# 支持向量机通俗导论（理解 SVM 的三层境界）

作者：July、pluskid ；致谢：白石、JerryLead
出处：结构之法算法之道 blog。

## 前言

动笔写这个支持向量机 (support vector machine) 是费了不少劲和困难的，原因很简单，一者这个东西本身就并不好懂，要深入学习和研究下去需花费不少时间和精力，二者这个东西也不好讲清楚，尽管网上已经有朋友写得不错了 (见文末参考链接)，但在描述数学公式的时候还是显得不够。得益于同学白石的数学证明，我还是想尝试写一下，希望本文在兼顾通俗易懂的基础上，真真正正能足以成为一篇完整概括和介绍支持向量机的导论性的文章。

本文在写的过程中，参考了不少资料，包括《支持向量机导论》、《统计学习方法》及网友 pluskid 的支持向量机系列等等，于此，还是一篇学习笔记，只是加入了自己的理解和总结，有任何不妥之处，还望海涵。全文宏观上整体认识支持向量机的概念和用处，微观上深究部分定理的来龙去脉，证明及原理细节，力保逻辑清晰 & 通俗易懂。

同时，阅读本文时建议大家尽量使用 chrome 等浏览器，如此公式才能更好的显示，再者，阅读时可拿张纸和笔出来，把本文所有定理. 公式都亲自推导一遍或者直接打印下来（可直接打印网页版或本文文末附的 PDF，享受随时随地思考、演算的极致快感），在文稿上演算。

Ok，还是那句原话，有任何问题，欢迎任何人随时不吝指正 & 赐教，感谢。

## 1 第一层、了解 SVM

### 1.1 什么是支持向量机 SVM

要明白什么是 SVM，便得从分类说起。

分类作为数据挖掘领域中一项非常重要的任务，它的目的是学会一个分类函数或分类模型 (或者叫做分类器)，而支持向量机本身便是一种监督式学习的方法 (至于具体什么是监督学习与非监督学习，请参见此系列 Machine L&Data Mining 第一篇)，它广泛的应用于统计分类以及回归分析中。

支持向量机（SVM）是 90 年代中期发展起来的基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。

通俗来讲，它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。

对于不想深究 SVM 原理的同学或比如就只想看看 SVM 是干嘛的，那么，了解到这里便足够了，不需上层。而对于那些喜欢深入研究一个东西的同学，甚至究其本质的，咱们则还有很长的一段路要走，万里长征，咱们开始迈第一步吧，相信你能走完。

### 1.2 线性分类

OK，在讲 SVM 之前，咱们必须先弄清楚一个概念：线性分类器 (也可以叫做感知机，这里的机表示的是一种算法，本文第三部分、证明 SVM 中会详细阐述)。

#### 1.2.1 分类标准

这里我们考虑的是一个两类的分类问题，数据点用 $x$ 来表示，这是一个 $n$ 维向量，$w^T$ 中的 $T$ 代表转置，而类别用 $y$ 来表示，可以取 1 或者 -1 ，分别代表两个不同的类。一个线性分类器的学习目标就是要在 $n$ 维的数据空间中找到一个分类超平面，其方程

可以表示为：

$$w^T x + b = 0 \tag{1}$$

上面给出了线性分类的定义描述，但或许读者没有想过：为何用 y 取 1 或者 -1 来表示两个不同的类别呢？其实，这个 1 或 -1 的分类标准起源于 logistic 回归，为了完整和过渡的自然性，咱们就再来看看这个 logistic 回归。

### 1.2.2 1 或 -1 分类标准的起源：logistic 回归

Logistic 回归目的是从特征学习出一个 0/1 分类模型，而这个模型是将特性的线性组合作为自变量，由于自变量的取值范围是负无穷到正无穷。因此，使用 logistic 函数（或称作 sigmoid 函数）将自变量映射到 $(0,1)$ 上，映射后的值被认为是属于 y=1 的概率。

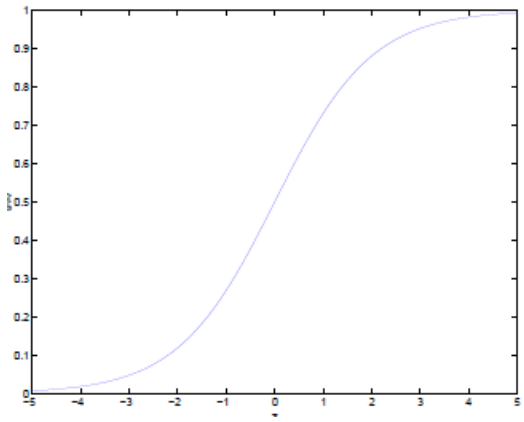
形式化表示就是

假设函数

$$h_0(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \tag{2}$$

其中 $x$ 是 $n$ 维特征向量，函数 $g$ 就是 logistic 函数。

而 $g(z) = \frac{1}{1+e^{-z}}$ 的图像是




可以看到，将无穷映射到了 $(0,1)$。

而假设函数就是特征属于 $y=1$ 的概率。

$$P(y=1|x;\theta) = h_\theta(x) P(y=0|x;\theta) = 1 - h_\theta(x) \tag{3}$$

当我们要判别一个新来的特征属于哪个类时，只需求，若大于 0.5 就是 $y=1$ 的类，反之属于 $y=0$ 类。

再审视一下 $h_\theta(x)$，发现 $h_\theta(x)$ 只和 $\theta^T$ 有关，$\theta^T x > 0$，那么 $h_\theta(x) > 0.5$，$g(z)$ 只不过是用来映射，真实的类别决定权还在 $\theta^T x$。还有当时 $\theta^T x >> 0$，$h_\theta(x) = 1$，反之 $h_\theta(x) = 0$。如果我们只从 $\theta^T x$ 出发，希望模型达到的目标无非就是让训练数据中 $y=1$ 的特征

$$\theta^T x >> 0$$

，而是 $y=0$ 的特征 $\theta^T x << 0$。Logistic 回归就是要学习得到 $\theta$，使得正例的特征远大于 0，负例的特征远小于 0，强调在全部训练实例上达到这个目标。

### 1.2.3 形式化标示

我们这次使用的结果标签是 $y=-1,y=1$，替换在 logistic 回归中使用的 $y=0$ 和 $y=1$。同时将 $\theta$ 替换成 $w$ 和 $b$。以前的 $\theta^T x=\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n$，其中认为 $x_0=1$。现在我们替换为 $b$，后面替换 $\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n$ 为 $w_1 x_1+w_2 x_2+\cdots+w_n x_n$（即 $x^T x$）。这样，我们让 $\theta^T x=w^T x+b$，进一步 $h_\theta(x)=g(\theta^T x)=g(w^T x+b)$。也就是说除了 $y$ 由 $y=0$ 变为 $y=-1$，只是标记不同外，与 logistic 回归的形式化表示没区别。

再明确下假设函数

$$h_{w,b}=g(w^T x+b) \tag{4}$$

上面提到过我们只需考虑 $\theta^T x$ 的正负问题，而不用关心 $g(z)$，因此我们这里将 $g(z)$ 做一个简化，将其简单映射到 $y=-1$ 和 $y=1$ 上。映射关系如下：

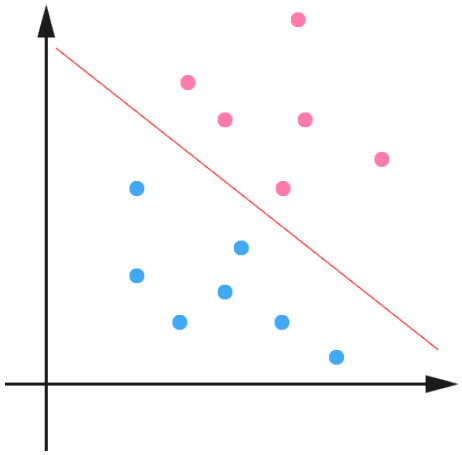
$$g(z)=\begin{cases}1, & z\geqslant 0\\ -1, & z<0\end{cases} \tag{5}$$

于此，想必已经解释明白了为何线性分类的标准一般用 1 或者 $-1$ 来标示。

注：上小节来自 jerrylead 所作的斯坦福机器学习课程的笔记。

## 1.3 线性分类的一个例子

下面举个简单的例子，一个二维平面 (一个超平面，在二维空间中的例子就是一条直线)，如下图所示，平面上有两种不同的点，分别用两种不同的颜色表示，一种为红颜色的点，另一种则为蓝颜色的点，红颜色的线表示一个可行的超平面。

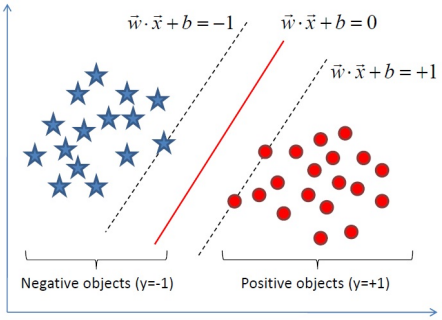


从上图中我们可以看出，这条红颜色的线把红颜色的点和蓝颜色的点分开来了。而这条红颜色的线就是我们上面所说的超平面，也就是说，这个所谓的超平面的的确确便把这两种不同颜色的数据点分隔开来，在超平面一边的数据点所对应的 $y$ 全是 $-1$，而在另一边全是 1。

接着，我们可以令分类函数（提醒：下文很大篇幅都在讨论着这个分类函数）：

$$f(x)=w^T x+b \tag{6}$$

显然，如果 $f(x)=0$，那么 $x$ 是位于超平面上的点。我们不妨要求对于所有满足 $f(x)<0$ 的点，其对应的 $y$ 等于 $-1$，而 $f(x)>0$ 则对应 $y=1$ 的数据点。




注：上图中，定义特征到结果的输出函数 $u=\vec{w}\cdot\vec{x}-b$，与我们之前定义的 $f(x)=w^Tx+b$ 实质是一样的。为什么？因为无论是 $u=\vec{w}\cdot\vec{x}-b$，还是 $f(x)=w^Tx+b$，不影响最终优化结果。下文你将看到，当我们转化到优化 $\max\frac{1}{\|w\|}, s.t., y_i(w^Tx_i+b)\geq 1, i=1,\cdots,n$ 的时候，为了求解方便，会把 $yf(x)$ 令为 1，即 $yf(x)$ 是 $y(w^x+b)$，还是 $y(w^x-b)$，对我们要优化的式子 $\max\frac{1}{\|w\|}$ 已无影响。

（有一朋友飞狗来自 Mare Desiderii，看了上面的定义之后，问道：请教一下 SVM functional margin 为 $\widehat{\gamma}=y(wTx+b)=yf(x)$ 中的 $\gamma$ 是只取 1 和 $-1$ 吗？$y$ 的唯一作用就是确保 functional margin 的非负性？真是这样的么？当然不是，详情请见本文评论下第 43 楼）

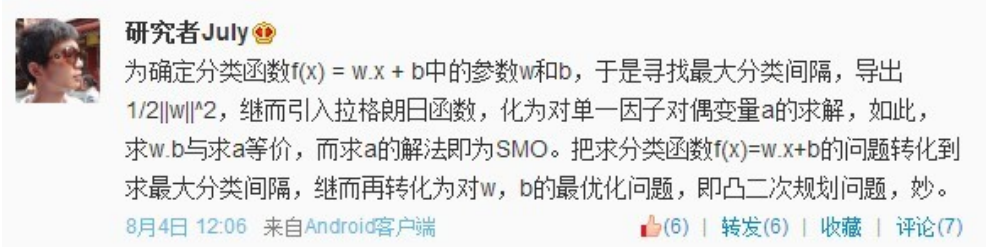
当然，有些时候，或者说大部分时候数据并不是线性可分的，这个时候满足这样条件的超平面就根本不存在 (不过关于如何处理这样的问题我们后面会讲)，这里先从最简单的情形开始推导，就假设数据都是线性可分的，亦即这样的超平面是存在的。

更进一步，我们在进行分类的时候，将数据点 $x$ 代入 $f(x)$ 中，如果得到的结果小于 0，则赋予其类别 $-1$，如果大于 0 则赋予类别 1。如果 $f(x)=0$，则很难办了，分到哪一类都不是。

请读者注意，下面的篇幅将按下述 3 点走：

1. 咱们就要确定上述分类函数 $f(x)=w\cdot x+b$（$w\cdot x$ 表示 w 与 x 的内积）中的两个参数 $w$ 和 $b$，通俗理解的话 $w$ 是法向量，$b$ 是截距（再次说明：定义特征到结果的输出函数 $u=|w\cdot|x-b$，与我们最开始定义的 $f(x)=w^Tx+b$ 实质是一样的）；
2. 那如何确定 $w$ 和 $b$ 呢？答案是寻找两条边界端或极端划分直线中间的最大间隔（之所以要寻最大间隔是为了能更好的划分不同类的点，下文你将看到：为寻最大间隔，导出 $1/2||w||^2$!!!!!!!!!!!!!!!!!!!!，继而引入拉格朗日函数和对偶变量 $a$，化为对单一因数对偶变量 $a$ 的求解，当然，这是后话），从而确定最终的最大间隔分类超平面 hyper plane 和分类函数；
3. 进而把寻求分类函数 $f(x)=w\cdot x+b$ 的问题转化为对 $w$，$b$ 的最优化问题，最终化为对偶因子的求解。

总结成一句话即是：从最大间隔出发（目的本就是为了确定法向量 $w$），转化为求对变量 $w$ 和 $b$ 的凸二次规划问题。亦或如下图所示（有点需要注意，如读者 @ 酱爆小八爪所说：从最大分类间隔开始，就一直是凸优化问题）：



## 1.4 函数间隔 Functional margin 与几何间隔 Geometrical margin

一般而言，一个点距离超平面的远近可以表示为分类预测的确信或准确程度。

- 在超平面 $w*x+b=0$ 确定的情况下，$|w*x+b|$ 能够相对的表示点 x 到距离超平面的远近，而 $w*x+b$ 的符号与类标记 $y$ 的符号是否一致表示分类是否正确，所以，可以用量 $y*(w*x+b)$ 的正负性来判定或表示分类的正确性和确信度。

