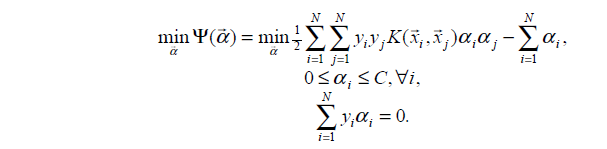
**【机器学习详解】SMO算法剖析**

<https://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：   
   
其中(xi,yi)(xi,yi)表示训练样本数据，xixi为样本特征,yi∈{−1,1}yi∈{−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(α1,α2,α3,...,αN)(α1,α2,α3,...,αN)，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(α1,α2,α3,...,αN)(α1,α2,α3,...,αN)，首先想到的是坐标上升的思路，例如求解α1α1,可以固定其他N-1个参数，可以看成关于α1α1的一元函数求解，但是注意到上述问题的等式约束条件∑Ni=1yiαi=0∑i=1Nyiαi=0,当固定其他参数时，参数α1α1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α1,α2α1,α2,固定其他参数α3,α4,...,αNα3,α4,...,αN,由于参数α3,α4,...,αNα3,α4,...,αN的固定，可以简化目标函数为只关于α1,α2α1,α2的二元函数，ConstantConstant表示常数项(不包含变量α1,α2α1,α2的项)。

min Ψ(α1,α2)=12K11α21+12K22α22+y1y2K12α1α2−(α1+α2)+y1v1α1+y2v2α2+Constant(1)min Ψ(α1,α2)=12K11α12+12K22α22+y1y2K12α1α2−(α1+α2)+y1v1α1+y2v2α2+Constant(1)

其中vi=∑Nj=3αjyjK(xi,xj),i=1,2vi=∑j=3NαjyjK(xi,xj),i=1,2

## ****2.2视为一元函数****

由等式约束得：α1y1+α2y2=−∑Ni=3αiyi=ζα1y1+α2y2=−∑i=3Nαiyi=ζ，可见ζζ为定值。   
等式α1y1+α2y2=ζα1y1+α2y2=ζ两边同时乘以y1,y1,且y21=1y12=1，得

α1=(ζ−y2α2)y1(2)α1=(ζ−y2α2)y1(2)

(2)式带回到(1)中得到只关于参数α2α2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ConstantConstant

min Ψ(α2)=12K11(ζ−α2y2)2+12K22α22+y2K12(ζ−α2y2)α2−(ζ−α2y2)y1−α2+v1(ζ−α2y2)+y2v2α2(3)min Ψ(α2)=12K11(ζ−α2y2)2+12K22α22+y2K12(ζ−α2y2)α2−(ζ−α2y2)y1−α2+v1(ζ−α2y2)+y2v2α2(3)

## ****2.3对一元函数求极值点****

上式中是关于变量α2α2的函数，对上式求导并令其为0得：   
∂Ψ(α2)∂α2=(K11+K22−2K12)α2−K11ζy2+K12ζy2+y1y2−1−v1y2+v2y2=0∂Ψ(α2)∂α2=(K11+K22−2K12)α2−K11ζy2+K12ζy2+y1y2−1−v1y2+v2y2=0

1.由上式中假设求得了α2α2的解，带回到(2)式中可求得α1α1的解，分别记为αnew1,αnew2α1new,α2new,优化前的解记为αold1,αold2α1old,α2old;由于参数α3,α4,...,αNα3,α4,...,αN固定，由等式约束∑Ni=1yiαi=0∑i=1Nyiαi=0有αold1y1+αold2y2=−∑Ni=3αiyi=αnew1y1+αnew2y2=ζα1oldy1+α2oldy2=−∑i=3Nαiyi=α1newy1+α2newy2=ζ

ζ=αold1y1+αold2y2(4)ζ=α1oldy1+α2oldy2(4)

2.假设SVM超平面的模型为f(x)=wTx+bf(x)=wTx+b,上一篇中已推导出ww的表达式,将其带入得f(x)=∑Ni=1αiyiK(xi,x)+b;f(xi)f(x)=∑i=1NαiyiK(xi,x)+b;f(xi)表示样本xixi的预测值,yiyi表示样本xixi的真实值，定义EiEi表示预测值与真实值之差为

