0、概述

SVM模型与logistic回归、感知机的思路都类似：对二类分类问题，都是求一个分类超平面，把样本尽量分隔成两部分，使每一部分样本属于一类。但SVM的特色是，在将样本准确分隔成两类的基础上，以最大化间隔（margin）为目标求最优的超平面。

最基本的SVM是二类分类模型，但它可以扩展到多类分类和回归问题。

由于“最大间隔”的目标，SVM的目标函数的优化问题是一个约束优化问题，其需要转换成对偶问题来求解，这部分内容属于拉格朗日对偶性理论。因此首先介绍拉格朗日对偶性理论。

一、拉格朗日对偶性理论

拉格朗日对偶问题是关于约束优化问题的理论。

大多数机器学习模型的目标函数优化问题属于无约束优化问题，对于无约束优化问题，常用的解析优化方法是令偏导数为0，构成方程组，然后解方程组（常用数值优化方法是梯度下降法、牛顿法等）。

而对于约束优化问题，常用的解析优化方法是拉格朗日乘数法。约束优化问题分为等式约束和等式+不等式约束两种情况。

**1、等式约束优化问题**



该问题的求解方法就是经典的拉格朗日乘数法，流程如下：

（1）构建拉格朗日函数：



其中是引入的变量，称为拉格朗日乘数。

（2）求的驻点处函数值：

求对x和的偏导数，并令它们等于0：



解上面的方程组，得的驻点。

（3）的驻点包含原约束优化问题的极值点，但不一定全是原约束优化问题的极值点。至于如何确定拉格朗日函数的驻点是否是原约束优化问题的极值点，在实际问题中往往可根据问题本身的性质来判定。

虽然拉格朗日乘数法引入了新的变量，但是将约束优化问题转化成了无约束的优化问题来求解，还是值得的。

**2、等式+不等式约束优化问题**



将该问题称为“原始问题”（primal problem）。

（1）广义拉格朗日函数：

针对对原始问题，可以构造广义拉格朗日函数：



其中。

（2）原始问题的对偶问题（dual problem）：

下面基于广义拉格朗日函数，构造原始问题的对偶问题，并说明原始问题和对偶问题的等价性。

考虑广义拉格朗日函数对的极大化，它是x的函数，记为：



注意，这里表示一个函数，而不是模型参数。

可以看出，对于不满足原始问题约束条件的x，有：



这是因为，对不满足约束条件的x，存在i使得或存在j使得，这时总可以给出系数，使。

而对于满足所有约束条件的x，有：



在的基础上，再对x极小化，即：



该问题称为广义拉格朗日函数的极小极大问题。极小极大问题的最优解和相应的最优值与原始问题相同。

那么这就是原始问题的对偶问题了吗？非也，还要将这个问题中极小和极大调换一下，变成极大极小问题，并写成严格的约束优化问题，才得到对偶问题，即：



注意，由于经过调换，导致对偶问题与极小极大问题的最优解不同，从而与原始问题的最优解不同。但可以保证，对偶问题的最优值不大于原始问题的最优值。并且当和是凸函数、是仿射函数，且不等式约束是严格可行的，即存在x，对所有j，使得成立，则对偶问题与原始问题同解。

注意，对偶问题并非将原始问题转化成无约束的优化问题，实际上，对偶问题仍然是约束优化问题。

（3）KKT条件：

KKT条件是拉格朗日对偶问题的最优解需满足的必要条件。

假设对偶问题的最优解为，则它需要满足以下几个条件：

（a）驻点条件（stationary）：与等式约束的拉格朗日乘数法类似，需要保证偏导数为0：



（b）原始可行性条件（primal feasibility）：就是原始问题中的约束条件：



（c）对偶可行性条件（dual feasibility）：



（d）互补松弛性条件（complementary slackness）：对于最优解，有两种情况：一是最优解在约束条件的可行域的内部，此时该约束条件不起作用，则；二是最优解在约束条件的可行域的边界上，此时有，因此总有：



KKT条件常用来求解具体的约束优化问题，此时不必考虑对偶问题的具体形式，只需根据KKT条件求解可能的最优解，然后在具体问题中验证这些可能的最优解哪个是真正的最优解。

