支持向量机（上）

JerryLead
csxulijie@gmail.com
2011 年 3 月 12 日星期六

1 简介

支持向量机基本上是最好的有监督学习算法了。最开始接触 SVM 是去年暑假的时候，老师要求交《统计学习理论》的报告，那时去网上下了一份入门教程，里面讲的很通俗，当时只是大致了解了一些相关概念。这次斯坦福提供的学习材料，让我重新学习了一些 SVM 知识。我看很多正统的讲法都是从 VC 维理论和结构风险最小原理出发，然后引出 SVM 什么的，还有些资料上来就讲分类超平面什么的。这份材料从前几节讲的 logistic 回归出发，引出了 SVM，既揭示了模型间的联系，也让人觉得过渡更自然。

2 重新审视 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。

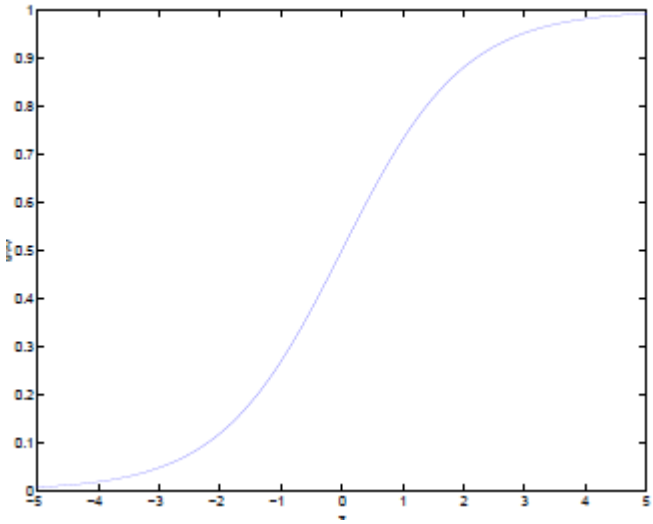
形式化表示就是

假设函数

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

其中 x 是 n 维特征向量，函数 g 就是 logistic 函数。

$g(z) = \frac{1}{1 + e^{-z}}$ 的图像是



可以看到，将无穷映射到了(0,1)。

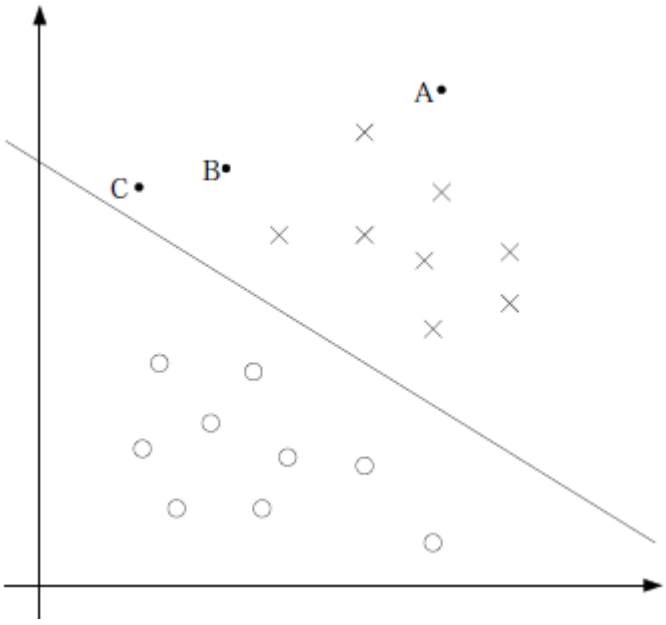
而假设函数就是特征属于 y=1 的概率。

$$P(y=1 \mid x;\theta) = h_\theta(x)$$
$$P(y=0 \mid x;\theta) = 1-h_\theta(x)$$

当我们要判别一个新来的特征属于哪个类时，只需求$h_\theta(x)$，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$\theta^T x$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$\theta^T x \gg 0$时，$h_\theta(x)=1$，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$\theta^T x \gg 0$，而是 y=0 的特征$\theta^T x \ll 0$。Logistic 回归就是要学习得到θ，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。

