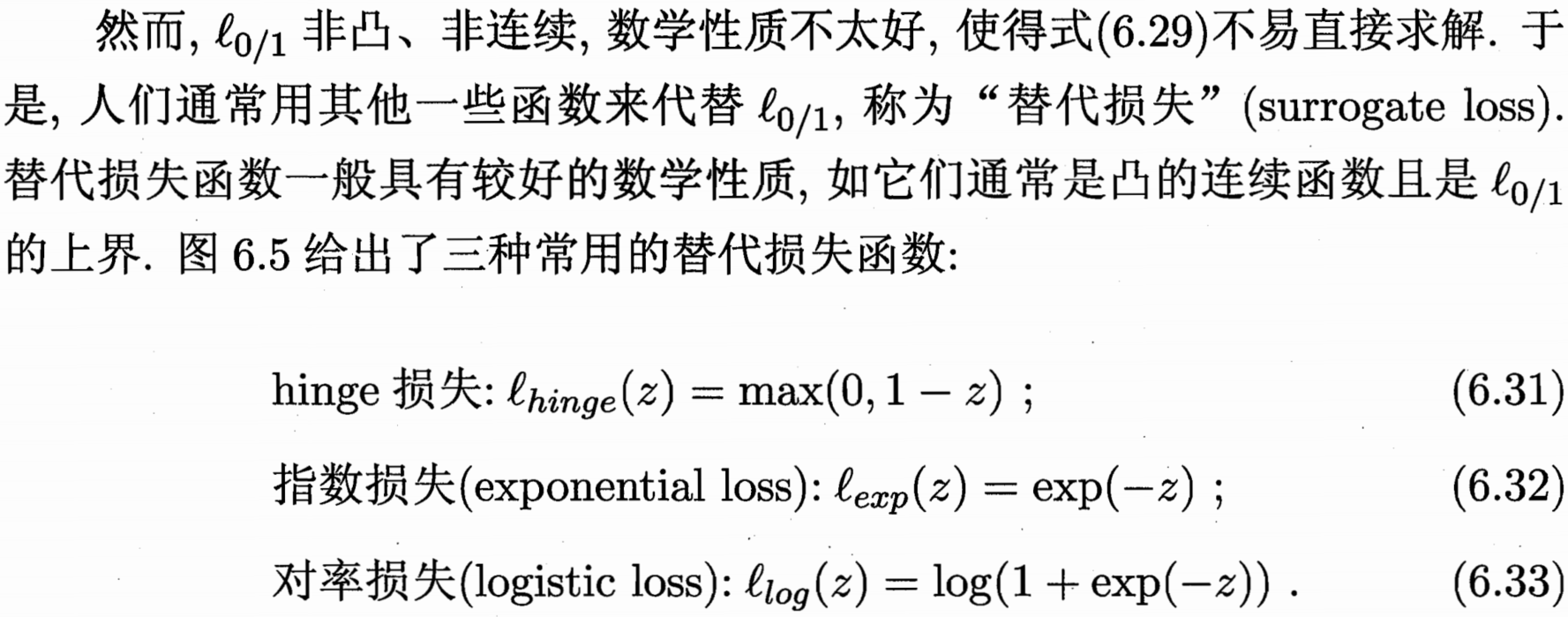
我们之前提到的硬间隔，是要求所有样本都满足约束条件，而软间隔则是允许某些样本不满足约束：