于此，我们便引出了定义样本到分类间隔距离的函数间隔 functional margin 的概念。

### 1.4.1 函数间隔 Functional margin

我们定义函数间隔 functional margin 为：

$$\widehat{\gamma} = y(w^T x + b) = yf(x) \tag{7}$$

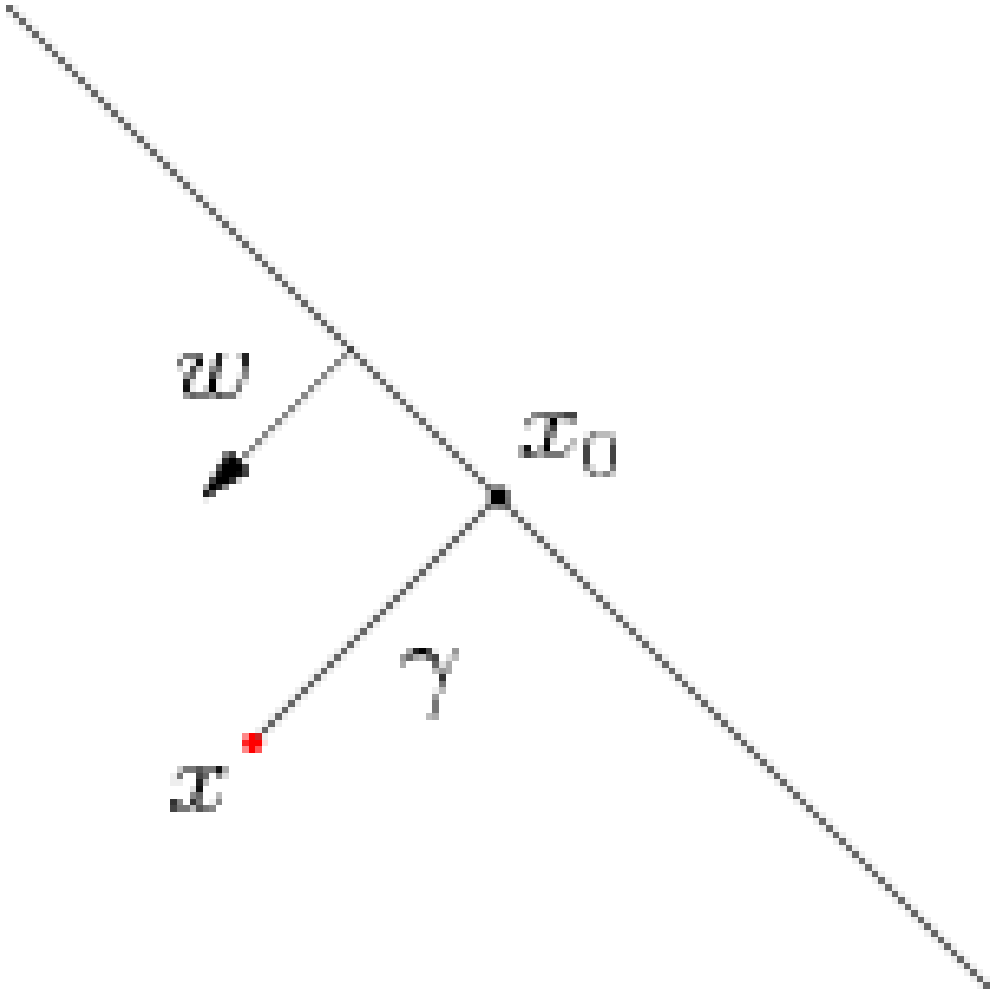
接着，我们定义超平面 $(w,b)$ 关于训练数据集 T 的函数间隔为超平面 $(w,b)$ 关于 T 中所有样本点 $(x_i,y_i)$ 的函数间隔最小值，其中，$x$ 是特征，$y$ 是结果标签，$i$ 表示第 $i$ 个样本，有：

$$\widehat{\gamma} = \min \widehat{\gamma}_i (i=1,\cdots,n) \tag{8}$$

然与此同时，问题就出来了。上述定义的函数间隔虽然可以表示分类预测的正确性和确信度，但在选择分类超平面时，只有函数间隔还远远不够，因为如果成比例的改变 $w$ 和 $b$，如将他们改变为 $2w$ 和 $2b$，虽然此时超平面没有改变，但函数间隔的值 $f(x)$ 却变成了原来的 2 倍。

其实，我们可以对法向量 $w$ 加些约束条件，使其表面上看起来规范化，如此，我们很快又将引出真正定义点到超平面的距离 --几何间隔 geometrical margin 的概念（很快你将看到，几何间隔就是函数间隔除以个 $\|w\|$，即 $yf(x)/\|w\|$）。

### 1.4.2 点到超平面的距离定义：几何间隔 Geometrical margin




在给出几何间隔的定义之前，咱们首先来看下，如上图所示，对于一个点 $x$，令其垂直投影到超平面上的对应的为 $x_0$，由于 $w$ 是垂直于超平面的一个向量，$\gamma$ 为样本 $x$ 到分类间隔的距离，我们有

$$x = x_0 + \gamma \frac{w}{\|w\|} \tag{9}$$

又由于 $x_0$ 是超平面上的点，满足 $f(x_0) = 0$，代入超平面的方程即可算出：

$$\gamma = \frac{w^T x + b}{\|w\|} = \frac{f(x)}{\|w\|} \tag{10}$$

（有的书上会写成把 $\|w\|$ 分开相除的形式，如本文参考文献及推荐阅读条目 11，其中，$\|w\|$ 为 $w$ 的二阶泛数）

不过这里的 $\gamma$ 是带符号的，我们需要的只是它的绝对值，因此类似地，也乘上对应的类别 $y$ 即可，因此实际上我们定义几何间隔 geometrical margin 为 (注：别忘了，上面 $\widehat{\gamma}$ 的定义，$\widehat{\gamma} = y(w^T x + b) = yf(x)$ )：

$$\widetilde{\gamma} = y\gamma = \frac{\widehat{\gamma}}{\|w\|} \tag{11}$$

（代入相关式子可以得出：yi*(w/||w|| + b/||w||)）

正如本文评论下读者 popol1991 留言：函数间隔 $y*(wx+b)=y*f(x)$ 实际上就是 $|f(x)|$，只是人为定义的一个间隔度量；而几何间隔 $|f(x)|/||w||$ 才是直观上的点到超平面距离。

想想二维空间里的点到直线公式：假设一条直线的方程为 $ax+by+c=0$, 点 $P$ 的坐标是 $(x_0,y_0)$，则点到直线距离为 $|ax_0+by_0+c|/sqrt(a^2+b^2)$。如下图所示：

点到平面的距离

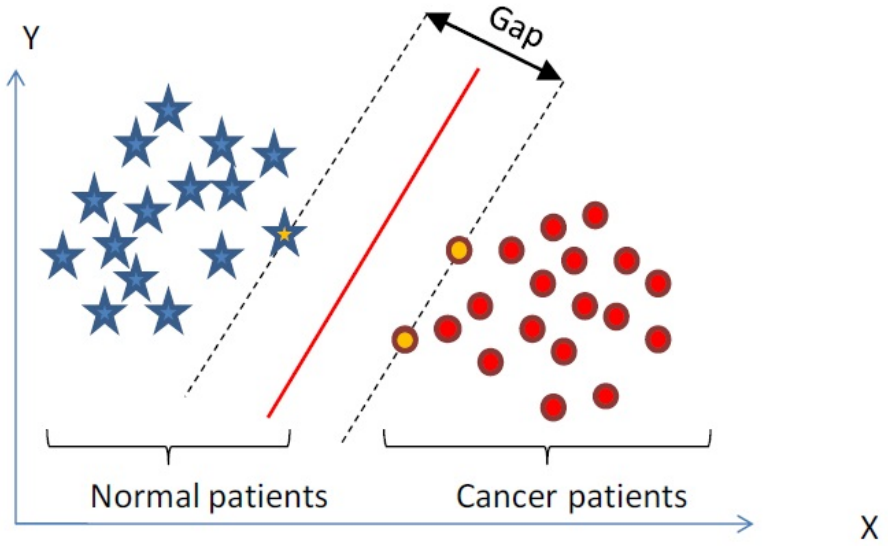
若点坐标为 $(x_0.y_0.z_0)$，平面为 $Ax+By+Cz+D=0$，则点到平面的距离为：

$$d=\left|\frac{Ax_0+By_0+Cz_0+D}{\sqrt{A^2+B^2+C^2}}\right| \tag{12}$$

那么如果用向量表示，设 $w=(a,b)$,$f(x)=wx+c$, 那么这个距离正是 $|f(p)|/||w||$。

## 1.5 最大间隔分类器 Maximum Margin Classifier 的定义

于此，我们已经很明显的看出，函数间隔 functional margin 和几何间隔 geometrical margin 相差一个的缩放因子。按照我们前面的分析，对一个数据点进行分类，当它的 margin 越大的时候，分类的 confidence 越大。对于一个包含 $n$ 个点的数据集，我们可以很自然地定义它的 margin 为所有这 $n$ 个点的 margin 值中最小的那个。于是，为了使得分类的 confidence 高，我们希望所选择的超平面 hyper plane 能够最大化这个 margin 值。




通过上节，我们已经知道：

1、functional margin 明显是不太适合用来最大化的一个量，因为在 hyper plane 固定以后，我们可以等比例地缩放 $w$ 的长度和 $b$ 的值，这样可以使得 $f(x)=w^Tx+b$ 的值任意大，亦即 functional margin$\hat{\gamma}$ 可以在 hyper plane 保持不变的情况下被取得任意大，

2、而 geometrical margin 则没有这个问题，因为除上了 $||w||$ 这个分母，所以缩放 $w$ 和 $b$ 的时候的值是不会改变的，它只随着 hyper plane 的变动而变动，因此，这是更加合适的一个 margin 。

这样一来，我们的 maximum margin classifier 的目标函数可以定义为：

$$\max \tilde{\gamma} \tag{13}$$

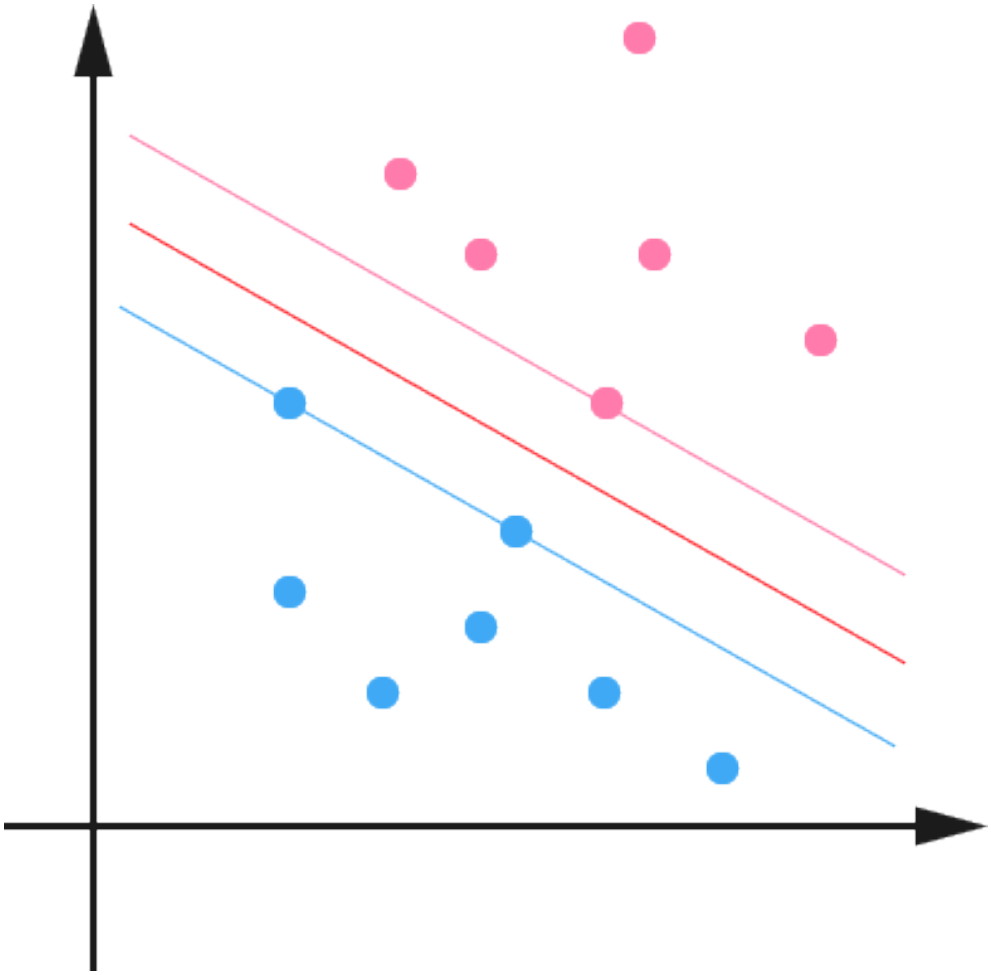
当然，还需要满足一些条件，根据 margin 的定义，我们有

$$y_i(w^T x_i + b) = \hat{\gamma}_i \geq \hat{\gamma}, i = 1, \cdots, n \tag{14}$$

其中 $\hat{\gamma} = \tilde{\gamma}\|w\|$ (等价于 $\tilde{\gamma} = \gamma/\|w\|$，故有稍后的 $\hat{\gamma} = 1$ 时，$\tilde{\gamma} = 1/\|w\|$)，处于方便推导和优化的目的，我们可以令 $\hat{\gamma} = 1$(对目标函数的优化没有影响，至于为什么，请见本文评论下第 42 楼回复)，此时，上述的目标函数 $\tilde{\gamma}$ 转化为 (其中，s.t.，即 subject to 的意思，它导出的是约束条件)：

$$\max \frac{1}{\|w\|}, w.t., y_i(w^T x_i + b) \geq 1, i = 1, \cdots, n \tag{15}$$