Ei=f(xi)−yi(5)Ei=f(xi)−yi(5)

3.由于vi=∑Nj=3αjyjK(xi,xj),i=1,2vi=∑j=3NαjyjK(xi,xj),i=1,2，因此

v1=f(x1)−∑j=12yjαjK1j−b(6)v1=f(x1)−∑j=12yjαjK1j−b(6)

v2=f(x2)−∑j=12yjαjK2j−b(7)v2=f(x2)−∑j=12yjαjK2j−b(7)

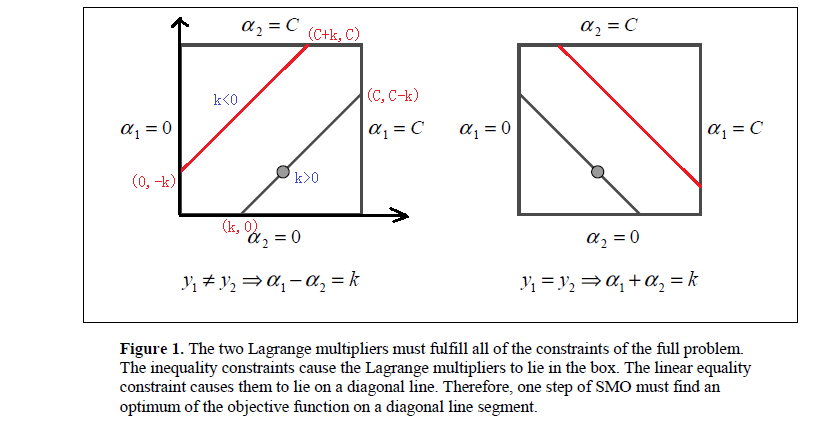
把(4)(6)(7)带入下式中：   
(K11+K22−2K12)α2−K11ζy2+K12ζy2+y1y2−1−v1y2+v2y2=0(K11+K22−2K12)α2−K11ζy2+K12ζy2+y1y2−1−v1y2+v2y2=0   
化简得: 此时求解出的αnew2α2new未考虑约束问题，先记为αnew,unclipped2α2new,unclipped：   
(K11+K22−2K12)αnew,unclipped2=(K11+K22−2K12)αold2+y2[y2−y1+f(x1)−f(x2)](K11+K22−2K12)α2new,unclipped=(K11+K22−2K12)α2old+y2[y2−y1+f(x1)−f(x2)]   
带入(5)式，并记η=K11+K22−2K12η=K11+K22−2K12得：

αnew,unclipped2=αold2+y2(E1−E2)η(8)α2new,unclipped=α2old+y2(E1−E2)η(8)

## ****2.4对原始解修剪****

上述求出的解未考虑到约束条件：

* 0≤αi=1,2≤C0≤αi=1,2≤C
* α1y1+α2y2=ζα1y1+α2y2=ζ

在二维平面上直观表达上述两个约束条件   
   
最优解必须要在方框内且在直线上取得，因此L≤αnew2≤HL≤α2new≤H;   
当y1≠y2y1≠y2时，L=max(0,αold2−αold1);H=min(C,C+αold2−αold1)L=max(0,α2old−α1old);H=min(C,C+α2old−α1old)   
当y1=y2y1=y2时，L=max(0,αold1+αold2−C);H=min(C,αold2+αold1)L=max(0,α1old+α2old−C);H=min(C,α2old+α1old)   
经过上述约束的修剪，最优解就可以记为αnew2α2new了。

αnew2=⎧⎩⎨⎪⎪ H ,αnew,unclipped2>Hαnew,unclipped2,L⩽αnew,unclipped2⩽H L ,αnew,unclipped2<Lα2new={ H ,α2new,unclipped>Hα2new,unclipped,L⩽α2new,unclipped⩽H L ,α2new,unclipped<L

## ****2.5求解****αnew1α1new

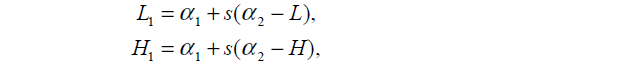
由于其他N-2个变量固定，因此αold1y1+αold2y2=αnew1y1+αnew2y2α1oldy1+α2oldy2=α1newy1+α2newy2所以可求得

