菊安酱的机器学习第5期

菊安酱的直播间：https://live.bilibili.com/14988341

每周一晚8:00 菊安酱和你不见不散哦~(^o^)/~

更新日期：2018-12-3

作者：菊安酱

课件内容说明：

- 本文为作者参考众多书籍和博客所写，转载请注明作者和出处
- 如果想获得此课件及录播视频，可扫描下方二维码，回复"k"进群
- 若有任何疑问，请给作者留言。

12期完整版课纲

直播时间：每周一晚8:00

直播内容：

时间	期数	算法
2018/11/05	第1期	k-近邻算法
2018/11/12	第2期	决策树
2018/11/19	第3期	朴素贝叶斯
2018/11/26	第4期	Logistic回归
2018/12/03	第5期	支持向量机
2018/12/10	第6期	AdaBoost 算法
2018/12/17	第7期	线性回归
2018/12/24	第8期	树回归
2018/12/31	第9期	K-均值聚类算法
2019/01/07	第10期	Apriori 算法
2019/01/14	第11期	FP-growth 算法
2019/01/21	第12期	奇异值分解SVD

支持向量机

一、 什么是SVM?

支持向量机（Support Vector Machine,SVM）是用于分类的一种算法，也属于有监督学习的范畴。

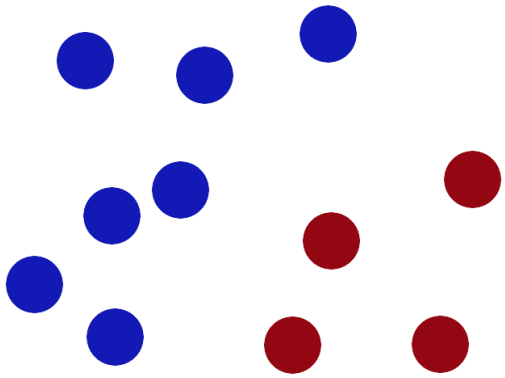
让我们先从一个大侠与反派的故事开始吧~

【该故事来源于https://www.reddit.com上的一个话题讨论：让5岁小孩也能看懂的SVM】

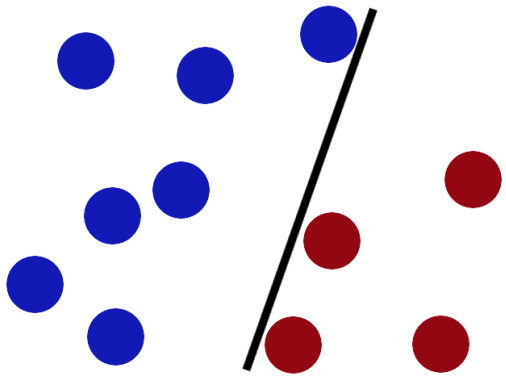
在很久以前，大侠的心上人被反派囚禁，大侠想要去救出他的心上人，于是便去和反派谈判。反派说只要你能顺利通过三关，我就放了你的心上人。

现在大侠的闯关正式开始：

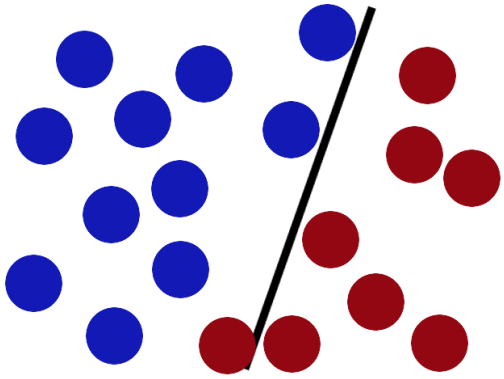
第一关：反派在桌子上似乎有规律地放了两种颜色的球，说：你用一根棍子分离开他们，要求是尽量再放更多的球之后，仍然适用。



