1.间隔与支持向量

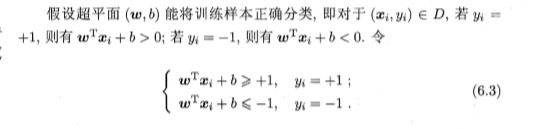
分类学习的基本思想是找到一个超平面这个超平面能够把不同类别的样本分开，我们要找的超平面应该具有良好的泛化性，在使用训练集以外的样本进行测试的时候，应该能够正确的分类，那么定义一个最优超平面，寻找两类样本最中间的超平面，这样对于极端样本例如接近最优平面的样本，其它的平面可能会误分类，那么最优平面则更能容忍这种样本，那么问题集中在然后找到最优超平面了

在样本空间中可将超平面描述为



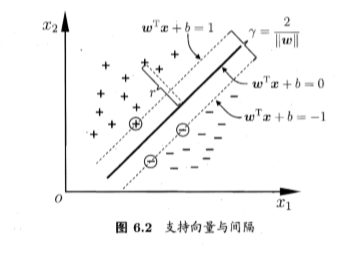
其中ω为超平面法向量，b为位移项，b为平面到原点的距离所以平面可以由两者确定，样本空间中任意点到平面的距离公式如下



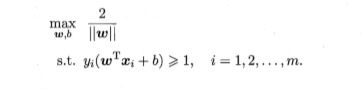


距离超平面最近的几个样本使得上式的等号成立，定义这几个向量为支持向量，两个异类支持向量到超平面的距离为

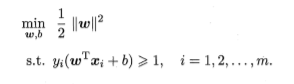
称为间隔



那么目的是要找到最大间隔的超平面

即

也可以写为



这就是基本的支持向量机

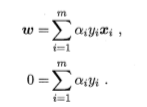
2.对偶问题

如何解上式呢，我们可以发现这是一个凸优化问题，必有唯一极值，那么我们可以用拉格朗日乘子法，对于每个约束给个大于零的拉格朗日乘子，则上式可写为

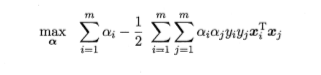


从式子可以看出其实这就是一个条件极值问题的求解过程

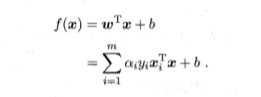
然后分别对ω和b，以及乘子求偏导数并令其为零就可以消去未知数即：



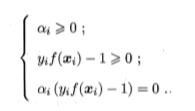
然后代入即可得



这就是一个对偶问题解出乘子即可得到ω和b从而得到模型

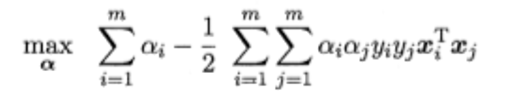


对于上述过程由于约束的作用所以必须满足KKT条件即



上面三个式子分别是三个约束条件，第一个是拉格朗日乘子法添加的约束条件，第二个是分类的约束条件及上一节式子的约束，最后一个为乘子法要满足的条件，这个式子也体现了模型仅与支持向量有关

对于



的求解施个二次规划问题，求解此类问题可用一般算法求解，但是求解问题的规模正比于训练样本数，为避免大开销，人们利用问题本身的特性，提出了很多高效算法，其中SMO就是其中一种

SMO方法的中心思想就是减少计算的变量，每次计算先固定除要求变量以外的所有变量比如求那么，由于存在约束那么可由其它固定变量导出，于是SMO每次选择两个变量固定其它参数

SMO采用一个启发式的算法，使选取两变量样本之间的间隔最大，两个变量之间差别比较大，这样带给函数值的变化更大

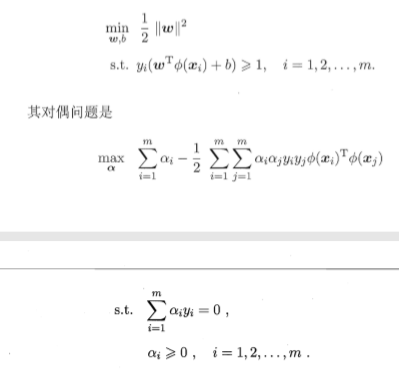
3.核函数

之前我们的讨论均是建立在样本集线性可分的前提之下，但是大部分实际情况是线性不可分的，找不到对应的超平面，对于这种问题我们可以报原始的样本空间映射到一个更高维的样本空间，那么在高维空间中就可以找到一个超平面进行划分，对于有线维度的样本空间，那么一定存在一个高维特征空间线性可分

