第21课：SVM——线性 SVM，间隔由硬到软

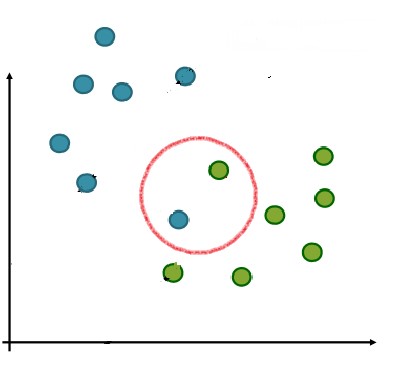
### 从线性可分 SVM 到线性 SVM

#### 从现实情况引出线性 SVM

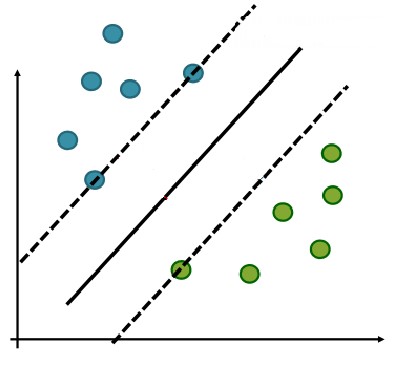
前面连续三篇讲得都是****线性可分 SVM****，这种 SVM 学习的训练数据本身就是线性可分的——可以很清晰地在特征向量空间里分成正集和负集。

线性可分 SVM 正负样本之间的间隔叫做“硬间隔”，也就是说在这个“隔离带”里面，肯定不会出现任何训练样本。

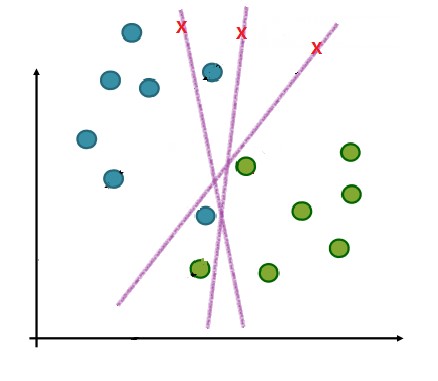
我们不难想到，这种情况在现实生活中其实是很少见的。更多的时候，可能是像下面这个样子：



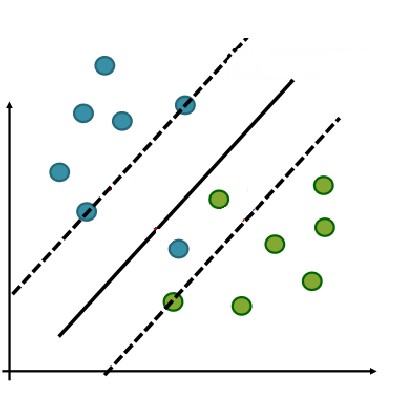
如果没有红圈里那两个点，本来可以很好的分割：



可是，偏偏多了那两个点！都找不到分隔超平面了！像下图这样，分来分去，怎么都分不开：



如果我们不那么“轴”，不是完全禁止两个辅助超平面之间有任何样本点。而是允许个别样本出现在“隔离带”里面，那样是不是会变得好分得多？比如像下面这样：



这样看起来也很合理啊。而且，一般情况下，怎么能保证样本就一定能够被分隔得清清楚楚呢？从直觉上我们也觉得，允许一部分样本存在于“隔离带”内更合理。

正是基于这种想法，相对于之前讲的线性可分 SVM 的****硬间隔（Hard Margin）****，人们提出了****软间隔（Soft Margin）**** 的概念。

相应的，对应于软间隔的 SVM，也就叫做****线性 SVM****。

下面我们对照来看一看它们：

#### 线性可分 SVM

****线性可分 SVM**** 成立的****前提****是训练样本在向量空间中****线性可分****，即存在一个超平面能够将不同类的样本完全彻底，且无一错漏地分开。

用数学式子表达，全部训练样本满足如下约束条件：

wxi+b⩾1,yi=1wxi+b⩾1,yi=1 wxi+b⩽1,yi=−1wxi+b⩽1,yi=−1

这时，wxi+b=1wxi+b=1 和 wxi+b=−1wxi+b=−1 这两个超平面之间的间隔叫做硬间隔。位于它们两个正中的 wxi+b=0wxi+b=0 是****最大分割超平面****。