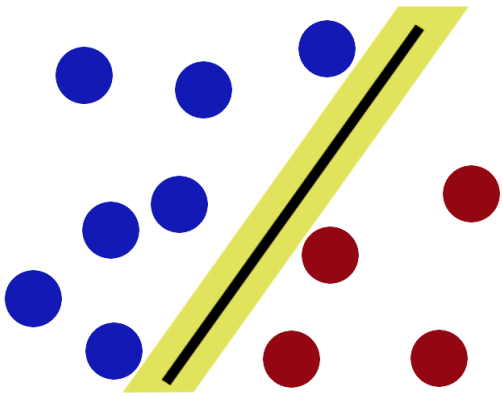
大侠很干净利索的放了一根棍子如下：



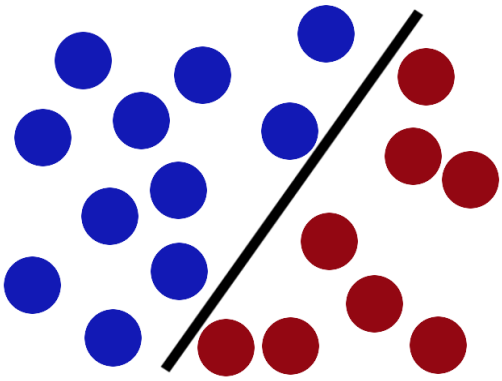
第二关：反派在桌子放上了更多的球，似乎有一个红球站错了阵营。



SVM就是试图把棍放在最佳位置，好让在棍的两边有尽可能大的间隙。

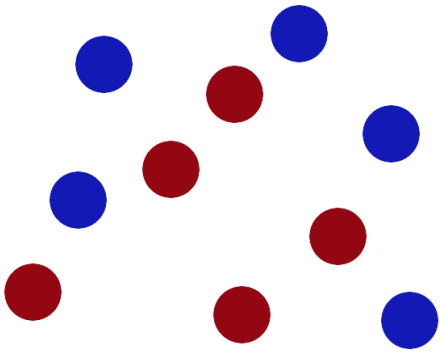


于是大侠将棍子调整如下，现在即使反派放入更多的球，棍子仍然是一个很好的分界线。

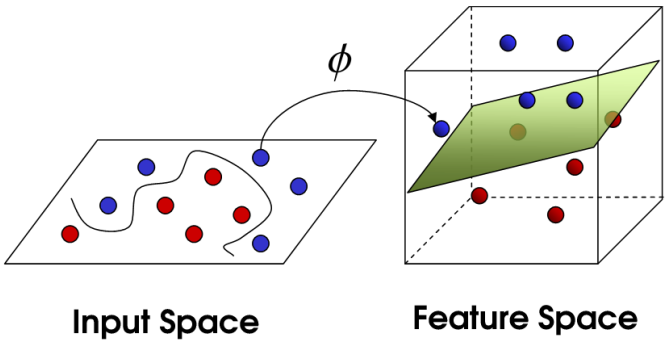


其实在SVM工具箱里还有另一个更加重要的**trick**。反派看到大侠已经学会了一个trick，于是心生一计，给大侠更难的一个挑战。

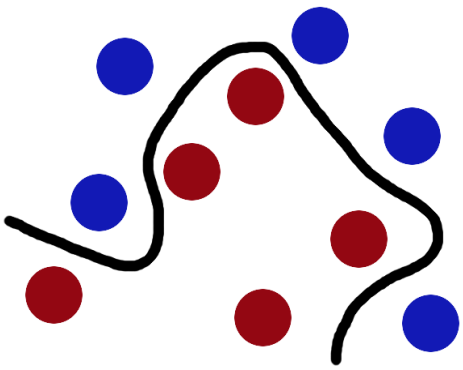
第三关：反派将球散乱地放在桌子上。



现在大侠已经没有方法用一根棍子将这些球分开了，怎么办呢？大侠灵机一动，使出三成内力拍向桌子，然后桌子上的球就被震到空中，说时迟那时快，大侠瞬间抓起一张纸，插到了两种球的中间。



现在从反派的角度看这些球，这些球像是被一条曲线分开了。于是反派乖乖地放了大侠的心上人。

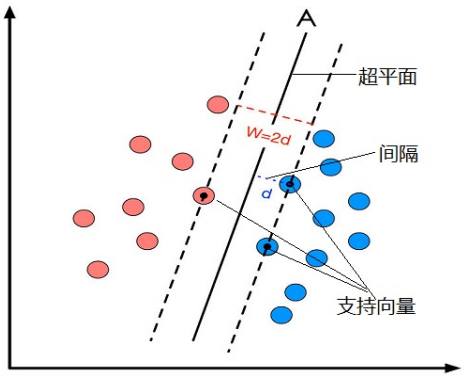


从此之后，江湖人便给这些分别起了名字，把这些球叫做 **「data」**，把棍子叫做 **「classifier」**，最大间隙trick 叫做**「optimization」**， 拍桌子叫做**「kernelling」**，那张纸叫做**「hyperplane」**。

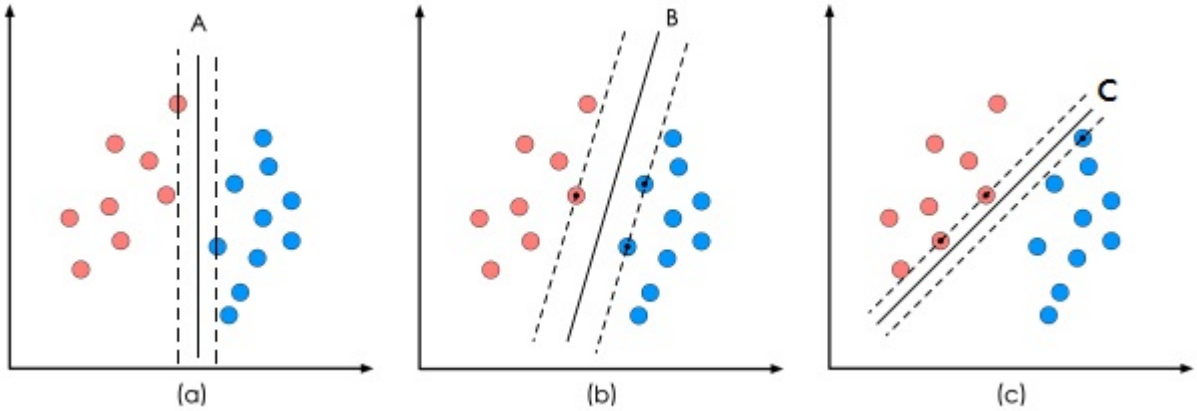
更为直观地感受一下吧（需翻墙）：https://www.youtube.com/watch?v=-Z4aojJ-pdg

