转载请注明出处：http://blog.csdn.net/luoshixian099/article/details/51227754 (http://blog.csdn.net/luoshixian099/article/details/51227754)
本文力求简化SMO的算法思想，毕竟自己理解有限，无奈还是要拿一堆公式推来推去，但是静下心看完本篇并随手推导，你会迎刃而解的。推荐参看SMO原文中的伪代码。

1.SMO概念

上一篇博客已经详细介绍了SVM原理 (http://blog.csdn.net/luoshixian099/article/details/51073885)，为了方便求解，把原始最优化问题转化成了其对偶问题，因为对偶问题是一个凸二次规划问题，这样的凸二次规划问题具有全局最优解，如下：

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

其中(x_i, y_i)表示训练样本数据，x_i为样本特征,$y_i \in \{-1, 1\}$为样本标签，C为惩罚系数由自己设定。上述问题是要求解N个参数$(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_N)$，其他参数均为已知，有多种算法可以对上述问题求解，但是算法复杂度均很大。但1998年，由Platt提出的序列最小最优化算法(SMO)可以高效的求解上述SVM问题，它把原始求解N个参数二次规划问题分解成很多个子二次规划问题分别求解，每个子问题只需要求解2个参数，方法类似于坐标上升，节省时间成本和降低了内存需求。每次启发式选择两个变量进行优化，不断循环，直到达到函数最优值。

2.SMO原理分析

2.1视为一个二元函数

为了求解N个参数$(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_N)$，首先想到的是坐标上升的思路，例如求解$\alpha_1$,可以固定其他N-1个参数，可以看成关于$\alpha_1$的一元函数求解，但是注意到上述问题的等式约束条件$\sum_{i=1}^{N} y_i\alpha_i = 0$,当固定其他参数时，参数$\alpha_1$也被固定，因此此种方法不可用。
SMO算法选择同时优化两个参数，固定其他N-2个参数，假设选择的变量为α_1, α_2,固定其他参数$\alpha_3, \alpha_4, \ldots, \alpha_N$,由于参数$\alpha_3, \alpha_4, \ldots, \alpha_N$的固定，可以简化目标函数为只关于$\alpha_1, \alpha_2$的二元函数，$Constant$表示常数项(不包含变量$\alpha_1, \alpha_2$的项)。

y的取值为-1或者1，所以y^2 = 1

$$min\ \Psi(\alpha_1, \alpha_2) = \frac{1}{2}K_{11}\alpha_1^2 + \frac{1}{2}K_{22}\alpha_2^2 + y_1y_2K_{12}\alpha_1\alpha_2 - (\alpha_1 + \alpha_2) + y_1v_1\alpha_1 + y_2v_2\alpha_2 + Constar$$

其中$v_i = \sum_{j=3}^{N}\alpha_j y_j K(x_i, x_j), i = 1, 2$

2.2视为一元函数

由等式约束得：$\alpha_1y_1 + \alpha_2y_2 = -\sum_{i=3}^{N}\alpha_i y_i = \zeta$，可见ζ为定值。
等式$\alpha_1y_1 + \alpha_2y_2 = \zeta$两边同时乘以$y_1$,且$y_1^2 = 1$，得

$$\alpha_1 = (\zeta - y_2\alpha_2)y_1 \qquad (2)$$

(2)式带回到(1)中得到只关于参数α_2的一元函数，由于常数项不影响目标函数的解，以下省略掉常数项$Constant$

$min\ \Psi(\alpha_2) = \frac{1}{2}K_{11}(\zeta-\alpha_2 y_2)^2 + \frac{1}{2}K_{22}\alpha_2^2 + y_2K_{12}(\zeta-\alpha_2 y_2)\alpha_2 - (\zeta-\alpha_2 y_2)y_1 - \alpha_2 + v_1(\zeta-\alpha$

2.3对一元函数求极值点

上式中是关于变量α_2的函数，对上式求导并令其为0得：
$\frac{\partial\Psi(\alpha_2)}{\partial\alpha_2} = (K_{11}+K_{22}-2K_{12})\alpha_2 - K_{11}\zeta y_2 + K_{12}\zeta y_2 + y_1y_2 - 1 - v_1y_2 + v_2y_2 = 0$