二、线性SVM目标函数

首先考虑对二类分类问题，线性可分数据情况下，SVM的目标函数和优化问题。

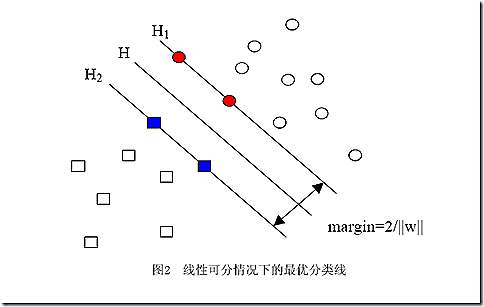


图1 线性可分SVM示意图

针对线性可分的二类分类问题，存在无穷多的分类超平面，它们都能将样本以100%的准确率分成两类。但是在这些超平面中，哪个超平面最好呢？

SVM认为，间隔（margin）最大的超平面是最好的超平面。所谓间隔，指的是在分类超平面两侧的样本中，距离超平面最近的样本点到超平面的距离。如图1所示，可以直观的看出，超平面的间隔越大，就相当于两类样本中间的“空白”越大，用这样的超平面去分类，“确信度”越大（或者说越稳定）。

下面将“间隔最大”的思想形式化，形成SVM的目标函数。

给定样本集，其中，（注意，这里标签y取值为，而不是，是为了推导的目标函数具有简洁的形式），样本点到超平面的距离为：



假设所有样本中，到超平面的距离最小的样本点到超平面的距离为，即：



则有



据此可以得到SVM的目标函数及优化问题：



可以看到，由于必然不小于0，因此约束条件能保证每个点都被正确分类。

该优化问题形式上有点别扭，因为最大化的参数是w和b，但最大化的目标并不能简单地写成w和b的函数。因此，我们需要将优化问题转换成不那么别扭的形式。

对于任意一组参数(w,b)对应的超平面，设其支持向量为，则，注意到同比例地缩放w和b，不会改变超平面，也不会改变样本点到超平面的距离，因此我们通过缩放，使，则支持向量到超平面的距离。

注意，对于不同的超平面（即不同的参数(w,b)），需要使用不同的缩放比例。

则优化问题变为：



可以看出，最大化间隔本质上是最小化w，这也是不难理解的：对于超平面来说，w越小（b不变的情况下），一个点到它的距离越大。

换个角度来看待这个问题转换的过程，我们寻找最优的超平面，表面上看需要优化w和b两个参数，但是由于同比例缩放w和b不会改变超平面，所以本质上需要优化的参数只有一个。

再将优化问题转换成最小化的形式，并将约束条件转化为小于等于0的形式：



该问题为不等式约束优化问题，不易直接求解，下面我们将它作为原始问题，并转换为对偶问题。

首先，构建相应的广义拉格朗日函数：



根据拉格朗日对偶性理论，对偶问题为：



可以看出，该对偶问题相当于先固定，求L关于w和b的极小值，然后再求L关于的极大值。根据L的具体形式，我们可以进一步简化对偶问题：

（1）求极小：

即令L对w和b的偏导为0，有：



将结果带入广义拉格朗日函数，得到消去w和b，只包含的结果：



以及条件：。

（2）求极大：

上一步求极小的结果，已经消去了w和b，得到的结果只包含，这一步求L关于的极大，（注意，还有一个条件），即约束优化问题：



或写成极小化的形式：



这样，我们得出了对偶问题的最终形式，它是一个约束优化问题，求解该约束优化问题，得到最优解，接着就可求出w和b的最优解和：





其中，是的一个任意选取的正分量。

三、软间隔SVM目标函数

对于线性不可分但是近似线性可分的二类分类问题，不能直接使用线性SVM，需要将“间隔最大化”的目标改为“软间隔最大化”。

线性不可分，意味着存在一些样本点，不满足优化问题中的约束条件，针对此，对每个样本点引入一个松弛变量，使约束条件变为：



同时，在目标函数中对松弛变量进行“惩罚”，得到新的目标函数：



该目标函数综合考虑了两个目标：一是间隔尽量大，二是误分类尽量少。

则优化问题变为：



与线性可分SVM类似，也可将该优化问题转化为对偶问题。

首先构造其拉格朗日函数：



接着给出其对偶问题的最终形式：



可以看出，软间隔最大化SVM的优化问题尽管引入了松弛变量，但最终的优化问题中并没有的身影。软间隔最大化SVM的优化问题与线性可分SVM的优化问题的对偶问题的最终形式十分相似（只有最后一个约束条件不一样）。显然，软间隔最大化的SVM包含硬间隔最大化的SVM。