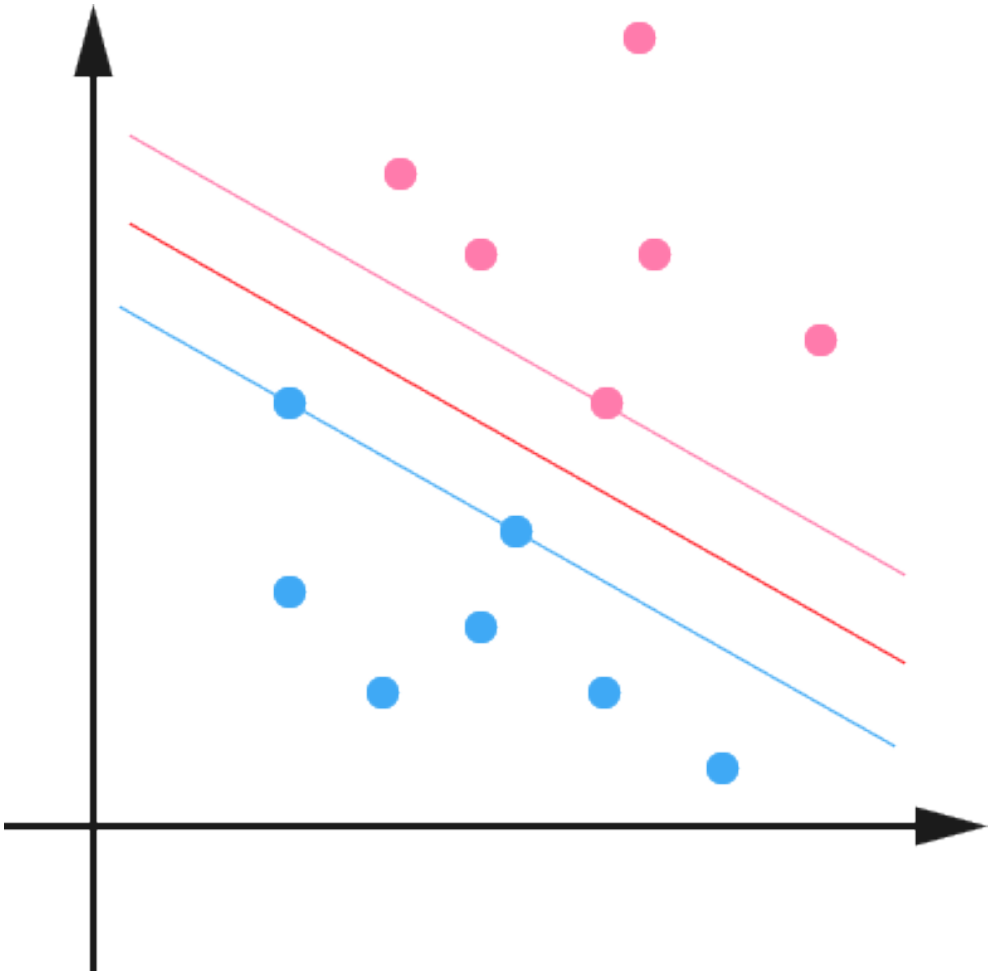
通过求解这个问题，我们就可以找到一个 margin 最大的 classifier，如下图所示，中间的红色线条是 Optimal Hyper Plane，另外两条线到红线的距离都是等于 $\tilde{\gamma}$ 的 ($\tilde{\gamma}$ 便是上文所定义的 geometrical margin，当令 $\hat{\gamma} = 1$ 时，$\tilde{\gamma}$ 便为 $1/\|w\|$，而我们上面得到的目标函数便是在相应的约束条件下，要最大化这个 $1/\|w\|$ 值)：



通过最大化 margin，我们使得该分类器对数据进行分类时具有了最大的 confidence，从而设计决策最优分类超平面。

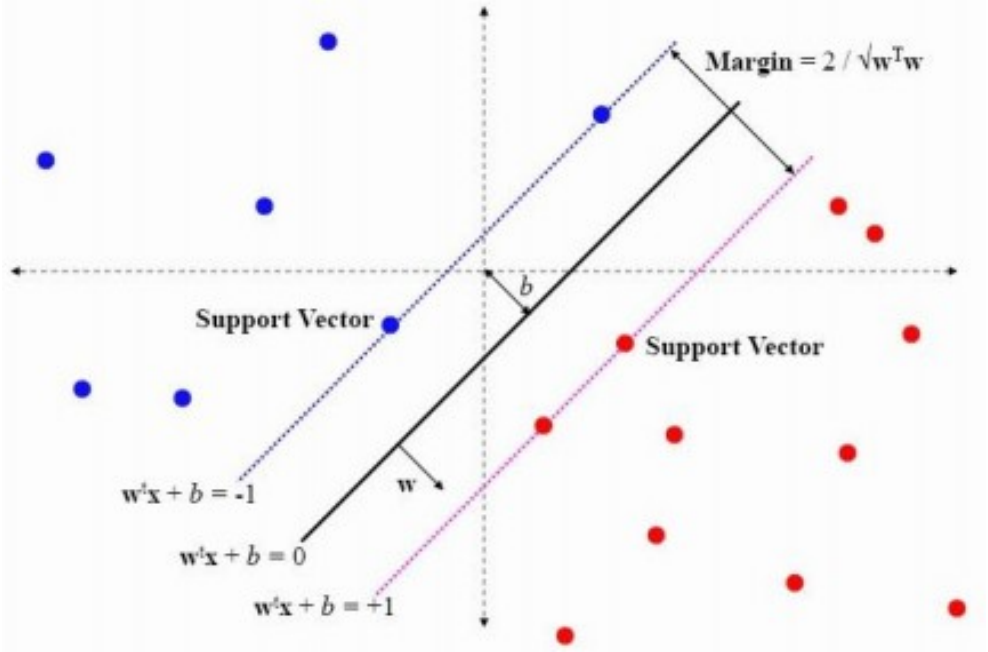
## 1.6 到底什么是 Support Vector

上节，我们介绍了 Maximum Margin Classifier，但并没有具体阐述到底什么是 Support Vector，本节，咱们来重点阐述这个概念。咱们不妨先来回忆一下上节 1.4 节最后一张图：



可以看到两个支撑着中间的 gap 的超平面，它们到中间的纯红线 separating hyper plane 的距离相等，即我们所能得到的最大的 geometrical margin$\tilde{\gamma}$，而“支撑”这两个超平面的必定会有一些点，而这些“支撑”的点便叫做支持向量 Support Vector。

或亦可看下来自此 PPT 中的一张图，Support Vector 便是那蓝色虚线和粉红色虚线上的点：




很显然，由于这些 supporting vector 刚好在边界上，所以它们满足 $y(w^Tx+b)=1$（还记得我们把 functional margin 定为 1 了吗？上节中：“处于方便推导和优化的目的，我们可以令 $\hat{\gamma}=1$”），而对于所有不是支持向量的点，也就是在“阵地后方”的点，则显然 $y(w^Tx+b)>1$ 有。当然，除了从几何直观上之外，支持向量的概念也可以从下文优化过程的推导中得到。

OK，到此为止，算是了解到了 SVM 的第一层，对于那些只关心怎么用 SVM 的朋友便已足够，不必再更进一层深究其更深的原理。

# 2 深入 SVM

## 2.1 从线性可分到线性不可分

### 2.1.1 从原始问题到对偶问题的求解

虽然上文 1.4 节给出了目标函数，却没有讲怎么来求解。现在就让我们来处理这个问题。回忆一下之前得到的目标函数（subject to 导出的则是约束条件）：

$$\max \frac{1}{\|w\|} \; s.t., y_i(w^Tx_i+b)\geq 1, i=1,\cdots,n \tag{16}$$

由于求 $\frac{1}{\|w\|}$ 的最大值相当于求 $\frac{1}{2}\|w\|^2$ 的最小值，所以上述问题等价于（$w$ 由分母变成分子，从而也有原来的 max 问题变为 min 问题，很明显，两者问题等价）：

$$\min \frac{1}{2}\|w\|^2 \; s.t., y_i(w^Tx_i+b)\geq 1, i=1,\cdots,n \tag{17}$$

- 转化到这个形式后，我们的问题成为了一个凸优化问题，或者更具体的说，因为现在的目标函数是二次的，约束条件是线性的，所以它是一个凸二次规划问题。这个问题可以用任何现成的 QP (Quadratic Programming) 的优化包进行求解，归结为一句话即是：在一定的约束条件下，目标最优，损失最小；
- 但虽然这个问题确实是一个标准的 QP 问题，但是它也有它的特殊结构，通过 Lagrange Duality 变换到对偶变量 (dual variable) 的优化问题之后，可以找到一种更

加有效的方法来进行求解，而且通常情况下这种方法比直接使用通用的 QP 优化包进行优化要高效得多。

也就说，除了用解决 QP 问题的常规方法之外，还可以通过求解对偶问题得到最优解，这就是线性可分条件下支持向量机的对偶算法，这样做的优点在于：一者对偶问题往往更容易求解；二者可以自然的引入核函数，进而推广到非线性分类问题。

至于上述提到，关于什么是 Lagrange duality？简单地来说，通过给每一个约束条件加上一个 Lagrange multiplier(拉格朗日乘值)，即引入拉格朗日乘子 $\alpha$ 如此我们便可以通过拉格朗日函数将约束条件融和到目标函数里去 (也就是说把条件融合到一个函数里头，现在只用一个函数表达式便能清楚的表达出我们的问题)：

$$\mathcal{L}(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i\left(y_i(w^Tx_i+b)-1\right) \tag{18}$$

然后我们令

$$\theta(w)=\max_{\alpha_i\geq 0}\mathcal{L}(w,b,\alpha) \tag{19}$$

容易验证，当某个约束条件不满足时，例如 $y_i(w^Tx_i+b)<1$，那么我们显然有 $\theta(w)=\inf$（只要令 $\alpha_i=\inf$ 即可）。而当所有约束条件都满足时，则有 $\theta(w)=\frac{1}{2}\|w\|^2$，亦即我们最初要最小化的量。因此，在要求约束条件得到满足的情况下最小化 $\frac{1}{2}\|w\|^2$，实际上等价于直接最小化 $\theta(w)$（当然，这里也有约束条件，就是 $\alpha_i\geq 0,i=1,\cdots,n$），因为如果约束条件没有得到满足，$\theta w$ 等于无穷大，自然不会是我们所要求的最小值。具体写出来，我们现在的目标函数变成了：

$$\min_{w,b}\theta(w)=\min_{w,b}\max_{\alpha_i\geq 0}\mathcal{L}(w,b,\alpha)=p^* \tag{20}$$

这里用 $p^*$ 表示这个问题的最优值，这个问题和我们最初的问题是等价的。不过，现在我们来把最小和最大的位置交换一下（稍后，你将看到，当下面式子满足了一定的条件之后，这个式子 $d$ 便是上式 $P$ 的对偶形式表示）：

$$\max_{\alpha_i\geq 0}\min_{w,b}\mathcal{L}(w,b,\alpha)=d^* \tag{21}$$

当然，交换以后的问题不再等价于原问题，这个新问题的最优值用 $d^*$ 来表示。并且，我们有 $d^*\leq p^*$，这在直观上也不难理解，最大值中最小的一个总也比最小值中最大的一个要大吧！总之，第二个问题的最优值 $d^*$ 在这里提供了一个第一个问题的最优值 $p^*$ 的一个下界，在满足某些条件的情况下，这两者相等，这个时候我们就可以通过求解第二个问题来间接地求解第一个问题。

也就是说，下面我们将先求 $L$ 对 $w$、$b$ 的极小，再求 $L$ 对 $\alpha$ 的极大。而且，之所以从 min max 的原始问题 $p^*$，转化为 max min 的对偶问题 $d^*$，一者因为 $d^*$ 是 $p^*$ 的近似解，二者，转化为对偶问题后，更容易求解。

### 2.1.2 KKT 条件

与此同时，上段说“在满足某些条件的情况下”，这所谓的“满足某些条件”就是要满足 KKT 条件。那 KKT 条件的表现形式是什么呢？据维基百科：KKT 条件的介绍，一般地，一个最优化数学模型能够表示成下列标准形式：

$$\begin{aligned}&\min .f(x)\\ &s.t.h_j(x)=0,j=1,\cdots,p,\\ &\quad g_k(x)\leq 0,k=1,\cdots,q,\\ &x\in X\subset\Re^n\end{aligned} \tag{22}$$

其中，$f(x)$ 是需要最小化的函数，$h(x)$ 是等式约束，$g(x)$ 是不等式约束，$p$ 和 $q$ 分别为等式约束和不等式约束的数量。同时，我们得明白以下两个定理：

- 凸优化的概念：$\mathcal{X} \subset \mathbb{R}^n$ 为一凸集，$f : \mathcal{X} \to \mathbb{R}$ 为一凸函数。凸优化就是要找出一点 $x^* \in \mathcal{X}$ ，使得每一 $x \in \mathcal{X}$ 满足 $f(x^*) \leq f(x)$。
- KKT 条件的意义：它是一个非线性规划（Nonlinear Programming）问题能有最优化解法的必要和充分条件。

那到底什么是所谓 Karush-Kuhn-Tucker 条件呢？KKT 条件就是指上面最优化数学模型的标准形式中的最小点 $x^*$ 必须满足下面的条件：

$$\begin{aligned} &1.\ h_j(x_*) = 0, j = 1, \cdots, p, g_k(x_*) \leq 0, k = 1, \cdots, q, \\ &2.\ \nabla f(x_*) + \sum_{j=1}^{p} \lambda_j \nabla h_j(x_*) + \sum_{k=1}^{q} \mu_k \nabla g_k(x_*) = 0, \\ &\qquad\qquad \lambda \neq 0 m \mu_k \geq 0, \mu_k g_k(x_*) = 0 \end{aligned} \tag{23}$$

经过论证，我们这里的问题是满足 KKT 条件的（首先已经满足 Slater condition，再者 $f$ 和 $g_i$ 也都是可微的，即 $L$ 对 $w$ 和 $b$ 都可导），因此现在我们便转化为求解第二个问题。也就是说，现在，咱们的原问题通过满足一定的条件，已经转化成了对偶问题。而求解这个对偶学习问题，分为 3 个步骤，首先要让 L(w，b，a) 关于 w 和 b 最小化，然后求对 α 的极大，最后利用 SMO 算法求解对偶因子。

### 2.1.3 对偶问题求解的 3 个步骤

（1）、首先固定，要让 $L$ 关于 $w$ 和 $b$ 最小化，我们分别对 $w$，$b$ 求偏导数，即令 $\partial \mathcal{L}/\partial w$ 和 $\partial \mathcal{L}/\partial b$ 等于零（对 $w$ 求导结果的解释请看本文评论下第 45 楼回复）：

$$\begin{aligned} \frac{\mathcal{L}}{\partial w} &= \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \frac{\mathcal{L}}{\partial b} &= \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{24}$$

以上结果代回上述的 $\mathcal{L}$：

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left( y_i (w^T x_i + b) - 1 \right) \tag{25}$$