令为x映射后的特征向量在特征空间中进行划分的模型可以表示为



类似之前的公式可有

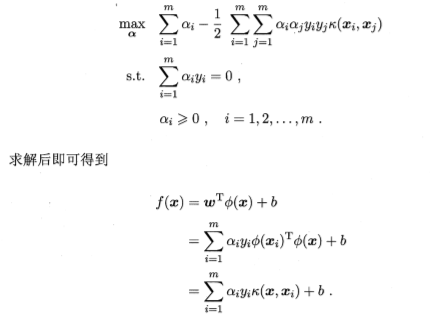


求解内积在维度比较高的情况下比较难，为了避开这个困难，我们可假设一个核函数



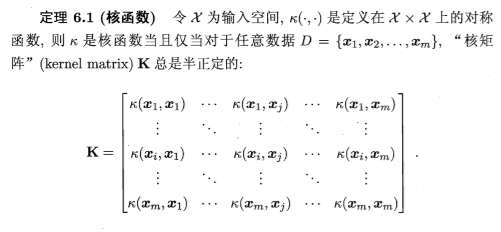
这样根据这个函数来表示内积计算，可以省略很多计算

所以对偶问题可以重写为



很显然如果映射已知可以写出对应的核函数，但是很多情况往往不知道映射，那么核函数是否存在，什么样的函数能做核函数呢?

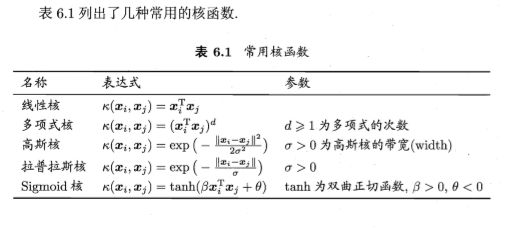
那么这里有一个定理

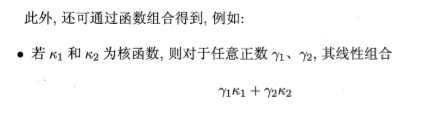


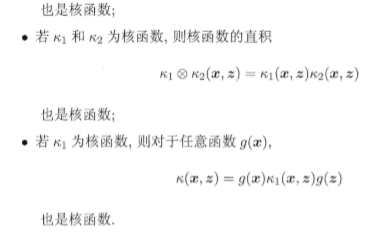
定理表明只要一个对称函数对应的和矩阵半正定，那么这个函数就可以作为核函数

那么核函数的选择就变得非常重要，因为不知映射时我们并不知道什么核函数合适，而且由于核函数只是隐式的定义了特征空间，因此合适的核函数选择是SVM的一个变数，那么选不好会影响性能，这里只有一些经验性的方法

下面是常用核函数



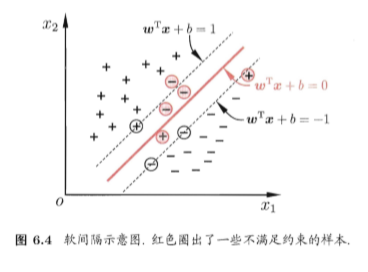




SMO?

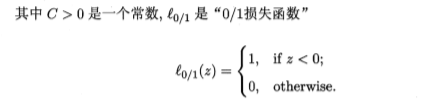
4.软间隔与正则化

那么对于线性不可分的问题，如果我们不能找到一个很好的核函数，使训练集在特征空间中线性可分，或者即使找到了核函数也无法确定结果是不是过拟合造成的，那么解决这个问题的方法就是允许SVM在一些样本上出错，那么这就是软间隔的思想，如下图：



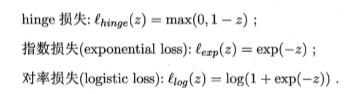
那么之前所说的都是硬间隔，在最大化间隔时不满足约束的样本应该尽可能的少，于是优化目标可写为