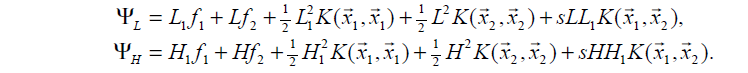
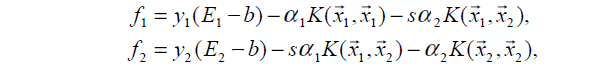
αnew1=αold1+y1y2(αold2−αnew2)(9)α1new=α1old+y1y2(α2old−α2new)(9)

## ****2.6取临界情况****

大部分情况下，有η=K11+K22−2K12>0η=K11+K22−2K12>0。但是在如下几种情况下，αnew2α2new需要取临界值L或者H.

1. η<0η<0,当核函数K不满足Mercer定理时，矩阵K非正定；
2. η=0η=0,样本x1x1与x2x2输入特征相同；

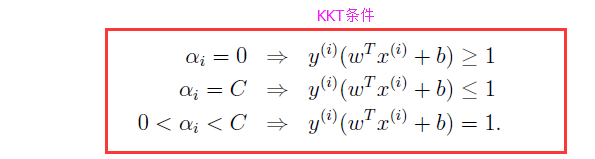
也可以如下理解，对(3)式求二阶导数就是η=K11+K22−2K12η=K11+K22−2K12,   
当η<0η<0时，目标函数为凸函数，没有极小值，极值在定义域边界处取得。   
当η=0η=0时，目标函数为单调函数，同样在边界处取极值。   
计算方法：   
即当αnew2=Lα2new=L和αnew2=Hα2new=H分别带入(9)式中，计算出αnew1=L1α1new=L1和αnew1=H1α1new=H1,其中s=y1y2s=y1y2   


带入目标函数(1)内，比较Ψ(α1=L1,α2=L)Ψ(α1=L1,α2=L)与Ψ(α1=H1,α2=H)Ψ(α1=H1,α2=H)的大小，α2α2取较小的函数值对应的边界点。   
   
其中   


# ****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αiαi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<αi<C0<αi<C)中违反KKT的αiαi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αiαi作为第一个变量。直到遍历整个样本集后，没有违反KKT条件αiαi，然后退出。   
边界上的样本对应的αi=0或者αi=Cαi=0或者αi=C，在优化过程中很难变化，然而非边界样本0<αi<C0<αi<C会随着对其他变量的优化会有大的变化。   


## ****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α1α1，第二个变量的选择希望能使α2α2有较大的变化，由于α2α2是依赖于|E1−E2||E1−E2|,当E1E1为正时，那么选择最小的EiEi作为E2E2,如果E1E1为负，选择最大EiEi作为E2E2，通常为每个样本的EiEi保存在一个列表中，选择最大的|E1−E2||E1−E2|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EiEi的计算。   
**1.**如果0<αnew1<C0<α1new<C，由KKT条件y1(wTx1+b)=1y1(wTx1+b)=1,得到∑Ni=1αiyiKi1+b=y1∑i=1NαiyiKi1+b=y1,由此得：

bnew1=y1−∑i=3NαiyiKi1−αnew1y1K11−αnew2y2K21b1new=y1−∑i=3NαiyiKi1−α1newy1K11−α2newy2K21

由(5)式得，上式前两项可以替换为：

y1−∑i=3NαiyiKi1=−E1+αold1y1K11+αold2y2K11+boldy1−∑i=3NαiyiKi1=−E1+α1oldy1K11+α2oldy2K11+bold

得出：

bnew1=−E1−y1K11(αnew1−αold1)−y2K21(αnew2−αold2)+boldb1new=−E1−y1K11(α1new−α1old)−y2K21(α2new−α2old)+bold

**2.**如果0<αnew2<C0<α2new<C，则

bnew2=−E2−y1K12(αnew1−αold1)−y2K22(αnew2−αold2)+boldb2new=−E2−y1K12(α1new−α1old)−y2K22(α2new−α2old)+bold

**3.**如果同时满足0<αnewi<C0<αinew<C,则bnew1=bnew2b1new=b2new   
**4.**如果同时不满足0<αnewi<C0<αinew<C，则bnew1与bnew2b1new与b2new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>