概述一下：

当一个分类问题，数据是**线性可分(linearly separable)**的，也就是用一根棍就可以将两种小球分开的时候，我们只要将棍的位置放在让小球距离棍的距离最大化的位置即可，寻找这个最大间隔的过程，就叫做**最优化**。但是，现实往往是很残酷的，一般的数据是线性不可分的，也就是找不到一个棍将两种小球很好的分类。这个时候，我们就需要像大侠一样，将小球拍起，用一张纸代替小棍将小球进行分类。想要让数据飞起，我们需要的东西就是**核函数(kernel)**，用于切分小球的纸，就是**超平面**(hyperplane)。如果数据集是N维的，那么超平面就是N-1维的。



把一个数据集正确分开的超平面可能有多个（如下图），而那个具有“最大间隔”的超平面就是SVM要寻找的最优解。而这个真正的最优解对应的两侧虚线所穿过的样本点，就是SVM中的支持样本点，称为“**支持向量(support vector)**”。支持向量到超平面的距离被称为**间隔(margin)**。



维基百科对SVM的介绍(需翻墙)：

https://zh.wikipedia.org/wiki/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA

二、线性SVM

一个最优化问题通常有两个最基本的因素：

1）目标函数，也就是你希望什么东西的什么指标达到最好；

2）优化对象，你期望通过改变哪些因素来使你的目标函数达到最优。

在线性SVM算法中，目标函数显然就是那个“间隔”，而优化对象则是超平面。

我们以线性可分的二分类问题为例。

1. 超平面方程

在线性可分的二分类问题中，超平面其实就是一条直线。相信直线方程大家都不陌生：

$y = ax + b$ (公式1)

现在我们做个小小的改变，让原来的x轴变成x_1轴，y变成x_2轴，于是公式(1)中的直线方程会变成下面的样子:

$x_2 = ax_1 + b$ (公式2)

$ax_1 + (-1)x_2 + b = 0$ （公式3)

向量形式可以写成：

$[a, -1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b = 0$ (公式4)

进一步可表示为：

$\boldsymbol{\omega}^T \boldsymbol{x} + b = 0$ (公式5)

看到变量$\boldsymbol{\omega}, \boldsymbol{x}$略显粗壮的身体了吗？他们是黑体，表示变量是个向量，$\boldsymbol{\omega} = [\omega_1, \omega_2]^T$，$\boldsymbol{x} = [x_1, x_2]^T$。一般我们提到向量的时候，都默认是列向量，所以对$\boldsymbol{\omega}$进行了转置。这里向量$\boldsymbol{\omega}$与直线是相互垂直的（感兴趣的小伙伴可以推导一下），也就是说$\boldsymbol{\omega}$控制了直线的方向，b就是截距，它控制了直线的位置。

2. 间隔的计算公式

“间隔”其实就是点到直线的距离，如果你在百度文库里面搜索 “点到直线距离推导公式”，那么你会得到至少6、7种推导方法。这里采用向量法：

$$d = \frac{|\boldsymbol{\omega}^T \boldsymbol{x} + b|}{||\boldsymbol{\omega}||} \quad (\text{公式}6)$$

这里$||\boldsymbol{\omega}||$是向量$\boldsymbol{\omega}$的模，假如$\boldsymbol{\omega} = [\omega_1, \omega_2]^T$，则$||\boldsymbol{\omega}|| = \sqrt{\omega_1^2 + \omega_2^2}$，表示在空间中向量的长度，$\boldsymbol{x} = [x_1, x_2]^T$就是支持向量样本点的坐标。$\boldsymbol{\omega}, b$就是超平面方程的参数。

我们的目标是找出一个分类效果好的超平面作为分类器。分类器的好坏评定依据是分类间隔的$W = 2d$的大小，即分类间隔W越大，我们认为这个超平面的分类效果越好。而追求分类间隔W的最大化也就是寻找d的最大化。

看起来我们已经找到了目标函数的数学形式。但问题当然不会这么简单，我们还需要面对一连串令人头疼的麻烦。

3. 约束条件

虽然我们找到了目标函数，但是：

（1）我们如何判断一条直线能够将所有的样本点都正确分类？

（2）超平面的位置应该是在间隔区域的中轴线上，所以确定超平面位置的b参数也不能随意的取值。

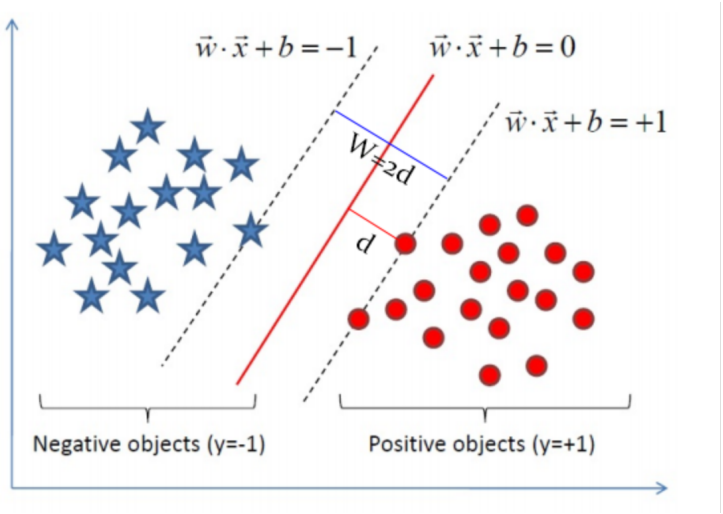
（3）对于一个给定的超平面，我们如何找到对应的支持向量，来计算距离d？

上述三个问题就是”约束条件“，也就是说，我们要优化的变量的取值范围收到了约束和限制。既然约束确实存在，那么就不得不用数学语言对它们进行描述。这里需要说明的是SVM可以通过一些小技巧，将这些约束条件糅合成一个不等式。请看下面糅合过程：

以下图为例，在平面空间中有红蓝两种点，对其分别标记为：

红色为正样本，标记为+1；

蓝色为负样本，标记为-1.