图形化表示如下：



中间那条线是$\theta^T x = 0$，logistic 回顾强调所有点尽可能地远离中间那条线。学习出的结

果也就中间那条线。考虑上面 3 个点 A、B 和 C。从图中我们可以确定 A 是×类别的，然而 C 我们是不太确定的，B 还算能够确定。这样我们可以得出结论，我们更应该关心靠近中间分割线的点，让他们尽可能地远离中间线，而不是在所有点上达到最优。因为那样的话，要使得一部分点靠近中间线来换取另外一部分点更加远离中间线。我想这就是支持向量机的思路和 logistic 回归的不同点，一个考虑局部（不关心已经确定远离的点），一个考虑全局（已经远离的点可能通过调整中间线使其能够更加远离）。这是我的个人直观理解。

3 形式化表示

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 θ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$\theta_0$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$w^T x$）。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

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

上一节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：

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

4 函数间隔（functional margin）和几何间隔（geometric margin）

给定一个训练样本$(x^{(i)}, y^{(i)})$，x 是特征，y 是结果标签。i 表示第 i 个样本。我们定义函数间隔如下：

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$$

可想而知，当$y^{(i)} = 1$时，在我们的 g(z)定义中，$w^T x^{(i)} + b \geq 0$，$\hat{\gamma}^{(i)}$的值实际上就是$|w^T x^{(i)} + b|$。反之亦然。为了使函数间隔最大（更大的信心确定该例是正例还是反例），当$y^{(i)} = 1$时，$w^T x^{(i)} + b$应该是个大正数，反之是个大负数。因此函数间隔代表了我们认为特征是正例还是反例的确信度。

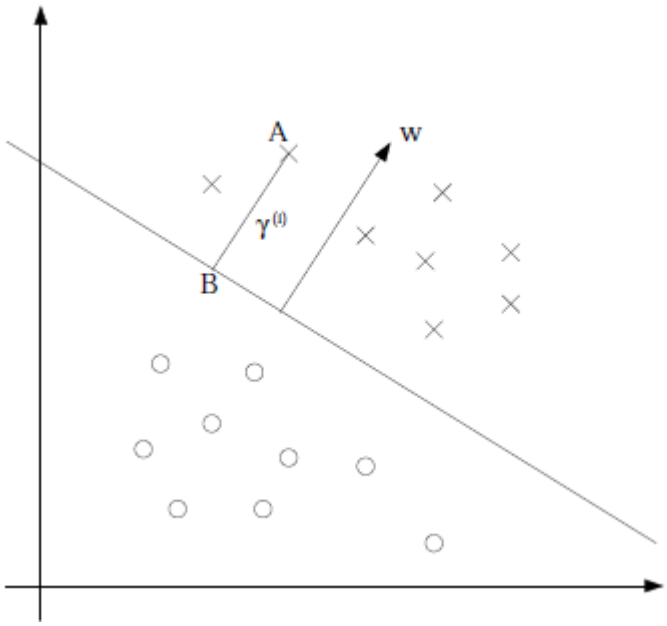
继续考虑 w 和 b，如果同时加大 w 和 b，比如在$(w^T x^{(i)} + b)$前面乘个系数比如 2，那么所有点的函数间隔都会增大二倍，这个对求解问题来说不应该有影响，因为我们要求解的是$w^T x + b = 0$，同时扩大 w 和 b 对结果是无影响的。这样，我们为了限制 w 和 b，可能需要加入归一化条件，毕竟求解的目标是确定唯一一个 w 和 b，而不是多组线性相关的向量。这个归一化一会再考虑。

刚刚我们定义的函数间隔是针对某一个样本的，现在我们定义全局样本上的函数间隔

$$\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}.$$

说白了就是在训练样本上分类正例和负例确信度最小那个函数间隔。

接下来定义几何间隔，先看图



假设我们有了 B 点所在的$w^T x + b = 0$分割面。任何其他一点，比如 A 到该面的距离以$\gamma^{(i)}$表示，假设 B 就是 A 在分割面上的投影。我们知道向量 BA 的方向是w（分割面的梯度），单位向量是$\frac{w}{||w||}$。A 点是$(x^{(i)}, y^{(i)})$，所以 B 点是 x=$x^{(i)} - \gamma^{(i)} \frac{w}{||w||}$（利用初中的几何知识），带入$w^T x + b = 0$得，

$$w^T(x^{(i)} - \gamma^{(i)} \frac{w}{||w||}) + b = 0$$

进一步得到

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||} = \left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||}.$$