得到：

$$\begin{aligned} \mathcal{L}(w, b, \alpha) &= \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \end{aligned} \tag{26}$$

提醒：有读者可能会问上述推导过程如何而来？说实话，其具体推导过程是比较复杂的，如下图所示：

$$
\begin{aligned}
\mathcal{L}(w,b,\alpha) &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m}\alpha_i[y^{(i)}(w^Tx^{(i)}+b)-1] \\
&= \frac{1}{2}w^Tw - \sum_{i=1}^{m}\alpha_i y^{(i)}w^Tx^{(i)} - \sum_{i=1}^{m}\alpha_i y^{(i)}b + \sum_{i=1}^{m}\alpha_i \\
&= \frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} - \sum_{i=1}^{m}\alpha_i y^{(i)}w^Tx^{(i)} - \sum_{i=1}^{m}\alpha_i y^{(i)}b + \sum_{i=1}^{m}\alpha_i \\
&= \frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} - w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} - \sum_{i=1}^{m}\alpha_i y^{(i)}b + \sum_{i=1}^{m}\alpha_i \\
&= -\frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} - \sum_{i=1}^{m}\alpha_i y^{(i)}b + \sum_{i=1}^{m}\alpha_i \\
&= -\frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} - b\sum_{i=1}^{m}\alpha_i y^{(i)} + \sum_{i=1}^{m}\alpha_i \\
&= -\frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}\right)^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} - b\sum_{i=1}^{m}\alpha_i y^{(i)} + \sum_{i=1}^{m}\alpha_i \\
&= -\frac{1}{2}\sum_{i=1}^{m}\alpha_i y^{(i)}\left(x^{(i)}\right)^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} - b\sum_{i=1}^{m}\alpha_i y^{(i)} + \sum_{i=1}^{m}\alpha_i \\
&= -\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i y^{(i)}\left(x^{(i)}\right)^T\alpha_j y^{(j)}x^{(j)} - b\sum_{i=1}^{m}\alpha_i y^{(i)} + \sum_{i=1}^{m}\alpha_i
\end{aligned}
\tag{27}
$$

最后，得到：

$$
\begin{aligned}
\mathcal{L}(w,b,\alpha) &= \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^Tx_j - \sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^Tx_j - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i \\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^Tx_j
\end{aligned}
\tag{28}
$$

如 jerrylead 所说：“倒数第 4 步”推导到“倒数第 3 步”使用了线性代数的转置运算，由于 ai 和 yi 都是实数，因此转置后与自身一样。“倒数第 3 步”推导到“倒数第 2 步”使 $(a+b+c+\ldots)(a+b+c+\cdots) = aa+ab+ac+ba+bb+bc+\cdots$ 的乘法运算法则。最后一步是上一步的顺序调整。

从上面的最后一个式子，我们可以看出，此时的拉格朗日函数只包含了一个变量，那就是 $\alpha_i$，然后下文的第 2 步，求出了 $\alpha_i$ 便能求出 $w$，和 $b$，由此可见，上文第 1.2 节提出来的核心问题：分类函数 $f(x) = w^Tx + b$ 也就可以轻而易举的求出来了。

（2）、求对 $\alpha$ 的极大，即是关于对偶问题的最优化问题，从上面的式子得到：

(不得不提醒下读者：经过上面第一个步骤的求 $w$ 和 $b$，得到的拉格朗日函数式子已经没有了变量 $w$，$b$，只有 $\alpha$，而反过来，求得的 $\alpha$ 将能导出 $w$，$b$ 的解，最终得出分离超平面和分类决策函数。为何呢？因为如果求出了 $\alpha_i$，根据 $w = \sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}$，即

可求出 $w$。然后通过 $b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T}x^{(i)} + \min_{i:y^{(i)}=1} w^{*T}x^{(i)}}{2}$，即可求出 $b$ )

$$\begin{aligned} \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ s.t, \alpha_i \geq 0, i = 1, \cdots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{29}$$

如前面所说，这个问题有更加高效的优化算法，即我们常说的 SMO 算法。

### 2.1.4 序列最小最优化 SMO 算法

细心的读者读至上节末尾处，怎么拉格朗日乘子的值可能依然心存疑惑。实际上，关于的求解可以用一种快速学习算法即 SMO 算法，这里先简要介绍下。

OK，当：

$$\begin{aligned} \max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ s.t.\ 0 \leq \alpha_i \leq C, i = 1, \cdots, m \\ \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \end{aligned} \tag{30}$$

要解决的是在参数 $\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ 上求最大值 $W$ 的问题，至于 $x^{(i)}$ 和 $y^{(i)}$ 都是已知数（其中 $C$ 是一个参数，用于控制目标函数中两项（“寻找 margin 最大的超平面”和“保证数据点偏差量最小”）之间的权重。和上文最后的式子对比一下，可以看到唯一的区别就是现在 dual variablei$\alpha$ 多了一个上限 $C$，关于 $C$ 的具体由来请查看下文第 2.3 节）。

要了解这个 SMO 算法是如何推导的，请跳到下文第 3.5 节、SMO 算法。

### 2.1.5 线性不可分的情况

OK，为过渡到下节 2.2 节所介绍的核函数，让我们再来看看上述推导过程中得到的一些有趣的形式。首先就是关于我们的 hyper plane ，对于一个数据点 $x$ 进行分类，实际上是通过把 $x$ 带入到 $f(x) = w^T x + b$ 算出结果然后根据其正负号来进行类别划分的。而前面的推导中我们得到

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{31}$$

因此分类函数为：

$$\begin{aligned} f(x) &= \left( \sum_{i=1}^{n} \alpha_i y_i x_i \right)^T x + b \\ &= \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \end{aligned} \tag{32}$$

这里的形式的有趣之处在于，对于新点 $x$ 的预测，只需要计算它与训练数据点的内积即可（$\langle *, * \rangle$ 表示向量内积），这一点至关重要，是之后使用 Kernel 进行非线性推广的基本前提。此外，所谓 Supporting Vector 也在这里显示出来——事实上，所有非 Supporting

Vector 所对应的系 $\alpha$ 数都是等于零的，因此对于新点的内积计算实际上只要针对少量的“支持向量”而不是所有的训练数据即可。

为什么非支持向量对应的 $\alpha$ 等于零呢？直观上来理解的话，就是这些“后方”的点——正如我们之前分析过的一样，对超平面是没有影响的，由于分类完全有超平面决定，所以这些无关的点并不会参与分类问题的计算，因而也就不会产生任何影响了。

回忆一下我们 2.1.1 节中通过 Lagrange multiplier 得到的目标函数：

$$\max_{\alpha_i \geq 0} \mathcal{L}(w, b, \alpha) = \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(w^T x_i + b) - 1\right) \tag{33}$$

注意到如果 $x_i$ 是支持向量的话，上式中红颜色的部分是等于 0 的（因为支持向量的 functional margin 等于 1 ），而对于非支持向量来说，functional margin 会大于 1 ，因此红颜色部分是大于零的，而 $\alpha_i$ 又是非负的，为了满足最大化，$\alpha_i$ 必须等于 0 。这也就是这些非 Supporting Vector 的点的局限性。

从 1.5 节到上述所有这些东西，便得到了一个 maximum margin hyper plane classifier，这就是所谓的支持向量机（Support Vector Machine）。当然，到目前为止，我们的 SVM 还比较弱，只能处理线性的情况，不过，在得到了对偶 dual 形式之后，通过 Kernel 推广到非线性的情况就变成了一件非常容易的事情了 (相信，你还记得本节开头所说的：“通过求解对偶问题得到最优解，这就是线性可分条件下支持向量机的对偶算法，这样做的优点在于：一者对偶问题往往更容易求解；二者可以自然的引入核函数，进而推广到非线性分类问题”)。

## 2.2 核函数 Kernel

### 2.2.1 特征空间的隐式映射：核函数

咱们首先给出核函数的来头：

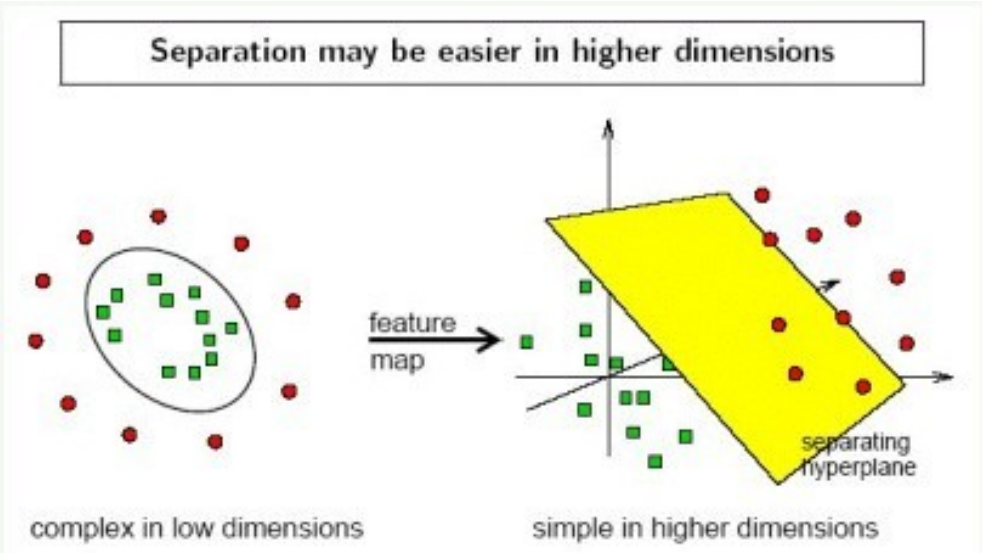
- 在上文中，我们已经了解到了 SVM 处理线性可分的情况，而对于非线性的情况，SVM 的处理方法是选择一个核函数 $k(*,*)$ ，通过将数据映射到高维空间，来解决在原始空间中线性不可分的问题。由于核函数的优良品质，这样的非线性扩展在计算量上并没有比原来复杂多少，这一点是非常难得的。当然，这要归功于核方法——除了 SVM 之外，任何将计算表示为数据点的内积的方法，都可以使用核方法进行非线性扩展。

也就是说，Minsky 和 Papert 早就在 20 世纪 60 年代就已经明确指出线性学习器计算能力有限。为什么呢？因为总体上来讲，现实世界复杂的应用需要有比线性函数更富有表达能力的假设空间，也就是说，目标概念通常不能由给定属性的简单线性函数组合产生，而是应该一般地寻找待研究数据的更为一般化的抽象特征。

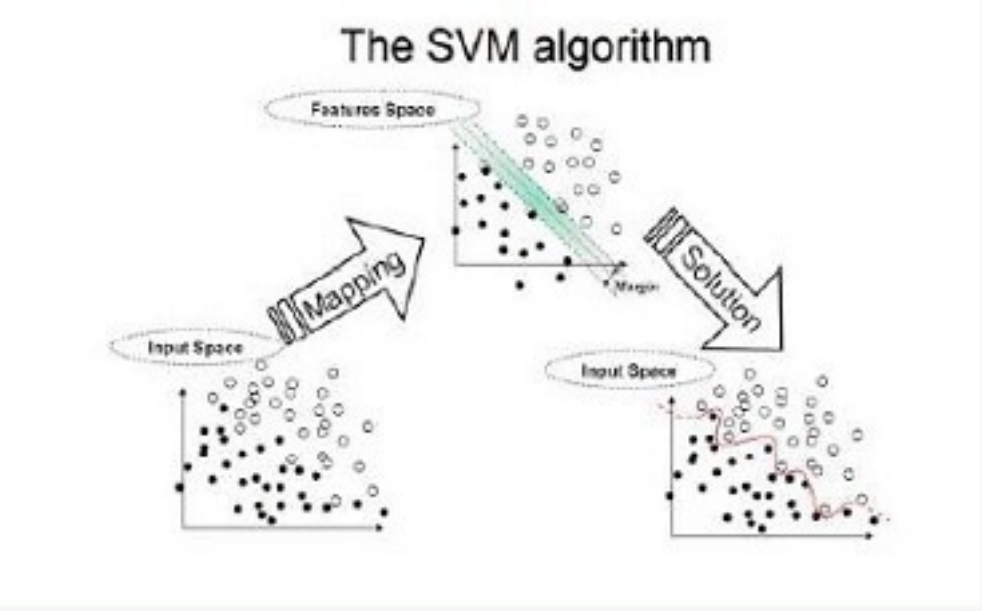
而下文我们将具体介绍的核函数则提供了此种问题的解决途径，从下文你将看到，核函数通过把数据映射到高维空间来增加第一节所述的线性学习器的能力，使得线性学习器对偶空间的表达方式让分类操作更具灵活性和可操作性。因为训练样例一般是不会独立出现的，它们总是以成对样例的内积形式出现，而用对偶形式表示学习器的优势在为在该表示中可调参数的个数不依赖输入属性的个数，通过使用恰当的核函数来替代内积，可以隐式得将非线性的训练数据映射到高维空间，而不增加可调参数的个数 (当然，前提是核函数能够计算对应着两个输入特征向量的内积)。

1、简而言之：在线性不可分的情况下，支持向量机通过某种事先选择的非线性映射 (核函数) 将输入变量映射到一个高维特征空间，在这个空间中构造最优分类超平面。我们使用 SVM 进行数据集分类工作的过程首先是同预先选定的一些非线性映射将输入

空间映射到高维特征空间 (下图很清晰的表达了通过映射到高维特征空间，而把平面上本身不好分的非线性数据分了开来)：




使得在高维属性空间中有可能最训练数据实现超平面的分割，避免了在原输入空间中进行非线性曲面分割计算。SVM 数据集形成的分类函数具有这样的性质：它是一组以支持向量为参数的非线性函数的线性组合，因此分类函数的表达式仅和支持向量的数量有关，而独立于空间的维度，在处理高维输入空间的分类时，这种方法尤其有效，其工作原理如下图所示：




2、具体点说：在我们遇到核函数之前，如果用原始的方法，那么在用线性学习器学习一个非线性关系，需要选择一个非线性特征集，并且将数据写成新的表达形式，这等价于应用一个固定的非线性映射，将数据映射到特征空间，在特征空间中使用线性学习器，因此，考虑的假设集是这种类型的函数：

$$f(x) = \sum_{i=1}^{N} w_i \phi_i(x) + b \tag{34}$$

这里 $\phi : x \to f$ 是从输入空间到某个特征空间的映射，这意味着建立非线性学习器分为两步：

1. 首先使用一个非线性映射将数据变换到一个特征空间 F，

2. 然后在特征空间使用线性学习器分类。

在上文我提到过对偶形式，而这个对偶形式就是线性学习器的一个重要性质，这意味着假设可以表达为训练点的线性组合，因此决策规则可以用测试点和训练点的内积来表示：

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i \langle \phi(x_i) \cdot \phi(x) \rangle + b \tag{35}$$