对每个样本点x_i加上类别标签y_i，则有

$$y_i = \begin{cases} +1 & \text{红} \\ -1 & \text{蓝} \end{cases}$$

如果我们的超平面能够完全将红蓝两种样本点分离开，那么则有

$$\begin{cases} |\boldsymbol{\omega}^T\boldsymbol{x}+b| > 0, y_i = 1 \\ |\boldsymbol{\omega}^T\boldsymbol{x}+b| < 0, y_i = -1 \end{cases} \quad (\text{公式}7)$$

如果要求在再高一点，假设超平面正好处于间隔区域的中轴线上，并且相应支持向量到超平面的距离为d,则公式可进一步写为：

$$\begin{cases} \frac{|\boldsymbol{\omega}^T\boldsymbol{x}+b|}{||\boldsymbol{\omega}||} \geq d, \ \forall\, y_i = +1 \\ \frac{|\boldsymbol{\omega}^T\boldsymbol{x}+b|}{||\boldsymbol{\omega}||} \leq -d, \ \forall\, y_i = -1 \end{cases} \quad (\text{公式}8)$$

符号∀是“对于所有满足条件的” 的缩写。也就是“任意一个”的意思。

对公式两边同时除以d，可得：

$$\begin{cases} \frac{|\boldsymbol{\omega}_d^T\boldsymbol{x}+b_d|}{||\boldsymbol{\omega}_d||} \geq 1, \ \forall\, y_i = +1 \\ \frac{|\boldsymbol{\omega}_d^T\boldsymbol{x}+b_d|}{||\boldsymbol{\omega}_d||} \leq -1, \ \forall\, y_i = -1 \end{cases} \quad (\text{公式}9)$$

其中，

$$\boldsymbol{\omega}_d = \frac{\boldsymbol{\omega}}{||\boldsymbol{\omega}||d}, \quad b_d = \frac{b}{||\boldsymbol{\omega}||d}$$

因为$||w||$和d都是标量。所以上述公式的两个矢量，依然描述一条直线的法向量和截距。所以下面两个公式，都是描述一条直线，数学模型代表的意义是一样的。

$$\boldsymbol{\omega}_d^T\boldsymbol{x} + b_d = 0$$
$$\boldsymbol{\omega}^T\boldsymbol{x} + b = 0$$

现在，让我们对w_d和b_d重新起个名字，就叫它们w和b，所以我们可得到：

$$\begin{cases} \boldsymbol{\omega}^T x_i + b \geq 1, \forall\, y_i = +1 \\ \boldsymbol{\omega}^T x_i + b \leq -1, \forall\, y_i = -1 \end{cases} \quad (\text{公式}10)$$

这个方程就是SVM最优化问题的约束条件。由于我们将标签定义为1和-1，所以此处我们可以将上述方程糅合成一个约束方程：

$$y_i(\boldsymbol{\omega}^T x_i + b) \geq 1, \forall x_i \quad (\text{公式}11)$$

4. 线性SVM优化问题基本描述

对于公式$\boldsymbol{\omega}^T x_i + b$ =1 or =-1，什么时候会发生呢？参考公式10 就会知道，只有当x_i是超平面的支持向量时，等于1或者-1的情况才会出现。无论是等于1还是-1，对于公式10来说，都有$|\boldsymbol{\omega}^T x_i + b| = 1$

所以对于这些支持向量来说：

$$d = \frac{|\boldsymbol{\omega}^T \boldsymbol{x}_i + b|}{||\boldsymbol{\omega}||} = \frac{1}{||\boldsymbol{\omega}||}, \forall \text{支持向量} x_i \quad (\text{公式}12)$$

我们原来的任务是找到一组参数$\boldsymbol{\omega}, b$使得分类间隔$W = 2d$最大化，根据公式12 就可以转变为$||\boldsymbol{\omega}||$的最小化问题，也等效于$\frac{1}{2}||\boldsymbol{\omega}||^2$的最小化问题。我们之所以要在$||\boldsymbol{\omega}||$上加上平方和1/2的系数，是为了以后进行最优化的过程中对目标函数求导时比较方便，但这绝不影响最优化问题最后的解。

所以，线性SVM最优化问题的数学描述就是：

$$\begin{aligned} &\min_{\boldsymbol{\omega},b} \ \frac{1}{2}||\boldsymbol{\omega}||^2 \\ &\text{s. t.} \ \ y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b) \geq 1, \ \ i = 1, 2, \ldots, n \end{aligned} \quad (\text{公式}13)$$