1.由上式中假设求得了α_2的解，带回到(2)式中可求得α_1的解，分别记为$\alpha_1^{new}, \alpha_2^{new}$,优化前的解记为$\alpha_1^{old}, \alpha_2^{old}$;由于参数$\alpha_3, \alpha_4, \ldots, \alpha_N$固定，由等式约束$\sum_{i=1}^N y_i\alpha_i = 0$有$\alpha_1^{old}y_1 + \alpha_2^{old}y_2 = -\sum_{i=3}^N \alpha_i y_i = \alpha_1^{new}y_1 + \alpha_2^{new}y_2 = \zeta$

$$\zeta = \alpha_1^{old}y_1 + \alpha_2^{old}y_2 \qquad (4)$$

2.假设SVM超平面的模型为$f(x) = w^Tx + b$,上一篇中已推导出w的表达式,将其带入得$f(x) = \sum_{i=1}^N \alpha_i y_i K(x_i, x) + b$; $f(x_i)$表示样本x_i的预测值,y_i表示样本x_i的真实值，定义E_i表示预测值与真实值之差为

$$E_i = f(x_i) - y_i \qquad (5)$$

3.由于$v_i = \sum_{j=3}^N \alpha_j y_j K(x_i, x_j), i = 1, 2$, 因此

$$v_1 = f(x_1) - \sum_{j=1}^2 y_j\alpha_j K_{1j} - b \qquad (6)$$

$$v_2 = f(x_2) - \sum_{j=1}^2 y_j\alpha_j K_{2j} - b \qquad (7)$$

把(4)(6)(7)带入下式中：
$(K_{11}+K_{22}-2K_{12})\alpha_2 - K_{11}\zeta y_2 + K_{12}\zeta y_2 + y_1y_2 - 1 - v_1y_2 + v_2y_2 = 0$
化简得: 此时求解出的α_2^{new}未考虑约束问题，先记为$\alpha_2^{new,unclipped}$：
$(K_{11}+K_{22}-2K_{12})\alpha_2^{new,unclipped} = (K_{11}+K_{22}-2K_{12})\alpha_2^{old} + y_2\left[y_2 - y_1 + f(x_1) - f(x_2)\right]$
带入(5)式，并记$\eta = K_{11}+K_{22}-2K_{12}$得：

$$\alpha_2^{new,unclipped} = \alpha_2^{old} + \frac{y_2(E_1-E_2)}{\eta} \qquad (8)$$

2.4对原始解修剪

上述求出的解未考虑到约束条件：

- $0 \le \alpha_{i=1,2} \le C$
- $\alpha_1y_1 + \alpha_2y_2 = \zeta$

在二维平面上直观表达上述两个约束条件

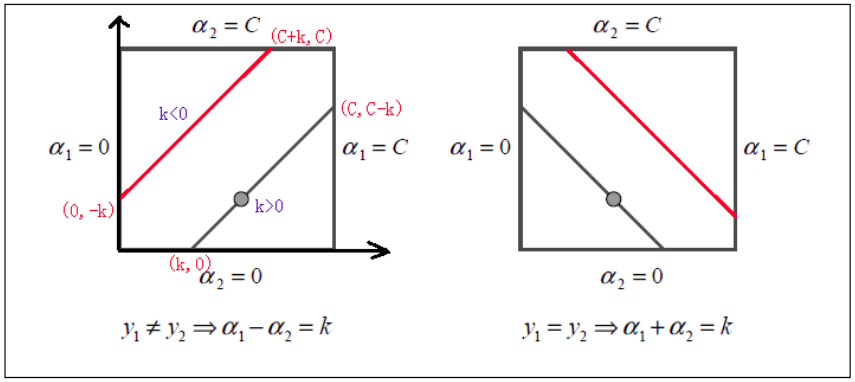

Figure 1. The two Lagrange multipliers must fulfill all of the constraints of the full problem. The inequality constraints cause the Lagrange multipliers to lie in the box. The linear equality constraint causes them to lie on a diagonal line. Therefore, one step of SMO must find an optimum of the objective function on a diagonal line segment.