如果有一种方式可以在特征空间中直接计算内积 $\langle \phi(x_i) \cdot \phi(x) \rangle$，就像在原始输入点的函数中一样，就有可能将两个步骤融合到一起建立一个非线性的学习器，这样直接计算法的方法称为核函数方法，于是，核函数便横空出世了。)

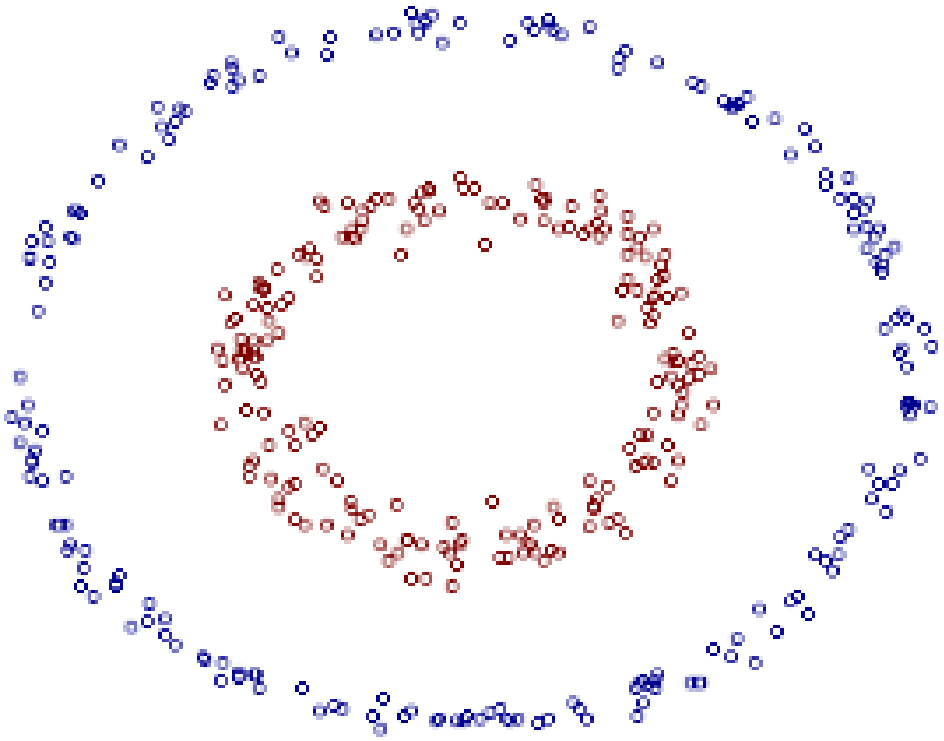
这里我直接给出一个定义：核是一个函数 $K$，对所有 $x$，$z \in X$，满足 $K(x, z) = \langle \phi(x) \cdot \phi(z) \rangle$，这里 $\phi$ 是从 $X$ 到内积特征空间 $F$ 的映射。)

3、总而言之，举个简单直接点的例子，如 @Wind 所说：如果不是用核技术，就会先计算线性映射 $\phi(x_1)$ 和 $\phi(x_2)$, 然后计算这两个特征的内积，使用了核技术之后，先把 $\phi(x_1)$ 和 $\phi(x_2)$ 的通用表达式子：$\langle \phi(x_1), \phi(x_2) \rangle \geq K(\langle x_1, x_2 \rangle)$ 计算出来，注意到这里的 $\langle *, * \rangle$ 表示内积，$K(*, *)$ 就是对应的核函数，这个表达往往非常简单，所以计算非常方便。

OK，接下来，咱们就进一步从外到里，来探探这个核函数的真面目。

### 2.2.2 核函数：如何处理非线性数据

在 2.1 节中我们介绍了线性情况下的支持向量机，它通过寻找一个线性的超平面来达到对数据进行分类的目的。不过，由于是线性方法，所以对非线性的数据就没有办法处理。举个例子来说，则是如下图所示的两类数据，分别分布为两个圆圈的形状，这样的数据本身就是线性不可分的，此时咱们该如何把这两类数据分开呢 (下文将会有一个相应的三维空间图)?



事实上，上图所述的这个数据集，是用两个半径不同的圆圈加上了少量的噪音生成得到的，所以，一个理想的分界应该是一个“圆圈”而不是一条线（超平面）。如果用 $X_1$ 和 $X_2$ 来表示这个二维平面的两个坐标的话，我们知道一条二次曲线（圆圈是二次曲线的一种特殊情况）的方程可以写作这样的形式：

$$a_1X_1 + a_2X_1^2 + a_3X_2 + a_4X_2^2 + a_5X_1X_2 + a_6 = 0 \tag{36}$$

注意上面的形式，如果我们构造另外一个五维的空间，其中五个坐标的值分别为 $Z_1 = X_1, Z_2 = X_1^2, Z_3 = X_2, Z_4 = X_2^2, Z_5 = X_1X_2$，那么显然，上面的方程在新的坐标系下可以写作：

$$\sum_{i=1}^{5} a_iZ_i + a_6 = 0 \tag{37}$$

于新的坐标 $Z$，这正是一个 hyper plane 的方程！也就是说，如果我们做一个映射 $\phi : \mathbb{R}^2 \to \mathbb{R}^5$，将 $X$ 按照上面的规则映射为 $Z$，那么在新的空间中原来的数据将变成线性可分的，从而使用之前我们推导的线性分类算法就可以进行处理了。这正是 Kernel 方法处理非线性问题的基本思想。

再进一步描述 Kernel 的细节之前，不妨再来看看这个例子映射过后的直观例子。当然，你我可能无法把 5 维空间画出来，不过由于我这里生成数据的时候就是用了特殊的情形，具体来说，我这里的超平面实际的方程是这个样子（圆心在 $X_2$ 轴上的一个正圆）：

$$\sum_{i=1}^{5} a_iZ_i + a_6 = 0 Ithinkthisequationiswrong \tag{38}$$

因此我只需要把它映射到 $Z_1 = X_1^2, Z_2 = X_2^2, Z_3 = X_2$ 这样一个三维空间中即可，下图即是映射之后的结果，将坐标轴经过适当的旋转，就可以很明显地看出，数据是可

以通过一个平面来分开的 (pluskid：下面的 gif 动画，先用 Matlab 画出一张张图片，再用 Imagemagick 拼贴成)：

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

现在让我们再回到 SVM 的情形，假设原始的数据时非线性的，我们通过一个映射 $\phi(\cdot)$ 将其映射到一个高维空间中，数据变得线性可分了，这个时候，我们就可以使用原来的推导来进行计算，只是所有的推导现在是在新的空间，而不是原始空间中进行。当然，推导过程也并不是可以简单地直接类比的，例如，原本我们要求超平面的法向量 $w$，但是如果映射之后得到的新空间的维度是无穷维的（确实会出现这样的情况，比如后面会提到的高斯核 Gaussian Kernel），要表示一个无穷维的向量描述起来就比较麻烦。于是我们不妨先忽略过这些细节，直接从最终的结论来分析，回忆一下，我们上一次 2.1 节中得到的最终的分类函数是这样的：

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \tag{39}$$

在则是在映射过后的空间，即：

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle + b \tag{40}$$

而其中的 $\alpha$ 也是通过求解如下 dual 问题而得到的：

$$\begin{aligned} \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle \\ s.t. \ , \alpha_i \geq 0, i = 1, \cdots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{41}$$

这样一来问题就解决了吗？似乎是的：拿到非线性数据，就找一个映射 $\phi(\cdot)$，然后一股脑把原来的数据映射到新空间中，再做线性 SVM 即可。不过事实上没有这么简单！其实刚才的方法稍想一下就会发现有问题：在最初的例子里，我们对一个二维空间做映射，选择的新空间是原始空间的所有一阶和二阶的组合，得到了五个维度；如果原始空间是三维，那么我们会得到 19 维的新空间，这个数目是呈爆炸性增长的，这给 $\phi(\cdot)$ 的计算带来了非常大的困难，而且如果遇到无穷维的情况，就根本无从计算了。所以就需要 Kernel 出马了。

不妨还是从最开始的简单例子出发，设两个向量 $x_1 = (\eta_1, \eta_2)^T$ 和 $x_2 = (\xi_1, \xi_2)^T$，而 $\phi(\cdot)$ 即是到前面 2.2.1 节说的五维空间的映射，因此映射过后的内积为：

$$\langle \phi(x_1), \phi(x_2) \rangle = \eta_1 \xi_1 + \eta_1^2 \xi_1^2 + \eta_2 \xi_2 + \eta_2^2 \xi_2^2 + \eta_1 \eta_2 \xi_1 \xi_2 \tag{42}$$

（公式说明：上面的这两个推导过程中，所说的前面的五维空间的映射，这里说的前面便是文中 2.2.1 节的所述的映射方式，仔细看下 2.2.1 节的映射规则，再看那第一个推导，其实就是计算 $x_1$，$x_2$ 各自的内积，然后相乘相加即可，第二个推导则是直接平方，去掉括号，也很容易推出来）

另外，我们又注意到：

$$(\langle x_1, x_2 \rangle + 1)^2 = 2\eta_1 \xi_1 + \eta_1^2 \xi_1^2 + 2\eta_2 \xi_2 + \eta_2^2 \xi_2^2 + 2\eta_1 \eta_2 \xi_1 \xi_2 + 1 \tag{43}$$

二者有很多相似的地方，实际上，我们只要把某几个维度线性缩放一下，然后再加上一个常数维度，具体来说，上面这个式子的计算结果实际上和映射

$$\varphi(X_1, X_2) = (\sqrt{2}X_1, X_1^2, \sqrt{2}X_2, X_2^2, \sqrt{2}X_1X_2, 1)^T \tag{44}$$

之后的内积 $\langle\varphi(x_1, \varphi(x_2))\rangle$ 的结果是相等的，那么区别在于什么地方呢？

1. 一个是映射到高维空间中，然后再根据内积的公式进行计算；

2. 而另一个则直接在原来的低维空间中进行计算，而不需要显式地写出映射后的结果。

（公式说明：上面之中，最后的两个式子，第一个算式，是带内积的完全平方式，可以拆开，然后，通过凑一个得到，第二个算式，也是根据第一个算式凑出来的）

回忆刚才提到的映射的维度爆炸，在前一种方法已经无法计算的情况下，后一种方法却依旧能从容处理，甚至是无穷维度的情况也没有问题。

我们把这里的计算两个向量在隐式映射过后的空间中的内积的函数叫做核函数 (Kernel Function)，例如，在刚才的例子中，我们的核函数为：

$$k(x_1, x_2) = (\langle x_1, x_2\rangle + 1)^2 \tag{45}$$

核函数能简化映射空间中的内积运算——刚好“碰巧”的是，在我们的 SVM 里需要计算的地方数据向量总是以内积的形式出现的。对比刚才我们上面写出来的式子，现在我们的分类函数为：

$$\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b \tag{46}$$

其中 $\alpha$ 由如下 dual 问题计算而得：

$$\begin{aligned} \max \alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ s.t.,\ \alpha_i \geq 0, i = 1, \cdots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{47}$$

这样一来计算的问题就算解决了，避开了直接在高维空间中进行计算，而结果却是等价的！当然，因为我们这里的例子非常简单，所以我可以手工构造出对应于 $\phi(\cdot)$ 的核函数出来，如果对于任意一个映射，想要构造出对应的核函数就很困难了。

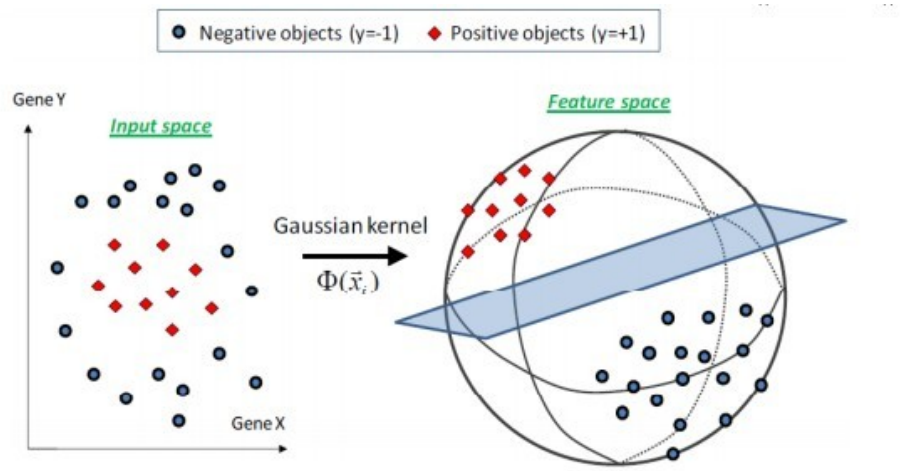
### 2.2.3 几个核函数

通常人们会从一些常用的核函数中选择（根据问题和数据的不同，选择不同的参数，实际上就是得到了不同的核函数），例如：

- 多项式核 $k(x_1, x_2) = (\langle x_1, x_2\rangle + R)^d$，显然刚才我们举的例子是这里多项式核的一个特例（$R = 1$，$d = 2$）。虽然比较麻烦，而且没有必要，不过这个核所对应的映射实际上是可以写出来的，该空间的维度是 $\binom{m+d}{d}$，其中 $m$ 是原始空间的维度。

- 高斯核 $k(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$，这个核就是最开始提到过的会将原始空间映射为无穷维空间的那个家伙。不过，如果 $\sigma$ 选得很大的话，高次特征上的权重实际上衰减得非常快，所以实际上（数值上近似一下）相当于一个低维的子空间；反

过来，如果 $\sigma$ 得很小，则可以将任意的数据映射为线性可分——当然，这并不一定是好事，因为随之而来的可能是非常严重的过拟合问题。不过，总的来说，通过调控参 $\sigma$ 数，高斯核实际上具有相当高的灵活性，也是使用最广泛的核函数之一。下图所示的例子便是把低维线性不可分的数据通过高斯核函数映射到了高维空间：