$\gamma^{(i)}$实际上就是点到平面距离。

再换种更加优雅的写法：

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||} \right).$$

当$||w|| = 1$时，不就是函数间隔吗？是的，前面提到的函数间隔归一化结果就是几何间隔。他们为什么会一样呢？因为函数间隔是我们定义的，在定义的时候就有几何间隔的色彩。

同样，同时扩大 w 和 b，w 扩大几倍，||w||就扩大几倍，结果无影响。同样定义全局的几何间隔 $\gamma = \min_{i=1,\dots,m} \gamma^{(i)}.$

5 最优间隔分类器（optimal margin classifier）

回想前面我们提到我们的目标是寻找一个超平面，使得离超平面比较近的点能有更大的间距。也就是我们不考虑所有的点都必须远离超平面，我们关心求得的超平面能够让所有点中离它最近的点具有最大间距。形象的说，我们将上面的图看作是一张纸，我们要找一条折线，按照这条折线折叠后，离折线最近的点的间距比其他折线都要大。形式化表示为：

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \dots, m \\ & ||w|| = 1. \end{aligned}$$

这里用||w||=1 规约 w，使得$w^T x + b$是几何间隔。

到此，我们已经将模型定义出来了。如果求得了 w 和 b，那么来一个特征 x，我们就能够分类了，称为最优间隔分类器。接下的问题就是如何求解 w 和 b 的问题了。

由于||w|| = 1不是凸函数，我们想先处理转化一下，考虑几何间隔和函数间隔的关系，$\gamma = \frac{\hat{\gamma}}{||w||}$，我们改写一下上面的式子：

$$\begin{aligned} \max_{\gamma,w,b} \quad & \frac{\hat{\gamma}}{||w||} \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \dots, m \end{aligned}$$

这时候其实我们求的最大值仍然是几何间隔，只不过此时的 w 不受||w|| = 1的约束了。然而这个时候目标函数仍然不是凸函数，没法直接代入优化软件里计算。我们还要改写。前面说到同时扩大 w 和 b 对结果没有影响，但我们最后要求的仍然是 w 和 b 的确定值，不是他们的一组倍数值，因此，我们需要对$\hat{\gamma}$做一些限制，以保证我们解是唯一的。这里为了简便我们取$\hat{\gamma} = 1$。这样的意义是将全局的函数间隔定义为 1，也即是将离超平面最近的点的距离定义为$\frac{1}{||w||}$。由于求$\frac{1}{||w||}$的最大值相当于求$\frac{1}{2}||w||^2$的最小值，因此改写后结果为：

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \dots, m \end{aligned}$$

这下好了，只有线性约束了，而且是个典型的二次规划问题（目标函数是自变量的二次函数）。代入优化软件可解。

到这里发现，这个讲义虽然没有像其他讲义一样先画好图，画好分类超平面，在图上标示出间隔那么直观，但每一步推导有理有据，依靠思路的流畅性来推导出目标函数和约束。

接下来介绍的是手工求解的方法了，一种更优的求解方法。

6 拉格朗日对偶（Lagrange duality）

先抛开上面的二次规划问题，先来看看存在等式约束的极值问题求法，比如下面的最优化问题：

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

目标函数是 f(w)，下面是等式约束。通常解法是引入拉格朗日算子，这里使用β来表示算子，得到拉格朗日公式为

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

L 是等式约束的个数。

然后分别对 w 和β求偏导，使得偏导数等于 0，然后解出 w 和β_i。至于为什么引入拉格朗日算子可以求出极值，原因是 f(w)的 dw 变化方向受其他不等式的约束，dw 的变化方向与 f(w)的梯度垂直时才能获得极值，而且在极值处，f(w)的梯度与其他等式梯度的线性组合平行，因此他们之间存在线性关系。（参考《最优化与 KKT 条件》）