这个问题属于凸二次规划问题，可以证明w的解是唯一的，但b的解不唯一，b的解存在于一个区间。

另外，软间隔最大化的优化问题等价于“合页损失（Hinge Loss）”目标函数的优化问题：



针对合页损失函数的优化问题，我们可以采用梯度下降法予以求解；针对对偶问题的最终形式，则有高效的SMO优化算法。

四、核化SVM目标函数

前面介绍了线性可分情况下的SVM（以硬间隔最大化为目标）和近似线性可分情况下的SVM（以软间隔最大化为目标）的目标函数和优化问题。后者包含前者，称为线性SVM。

而对于非线性可分的情况，又如何通过扩展SVM而求解呢？这就要用到核技巧（Kernel Trick）。

对于二类分类问题，非线性可分问题指的是存在一个超曲面（而不是超平面）可以将样本集中的正例和负例分开。

基于SVM模型解决非线性可分的二类分类问题，我们并不是求解一个超曲面，而是考虑先将原数据映射到更高维的特征空间，使高维特征空间中数据变得线性可分，故可以在高维空间中应用线性SVM模型。

设对任一个样本x，到高维特征空间中的映射为：



然后对高维特征空间中的数据应用线性SVM模型，其优化问题的推导过程与线性SVM类似，最终得到其目标函数为：



这就是非线性SVM的目标函数。

可以看出，它跟对数据直接应用SVM模型的区别仅在于第一项中求两个样本向量的内积，变成了求两个样本向量映射到高维特征空间后的向量的内积。

实际中，我们很难确定映射的形式，因此，对该目标函数的优化需要绕过计算这一步。

对于映射，我们引出它的“核函数”：



核函数输入任意两个样本的特征向量，输出它们映射到高维空间的向量的内积。

此时，非线性SVM的目标函数变为：



则最终优化问题为：



实际中，核函数比映射更容易给出，如果给定映射的核函数，就可以实现非线性SVM。常用的核函数有：多项式核、高斯核、字符串核。

为什么越高维的空间，越可能线性可分？因为稀疏吗？

五、SVM的优化方法

以上给出了线性SVM和非线性SVM的优化问题，并推导出它们的对偶问题的最终形式——是一个关于的优化问题。它属于凸二次规划的问题，有许多算法可以求解该问题，但是其中效率最高的算法是SMO算法（序列最小最优化算法）。

SMO算法利用了KKT条件。

对于优化问题：



其优化参数为，每个分量对应于一个样本。

SMO算法的基本思想是：迭代，每次选取的两个分量和，固定的其他分量，构成一个两分量的二次规划问题，对该问题易解析地求出最优解。这样使目标函数逐步缩小。

如何在每次迭代中选取分量和呢？SMO使用启发式的策略选取它们。选取违反KKT条件最严重的的分量作为第一个分量；然后选取更新前后变化最大的分量作为第二个分量。

SMO的具体算法流程不再详述。

六、SVM回归

SVM分类模型（SVC）是以最大化间隔的目标实现线性分类（即求最优的分类超平面）；SVM的回归模型（SVR）则是以最小化“不敏感损失”为目标实现的线性回归。

“不敏感损失”的意思是，常规线性回归模型通常使用MSE损失函数，它直接计算样本的预测值与真实值之间的误差，而不敏感损失是给定一个“容忍度”，当样本预测值和真实值的误差绝对值大于时，才开始计算损失。

SVR的优化问题为：



其中为“不敏感损失”函数，其形式为：



SVR的对偶问题的推导及优化方法在此不作详细介绍。