- 线性核 $k(x_1, x_2) = \langle x_1, x_2 \rangle$，这实际上就是原始空间中的内积。这个核存在的主要目的是使得“映射后空间中的问题”和“映射前空间中的问题”两者在形式上统一起来了 (意思是说，咱们有的时候，写代码，或写公式的时候，只要写个模板或通用表达式，然后再代入不同的核，便可以了，于此，便在形式上统一了起来，不用再分别写一个线性的，和一个非线性的)。

### 2.2.4 核函数的本质

上面说了这么一大堆，读者可能还是没明白核函数到底是个什么东西？我再简要概括下，即以下三点：

1. 实际中，我们会经常遇到线性不可分的样例，此时，我们的常用做法是把样例特征映射到高维空间中去 (如上文 2.2 节最开始的那幅图所示，映射到高维空间后，相关特征便被分开了，也就达到了分类的目的)；

2. 但进一步，如果凡是遇到线性不可分的样例，一律映射到高维空间，那么这个维度大小是会高到可怕的 (如上文中 19 维乃至无穷维的例子)。那咋办呢？

3. 此时，核函数就隆重登场了，核函数的价值在于它虽然也是讲特征进行从低维到高维的转换，但核函数绝就绝在它事先在低维上进行计算，而将实质上的分类效果表现在了高维上，也就如上文所说的避免了直接在高维空间中的复杂计算。

经过前面内容的讲解，我们已经知道，当把内积 $\langle x_i, x_j \rangle$ 变成 $\langle \Phi(x_i), \Phi(x_j) \rangle$ 之后，求 $\langle \Phi(x_i), \Phi(x_j) \rangle$ 将有两种方法：

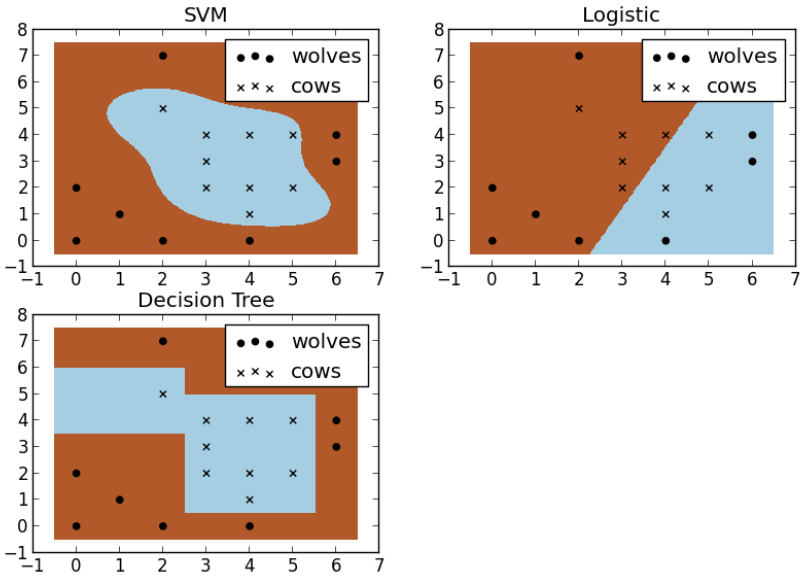
1. 先找到这种映射，然后将输入空间中的样本映射到新的空间中，最后在新空间中去求内积 $\langle \Phi(x_i), \Phi(x_j) \rangle$。以多项式 $x_1 + x_2 + x_1^2 + x_2^2 + c = 0$ 为例，对其进行变换，$c_1 = x_1$，$c_2 = x_2$，$c_3 = x_1^2$，$c_4 = x_2^2$，得到：$c_1 + c_2 + c_3 + c_4 + c = 0$，也就是说通过把输入空间从二维向四维映射后，样本由线性不可分变成了线性可分，但是这种转化带来的直接问题是维度变高了，这意味着，首先可能导致后续计算变复杂，其

次可能出现维度之咒，对于学习器而言就是：特征空间维数可能最终无法计算，而它的泛化能力 (学习器对训练样本以外数据的适应性) 会随着维度的增长而大大降低，这也违反了“奥坎姆的剃刀”，最终可能会使得积 $\langle \Phi(x_i), \Phi(x_j) \rangle$ 无法求出，于是也就失去了这种转化的优势了；

2. 或者是找到某种方法，它不需要显式的将输入空间中的样本映射到新的空间中而能够在输入空间中直接计算出内积 $\langle \Phi(x_i), \Phi(x_j) \rangle$。它其实是对输入空间向高维空间的一种隐式映射，它不需要显式的给出那个映射，在输入空间就可以计算 $\langle \Phi(x_i), \Phi(x_j) \rangle$，这就是传说中的核函数方法。

最后引用这里的一个例子举例说明下核函数解决非线性问题的直观效果。

假设现在你是一个农场主，圈养了一批羊群，但为预防狼群袭击羊群，你需要搭建一个篱笆来把羊群围起来。但是篱笆应该建在哪里呢？你很可能需要依据牛群和狼群的位置建立一个“分类器”，比较下图这几种不同的分类器，我们可以看到 SVM 完成了一个很完美的解决方案。



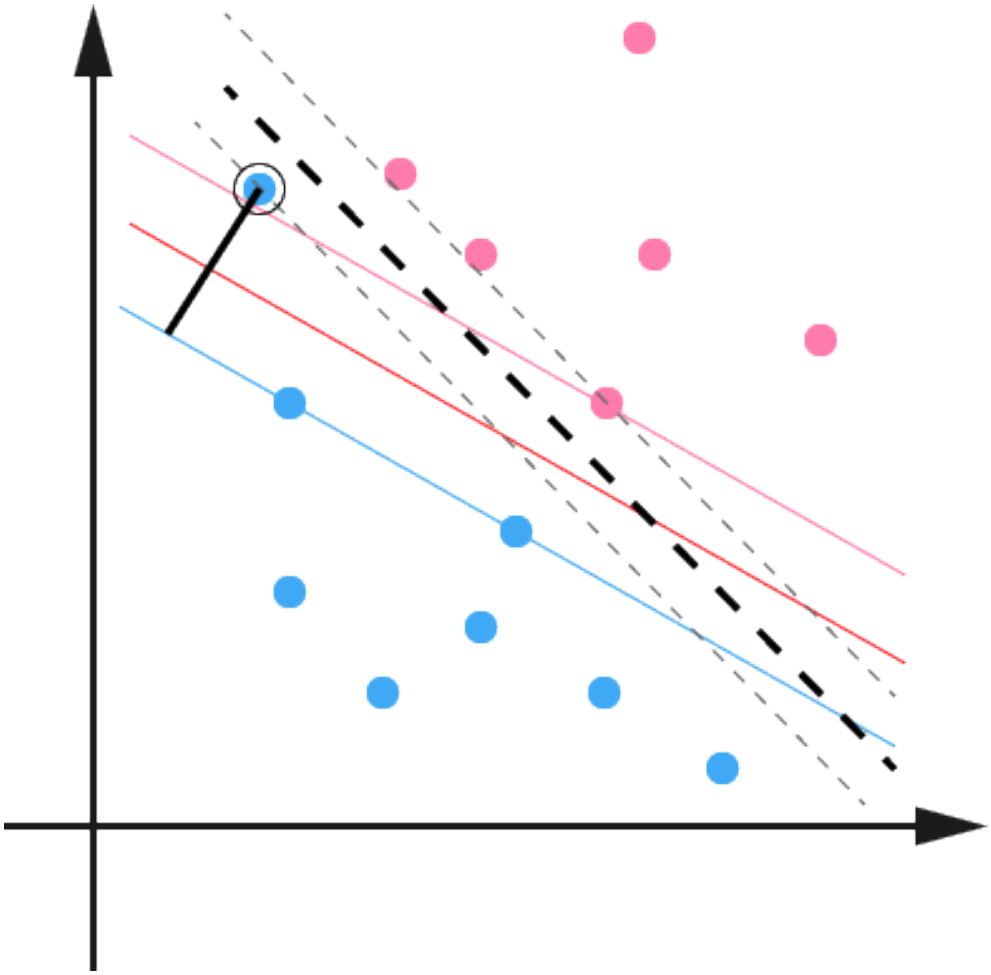

这个例子从侧面简单说明了 SVM 使用非线性分类器的优势，而逻辑模式以及决策树模式都是使用了直线方法。

OK，不再做过多介绍了，对核函数有进一步兴趣的，还可以看看此文。

## 2.3 使用松弛变量处理 outliers 方法

在本文第一节最开始讨论支持向量机的时候，我们就假定，数据是线性可分的，亦即我们可以找到一个可行的超平面将数据完全分开。后来为了处理非线性数据，在上文 2.2 节使用 Kernel 方法对原来的线性 SVM 进行了推广，使得非线性的的情况也能处理。虽然通过映射 $\phi(\cdot)$ 将原始数据映射到高维空间之后，能够线性分隔的概率大大增加，但是对于某些情况还是很难处理。

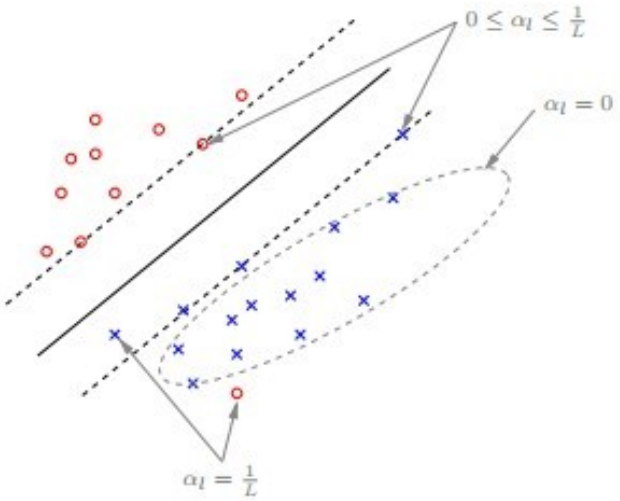
例如可能并不是因为数据本身是非线性结构的，而只是因为数据有噪音。对于这种偏离正常位置很远的数据点，我们称之为 outlier ，在我们原来的 SVM 模型里，outlier 的存在有可能造成很大的影响，因为超平面本身就是只有少数几个 support vector 组成的，如果这些 support vector 里又存在 outlier 的话，其影响就很大了。例如下图：



用黑圈圈起来的那个蓝点是一个 outlier，它偏离了自己原本所应该在的那个半空间，如果直接忽略掉它的话，原来的分隔超平面还是挺好的，但是由于这个 outlier 的出现，导致分隔超平面不得不被挤歪了，变成途中黑色虚线所示（这只是一个示意图，并没有严格计算精确坐标），同时 margin 也相应变小了。当然，更严重的情况是，如果这个 outlier 再往右上移动一些距离的话，我们将无法构造出能将数据分开的超平面来。

为了处理这种情况，SVM 允许数据点在一定程度上偏离一下超平面。例如上图中，黑色实线所对应的距离，就是该 outlier 偏离的距离，如果把它移动回来，就刚好落在原来的超平面上，而不会使得超平面发生变形了。

插播下一位读者 @Copper_PKU 的理解：“换言之，在有松弛的情况下 outline 点也属于支持向量 SV，同时，对于不同的支持向量，拉格朗日参数的值也不同，如此篇论文《Large Scale Machine Learning》中的下图所示：




对于远离分类平面的点值为 0；对于边缘上的点值在 $[0, 1/L]$ 之间，其中，L 为训练数据集个数，即数据集大小；对于 outline 数据和内部的数据值为 1/L。更多请参看本文文末参考条目第 51 条。”

OK，继续回到咱们的问题。我们，原来的约束条件为：

$$y_i(w^T x_i + b) \ge 1,\ i = 1, \cdots, n \tag{48}$$

现在考虑到 outlier 问题，约束条件变成了：

$$y_i(w^T x_i + b) \ge 1 - \xi_i,\ i = 1, \cdots, n \tag{49}$$

其中 $\xi_i \ge 0$ 称为松弛变量 (slack variable)，对应数据点允许偏离的 functional margin 的量。当然，如果我们运行 $\xi_i$ 任意大的话，那任意的超平面都是符合条件的了。所以，我们在原来的目标函数后面加上一项，使得这些 $\xi_i$ 的总和也要最小：

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{50}$$

其中 $C$ 是一个参数，用于控制目标函数中两项（“寻找 margin 最大的超平面”和“保证数据点偏差量最小”）之间的权重。注意，$\xi$ 其中是需要优化的变量（之一），而 $C$ 是一个事先确定好的常量。完整地写出来是这个样子：

$$\begin{aligned} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ s.t., y_i(w^T + b) \ge 1 - \xi_i, i = 1, \cdots, n \\ \xi_i \ge 0, i = 1, \cdots, n \end{aligned} \tag{51}$$

用之前的方法将限制或约束条件加入到目标函数中，得到新的拉格朗日函数，如下所示：

$$\mathcal{L}(w, b, \xi, \alpha, r) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left(y_i(w^T x_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{n} r_i\xi_i \tag{52}$$

分析方法和前面一样，转换为另一个问题之后，我们先让 $\mathcal{L}$ 针对 $w$、$b$ 和 $\xi$ 最小化：

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} &= 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i \\
\frac{\partial \mathcal{L}}{\partial b} &= 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \\
\frac{\partial \mathcal{L}}{\partial \xi_i} &= 0 \Rightarrow C - \alpha - i - r_i = 0, i = 1, \cdots, n
\end{aligned} \tag{53}$$

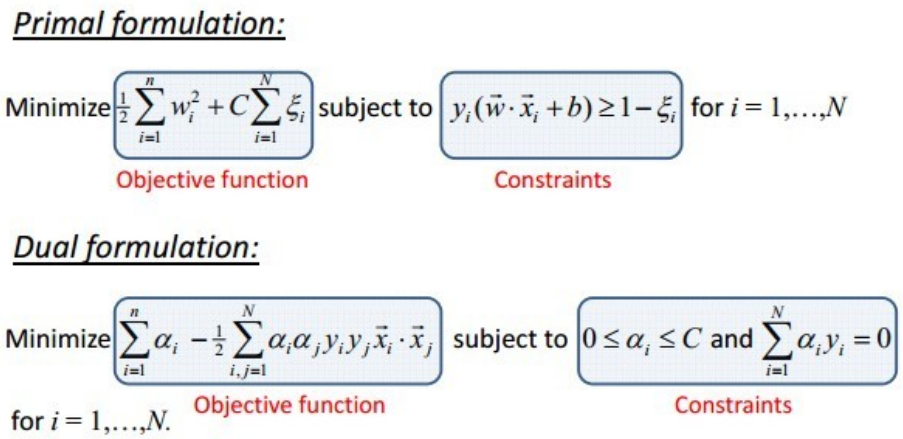
将 $w$ 带回 $\mathcal{L}$ 并化简，得到和原来一样的目标函数：