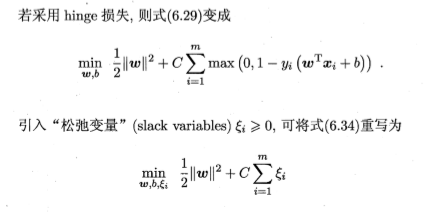


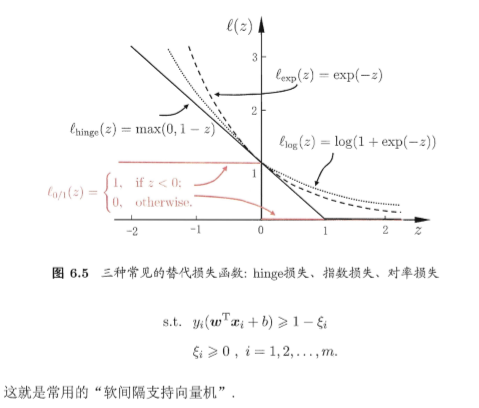


C趋于正无穷时，要求所有的样本均能被正确分类也就退化为了硬间隔，C取有限值时可以允许一些样本不满足约束即

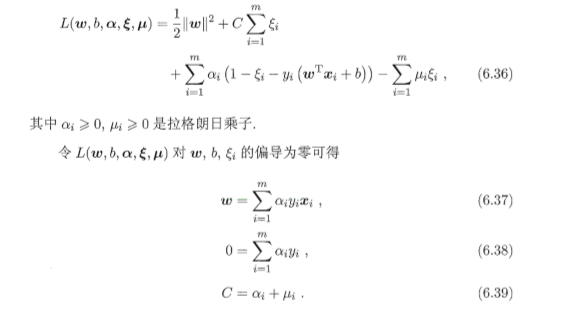
上述的损失函数数学性质不好，非凸非连续，所以人们用一些比较好的函数代替如下：

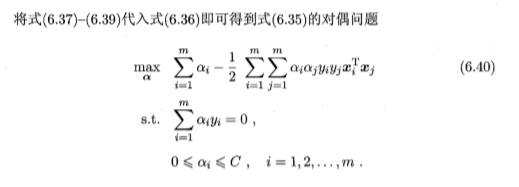






对于每一个样本都有一个松弛变量，表示该样本不满足约束的程度，这也是个二次规划问题，那么依然可以用拉格朗日乘子法，如下式

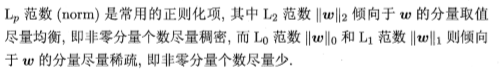




简而言之上述模型的意义就在于允许误差的发生，这样简化了超平面的求解，损失函数的作用就是用来控制损失的大小，那么可将最大化间隔公式改写为



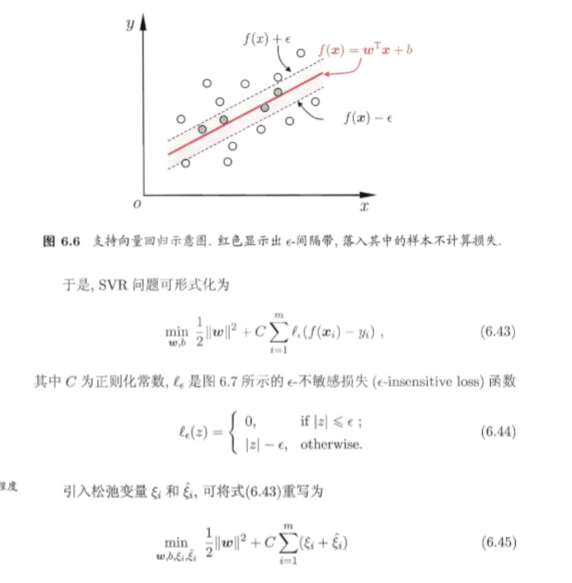
对于不同的损失函数，模型有不同的特性，对于所有的损失函数都有的一个共性就是，上述公式第一项用来描述划分超平面的间隔大小将其称为结构风险，第二项来描述训练集上的误差是经验风险，C对两者进行折中，第一项表述了我们想要什么性质的模型，我们最小化的目标是什么，一般用的是Lp范数不同的范数具有不同特性对于模型影响也不一样，称上述问题为正则化问题，第一项为正则项，C为正则化常数