这里n是样本点的总个数，缩写s. t. 表示“Subject to”，是“服从某某条件”的意思。公式13 描述的是一个典型的不等式约束条件下的二次型函数优化问题，同时也是支持向量机的基本数学模型。

5. 最优化问题的求解

通常我们需要求解的最优化问题有如下几类：

- 无约束优化问题，可以写为：

$$min\ f(x)$$

- 有等式约束的优化问题，可以写为：

$$\begin{aligned} &min\ f(x) \\ &s.t.\ h_i(x) = 0, \ \ i = 1, 2, \ldots, n \end{aligned}$$

- 有不等式约束的优化问题，可以写为：

$$\begin{aligned} &min\ f(x) \\ &s.t.\ g_i(x) \leq 0, \ \ i = 1, 2, \ldots, n \\ &h_j(x) = 0, \ \ j = 1, 2, \ldots, m \end{aligned}$$

对于第(1)类的优化问题，尝试使用的方法就是**费马大定理**(Fermat)，即使用求取函数f(x)的导数，然后令其为零，可以求得候选最优值，再在这些候选值中验证；如果是凸函数，可以保证是最优解。这也就是我们高中经常使用的求函数的极值的方法。

对于第(2)类的优化问题，常常使用的方法就是**拉格朗日乘子法**（Lagrange Multiplier），即把等式约束$h_i(x)$用一个系数与f(x)写为一个式子，称为拉格朗日函数，而系数称为拉格朗日乘子。通过拉格朗日函数对各个变量求导，令其为零，可以求得候选值集合，然后验证求得最优值。

对于第(3)类的优化问题，常常使用的方法就是**KKT条件**(Karush-Kuhn-Tucker conditions)。同样地，我们把所有的等式、不等式约束与f(x)写为一个式子，也叫拉格朗日函数，系数也称拉格朗日乘子，通过一些条件，可以求出最优值的**必要条件**，这个条件称为KKT条件。对KTT条件感兴趣的可参考：

https://blog.csdn.net/james_616/article/details/72869015

必要条件和充要条件如果不理解，可以看下面这句话：

- A的**必要条件**就是A可以推出的**结论**
- A的**充分条件**就是可以推出A的**前提**

对于我们的线性SVM最优化问题：

$$\begin{aligned} &\min_{\boldsymbol{\omega},b}\ \frac{1}{2}||\boldsymbol{\omega}||^2 \\ &\text{s. t. }\ y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)\geq 1,\ \ i=1,2,\dots,n \end{aligned} \qquad (公式13)$$

显然，它属于第(3)类的优化问题。那么在求解这类优化问题之前，我们还需要了解两个概念——拉格朗日函数和KKT条件。

6. 拉格朗日函数

首先，我们先要从宏观的视野上了解一下**拉格朗日对偶问题出现的原因和背景。**

我们知道我们要求解的是最小化问题，所以一个直观的想法是如果我能够构造一个函数，使得该函数在可行解区域内与原目标函数完全一致，而在可行解区域外的数值非常大，甚至是无穷大，那么这个**没有约束条件的新目标函数的优化问题**就与原来**有约束条件的原始目标函数的优化问题**是等价的问题。这就是使用拉格朗日方程的目的，它**将约束条件放到目标函数**中，从而将有约束优化问题转换为无约束优化问题。

但是对于拉格朗日函数，直接使用求导的方式求解仍然很困难，所以便有了**拉格朗日对偶**的诞生。

所以，显而易见的是，我们在拉格朗日优化我们的问题这个道路上，**需要进行下面二个步骤：**

- 将有约束的原始目标函数转换为无约束的新构造的拉格朗日目标函数
- 使用拉格朗日对偶性，将不易求解的优化问题转化为易求解的优化

第一步：**将有约束的原始目标函数转换为无约束的新构造的拉格朗日目标函数**

原始目标函数：

$$\begin{aligned} &\min_{\boldsymbol{\omega},b}\ \frac{1}{2}||\boldsymbol{\omega}||^2 \\ &\text{s. t. }\ y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)\geq 1,\ \ i=1,2,\dots,n \end{aligned} \qquad (公式13)$$

新构造的目标函数：

$$L(\boldsymbol{\omega},b,\boldsymbol{\alpha})=\frac{1}{2}||\boldsymbol{\omega}||^2-\sum_{i=1}^{n}\alpha_i(y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)-1) \quad 公式（14）$$

其中α_i是拉格朗日乘子，且$\alpha_i\geq 0$，是我们人为设定的参数。

大家知道我们的目标是追求$\frac{1}{2}||\boldsymbol{\omega}||^2$的最小化，又因为

$$\alpha_i \geq 0$$
$$y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b) \geq 1$$
$$y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)-1 \geq 0$$
$$\sum_{i=1}^{n}\alpha_i(y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)-1) \geq 0$$

所以我们的新目标函数：

$$\min_{\boldsymbol{\omega},b}\left[\max_{\boldsymbol{\alpha}:\alpha_j\geq 0} L(\boldsymbol{\omega},b,\boldsymbol{\alpha})\right] \text{（公式15）}$$

第二步：**拉格朗日对偶函数**

对偶后的目标函数：

$$\max_{\boldsymbol{\alpha}:\alpha_j\geq 0}\left[\min_{\boldsymbol{\omega},b} L(\boldsymbol{\omega},b,\boldsymbol{\alpha})\right] \text{（公式16）}$$