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \tag{54}$$

不过，由于我们得到 $C - \alpha_i - r_i = 0$ 而又有 $r_i \geq 0$（作为 Lagrange multiplier 的条件），因此有 $\alpha_i \leq C$，所以整个 dual 问题现在写作：

$$\begin{aligned}
\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
s.t., 0 \leq \alpha\alpha_i \leq C, i = 1, \cdots, n \\
\sum_{i=1}^{n} \alpha_i y_i = 0
\end{aligned} \tag{55}$$

把前后的结果对比一下（错误修正：图中的 Dual formulation 中的 Minimize 应为 maxmize）：




可以看到唯一的区别就是现在 dual variable $\alpha$ 多了一个上限 $C$ 。而 Kernel 化的非线性形式也是一样的，只要把 $\langle x_i, x_j \rangle$ 换成 $k(x_i, x_j)$ 即可。这样一来，一个完整的，可以处理线性和非线性并能容忍噪音和 outliers 的支持向量机才终于介绍完毕了。

行文至此，可以做个小结，不准确的说，SVM 它本质上即是一个分类方法，用 $w^T + b$ 定义分类函数，于是求 $w$、$b$，为寻最大间隔，引出 $1/2\|w\|^2$，继而引入拉格朗日因子，化为对拉格朗日乘子 a 的求解（求解过程中会涉及到一系列最优化或凸二次规划

等问题)，如此，求 $w.b$ 与求 $a$ 等价，而 $a$ 的求解可以用一种快速学习算法 SMO，至于核函数，是为处理非线性情况，若直接映射到高维计算恐维度爆炸，故在低维计算，等效高维表现。

OK，理解到这第二层，已经能满足绝大部分人一窥 SVM 原理的好奇心，然对于那些想在证明层面理解 SVM 的则还很不够，但进入第三层理解境界之前，你必须要有比较好的数理基础和逻辑证明能力，不然你会跟我一样，吃不少苦头的。

# 3 证明 SVM

说实话，凡是涉及到要证明的东西. 理论，便一般不是怎么好惹的东西。绝大部分时候，看懂一个东西不难，但证明一个东西则需要点数学功底，进一步，证明一个东西也不是特别难，难的是从零开始发明创造这个东西的时候，则显艰难 (因为任何时代，大部分人的研究所得都不过是基于前人的研究成果，前人所做的是开创性工作，而这往往是最艰难最有价值的，他们被称为真正的先驱。牛顿也曾说过，他不过是站在巨人的肩上。你，我则更是如此)。

正如陈希孺院士在他的著作《数理统计学简史》的第 4 章、最小二乘法中所讲：在科研上诸多观念的革新和突破是有着很多的不易的，或许某个定理在某个时期由某个人点破了，现在的我们看来一切都是理所当然，但在一切没有发现之前，可能许许多多的顶级学者毕其功于一役，耗尽一生，努力了几十年最终也是无功而返。

话休絮烦，要证明一个东西先要弄清楚它的根基在哪，即构成它的基础是哪些理论。OK，以下内容基本是上文中未讲到的一些定理的证明，包括其背后的逻辑、来源背景等东西，还是读书笔记。

本部分导述

- 3.1 节线性学习器中，主要阐述感知机算法；
- 3.2 节非线性学习器中，主要阐述 mercer 定理；
- 3.3 节、损失函数；
- 3.4 节、最小二乘法；
- 3.5 节、SMO 算法；
- 3.6 节、简略谈谈 SVM 的应用；

## 3.1 线性学习器

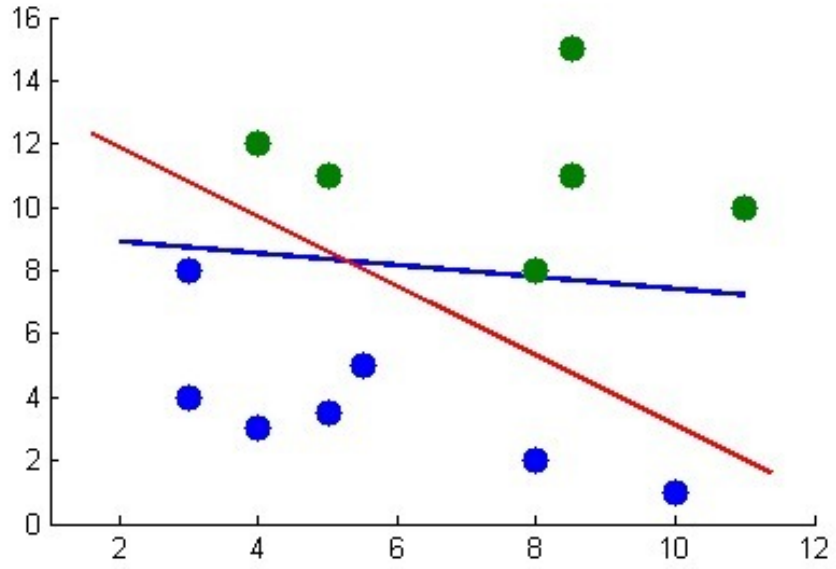
### 3.1.1 感知机算法

这个感知机算法是 1956 年提出的，年代久远，依然影响着当今，当然，可以肯定的是，此算法亦非最优，后续会有更详尽阐述。不过，有一点，你必须清楚，这个算法是为了干嘛的：不断的训练试错以期寻找一个合适的超平面 (是的，就这么简单)。

表 2.1 感知机算法（原始形式）

给定线性可分的数据集$S$和学习率$\eta \in \mathbb{R}^+$
$\mathbf{w}_0 \leftarrow \mathbf{0}; b_0 \leftarrow 0; k \leftarrow 0$
$R \leftarrow \max_{1 \leqslant i \leqslant l} \|\mathbf{x}_i\|$
重复
  for $i = 1$ to $l$
    if $y_i(\mathbf{w}_k \cdot \mathbf{x}_i) + b_k) \leqslant 0$ then
      $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \eta y_i \mathbf{x}_i$
      $b_{k+1} \leftarrow b_k + \eta y_i R^2$
      $k \leftarrow k + 1$
    end if
  end for
直到在 for 循环中没有错误发生
返回$(\mathbf{w}_k, b_k)$，这里$k$是错误次数

下面，举个例子。如下图所示，凭我们的直觉可以看出，图中的红线是最优超平面，蓝线则是根据感知机算法在不断的训练中，最终，若蓝线能通过不断的训练移动到红线位置上，则代表训练成功。

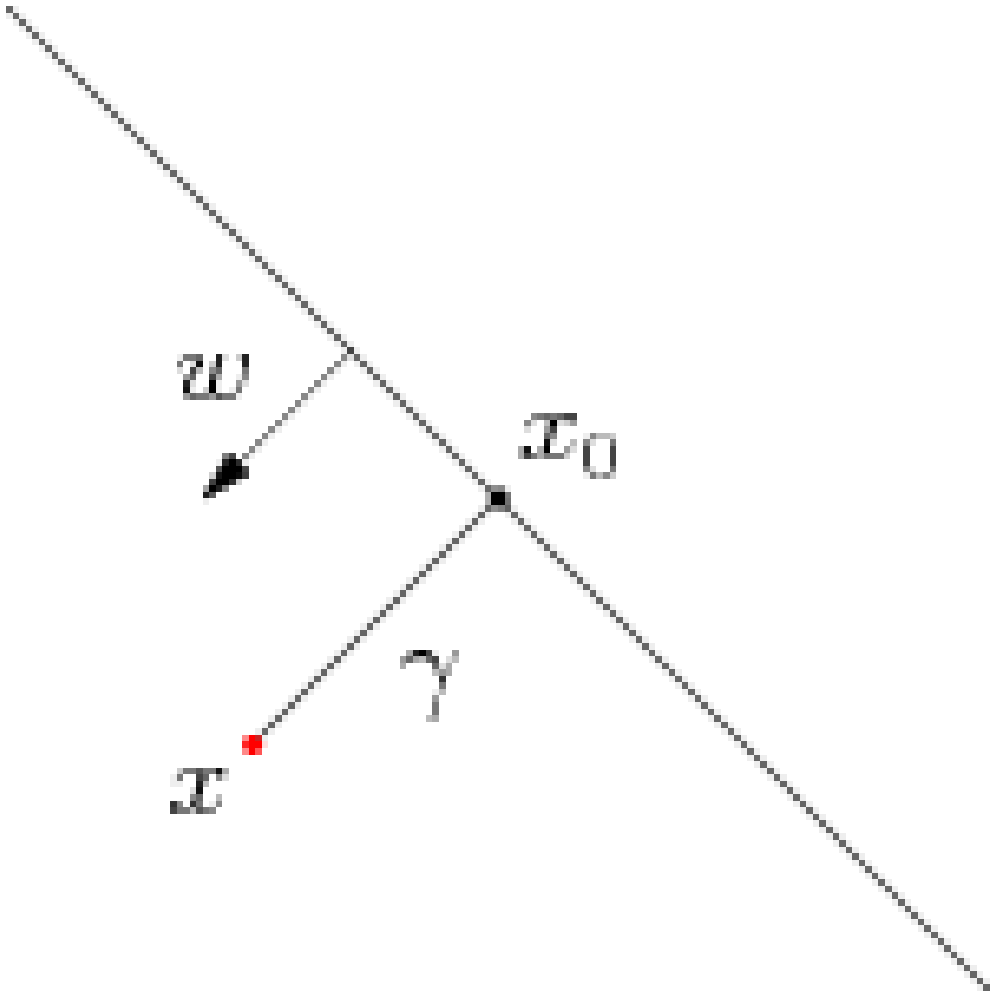



既然需要通过不断的训练以让蓝线最终成为最优分类超平面，那么，到底需要训练多少次呢？Novikoff 定理告诉我们当间隔是正的时候感知机算法会在有限次数的迭代中收敛，也就是说 Novikoff 定理证明了感知机算法的收敛性，即能得到一个界，不至于无穷循环下去。

- Novikoff 定理：如果分类超平面存在, 仅需在序列 $S$ 上迭代几次，在界为 $\left(\frac{2R}{\gamma}\right)^2$ 的错误次数下就可以找到分类超平面，算法停止。

这里 $R = \max_{1 \leq i \leq l} \|x_i\|$，$\gamma$ 为扩充间隔。根据误分次数公式可知, 迭代次数与对应于扩充 (包括偏置) 权重的训练集的间隔有关。

顺便再解释下这个所谓的扩充间隔 $\gamma$，$\gamma$ 即为样本到分类间隔的距离，即从 $\gamma$ 引出的最大分类间隔。OK，还记得上文第 1.3.2 节开头的内容么？如下：“




在给出几何间隔的定义之前，咱们首先来看下，如上图所示，对于一个点 $x$，令其垂直投影到超平面上的对应的为 $x_0$，由于 $w$ 是垂直于超平面的一个向量，$\gamma$ 为样本 $x$ 到分类间隔的距离，我们有

$$x = x_0 + \gamma \frac{w}{\|w\|} \tag{56}$$

然后后续怎么推导出最大分类间隔请回到本文第一、二部分，此处不重复板书。

同时有一点得注意：感知机算法虽然可以通过简单迭代对线性可分数据生成正确分类的超平面，但不是最优效果，那怎样才能得到最优效果呢，就是上文中第一部分所讲的寻找最大分类间隔超平面。此外，Novikoff 定理的证明请见这里。

## 3.2 非线性学习器

### 3.2.1 Mercer 定理

Mercer 定理：如果函数 K 是上的映射（也就是从两个 n 维向量映射到实数域）。那么如果 K 是一个有效核函数（也称为 Mercer 核函数），那么当且仅当对于训练样例，其相应的核函数矩阵是对称半正定的。

要理解这个 Mercer 定理，先要了解什么是半正定矩阵，要了解什么是半正定矩阵，先得知道什么是正定矩阵（矩阵理论“博大精深”，我自己也未能彻底理清，等我理清

了再续写此节，顺便推荐我正在看的一本《矩阵分析与应用》)。然后这里有一个此定理的证明，可以看下。

正如 @Copper_PKU 所说：核函数在 SVM 的分类效果中起了重要的作用，最后这里有个 tutorial 可以看看。

## 3.3 损失函数

在本文 1.0 节有这么一句话“支持向量机 (SVM) 是 90 年代中期发展起来的基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。”但初次看到的读者可能并不了解什么是结构化风险，什么又是经验风险。要了解这两个所谓的“风险”，还得又从监督学习说起。

监督学习实际上就是一个经验风险或者结构风险函数的最优化问题。风险函数度量平均意义下模型预测的好坏，模型每一次预测的好坏用损失函数来度量。它从假设空间 F 中选择模型 f 作为决策函数，对于给定的输入 X，由 f(X) 给出相应的输出 Y，这个输出的预测值 f(X) 与真实值 Y 可能一致也可能不一致，用一个损失函数来度量预测错误的程度。损失函数记为 L(Y, f(X))。

常用的损失函数有以下几种（基本引用自《统计学习方法》)：

(1) $0-1$ 损失函数

$$L(Y, f(X)) = \begin{cases} 1, Y \neq f(X) \\ 0, Y = f(X) \end{cases}$$

(2) 平方损失函数

$$L(Y, f(X)) = (Y - f(X))^2$$

(3) 绝对损失函数

$$L(Y, f(X)) = |Y - f(X)|$$

(4) 对数损失函数

$$L(Y, P(Y|X)) = -\log P(Y|X)$$

模型 $f(X)$ 关于训练数据集的平均损失称为经验风险，如下：

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) \tag{57}$$

关于如何选择模型，监督学习有两种策略：经验风险最小化和结构风险最小化。

经验风险最小化的策略认为，经验风险最小的模型就是最优的模型，则按照经验风险最小化求解如下最优化问题：

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) \tag{58}$$

<F4>

当样本容量很小时，经验风险最小化的策略容易产生过拟合的现象。结构风险最小化可以防止过拟合。结构风险是在经验风险的基础上加上白哦是模型复杂度的正则化项，结构风险定义如下：

$$R_{srm}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i.f(x_i)) + \lambda J(f) \tag{59}$$

