

第18课：SVM——线性可分 SVM 原理

****注意****：从今天开始，我们的学习之旅进入了一个新的阶段，也就是我们第四个主题：有监督学习 Ⅱ。随着课程难度的加深，相信很多同学会感觉学习起来压力很大。在此我也想告诉大家：虽然我们制定的是 21 天的学习计划，但是想真正吃透这个课程的内容，仔细学习一个学期也不为过，我们在训练营里主要还是为了带领大家先过一遍知识点，对每个概念有一个大致的认知，接下来还是需要同学们用更多的时间去深入学习。所以跟不上进度并不代表你就是学不会学不懂，持之以恒地努力，终将收获成功。下面开始本章的内容：

之前我们讲过的几个模型：线性回归、朴素贝叶斯、逻辑回归和决策树，其背后数学原理的难度属于初级。而接下来要讲的 SVM 和 SVR 则上升到了中级。

在下面的 7 篇课程中，我们将不会再看到可爱、搞笑的图片和轻松的调剂，而要一步步剥离清楚 SVM/SVR 的出发点与运作过程。

### 线性可分和超平面

#### 二分类问题

在机器学习的应用中，至少现阶段，分类是一个非常常见的需求。特别是****二分类****，它是一切分类的基础。而且，很多情况下，多分类问题可以转化为二分类问题来解决。

所谓****二分类****问题就是：给定的各个样本数据分别属于两个类之一，而目标是确定新数据点将归属到哪个类中。

#### 特征的向量空间模型

一个个具体的样本，在被机器学习算法处理时，由其****特征****来表示。换言之，每个现实世界的事物，在用来进行机器学习训练或预测时，需要转化为一个****特征向量****。

假设样本的特征向量为 n 维，那么我们说这些样本的特征向量处在 n 维的特征空间中。

注意：我们在前几课中经常会用某个欧几里得空间（一般是二维或三维）来描绘样本的特征向量和最终标签之间的关系。例如在一个二维的坐标系中，用 Y 轴对应的值表示标签，用 X 轴对应的值表示1维的特征向量。

此处我们说的 n 维，则仅仅限于特征向量自身的维度。如果对应到最早线性回归中工资和工作经验关系的例子里，该例子的特征空间维度为1，而不是2。

一般来说，特征空间可以是欧氏空间，也可以是希尔伯特空间，不过为了便于理解，我们在以后的所有例子中都使用欧氏空间。

直观上，当我们把一个 n 维向量表达在一个 n 维欧氏空间中的时候，能够“看到”的一个个向量对应为该空间中的一个个点。

这样来想象一下：我们把若干样本的特征向量放到特征空间里去，就好像在这个 n 维空间中撒了一把“豆”。

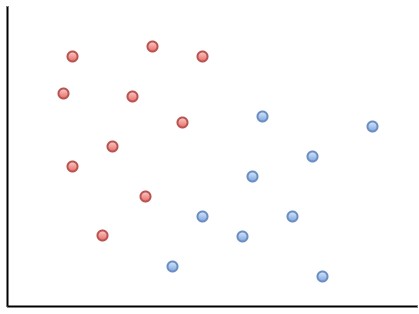
当 n=1 时，这些“豆”是一条直线上的若干点；当 n=2 时，这些“豆”是一个平面上的若干点；当 n=3 时，这些“豆”是一个几何体里面的若干点……

#### 线性可分

现在再想想我们****选取特征的目的****：我们将一个事物的某些属性数字化，再映射为特征空间中的点，其目的当然是为了对其进行计算。

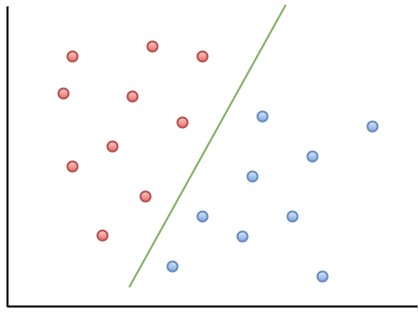
但是如果这些点在特征空间中就能够对应它们预期的二分类分为两个部分，那不是最理想的情况吗？

比如，我们的特征向量是2维的，下面图中的红蓝两色点都是样本的特征向量，不过红色点对应的是正类，而蓝色点对应的是负类：



我们发现，哎？在当前的特征空间（上面二维坐标系）中，正负两类样本各自和自己的“同伙”“站在”一个阵营里，而这两个“阵营”之间，则已经有了一条隐隐的楚河汉界。

我们可以把这条楚河汉界（分割线）画出来，见下图中绿色线：



这样，两类样本完美地被绿线分隔开。此时，我们说这两类样本在其特征空间里****线性可分****。

上面的表述很不严谨，我们来看看****线性可分****严格的数学定义：

D0D0 和 D1D1 是 nn 维欧氏空间中的两个点集（点的集合）。如果存在 nn 维向量 ww 和实数 bb，使得所有属于 D0D0 的点 xixi 都有 wxi+b>0wxi+b>0，而对于所有属于 D1D1 的点 xjxj 则有 wxj+b<0wxj+b<0。则我们称 D0D0 和 D1D1 线性可分。