接下来，我们就可以求解拉格朗日对偶函数了，求解出来的值就是我们最优化问题的结果，也就是可以得到最大间隔。

7. 对偶问题求解

第一步

根据公式16，我们可以先求$\min_{\boldsymbol{\omega},b} L(\boldsymbol{\omega},b,\boldsymbol{\alpha})$：

$$\min_{\boldsymbol{\omega},b} L(\boldsymbol{\omega},b,\boldsymbol{\alpha}) = \min_{\boldsymbol{\omega},b}\left[\frac{1}{2}||\boldsymbol{\omega}||^2-\sum_{i=1}^{n}\alpha_i(y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)-1)\right] \text{公式}(17)$$

分别令函数 $L(\boldsymbol{\omega},b,\boldsymbol{\alpha})$对 $\boldsymbol{\omega},b$求偏导，并使其等于0。

$$\frac{\partial L}{\partial \boldsymbol{\omega}}=0 \implies \boldsymbol{\omega}=\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i \text{（公式18）}$$

$$\frac{\partial L}{\partial b}=0 \implies \sum_{i=1}^{n}\alpha_i y_i = 0 \text{（公式19）}$$

将公式18和公式19带入到公式17中：

$$
\begin{aligned}
L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) &= \frac{1}{2}||\boldsymbol{\omega}||^2 - \sum_{i=1}^{n}\alpha_i(y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i + b) - 1) \\
&= \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} - \boldsymbol{\omega}^T\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i \\
&= \frac{1}{2}\boldsymbol{\omega}^T\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i - \boldsymbol{\omega}^T\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i - b*0 + \sum_{i=1}^{n}\alpha_i \\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\boldsymbol{\omega}^T\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i \\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right)^T\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i \\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j \quad \text{公式}(20)
\end{aligned}
$$

从上面的最后一个式子，我们可以看出，此时的L(w,b,α)函数只含有一个变量，即α_i。

第二步

现在内侧的最小值求解完成，我们求解外侧的最大值，从上面的式子得到：

$$
\max_{\boldsymbol{\alpha}:\alpha_j\geq 0}\left[\min_{\boldsymbol{\omega},b} L(\boldsymbol{\omega}, b, \boldsymbol{\alpha})\right]
$$

$$
\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j\right]
$$

$$
s.t.\ \alpha_i \geq 0,\ i = 1, 2, \ldots, n
$$

$$
\sum_{i=1}^{n}\alpha_i y_i = 0
$$

现在我们的优化问题变成了如上的形式。至此，一切都很完美。但是这里有个假设：数据必须100%线性可分。但是，目前为止，我们知道几乎所有数据都不那么"干净"。这时我们就可以通过引入所谓的松弛变量C，来允许有些数据点可以处于超平面的错误的一侧。此时，我们的目标函数不变，约束条件变为：

$$
s.t.\ C \geq \alpha_i \geq 0,\ i = 1, 2, \ldots, n
$$

$$
\sum_{i=1}^{n}\alpha_i y_i = 0
$$

我们为什么要费这么大劲把优化问题转化成这样呢？实际上，是为了使用高效优化算法SMO算法。

三、SMO算法

1. 什么是SMO算法?

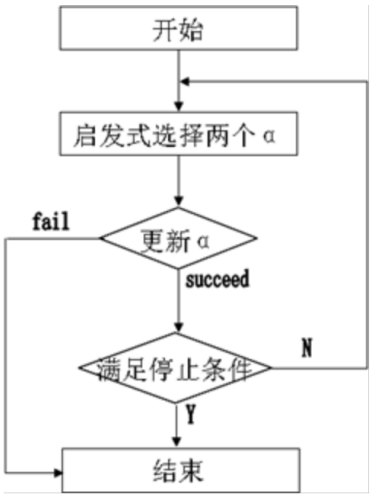
SMO算法就是序列最小优化（Sequential Minimal Optimization），它是由 John Platt 于1996年发布的专门用于训练SVM的一个强大算法。SMO算法的目的是将大优化问题分解为多个小优化问题来求解。这些小优化问题往往很容易求解，并且对它们进行顺序求解的结果与将它们作为整体来求解的结果完全一致的。在结果完全相同的同时，SMO算法的求解时间短很多。

SMO算法的目标是求出一系列α和b，一旦求出了这些α，就很容易计算出权重向量w并得到分隔超平面。

SMO算法的工作原理是：每次循环中选择两个α进行优化处理。一旦找到了一对合适的α，那么就增大其中一个同时减小另一个。这里所谓的"合适"就是指两个α必须符合以下两个条件:

- 两个α必须要在间隔边界之外
- 这两个α还没有进行过区间化处理或者不在边界上。

2. SMO算法流程



关于SMO算法流程这一块内容，李航的《统计学习方法》中介绍的比较详细（$P_{124}-P_{133}$）,感兴趣的小伙伴可自行翻阅。这里我们就不再赘述一系列的推导过程，只给大家梳理一下SMO算法的实施步骤：

步骤1：计算误差：

$$E_i = f(x_i) - y_i = \left(\sum_{j=1}^{n} \alpha_j y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + b\right) - y_i$$