其中 $J(f)$ 为模型的复杂度，模型 $f$ 越复杂，$J(f)$ 值就越大，模型越简单，$J(f)$ 值就越小，也就是说 $J(f)$ 是对复杂模型的惩罚。$\lambda \geq 0$ 是系数，用以权衡经验风险和模型复杂度。结构风险最小化的策略认为结构风险最小的模型是最优的模型，所以求最优的模型就是求解下面的最优化问题：

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{60}$$

这样，监督学习问题就变成了经验风险或结构风险函数的最优化问题，如式58和式60

如此，SVM 有第二种理解，即最优化 + 损失最小，或如 @ 夏粉 _ 百度所说“可从损失函数和优化算法角度看 SVM，boosting，LR 等算法，可能会有不同收获”。

OK，关于更多统计学习方法的问题，请参看此文。

关于损失函数，如下文读者评论中所述：可以看看张潼的这篇《Statistical behavior and consistency of classification methods based on convex risk minimization》。各种算法中常用的损失函数基本都具有 fisher 一致性，优化这些损失函数得到的分类器可以看作是后验概率的“代理”。

此外，他还有另外一篇论文《Statistical analysis of some multi-category large margin classification methods》，在多分类情况下 margin loss 的分析，这两篇对 Boosting 和 SVM 使用的损失函数分析的很透彻。

## 3.4 最小二乘法

### 3.4.1 什么是最小二乘法？

既然本节开始之前提到了最小二乘法，那么下面引用《正态分布的前世今生》里的内容稍微简单阐述下。

我们口头中经常说：一般来说，平均来说。如平均来说，不吸烟的健康优于吸烟者，之所以要加“平均”二字，是因为凡事皆有例外，总存在某个特别的人他吸烟但由于经常锻炼所以他的健康状况可能会优于他身边不吸烟的朋友。而最小二乘法的一个最简单的例子便是算术平均。

最小二乘法（又称最小平方法）是一种数学优化技术。它通过最小化误差的平方和寻找数据的最佳函数匹配。利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小。用函数表示为：

$$\min_{\vec{x}} \sum_{i=1}^{n} (y_m - y_i)^2 \tag{61}$$

使误差「所谓误差，当然是观察值与实际真实值的差量」平方和达到最小以寻求估计值的方法，就叫做最小二乘法，用最小二乘法得到的估计，叫做最小二乘估计。当然，取平方和作为目标函数只是众多可取的方法之一。

最小二乘法的一般形式可表示为：

$$\min_{\vec{x}} \|\vec{y}_m(\vec{x}) - \vec{y}\|_2 \tag{62}$$

有效的最小二乘法是勒让德在 1805 年发表的，基本思想就是认为测量中有误差，所以所有方程的累积误差为

$$\text{累积误差} = \sum (\text{观测值} - \text{理论值})^2 \tag{63}$$

我们求解出导致累积误差最小的参数即可：

$$\begin{aligned}\widehat{\beta} &= argmin_{\beta}\sum_{i=1}^{n} e_i^2 \\ &= argmin_{\beta}\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})]^2\end{aligned} \tag{64}$$

勒让德在论文中对最小二乘法的优良性做了几点说明：

- 最小二乘使得误差平方和最小，并在各个方程的误差之间建立了一种平衡，从而防止某一个极端误差取得支配地位
- 计算中只要求偏导后求解线性方程组，计算过程明确便捷
- 最小二乘可以导出算术平均值作为估计值

对于最后一点，从统计学的角度来看是很重要的一个性质。推理如下：假设真值为 $\theta, x_q, \cdots, x_n$ 为 $n$ 次测量值, 每次测量的误差为 $e_i = x_i - \theta$，按最小二乘法，误差累积为

$$L(\theta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(x_i - \theta)^2 \tag{65}$$

求解 $\theta$ 使 $L(\theta)$ 达到最小，正好是算术平均 $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$。

由于算术平均是一个历经考验的方法，而以上的推理说明，算术平均是最小二乘的一个特例，所以从另一个角度说明了最小二乘方法的优良性，使我们对最小二乘法更加有信心。

最小二乘法发表之后很快得到了大家的认可接受，并迅速的在数据分析实践中被广泛使用。不过历史上又有人把最小二乘法的发明归功于高斯，这又是怎么一回事呢。高斯在 1809 年也发表了最小二乘法，并且声称自己已经使用这个方法多年。高斯发明了小行星定位的数学方法，并在数据分析中使用最小二乘方法进行计算，准确的预测了谷神星的位置。

说了这么多，貌似跟本文的主题 SVM 没啥关系呀，别急，请让我继续阐述。本质上说，最小二乘法即是一种参数估计方法，说到参数估计，咱们得从一元线性模型说起。

### 3.4.2 最小二乘法的解法

什么是一元线性模型呢？请允许我引用这里的内容，先来梳理下几个基本概念：

- 监督学习中，如果预测的变量是离散的，我们称其为分类（如决策树，支持向量机等），如果预测的变量是连续的，我们称其为回归。
- 回归分析中，如果只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。
- 如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。
- 对于二维空间线性是一条直线；对于三维空间线性是一个平面，对于多维空间线性是一个超平面

对于一元线性回归模型, 假设从总体中获取了 $n$ 组观察值 $(X_1, Y_1), (X_2.Y_2), \cdots, (X_n, Y_n)$。对于平面中的这 $n$ 个点，可以使用无数条曲线来拟合。要求样本回归函数尽可能好地拟合这组值。综合起来看，这条直线处于样本数据的中心位置最合理。

选择最佳拟合曲线的标准可以确定为：使总的拟合误差（即总残差）达到最小。有以下三个标准可以选择：

1. 用“残差和最小”确定直线位置是一个途径。但很快发现计算“残差和”存在相互抵消的问

2. 用“残差绝对值和最小”确定直线位置也是一个途径。但绝对值的计算比较麻。

3. 最小二乘法的原则是以“残差平方和最小”确定直线位置。用最小二乘法除了计算比较方便外，得到的估计量还具有优良特性。这种方法对异常值非常敏。

最常用的是普通最小二乘法（Ordinary Least Square，OLS）：所选择的回归模型应该使所有观察值的残差平方和达到最小，即采用平方损失函数。

我们定义样本回归模型为：

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i \Rightarrow e_i = Y_i - \widehat{\beta}_0 = \widehat{\beta}_1 X_i \tag{66}$$

其中 $e_i$ 为样本 $(X_i, Y_i)$ 的误差。

接着，定义平方损失函数 $Q$：

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 = \widehat{\beta}_1 X_i)^2 \tag{67}$$

则通过 $Q$ 最小确定这条直线，即确定 $\widehat{\beta}_0, \widehat{\beta}_1$，以 $\widehat{\beta}_0, \widehat{\beta}_1$ 为变量，把它们看作是 $Q$ 的函数，就变成了一个求极值的问题，可以通过求导数得到。

求 $Q$ 对两个待估参数的偏导数：

$$\begin{cases} \frac{\partial Q}{\partial \widehat{\beta}_0} = 2\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = 0 \\ \frac{\partial Q}{\partial \widehat{\beta}_1} = 2\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = 0 \end{cases} \tag{68}$$

根据数学知识我们知道，函数的极值点为偏导为 0 的点。

解得：

$$\begin{aligned} \widehat{\beta}_2 &= \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ \widehat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} \end{aligned} \tag{69}$$

这就是最小二乘法的解法，就是求得平方损失函数的极值点。自此，你看到求解最小二乘法与求解 SVM 问题何等相似，尤其是定义损失函数，而后通过偏导求得极值。

OK，更多请参看陈希孺院士的《数理统计学简史》的第 4 章、最小二乘法，和本文参考条目第 59 条《凸函数》。

## 3.5 SMO 算法

在上文 2.1.2 节中，我们提到了求解对偶问题的序列最小最优化 SMO 算法，但并未提到具体解法。

事实上，SMO 算法是由 Microsoft Research 的 John C. Platt 在 1998 年发表的一篇论文《Sequential Minimal Optimization A Fast Algorithm for Training Support Vector Machines》中提出，它很快成为最快的二次规划优化算法，特别针对线性 SVM 和数据稀疏时性能更优。

接下来，咱们便参考 John C. Platt 的这篇文章来看看 SMO 的解法是怎样的。

### 3.5.1 SMO 算法的解法

咱们首先来定义特征到结果的输出函数为

$$u = \vec{w} \cdot \vec{x} - b \tag{70}$$

再三强调，这个 $u$ 与我们之前定义的 $f(x) = w^T x + b$ 实质是一样的。

接着，咱们重新定义咱们原始的优化问题，权当重新回顾，如下：

$$\min_{w,b} \frac{1}{2}\|\vec{w}\|^2 subjecttoy_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \forall i \tag{71}$$

求导得到：

$$\vec{w} = \sum_{i=1}^{N} y_i \alpha_i \vec{X}_i, b = \vec{w} \cdot \vec{x}_k - y_k forsome\alpha_k > 0 \tag{72}$$

代入 $u = \vec{w} \cdot \vec{x} - b$ 中，可得 $u = \sum_{j=1} N y_j \alpha_j K(\vec{x}_j, \vec{x}) - b$。

引入对偶因子后，得：

$$\begin{aligned} \min_{\alpha} \Psi(\vec{\alpha}) = \min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j (\vec{x}_i \cdot \vec{x}_j) \alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i \\ s.t.: \ \alpha_i \geq 0, \forall i, \\ 且 \sum_{i=1}^{N} y_i \alpha_i = 0 \end{aligned} \tag{73}$$

注：这里得到的 min 函数与我们之前的 max 函数实质也是一样，因为把符号变下，即有 min 转化为 max 的问题，且 $y_i$ 也与之前的 $y^{(i)}$ 等价，$y_j$ 亦如此。

经过加入松弛变量后，模型修改为：

$$\begin{aligned} \min_{\vec{w},b,\vec{\xi}} \frac{1}{2}\|\vec{w}\|^2 + C \sum_{i=1}^{N} \xi_i subjecttoy_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 - \xi_i, \forall i \\ 0 \leq \alpha_i \leq C, \forall i \end{aligned} \tag{74}$$

从而最终我们的问题变为：

$$\begin{aligned} \min_{\alpha} \Psi(\vec{\alpha}) = \min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j K(\vec{x}_i \cdot \vec{x}_j) \alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i \\ 0 \leq \alpha_i \leq C, \forall i \sum_{i=1}^{N} y_i \alpha_i = 0 \end{aligned} \tag{75}$$

继而，根据 KKT 条件可以得出其中 $\alpha_i$ 取值的意义为：

$$\begin{aligned} \alpha_i = 0 &\Leftrightarrow y_i u_i \leq 1, \\ 0 < \alpha_i < C &\Leftrightarrow y_i u_i = 1, \\ \alpha_i = C &\Leftrightarrow y_i u_i \leq 1. \end{aligned} \tag{76}$$

这里的 $\alpha_i$ 还是拉格朗日乘子 (问题通过拉格朗日乘法数来求解)

1. 对于第 1 种情况，表明是正常分类，在边界内部（我们知道正确分类的点 yi*f(xi)>=0）；

2. 对于第 2 种情况，表明了是支持向量，在边界上；

3. 对于第 3 种情况，表明了是在两条边界之间；

而最优解需要满足 KKT 条件，即上述 3 个条件都得满足，以下几种情况出现将会出现不满足：

- $y_i u_i \leq 1$ 但是 $\alpha < C$ 则是不满足的, 而原本 $\alpha_i = C$
- $y_i u_i \geq 1$ 但是 $\alpha > 0$ 则是不满足的而原本 $\alpha_i = 0$
- $y_i u_i = 1$ 但是 $\alpha_i = 0$ 或者 $alpha_i = C$ 则表明不满足的，而原本应该是 $0 < \alpha_i < C$

所以要找出不满足 KKT 条件的这些 $\alpha_i$，并更新这些 $\alpha_i$，但这些 $\alpha_i$ 又受到另外一个约束，即

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \tag{77}$$

注：别忘了 2.1.1 节中，$L$ 对 $a$、$b$ 求偏导，得到：

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{78}$$

因此，我们通过另一个方法，即同时更新 ai 和 aj，要求满足以下等式：

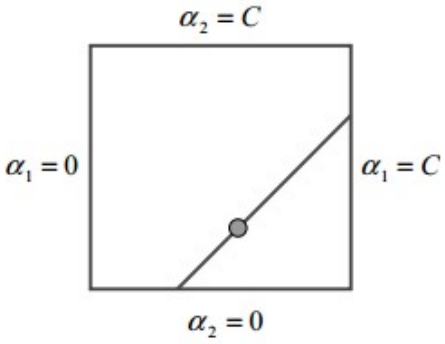
$$a_i^{new} y_i + a_j^{new} y_j = a_i^{old} y_i + a_j^{old} y_j = 常数 \tag{79}$$

就能保证和为 0 的约束。

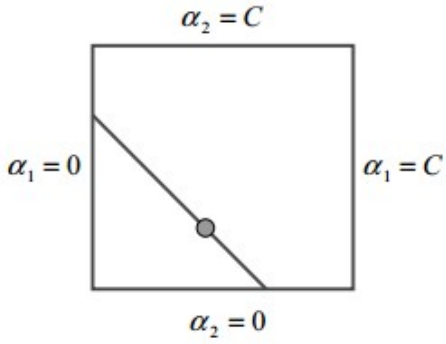
利用 $y_i a_i + y_j a_j =$ 常数，消去 $a_i$，可得到一个关于单变量 $a_j$ 的一个凸二次规划问题，不考虑其约束 $0 \leq aj \leq C$, 可以得其解为：

$$\alpha_j^{new,clipped} = \begin{cases} H & if \alpha_j^{new} \geq H \\ \alpha_j^{new} & if L < \alpha_j^{new} < H \\ L & if \alpha_j^{new} \leq L \end{cases} \tag{80}$$

把 SMO 中对于两个参数求解过程看成线性规划来理解来理解的话，那么下图所表达的便是约束条件 $a_i^{new} y_i + a_j^{new} y_j = a_i^{old} y_i + a_j^{old} y_j =$ 常数:




$y_1 \neq y_2 \Rightarrow \alpha_1 - \alpha_2 = k$




$y_1 = y_2 \Rightarrow \alpha_1 + \alpha_2 = k$