然后我们探讨有不等式约束的极值问题求法，问题如下：

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \le 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

我们定义一般化的拉格朗日公式

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

这里的α_i和β_i都是拉格朗日算子。如果按这个公式求解，会出现问题，因为我们求解的是最小值，而这里的$g_i(w) \le 0$，我们可以将α_i调整成很大的正值，来使最后的函数结果是负无穷。因此我们需要排除这种情况，我们定义下面的函数：

$$\theta_{\mathcal{P}}(w) = \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta).$$

这里的 P 代表 primal。假设$g_i(w) > 0$或者$h_i(w) \neq 0$，那么我们总是可以调整α_i和β_i来使得$\theta_{\mathcal{P}}$(w)有最大值为正无穷。而只有 g 和 h 满足约束时，$\theta_{\mathcal{P}}$(w)为 f(w)。这个函数的精妙之处在于$\alpha_i \geq 0$，而且求极大值。

因此我们可以写作

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$$

这样我们原来要求的 min f(w)可以转换成求$\min_w \theta_{\mathcal{P}}$(w)了。

$$\min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta),$$

我们使用p^*来表示$\min_w \theta_{\mathcal{P}}$(w)。如果直接求解，首先面对的是两个参数，而$\alpha_i$也是不等式约束，然后再在 w 上求最小值。这个过程不容易做，那么怎么办呢？

我们先考虑另外一个问题 $\theta_{\mathcal{D}}(\alpha,\beta) = \min_w \mathcal{L}(w,\alpha,\beta).$

D 的意思是对偶，$\theta_{\mathcal{D}}(\alpha,\beta)$将问题转化为先求拉格朗日关于 w 的最小值，将α和β看作是固定值。之后在 $\theta_{\mathcal{D}}(\alpha,\beta)$求最大值的话：

$$\max_{\alpha,\beta:\alpha_i\geq 0} \theta_{\mathcal{D}}(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0} \min_w \mathcal{L}(w,\alpha,\beta).$$

这个问题是原问题的对偶问题，相对于原问题只是更换了 min 和 max 的顺序，而一般更换顺序的结果是 Max Min(X) <= Min Max(X)。然而在这里两者相等。用d^*来表示对偶问题如下：

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \min_w \mathcal{L}(w,\alpha,\beta) \leq \min_w \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta) = p^*.$$

下面解释在什么条件下两者会等价。假设 f 和 g 都是凸函数，h 是仿射的（affine，there exists a_i, b_i, so that $h_i(w) = a_i^T w + b_i$）。并且存在 w 使得对于所有的 i，$g_i(w) < 0$。在这种假设下，一定存在w^*, α^*, β^*使得w^*是原问题的解，α^*, β^*是对偶问题的解。还有

$p^* = d^* = \mathcal{L}(w^*, \alpha^*, \beta^*)$.另外，$w^*, \alpha^*, \beta^*$满足库恩-塔克条件（Karush-Kuhn-Tucker, KKT condition），该条件如下：

$$\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, n \tag{3}$$

$$\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, l \tag{4}$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k \tag{5}$$

$$g_i(w^*) \leq 0, \quad i = 1, \ldots, k \tag{6}$$

$$\alpha^* \geq 0, \quad i = 1, \ldots, k \tag{7}$$

所以如果w^*, α^*, β^*满足了库恩-塔克条件，那么他们就是原问题和对偶问题的解。让我们再次审视公式（5），这个条件称作是 KKT dual complementarity 条件。这个条件隐含了如果$\alpha^* > 0$，那么$g_i(w^*) = 0$。也就是说，$g_i(w^*) = 0$时，w 处于可行域的边界上，这时才是起作用的约束。而其他位于可行域内部（$g_i(w^*) < 0$的）点都是不起作用的约束，其$\alpha^* = 0$。这个 KKT 双重补足条件会用来解释支持向量和 SMO 的收敛测试。

这部分内容思路比较凌乱，还需要先研究下《非线性规划》中的约束极值问题，再回头看看。KKT 的总体思想是认为极值会在可行域边界上取得，也就是不等式为 0 或等式约束里取得，而最优下降方向一般是这些等式的线性组合，其中每个元素要么是不等式为 0 的约束，要么是等式约束。对于在可行域边界内的点，对最优解不起作用，因此前面的系数为 0。