#### 超平面

上面提到的，将 D0D0 和 D1D1 完全正确地划分开的 wx+b=0wx+b=0 ，就是超平面（Hyperplane）。

****超平面****：n 维欧氏空间中维度等于 n-1 的线性子空间。

1维欧氏空间（直线）中的超平面为0维（点），2维欧氏空间中的超平面为1维（直线）；3维欧氏空间中的超平面为2维（平面）；以此类推……

在数学意义上，将线性可分的样本用超平面分隔开的分类模型，叫做****线性分类模型****，或****线性分类器****。

我们可以想象，在一个样本特征向量线性可分的特征空间里，可能有许多超平面可以把两类样本分开。在这种情况下，我们当然要找最佳超平面。

什么样的****超平面是最佳****的呢，一个合理的****策略****是：以最大间隔把两类样本分开的超平面，是最佳超平面！

因此，我们将这样的超平面作为最佳超平面：

1. 两类样本分别分隔在该超平面的两侧；
2. 两侧距离超平面最近的样本点到超平面的距离被最大化了。

这样的超平面又叫做****最大间隔超平面****。

### 线性可分支持向量机

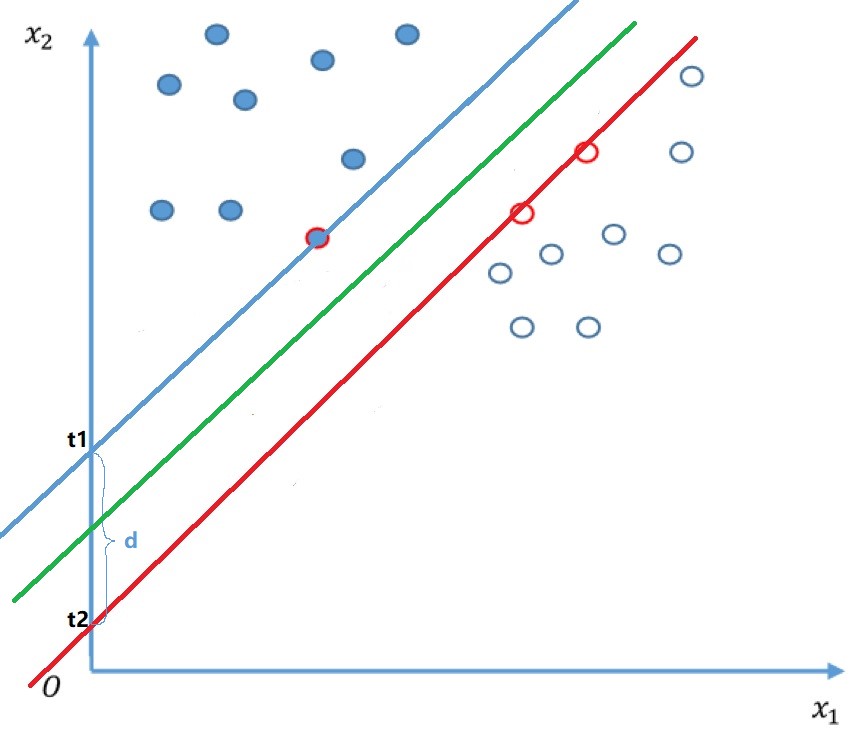
#### 找到最大间隔超平面

****线性可分支持向量机****就是：以找出线性可分的样本在特征空间中的最大间隔超平面为学习目的的分类模型。

怎么能找到最大间隔超平面呢？

我们可以先找到两个平行的，能够分离正负例的辅助超平面，然后将它们分别推向正负例两侧，使得它们之间的距离尽可能大，一直到有至少一个正样本或者负样本通过对应的辅助超平面为止——推到无法再推，再推就“过界”为止。

下图是二维坐标系里，两个辅助超平面（蓝、红两条直线）的例子：



这两个超平面互相平行，它们范围内的区域称为“****间隔****”，最大间隔超平面位于这两个辅助超平面正中的位置与它们平行的超平面——图中绿线为最大间隔超平面。

用我们初中时就熟悉的表示方法来表示这两条直线，则：

蓝线：x2=sx1+t1=>sx1–x2+t1=0(1)x2=sx1+t1=>sx1–x2+t1=0(1)

红线：x2=sx1+t2=>sx1–x2+t2=0(2)x2=sx1+t2=>sx1–x2+t2=0(2)