步骤2：计算上下界H和L：

$$\begin{cases} L = max(0, \alpha_j^{old} - \alpha_i^{old}), H = min(C, C + \alpha_j^{old} - \alpha_i^{old}) \ \ if \ y_i \neq y_j \\ L = max(0, \alpha_j^{old} + \alpha_i^{old} - C), H = min(C, \alpha_j^{old} + \alpha_i^{old}) \ \ if \ y_i = y_j \end{cases}$$

步骤3：计算学习率η:

$$\eta = \boldsymbol{x}_i^T \boldsymbol{x}_i + \boldsymbol{x}_j^T \boldsymbol{x}_j - 2\boldsymbol{x}_i^T \boldsymbol{x}_j$$

步骤4：更新α_j:

$$\alpha_j^{new} = \alpha_j^{old} + \frac{y_i(E_i - E_j)}{\eta}$$

步骤5：根据取值范围修剪α_j:

$$\alpha^{new,clipped} = \begin{cases} H, & if \ \alpha_j^{new} \geq H \\ \alpha_j^{new}, & if \ L \leq \alpha_j^{new} \leq H \\ L, & if \ \alpha_j^{new} \leq L \end{cases}$$

步骤6：更新α_i：

$$\alpha_i^{new} = \alpha_i^{old} + y_i y_j(\alpha_j^{old} - \alpha_j^{new,clipped})$$

步骤7：更新b1和b2:

$$b_1^{new} = b^{old} - E_i - y_i(\alpha_i^{new} - \alpha_i^{old})\boldsymbol{x}_i^T\boldsymbol{x}_i - y_j(\alpha_j^{new} - \alpha_j^{old})\boldsymbol{x}_j^T\boldsymbol{x}_i$$
$$b_2^{new} = b^{old} - E_j - y_i(\alpha_i^{new} - \alpha_i^{old})\boldsymbol{x}_i^T\boldsymbol{x}_j - y_j(\alpha_j^{new} - \alpha_j^{old})\boldsymbol{x}_j^T\boldsymbol{x}_j$$

步骤8：根据b1和b2更新b:

$$b = \begin{cases} b_1 & 0 < \alpha_1^{new} < C \\ b_2 & 0 < \alpha_2^{new} < C \\ \frac{1}{2}(b_1 + b_2) & otherwise \end{cases}$$

3. 简化版SMO算法

SMO算法的完整版实现需要大量的代码。这里我们先讨论SMO算法的简化版，主要用来正确理解这个算法的工作流程。然后会对这个简化版的SMO算法进行优化，加快它的运行速度。

3.1 SMO算法的伪代码

SMO函数的伪代码：

```
创建一个α向量并初始化为0向量
当迭代次数 < 最大迭代次数时（外循环）：
    对数据集中每个数据向量（内循环）：
        如果该数据向量可以被优化：
            随机选择另外一个数据向量
            同时优化这两个向量
            如果两个向量都不能被优化，则退出内循环
    如果所有向量都没被优化，迭代次数+1，继续下一次循环
```

3.2 构建辅助函数

生成特征向量和标签向量

```
import numpy as np
import pandas as pd
"""
函数功能：创建特征向量和标签向量
参数说明：
    file：原始文件路径
返回：
    xMat：特征向量
    yMat：标签向量
"""
def loadDataSet(file):
    dataSet= pd.read_table(file,header = None)
    xMat=np.mat(dataSet.iloc[:,:-1].values)
    yMat=np.mat(dataSet.iloc[:,-1].values).T
```

```
    return xMat,yMat
```

数据集可视化

```
import matplotlib.pyplot as plt
%matplotlib inline

def showDataSet(xMat, yMat):
    data_p = []                                     #正样本
    data_n = []                                     #负样本
    m = xMat.shape[0]                               #样本总数
    for i in range(m):
        if yMat[i] > 0:
            data_p.append(xMat[i])
        else:
            data_n.append(xMat[i])
    data_p_ = np.array(data_p)                      #转换为numpy矩阵
    data_n_ = np.array(data_n)                      #转换为numpy矩阵
    plt.scatter(data_p_.T[0], data_p_.T[1])         #正样本散点图
    plt.scatter(data_n_.T[0], data_n_.T[1])         #负样本散点图
    plt.show()
```

随机选择alpha对：

```
import random
"""
函数功能：随机选择一个索引
参数说明：
    i:第一个alpha索引
    m:数据集总行数
返回：
    j:随机选择的不与i相等的值
"""
def selectJrand(i,m):
    j=i
    while (j==i):
        j=int(random.uniform(0,m))
    return j
```

α_j的修剪函数：

```
"""
函数功能：修剪alpha_j
"""
def clipAlpha(aj,H,L):
    if aj>H:
        aj=H
    if L>aj:
        aj=L
    return aj
```