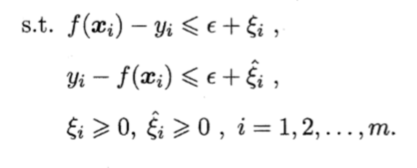


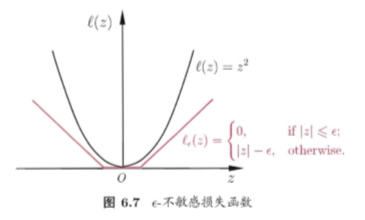
5.支持向量回归

支持向量回归，就是利用之前的理论解决回归问题，那么还是用支持向量的公式只是这里的损失函数换为了不敏感损失函数，这个不敏感损失函数的作用就是可以容忍一定的偏差，就是f（x）与y的差。

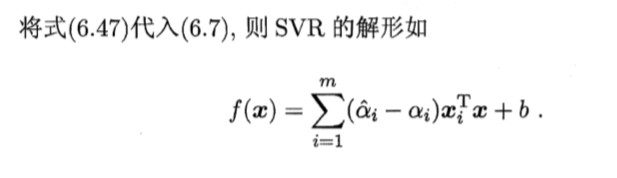
如图所示构造了一个2ε的区域容忍落入区域内但是没在回归曲线上的点

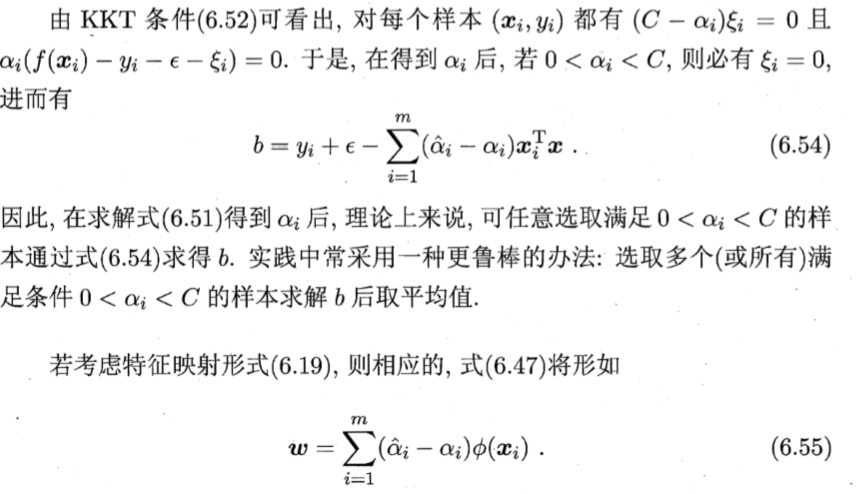


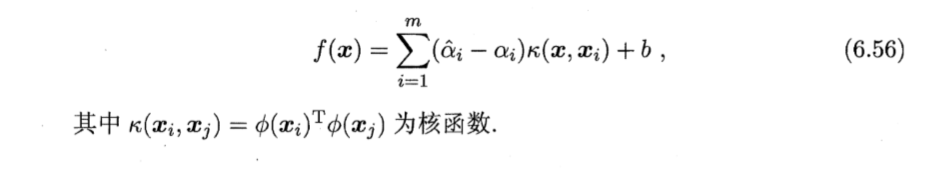




一系列的拉格朗日乘子法求解后最后可得

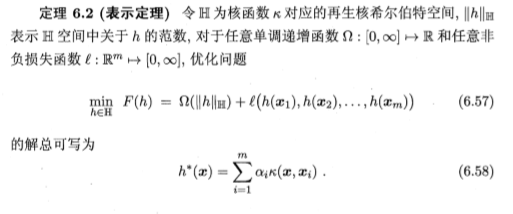






6.核方法

我们可发现无论是SVM还是SVR结果1都能表示为核函数的线性组合，那么这个我们可以用一个定理来阐述



那么我们发现优化问题6.57的最优解都可以表示为核函数的线性组合的形式，那么有人就利用核函数的性质开发出了各种学习方法，统称核方法，例如最常见的就是通过“核化”来将线性学习器扩展为非线性学习器，将低维向量映射到高维来解决在低维无法线性划分的问题