#### 线性 SVM

##### ****硬间隔到软间隔****

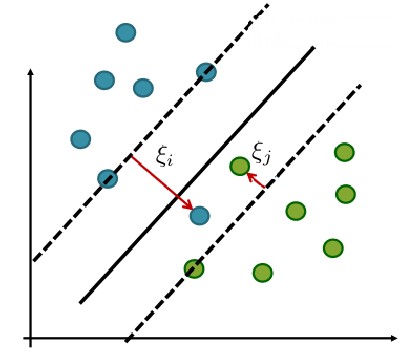
由于样本线性可分的情况在现实当中出现很少，为了更有效地应对实际问题，我们不再要求所有不同类的样本全部线性可分，也就是不再要求硬间隔存在。

取而代之的是将不同类样本之间的硬间隔变成****软间隔****，即****允许部分样本不满足约束条件：****yi(wxi+b)⩾1yi(wxi+b)⩾1。

当然，我们还是希望不满足硬间隔条件的样本尽量少，还能够是一个“软”间隔，而非间隔根本不存在。

为了度量这个间隔“软”到何种程度，我们针对每一个样本 (xi,yi)(xi,yi)，引入一个松弛变量 ξiξi，令 ξi⩾0ξi⩾0，且 yi(wxi+b)⩾1−ξiyi(wxi+b)⩾1−ξi。

对应到图形上是这样：



这样看起来，确实比硬间隔合理多了。

##### ****优化目标****

于是，我们的优化目标就从原来的：

minw,b||w||22minw,b||w||22

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

变成了：

minw,b,ξ12||w||2+C∑mi=1ξiminw,b,ξ12||w||2+C∑i=1mξi

s.t.yi(wxi+b)⩾1−ξi,i=1,2,...,m;ξi⩾0,i=1,2,...,ms.t.yi(wxi+b)⩾1−ξi,i=1,2,...,m;ξi⩾0,i=1,2,...,m

其中 CC 是一个大于0的常数，若 CC 为无穷大，则 ξiξi 必然为无穷小，否则将无法最小化主问题。如此一来，线性 SVM 就又变成了线性可分 SVM。

当 CC 为有限值的时候，才能允许部分样本不遵守约束条件 1–yi(wxi+b)⩽01–yi(wxi+b)⩽0。

这就是****线性 SVM 的主问题****！

### 对偶法最优化线性 SVM 主问题

#### 算法思路

上面我们得出了线性 SVM 的主问题。

现在来回顾一下上节课我们讲解的，用对偶法求解线性可分 SVM 的主问题的思路——当时一共分了7步，不过这7步再抽象一下，大致可以分为4个阶段：

****Stage-1****：根据主问题构建拉格朗日函数，由拉格朗日函数的对偶性，将主问题转化为极大极小化拉格朗日函数的对偶问题。

****Stage-2****：分步求解极大极小问题。

在每次求解极值的过程中都是先对对应的函数求梯度，再令梯度为0。以此来推导出主问题参数和拉格朗日乘子之间的关系。

再将用拉格朗日乘子表达的主问题参数带回到拉格朗日函数中，最终一步步将整个对偶问题推导为拉格朗日乘子和样本 (xi,yi)(xi,yi) 之间的关系。

****Stage-3****：通过最小化拉格朗日乘子与样本量组成的函数（也就是 Stage-2 的结果），求出拉格朗日乘子的值。

这里，可以用 SMO 算法进行求解。

****Stage-4****：将 Stage-3 求出的拉格朗日乘子的值带回到 Stage-2 中确定的乘子与主问题参数关系的等式中，求解主问题参数。

再根据主问题参数构造最终的分隔超平面和决策函数。

#### 主问题求解

现在我们就按这个思路来对线性 SVM 主问题进行求解。

首先，将主问题写成我们熟悉的约束条件小于等于0的形式，如下：

minw,b,ξ12||w||2+C∑mi=1ξiminw,b,ξ12||w||2+C∑i=1mξi

s.t.1−ξi−yi(wxi+b)⩽0,i=1,2,...,m;−ξi⩽0,i=1,2,...,ms.t.1−ξi−yi(wxi+b)⩽0,i=1,2,...,m;−ξi⩽0,i=1,2,...,m