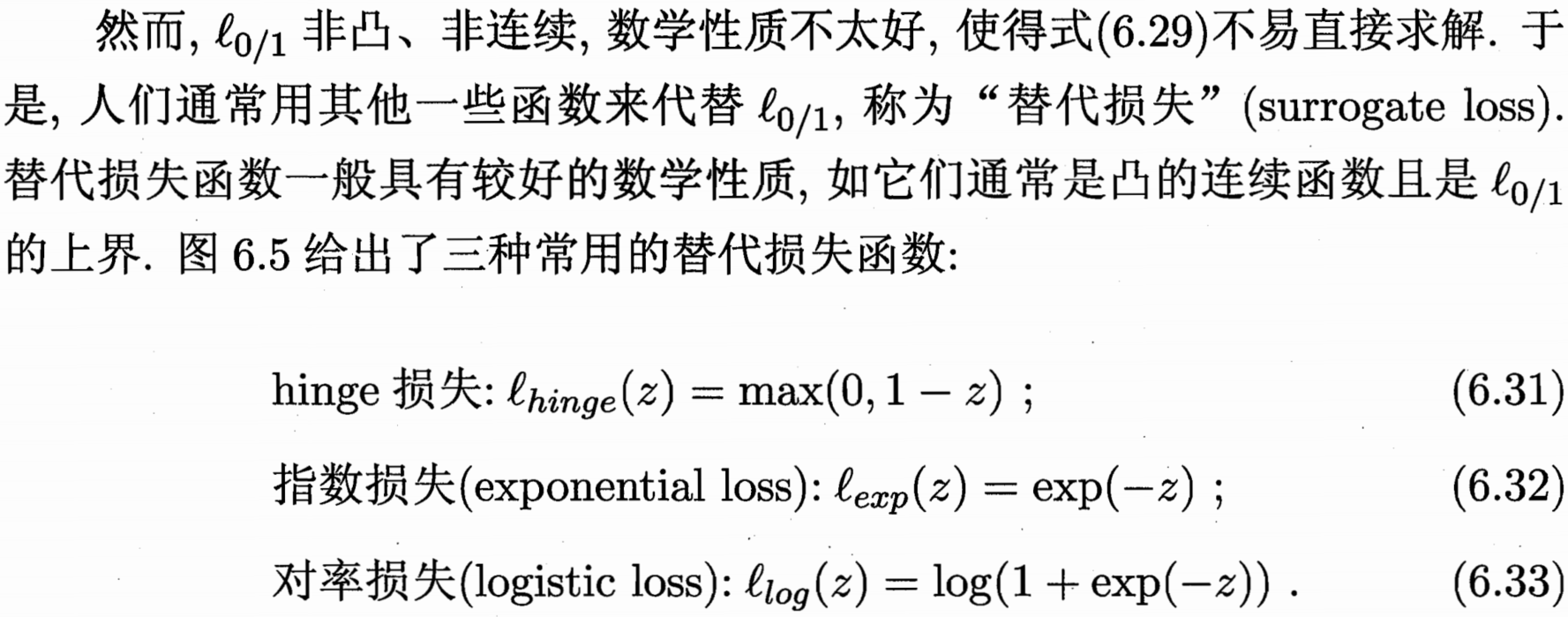


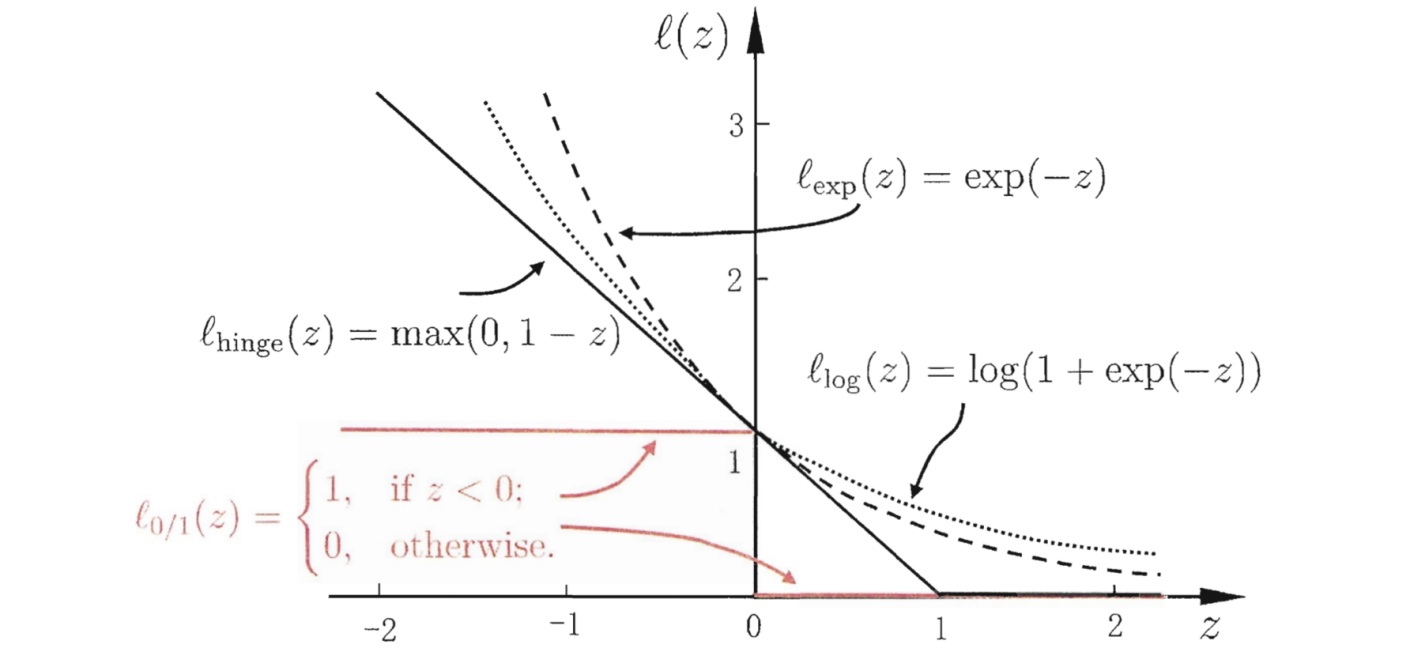
当然，在最大化间隔的同时，不满足约束的样本应尽可能少.于是，优化目标可写为



显然，当C为无穷大时，式（6.29）迫使所有样本均满足约束（6.28），于是式（6.29）等价于原始的硬间隔问题；当C取有限值时，式（6.29）允许一些样本不满足约束.









采用hinge损失为例，为了解决某些样本不满足约束条件，可以对每个样本点引入一个松弛变量，这样，约束条件就变为：



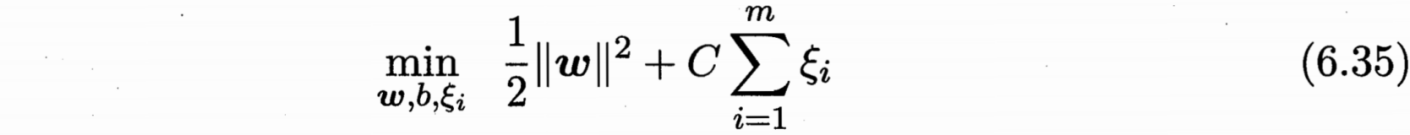