最优解必须要在方框内且在直线上取得，因此$L \leq \alpha_2^{new} \leq H$;
当$y_1 \neq y_2$时，$L = max(0, \alpha_2^{old} - \alpha_1^{old}); H = min(C, C + \alpha_2^{old} - \alpha_1^{old})$
当$y_1 = y_2$时，$L = max(0, \alpha_1^{old} + \alpha_2^{old} - C); H = min(C, \alpha_2^{old} + \alpha_1^{old})$
经过上述约束的修剪，最优解就可以记为α_2^{new}了。

$$\alpha_2^{new} = \begin{cases} \mathrm{H}, \alpha_2^{new,unclipped} > H \\ \alpha_2^{new,unclipped}, L \leqslant \alpha_2^{new,unclipped} \leqslant H \\ \mathrm{L}, \alpha_2^{new,unclipped} < L \end{cases}$$

2.5求解α_1^{new}

由于其他N-2个变量固定，因此$\alpha_1^{old} y_1 + \alpha_2^{old} y_2 = \alpha_1^{new} y_1 + \alpha_2^{new} y_2$所以可求得

$$\alpha_1^{new} = \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^{new}) \qquad (9)$$

2.6取临界情况

大部分情况下，有$\eta = K_{11} + K_{22} - 2K_{12} > 0$。但是在如下几种情况下，$\alpha_2^{new}$需要取临界值L或者H.

1. $\eta < 0$,当核函数K不满足Mercer定理时，矩阵K非正定；
2. $\eta = 0$,样本x_1与x_2输入特征相同；

也可以如下理解，对(3)式求二阶导数就是$\eta = K_{11} + K_{22} - 2K_{12}$,
当$\eta < 0$时，目标函数为凸函数，没有极小值，极值在定义域边界处取得。
当$\eta = 0$时，目标函数为单调函数，同样在边界处取极值。
计算方法：
即当$\alpha_2^{new} = L$和$\alpha_2^{new} = H$分别带入(9)式中，计算出$\alpha_1^{new} = L_1$和$\alpha_1^{new} = H_1$,其中$s = y_1 y_2$

$$L_1 = \alpha_1 + s(\alpha_2 - L),$$
$$H_1 = \alpha_1 + s(\alpha_2 - H),$$

带入目标函数(1)内，比较$\Psi(\alpha_1 = L_1, \alpha_2 = L)$与$\Psi(\alpha_1 = H_1, \alpha_2 = H)$的大小，$\alpha_2$取较小的函数值对应的边界点。

$$\Psi_L = L_1 f_1 + L f_2 + \tfrac{1}{2} L_1^2 K(\vec{x}_1, \vec{x}_1) + \tfrac{1}{2} L^2 K(\vec{x}_2, \vec{x}_2) + s L L_1 K(\vec{x}_1, \vec{x}_2),$$
$$\Psi_H = H_1 f_1 + H f_2 + \tfrac{1}{2} H_1^2 K(\vec{x}_1, \vec{x}_1) + \tfrac{1}{2} H^2 K(\vec{x}_2, \vec{x}_2) + s H H_1 K(\vec{x}_1, \vec{x}_2).$$

其中

$$f_1 = y_1(E_1 - b) - \alpha_1 K(\vec{x}_1, \vec{x}_1) - s\alpha_2 K(\vec{x}_1, \vec{x}_2),$$
$$f_2 = y_2(E_2 - b) - s\alpha_1 K(\vec{x}_1, \vec{x}_2) - \alpha_2 K(\vec{x}_2, \vec{x}_2),$$

