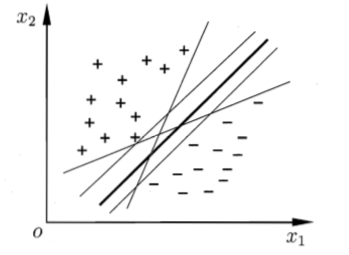
支持向量机

1、间隔与支持向量

给定训练样本集，分类学习最基本的想法就是基于训练集*D*在样本空间中找到一个划分超平面，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？



直观上看，应该去找位于两类训练样本“正中间”的划分超平面。即图中粗线，因为该划分超平面对训练样本局部扰动的“容忍”性最好。例如，由于训练集的局限性或噪声的因素，训练集外的样本可能比图中的训练样本更接近两个类的分隔界，这将使许多划分超平面出现错误，而粗线的超平面受影响最小。换言之，这个划分超平面所产生的分类结果是最鲁棒的，对未见示例的泛化能力最强。

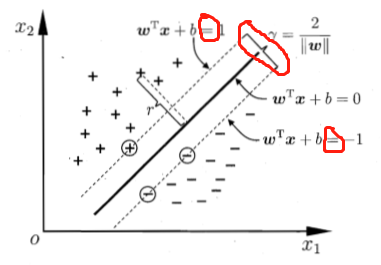
在样本空间中，划分超平面可通过如下线性方程来描述：

其中为法向量，决定了超平面的方向；为位移项，决定了超平面与原点之间的距离。显然，划分超平面可被法向量和位移确定，下面我们将其记为。样本空间中任意点到超平面的距离可写为：

假设超平面能将训练样本正确分类，即对于，若，则有；若，则有。令

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

它被称为“间隔”。



欲找到具有“最大间隔”的划分超平面，也就是要找到能满足上式中约束的参数和，使得最大，即

显然，为了最大化间隔，仅需最大化，这等价于最小化，于是上式可转化为：

这就是支持向量机（SVM）的基本型。

2、对偶问题（拉格朗日乘子）

我们希望求解上式来得到大间隔划分超平面所对应的模型：

其中，W和b是模型参数。注意到本身是一个凸两次规划问题，能直接用现成的优化计算包求解，但我们可以有更高效的办法。

对使用拉格朗日乘子法可得到其“对偶问题”。具体来说，对式子的每条约束添加拉格朗日乘子，则该问题的拉格朗日函数可写为：

其中。令对W和b的偏导为0可得：

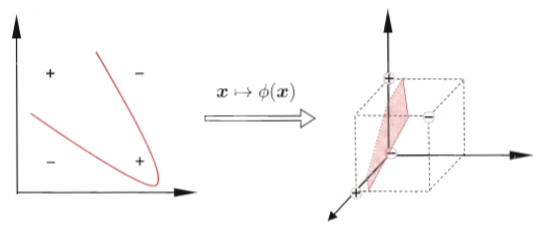
将这两个式子带入中，可将W和b消除，即可得到的对偶问题：

解出后，求出W和b即可得到模型：

从对偶问题

3、核函数

在本章前面的讨论中，我们假设训练样本是线性可分的，即存在一个划分超平面能将训练样本正确分类。然而在现实任务中，原始样本空间内也许并不存在一个能正确划分两类样本的超平面。比如“异或”问题就不是线性可分的（下图左）。



对这样的问题，可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。例如上图（右），若将原始的二维空间映射到一个合适的三维空间，就能找到一个合适的划分超平面。幸运的是，如果原始空间是有限维，即属性数有限，那么一定存在一个高维度特征使样本可分。

令表示将映射后的特征向量，于是，在特征空间中划分超平面所对应的模型可表示为

其中W和b是模型参数，则目标函数有：

其对偶问题是：

求解过程不懂？？？

通过前面的讨论可知，我们希望样本在特征空间内线性可分，因此特征空间的好坏对支持向量机的性能至关重要。需注意的是，在不知道特征映射的形式时，我们并不知道什么样的核函数是合适的，而核函数也仅是隐式地定义了这个特征空间。于是，“核函数选择”成为支持向量机的最大变数。若核函数选择不合适，则意味着将样本映射到了一个不合适的特征空间，很可能导致性能不佳。

常用的核函数

|  |  |  |
| --- | --- | --- |
| 名称 | 表达式 | 参数 |
| 线性核 |  |  |
| 多项式核 |  |  |
| 高斯核 |  |  |
| 拉普拉斯核 |  |  |
| Sigmoid核 |  |  |

啊

4、软间隔与正则化

啊

5、支持向量回归

啊

6、核方法

啊

7、阅读材料

啊