设 d=t1–t2d=t1–t2，位于（1）和（2）正中的绿线（所求最大分割超平面）则表示为：

x2=sx1+t1–d2=>sx1–x2+t1−d2=0x2=sx1+t1–d2=>sx1–x2+t1−d2=0 (3)

将 t2=t1−dt2=t1−d 带入（2），然后将（1）两侧同时减去 d2d2，（2）两侧同时加上 d2d2，则有:

sx1–x2+(t1−d2)=−d2(1′)sx1–x2+(t1−d2)=−d2(1′) sx1–x2+(t1−d2)=d2(2′)sx1–x2+(t1−d2)=d2(2′)

将（1'），（2'）和（3）两侧同时除以 −d2−d2，则有：

−2sdx1+2dx2+(1−2t1d)=1(1′′)−2sdx1+2dx2+(1−2t1d)=1(1″)

−2sdx1+2dx2+(1−2t1d)=−1(2′′)−2sdx1+2dx2+(1−2t1d)=−1(2″)

−2sdx1+2dx2+(1−2t1d)=0(3′′)−2sdx1+2dx2+(1−2t1d)=0(3″)

设 x=(x1,x2)，w=(−2sd,2d)，b=1−2t1dx=(x1,x2)，w=(−2sd,2d)，b=1−2t1d，则这三个超平面可以分别表示为：

蓝线： wx+b=1wx+b=1

红线：wx+b=−1wx+b=−1

绿线（最大分割超平面）：wx+b=0wx+b=0

#### 定义

上一节我们从直观角度介绍了找到最大间隔超平面的方法。

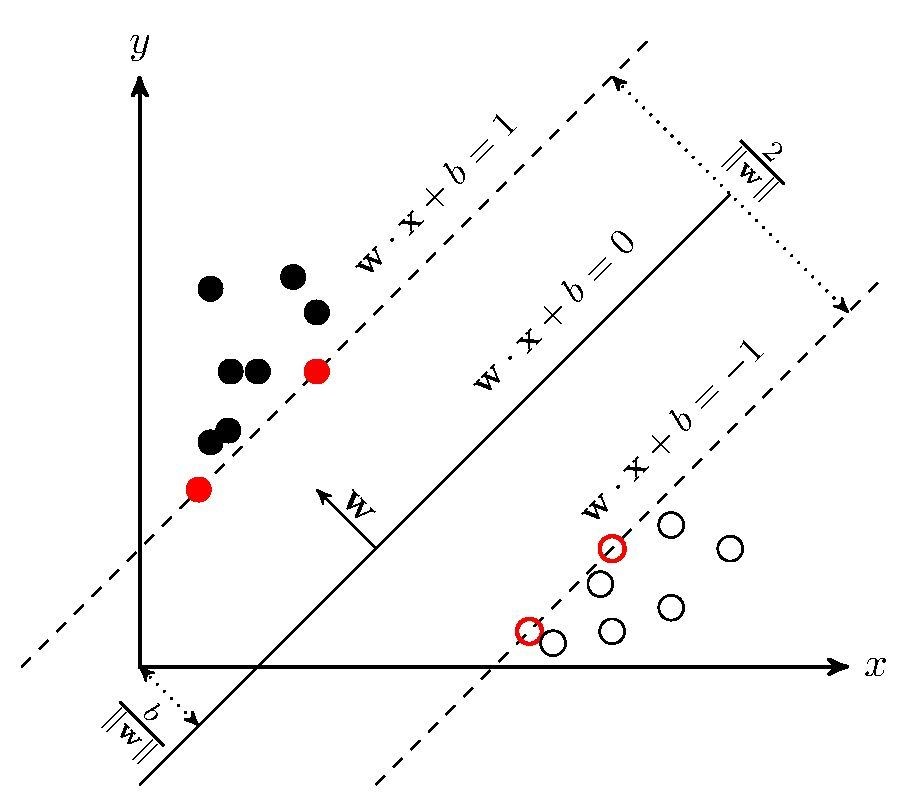
现在我们先来熟悉一组符号：用来训练线性可分支持向量机（Support Vector Machine，SVM）的样本记作：

T={(x1,y1),(x2,y2),…,(xm,ym)}T={(x1,y1),(x2,y2),…,(xm,ym)}

其中，xixi 为 nn 维实向量，而 yiyi 的取值要么是1，要么是-1， i=1,2,…,mi=1,2,…,m，yiyi为 xixi 的标签，当 yi=1yi=1 时， xixi 为正例；当 yi=−1yi=−1 时， xixi 为负例。