3.启发式选择变量

上述分析是在从N个变量中已经选出两个变量进行优化的方法，下面分析如何高效地选择两个变量进行优化，使得目标函数下降的最快。

第一个变量的选择

第一个变量的选择称为外循环，首先遍历整个样本集，选择违反KKT条件的α_i作为第一个变量，接着依据相关规则选择第二个变量(见下面分析),对这两个变量采用上述方法进行优化。当遍历完整个样本集后，遍历非边界样本集$(0 < \alpha_i < C)$中违反KKT的α_i作为第一个变量，同样依据相关规则选择第二个变量，对此两个变量进行优化。当遍历完非边界样本集后，再次回到遍历整个样本集中寻找，即在整个样本集与非边界样本集上来回切换，寻找违反KKT条件的α_i作为第一个变量。直到遍历整个样本集后，没有违反KKT条件α_i，然后退出。
边界上的样本对应的$\alpha_i = 0$或者$\alpha_i = C$，在优化过程中很难变化，然而非边界样本$0 < \alpha_i < C$会随着对其他变量的优化会有大的变化。

KKT条件

$$\begin{aligned} \alpha_i = 0 &\Rightarrow y^{(i)}(w^T x^{(i)} + b) \geq 1 \\ \alpha_i = C &\Rightarrow y^{(i)}(w^T x^{(i)} + b) \leq 1 \\ 0 < \alpha_i < C &\Rightarrow y^{(i)}(w^T x^{(i)} + b) = 1. \end{aligned}$$

第二个变量的选择

SMO称第二个变量的选择过程为内循环，假设在外循环中找个第一个变量记为α_1，第二个变量的选择希望能使α_2有较大的变化，由于α_2是依赖于$|E_1 - E_2|$,当E_1为正时，那么选择最小的E_i作为E_2,如果E_1为负，选择最大E_i作为E_2，通常为每个样本的E_i保存在一个列表中，选择最大的$|E_1 - E_2|$来近似最大化步长。
有时按照上述的启发式选择第二个变量，不能够使得函数值有足够的下降，这时按下述步骤:

> 首先在非边界集上选择能够使函数值足够下降的样本作为第二个变量，
> 如果非边界集上没有，则在整个样本集上选择第二个变量，
> 如果整个样本集依然不存在，则重新选择第一个变量。

4.阈值b的计算

每完成对两个变量的优化后，要对b的值进行更新，因为b的值关系到f(x)的计算，即关系到下次优化时E_i的计算。
1.如果$0 < \alpha_1^{new} < C$，由KKT条件$y_1(w^T x_1 + b) = 1$,得到$\sum_{i=1}^{N} \alpha_i y_i K_{i1} + b = y_1$,由此得：

$$b_1^{new} = y_1 - \sum_{i=3}^{N} \alpha_i y_i K_{i1} - \alpha_1^{new} y_1 K_{11} - \alpha_2^{new} y_2 K_{21}$$

由(5)式得，上式前两项可以替换为：

$$y_1-\sum_{i=3}^{N}\alpha_i y_i K_{i1}=-E_1+\alpha_1^{old}y_1K_{11}+\alpha_2^{old}y_2K_{11}+b^{old}$$

得出：

$$b_1^{new}=-E_1-y_1K_{11}(\alpha_1^{new}-\alpha_1^{old})-y_2K_{21}(\alpha_2^{new}-\alpha_2^{old})+b^{old}$$

2.如果$0<\alpha_2^{new}<C$，则

$$b_2^{new}=-E_2-y_1K_{12}(\alpha_1^{new}-\alpha_1^{old})-y_2K_{22}(\alpha_2^{new}-\alpha_2^{old})+b^{old}$$

3.如果同时满足$0<\alpha_i^{new}<C$,则$b_1^{new}=b_2^{new}$
4.如果同时不满足$0<\alpha_i^{new}<C$，则b_1^{new}与b_2^{new}以及它们之间的数都满足KKT阈值条件，这时选择它们的中点。（关于这个我不理解…）

建议参看SMO原文的伪代码

参考：
统计学习方法，李航
Sequential Minimal Optimization:A Fast Algorithm for Training Support Vector Machines，John C. Platt
http://www.cnblogs.com/jerrylead/archive/2011/03/18/1988419.html (http://www.cnblogs.com/jerrylead/archive/2011/03/18/1988419.html)