当样本分类正确且在间隔两边时，，

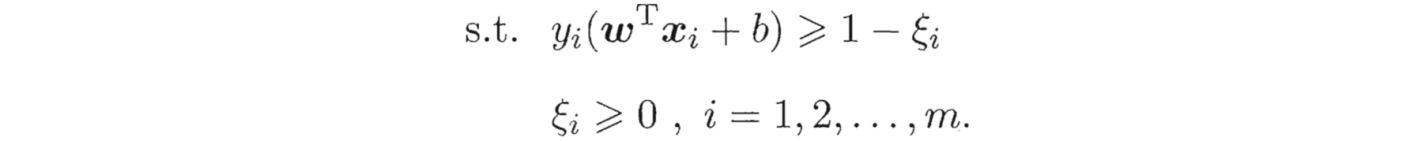
当样本分类正确且在间隔之间时，，

当样本在划分超平面上时，，

当样本分类错误时，，

引入松弛变量后，优化目标可写为

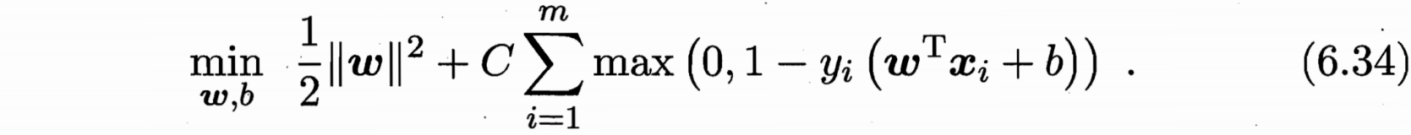




其中，可转换为，因此有



因此，采用hinge损失形式，即



hinge损失函数图像也称为合页图

这里C称为惩罚系数，C值大时对误分类的惩罚增加，容易导致过拟合，C值小时对误分类的惩罚减小，容易导致欠拟合。

有了上面的思路，可以和线性可分支持向量机一样来考虑当前问题，通过拉格朗日乘子法可得到式（6.35）的拉格朗日函数