然后开始逐步求解：

##### ****1. 构建拉格朗日函数如下：****

L(w,b,ξ,α,μ)=12||w||2+C∑mi=1ξi+∑mi=1αi[1−ξi−yi(wxi+b)]+∑mi=1(−μiξi)L(w,b,ξ,α,μ)=12||w||2+C∑i=1mξi+∑i=1mαi[1−ξi−yi(wxi+b)]+∑i=1m(−μiξi)

αi⩾0,μi⩾0αi⩾0,μi⩾0

其中 αiαi 和 μiμi 是拉格朗日乘子，而 w，bw，b 和 ξiξi 是****主问题参数****。

根据主问题的对偶性，主问题的****对偶问题****是：

maxα,μminw,b,ξL(w,b,ξ,α,μ)maxα,μminw,b,ξL(w,b,ξ,α,μ)

##### ****2. 极大极小化拉格朗日函数****

（1）极小化

首先 对 w，bw，b 和 ξξ 极小化 L(w,b,ξ,α,μ)L(w,b,ξ,α,μ)——分别对 w，b和ξiw，b和ξi 求偏导，然后令导数为0，得出如下关系：

w=∑mi=1αiyixiw=∑i=1mαiyixi

0=∑mi=1αiyi0=∑i=1mαiyi

C=αi+μiC=αi+μi

将这些关系带入线性 SVM 主问题的拉格朗日函数，得到：

minw,b,ξL(w,b,ξ,α,μ)=∑mi=1αi−12∑mi=1∑mj=1αiαjyiyj(xi⋅xj)minw,b,ξL(w,b,ξ,α,μ)=∑i=1mαi−12∑i=1m∑j=1mαiαjyiyj(xi⋅xj)

（2）极大化

然后，就要对 αα 和 μμ 进行极大化。

因为上面极小化的结果中只有 αα 而没有 μμ，所以现在只需要极大化 αα 就好：

maxα,μminw,b,ξL(w,b,ξ,α,μ)=maxα(∑mi=1αi−12∑mi=1∑mj=1αiαjyiyj(xi⋅xj))maxα,μminw,b,ξL(w,b,ξ,α,μ)=maxα(∑i=1mαi−12∑i=1m∑j=1mαiαjyiyj(xi⋅xj))

s.t.∑mi=1αiyi=0;C−αi−μi=0;αi⩾0;μi⩾0;i=1,2,...,ms.t.∑i=1mαiyi=0;C−αi−μi=0;αi⩾0;μi⩾0;i=1,2,...,m

##### ****3. SMO 算法求解对偶问题****

我们将上面极大化目标约束条件中的 μμ 用 αα 替换掉，并将极大化目标求负转为极小化问题，得到：

maxα(∑mi=1αi−12∑mi=1∑mj=1αiαjyiyj(xi⋅xj))=min(12∑mi=1∑mj=1αiαjyiyj(xi⋅xj)−∑mi=1αi)maxα(∑i=1mαi−12∑i=1m∑j=1mαiαjyiyj(xi⋅xj))=min(12∑i=1m∑j=1mαiαjyiyj(xi⋅xj)−∑i=1mαi)

s.t.∑mi=1αiyi=0;0⩽αi⩽C;i=1,2,...,ms.t.∑i=1mαiyi=0;0⩽αi⩽C;i=1,2,...,m

我们对照一下上一篇线性可分 SVM 最优化过程中步骤3的结果，不难发现，两者的极小化目标是一样的，所不同的就是约束条件而已。

所以，在上一篇我们用到的 SMO 算法，同样可以用于此处。运用 SMO 求解出拉格朗日乘子 α1,α2,…,αmα1,α2,…,αm。

##### ****4. 根据拉格朗日乘子与主问题参数的关系求解分隔超平面和决策函数****

由 w=∑mi=1αiyixiw=∑i=1mαiyixi 求出 ww。

因为最终要求得的超平面满足 wx+b=0wx+b=0，这一点是和线性可分 SVM 的超平面一样的，因此求解 b 的过程也可以照搬：

b=1|S|∑s∈S(ys−wxs)b=1|S|∑s∈S(ys−wxs)