7 最优间隔分类器（optimal margin classifier）

重新回到 SVM 的优化问题：

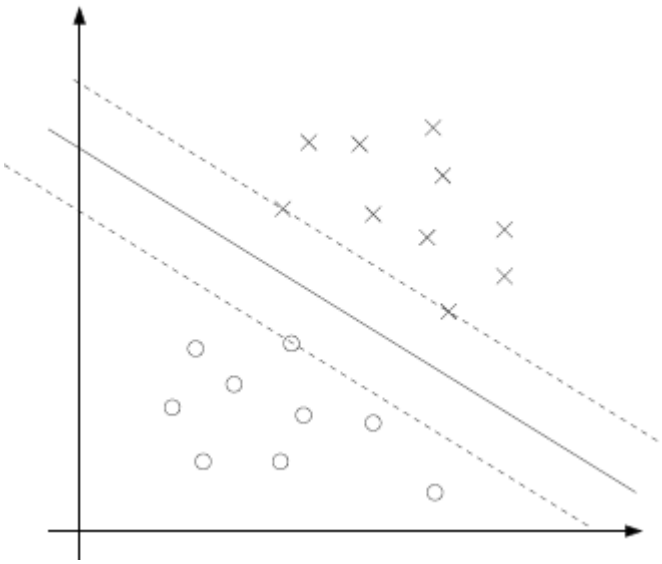
$$\begin{aligned} \min_{\gamma, w, b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

我们将约束条件改写为：

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0.$$

从 KKT 条件得知只有函数间隔是 1（离超平面最近的点）的线性约束式前面的系数$\alpha_i > 0$，也就是说这些约束式$g_i(w) = 0$，对于其他的不在线上的点($g_i(w) < 0$)，极值不会在他们所在的范围内取得，因此前面的系数$\alpha_i = 0$.注意每一个约束式实际就是一个训练样本。

看下面的图：



实线是最大间隔超平面，假设×号的是正例，圆圈的是负例。在虚线上的点就是函数间隔是 1 的点，那么他们前面的系数$\alpha_i > 0$，其他点都是$\alpha_i = 0$。这三个点称作支持向量。构造拉格朗日函数如下：

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1\right].$$

注意到这里只有α_i没有β_i是因为原问题中没有等式约束，只有不等式约束。

下面我们按照对偶问题的求解步骤来一步步进行，

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

首先求解 $\mathcal{L}(w,b,\alpha)$ 的最小值，对于固定的α_i， $\mathcal{L}(w,b,\alpha)$ 的最小值只与 w 和 b 有关。对 w 和 b 分别求偏导数。

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

$$\frac{\partial}{\partial b}\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

并得到

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}.$$

将上式带回到拉格朗日函数中得到，此时得到的是该函数的最小值（目标函数是凸函数）

化简过程如下：

$$\mathcal{L}(\mathrm{w},\mathrm{b},\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{m}\alpha_i\left[y^{(i)}\left(w^Tx^{(i)}+b\right)-1\right]$$

$$=\frac{1}{2}\mathrm{w}^T\mathrm{w}-\sum_{i=1}^{m}\alpha_iy^{(i)}w^Tx^{(i)}-\sum_{i=1}^{m}\alpha_iy^{(i)}b+\sum_{i=1}^{m}\alpha_i$$

$$=\frac{1}{2}\mathrm{w}^T\sum_{i=1}^{m}\alpha_iy^{(i)}x^{(i)}-\sum_{i=1}^{m}\alpha_iy^{(i)}w^Tx^{(i)}-\sum_{i=1}^{m}\alpha_iy^{(i)}b+\sum_{i=1}^{m}\alpha_i$$