3.3 简化版SMO算法

```
"""
函数功能:
参数说明:
    xMat:特征向量
    yMat:标签向量
    C:常数
    toler:容错率
    maxIter:最大迭代次数
返回:
    b、alpha
"""
def smoSimple(xMat,yMat,C,toler,maxIter):
    b=0                                         #初始化b参数
    m,n=xMat.shape                              #m为数据集的总行数，n为特征的数量
    alpha = np.mat(np.zeros((m,1)))             #初始化alpha参数，设为0
    iters =0                                    #初始化迭代次数
    while (iters<maxIter):
        alpha_ = 0                              #初始化alpha优化次数
        for i in range(m):
            #步骤1：计算误差Ei
            fXi=np.multiply(alpha,yMat).T*(xMat*xMat[i,:].T)+b
            Ei=fXi-yMat[i]
            #优化alpha，设定容错率
            if ((yMat[i]*Ei<-toler)and(alpha[i]<C)) or
((yMat[i]*Ei>toler)and(alpha[i]>0)):
                #随机选择一个与alpha_i成对优化的alpha_j
                j=selectJrand(i,m)
                #步骤1：计算误差Ej
                fXj=np.multiply(alpha,yMat).T*(xMat*xMat[j,:].T)+b
                Ej=fXj-yMat[j]
                #保存更新前的alpha_i和alpha_j
                alphaIold=alpha[i].copy()
                alphaJold=alpha[j].copy()
                #步骤2：计算上下界H和L
                if (yMat[i]!=yMat[j]):
                    L=max(0,alpha[j]-alpha[i])
                    H=min(C,C+alpha[j]-alpha[i])
                else:
                    L=max(0,alpha[j]+alpha[i]-C)
                    H=min(C,C+alpha[j]+alpha[i])
                if L==H:
                    #print('L==H')
                    continue
                #步骤3：计算学习率eta(eta是alpha_j的最优修改量)
                eta=2*xMat[i,:]*xMat[j,:].T-xMat[i,:]*xMat[i,:].T-xMat[j,:]*xMat[j,:].T
                if eta>=0:
                    #print('eta>=0')
                    continue
                #步骤4：更新alpha_j
                alpha[j]-= yMat[j]*(Ei-Ej)/eta
                #步骤5：修剪alpha_j
```

```
                alpha[j]=clipAlpha(alpha[j],H,L)
                if abs(alpha[j]-alphaJold)<0.00001:
                    #print('alpha_j 变化太小')
                    continue
                #步骤6：更新alpha_i
                alpha[i]+=yMat[j]*yMat[i]*(alphaJold-alpha[j])
                #步骤7：更新b_1和b_2
                b1=b-Ei-yMat[i]*(alpha[i]-alphaIold)*xMat[i,:]*xMat[i,:].T-yMat[j]*
(alpha[j]-alphaJold)*xMat[i,:]*xMat[j,:].T
                b2=b-Ej-yMat[i]*(alpha[i]-alphaIold)*xMat[i,:]*xMat[j,:].T-yMat[j]*
(alpha[j]-alphaJold)*xMat[j,:]*xMat[j,:].T
                #步骤8：根据b_1和b_2更新b
                if (0<alpha[i])and(C>alpha[i]): b=b1
                elif (0<alpha[j])and(C>alpha[j]): b=b2
                else: b=(b1+b2)/2
                #统计优化次数
                alpha_+=1
                #print(f'第{iters}次迭代 样本{i},alpha优化次数:{alpha_}')
        #更新迭代次数
        if alpha_==0: iters+=1
        else: iters=0
        #print(f'迭代次数为:{iters}')
    return b,alpha
```

查看代码运行时间及结果：

```
%time b,alpha=smoSimple(xMat,yMat,0.6,0.001,5)
```

3.4 支持向量的可视化

```
def get_sv(xMat,yMat,alpha):
    m=xMat.shape[0]
    sv_x=[]
    sv_y=[]
    for i in range(m):
        if alpha[i]>0:
            sv_x.append(xMat[i])
            sv_y.append(yMat[i])
    sv_x1 =np.array(sv_x).T
    sv_y1 = np.array(sv_y).T
    return sv_x1,sv_y1
```

```
def showPlot(xMat, yMat,alpha):
    data_p = []                                    #正样本
    data_n = []                                    #负样本
    m = xMat.shape[0]                              #样本总数
    for i in range(m):
        if yMat[i] > 0:
```

```
            data_p.append(xMat[i])
        else:
            data_n.append(xMat[i])
    data_p_ = np.array(data_p)                    #转换为numpy矩阵
    data_n_ = np.array(data_n)                    #转换为numpy矩阵
    #样本散点图
    plt.scatter(data_p_.T[0], data_p_.T[1])      #正样本散点图
    plt.scatter(data_n_.T[0], data_n_.T[1])      #负样本散点图
    #绘制支持向量
    sv_x,sv_y=get_sv(xMat,yMat,alpha)
    plt.scatter(sv_x[0], sv_x[1], s=150, c='none', alpha=0.7, linewidth=1.5,
edgecolor='red')
```

其他

- 菊安酱的直播间：https://live.bilibili.com/14988341
- 下周一（2018/12/10）将讲解**Adaboost算法**，欢迎各位进入菊安酱的直播间观看直播
- 如有问题，可以给我留言哦~

参考资料：

[1] 5岁小孩也能看懂的SVM：

http://bytesizebio.net/2014/02/05/support-vector-machines-explained-well/

[2] 大侠与魔鬼的故事：

https://www.zhihu.com/question/21094489/answer/86273196

[3] 陈老师的知乎专栏：

https://zhuanlan.zhihu.com/p/24638007

[4] 深入理解拉格朗日乘子法和KKT条件

https://blog.csdn.net/xianlingmao/article/details/7919597

[5] Jack大佬的博客：

https://cuijiahua.com/blog/2017/11/ml_8_svm_1.html