其中 SS 是支持向量的集合。

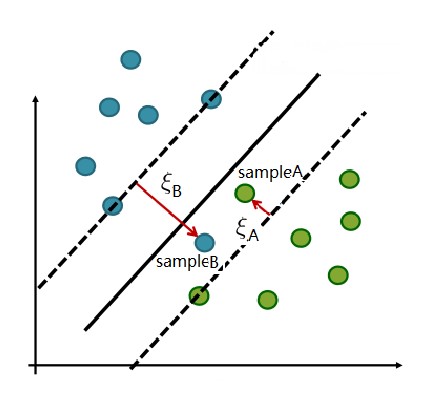
### 线性 SVM 的支持向量

这里有个问题，到底哪些样本算是****线性 SVM 的支持向量****？

对于线性可分 SVM，支持向量本身是很明确的，就是那些落在最大分隔超平面两侧的两个辅助超平面上的样本。因为样本线性可分，所以这两个辅助超平面中间的硬间隔里，是没有任何样本存在的。

但是，对于线性 SVM，有些不同，这两个辅助超平面中间是软间隔，软间隔的区域内也存在若干样本。这些样本是和辅助超平面上的样本一样算作支持向量呢？还是不算作支持向量？

比如下图中的 sampleA 和 sampleB，前者还好，只是“分得不够清楚”， 后者根本就“跨界”到了“对方的地盘”。它们两个到底算不算支持向量呢？



我们先来看看****线性 SVM（又名软间隔 SVM）****主问题拉格朗日函数的 ****KKT 条件****：

αi⩾0,μi⩾0αi⩾0,μi⩾0

yif(xi)–1+ξi⩾0yif(xi)–1+ξi⩾0

αi(yif(xi)–1+ξi)=0αi(yif(xi)–1+ξi)=0

ξi⩾0ξi⩾0

μiξi=0μiξi=0

其中 f(x)=wx+b,i=1,2,…,mf(x)=wx+b,i=1,2,…,m

对于任意样本 (xi,yi)(xi,yi)，要么 αi=0αi=0， 要么 yif(xi)–1+ξi=0yif(xi)–1+ξi=0。

我们又知道 ww 的计算公式为：

w=∑mi=1αiyixiw=∑i=1mαiyixi

其中拉格朗日乘子为0（即 αi=0αi=0）的项，对于 ww 的值是没有影响的，能够影响 ww 的，一定是对应拉格朗日乘子大于0的样本。

根据 KKT 条件，这样的样本一定同时满足 yif(xi)–1+ξi=0yif(xi)–1+ξi=0，也就是 yif(xi)=1–ξiyif(xi)=1–ξi。所有这样的样本，都是线性 SVM 的****支持向量****。

在满足 yif(xi)=1–ξiyif(xi)=1–ξi 的前提之下，我们来看 ξiξi。

若 ξi=0ξi=0, 则 yif(xi)=1yif(xi)=1，此时，样本正好落在两个辅助超平面上。所以，两个辅助超平面上的样本，肯定是支持向量。

若 ξi≠0ξi≠0：

当 ξi⩽1ξi⩽1 时（例如上图中的 ξAξA），1−ξi>01−ξi>0， yif(xi)>0yif(xi)>0。也就是说 yiyi 和 f(xi)f(xi) 的结果相乘虽然不为1，但至少这个样本还没有被归错类。

当 ξi>1ξi>1时（例如上图中的 ξBξB），1−ξi<01−ξi<0，则 yif(xi)<0yif(xi)<0，这时，样本根本就被归错了类。但是，即使如此，毕竟这样的样本也影响了最终 ww 的取值，所以，它也是支持向量。

也就是说，对于****线性 SVM**** 而言，除了落在两个辅助超平面上的样本，落在软间隔之内的样本也是它的****支持向量****。