$$=\frac{1}{2}\mathrm{w}^T\sum_{i=1}^{m}\alpha_iy^{(i)}x^{(i)}-w^T\sum_{i=1}^{m}\alpha_iy^{(i)}x^{(i)}-\sum_{i=1}^{m}\alpha_iy^{(i)}b+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\mathrm{w}^T\sum_{i=1}^{m}\alpha_iy^{(i)}x^{(i)}-\sum_{i=1}^{m}\alpha_iy^{(i)}b+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\mathrm{w}^T\sum_{i=1}^{m}\alpha_iy^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_iy^{(i)}+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\left(\sum_{i=1}^{m}\alpha_iy^{(i)}x^{(i)}\right)^T\sum_{i=1}^{m}\alpha_iy^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_iy^{(i)}+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\sum_{i=1}^{m}\alpha_iy^{(i)}\left(x^{(i)}\right)^T\sum_{i=1}^{m}\alpha_iy^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_iy^{(i)}+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\sum_{i=1,j=1}^{m}\alpha_iy^{(i)}\left(x^{(i)}\right)^T\alpha_jy^{(j)}x^{(j)}-b\sum_{i=1}^{m}\alpha_iy^{(i)}+\sum_{i=1}^{m}\alpha_i$$

$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1,j=1}^{m}y^{(i)}y^{(i)}\alpha_i\alpha_j(x^{(i)})^Tx^{(j)}-b\sum_{i=1}^{m}\alpha_iy^{(i)}$$

“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α_i和$y^{(i)}$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”使用了(a+b+c+…)(a+b+c+…)=aa+ab+ac+ba+bb+bc+…的乘法运算法则。最后一步是上一步的顺序调整。

最后得到：

$$\mathcal{L}(w,b,\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j(x^{(i)})^Tx^{(j)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}.$$

由于最后一项是 0，因此简化为

$$\mathcal{L}(w,b,\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j(x^{(i)})^Tx^{(j)}.$$

这里我们将向量内积$(x^{(i)})^Tx^{(j)}$表示为$\langle x^{(i)},x^{(j)}\rangle$.

此时的拉格朗日函数只包含了变量α_i。然而我们求出了α_i才能得到 w 和 b。

接着是极大化的过程 $d^*=\max_{\alpha,\beta:\alpha_i\geq 0}\min_w\mathcal{L}(w,\alpha,\beta)$，

$$\begin{aligned}\max_\alpha\quad & W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)},x^{(j)}\rangle.\\ \text{s.t.}\quad & \alpha_i\geq 0,\quad i=1,\dots,m\\ & \sum_{i=1}^{m}\alpha_i y^{(i)}=0,\end{aligned}$$

前面提到过对偶问题和原问题满足的几个条件，首先由于目标函数和线性约束都是凸函数，而且这里不存在等式约束 h。存在 w 使得对于所有的 i，$g_i(w)<0$。因此，一定存在w^*,α^*使得w^*是原问题的解，α^*是对偶问题的解。在这里，求α_i就是求α^*了。

如果求出了α_i，根据 $w=\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}$. 即可求出 w（也是$w^*$，原问题的解）。然后

$$b^*=-\frac{\max_{i:y^{(i)}=-1}w^{*T}x^{(i)}+\min_{i:y^{(i)}=1}w^{*T}x^{(i)}}{2}.$$

即可求出 b。即离超平面最近的正的函数间隔要等于离超平面最近的负的函数间隔。

关于上面的对偶问题如何求解，将留给下一篇中的 SMO 算法来阐明。

这里考虑另外一个问题，由于前面求解中得到

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

我们通篇考虑问题的出发点是$w^T x + b$，根据求解得到的α_i，我们代入前式得到

$$\begin{aligned} w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b. \end{aligned}$$

也就是说，以前新来的要分类的样本首先根据 w 和 b 做一次线性运算，然后看求的结果是大于 0 还是小于 0,来判断正例还是负例。现在有了α_i，我们不需要求出 w，只需将新来的样本和训练数据中的所有样本做内积和即可。那有人会说，与前面所有的样本都做运算是不是太耗时了？其实不然，我们从 KKT 条件中得到，只有支持向量的$\alpha_i > 0$，其他情况$\alpha_i = 0$。因此，我们只需求新来的样本和支持向量的内积，然后运算即可。这种写法为下面要提到的核函数（kernel）做了很好的铺垫。这是上篇，先写这么多了。