我们要找到将上面 mm 个样本完整正确地分隔为正负两类的最大间隔超平面 w⋅x+b=0w·x+b=0。

这个超平面由其法向量 ww 和截距 bb 确定，可用 （w,b）（w,b） 表示。



这 mm 个样本在特征空间是线性可分的，因此我们可以找到两个将正负两类样本分离到距离尽可能大的超平面，它们分别是：

w⋅x+b=1w·x+b=1 （超平面1）

和

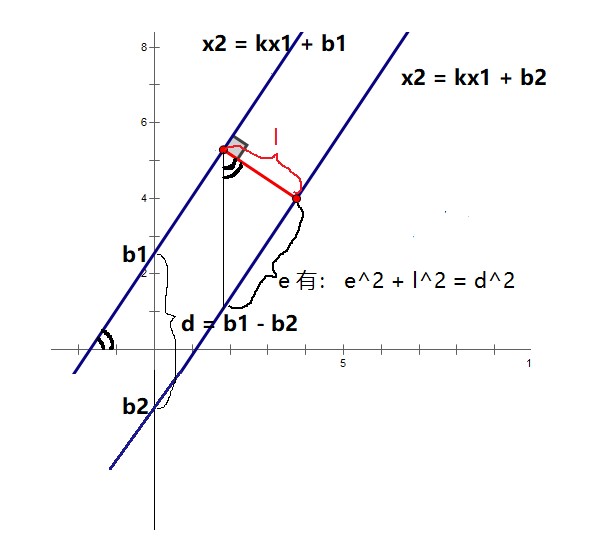
w⋅x+b=−1w·x+b=−1 （超平面2）

通过几何不难得到这两个超平面之间的距离是 2||w||2||w||，因此要使两平面间的距离最大，我们需要最小化 ||w||||w||。

此处如果对 2||w||2||w|| 有疑问的话，我们再来详细分析一下，还是以二维坐标系为例。

设 w=(k,−1)，b1=b+1，b2=b−1w=(k,−1)，b1=b+1，b2=b−1，则超平面1又可以表示为：x2=kx1+b1x2=kx1+b1；超平面2则为：x2=kx1+b2x2=kx1+b2。

设超平面1和超平面2之间的距离为 ll， 参见下图



因为有：el=kel=k

则有：

√(d2–l2)l=k(d2–l2)l=k

(d2–l2)l2=k2(d2–l2)l2=k2

(b1−b2)2l2–1=k2(b1−b2)2l2–1=k2

因为 b1−b2=2b1−b2=2，所以：

4l2=k2+14l2=k2+1

l=2√(k2+1)l=2(k2+1)

又因为 w=(k,−1)=>||w||=√(k2+(−1)2)=√(k2+1)w=(k,−1)=>||w||=(k2+(−1)2)=(k2+1)

所以：

l=2||w||l=2||w||

#### 学习目标

依据定义，我们知道样本数据需满足下列两个条件之一：

若 yi=1yi=1，则 wxi+b≥1wxi+b≥1

若 yi=−1yi=−1，则 wxi+b≤−1wxi+b≤−1

也就是说没有样本点处在超平面1和超平面2中间，且所有点都在正确的那一侧。

将上面两个式子合并一下就是 yi(wxi+b)≥1yi(wxi+b)≥1。

也就是说，我们要最小化 ||w||||w|| 是有约束条件的，条件就是：

yi(wxi+b)≥1，i=1，2，…,myi(wxi+b)≥1，i=1，2，…,m

因此，求最大分割超平面问题其实是一个****约束条件下的最优化问题****，我们要的是：

min(||w||)min(||w||)

s.t.yi(wxi+b)≥1，i=1，2，…,ms.t.yi(wxi+b)≥1，i=1，2，…,m

为了后面好算，我们用 (||w||2)2(||w||2)2 来代替 ||w||||w||，并将约束条件变一下形式：

min(||w||22)min(||w||22)

s.t.1−yi(wxi+b)≤0，i=1，2，…,ms.t.1−yi(wxi+b)≤0，i=1，2，…,m

这就是支持向量机的****学习目标****，其中 min(||w||22)min(||w||22) 是目标函数，我们要对它进行最优化。

对这样一个最优化问题，需要通过优化算法来求解。关于优化算法，我们之前学习过梯度下降法，简单直接，是不是可以应用到这里呢？可惜不行。因为，梯度下降是优化没有约束条件问题的算法。

此处我们遇到的，是一个****有约束条件的最优化问题****。

那么该如何求解呢？在此我们要学习一个新的算法：****拉格朗日乘子法****。

从拉格朗日乘子法开始，本课程的有监督学习模型部分对数学工具的要求上了一个台阶。整体难度比之前提升了一个量级。

因此，我们需要好好讲解一下拉格朗日乘子法。请看下一课！

[邀请好友一起学，获得 25% 返现奖励](https://gitbook.cn/m/mazi/columns/5bc6ac7442d7d32f50f19a98/topics/5bf28278fd72950cafdccd99)

[IMG_262](https://gitbook.cn/m/mazi/comp/column?columnId=5bc6ac7442d7d32f50f19a98%26tag=2#catalog)