

支持向量机：SMO算法剖析

勿在浮砂筑高台 机器学习算法那些事 2018-11-22

作者：勿在浮砂筑高台

来源：CSDN

链接：

<https://blog.csdn.net/luoshixian099/article/details/51227754>

SMO算法概括：SMO算法是序列最小最优化算法，如下图，首先选择两个变量，固定其他变量，针对这两个变量构建一个二次规划问题，求解目标函数最小值对应的拉格朗日乘子 α_1 和 α_2 ；然后，选择两个变量的方法：第一个变量遍历所有样本的拉格朗日乘子 α ，选择违反KKT条件最严重的变量 α_1 ；第二个变量遍历所有样本，选择与第一个变量变化值最大（ $|E_2 - E_1|$ ）的拉格朗日乘子 α_2 。

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^N \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \end{aligned}$$

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

上一篇文章已经详细介绍了SVM原理，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：

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

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 $(\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_N)$ ，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 $(\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_N)$ ，首先想到的是坐标上升的思路，例如求解 α_1 ，可以固定其他N-1个参数，可以看成关于 α_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 α_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 $\alpha_3, \alpha_4, \dots, \alpha_N$ ，由于参数 $\alpha_3, \alpha_4, \dots, \alpha_N$ 固定，可以简化目标函数为只关于 α_1, α_2 的二元函数，Constant表示常数项(不包含变量 α_1, α_2 的项)，如下：

$$\min \Psi(\alpha_1, \alpha_2) = \frac{1}{2} K_{11} \alpha_1^2 + \frac{1}{2} K_{22} \alpha_2^2 + y_1 y_2 K_{12} \alpha_1 \alpha_2 - (\alpha_1 + \alpha_2) + y_1 v_1 \alpha_1 + y_2 v_2 \alpha_2 + \text{Constant} \quad (1)$$

$$\text{其中 } K_{ij} = K(x_i, x_j)$$

2.2 视为一元函数

由等式约束得：

$$\alpha_1 y_1 + \alpha_2 y_2 = - \sum_{i=3}^N y_i \alpha_i = \zeta$$

其中， ζ 为定值。

等式 $\alpha_1 y_1 + \alpha_2 y_2 = \zeta$ 两边同时乘以 y_1 ，且 $y_1^2 = 1$ ，得

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

(2)式代入(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

$$\min \Psi(\alpha_2) = \frac{1}{2} K_{11} (\zeta - \alpha_2 y_2)^2 + \frac{1}{2} K_{22} \alpha_2^2 + y_2 K_{12} (\zeta - \alpha_2 y_2) \alpha_2 - (\zeta - \alpha_2 y_2) y_1 - \alpha_2 + v_1 (\zeta - \alpha_2 y_2) + y_2 v_2 \alpha_2 \quad (3)$$

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：

$$\frac{\partial \Psi(\alpha_2)}{\partial \alpha_2} = (K_{11} + K_{22} - 2K_{12})\alpha_2 - K_{11}\zeta y_2 + K_{12}\zeta y_2 + y_1 y_2 - 1 - v_1 y_2 + v_2 y_2 = 0, (2.3)$$

1. 由上式中假设求得 α_2 的解，代入(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 $\alpha_3, \alpha_4, \dots, \alpha_N$ 固定，由等式约束 $\sum_{i=1}^N y_i \alpha_i = 0$ ，得

$$\alpha_1^{old} y_1 + \alpha_2^{old} y_2 = -\sum_{i=3}^N y_i \alpha_i = \alpha_1^{new} y_1 + \alpha_2^{new} y_2 = \zeta$$

即：

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

2. 假设SVM超平面的模型为 $f(x) = w^T x + b$ ，支持向量机（二）已推导出 w 的表达式，将其代入模型得：

$$f(x) = \sum_{i=1}^N \alpha_i y_i K(x_i, x) + b$$

$f(x_i)$ 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：

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

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

3. 由于，因此，由（5）式得：

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

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

把（4）（6）（7）代入（2.3）式得：

$$(K_{11} + K_{22} - 2K_{12})\alpha_2 - K_{11}\zeta y_2 + K_{12}\zeta y_2 + y_1 y_2 - 1 - v_1 y_2 + v_2 y_2 = 0$$

化简求解 α_2^{new} ，未考虑约束问题，先记为 $\alpha_2^{new, unclipped}$ ，得：

$$(K_{11} + K_{22} - 2K_{12})\alpha_2^{new, unclipped} = (K_{11} + K_{22} - 2K_{12})\alpha_2^{old} + y_2[y_2 - y_1 + f(x_1) - f(x_2)]$$

代入 (5) 式，并记 $\eta = K_{11} + K_{22} - 2K_{12}$ ，得：

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

2.4 对原始解修剪

上述求出的解未考虑到约束条件：

- $0 \leq \alpha_{i-1,2} \leq C$
- $\alpha_1 y_1 + \alpha_2 y_2 = \zeta$

在二维平面上直观表达上述两个约束条件

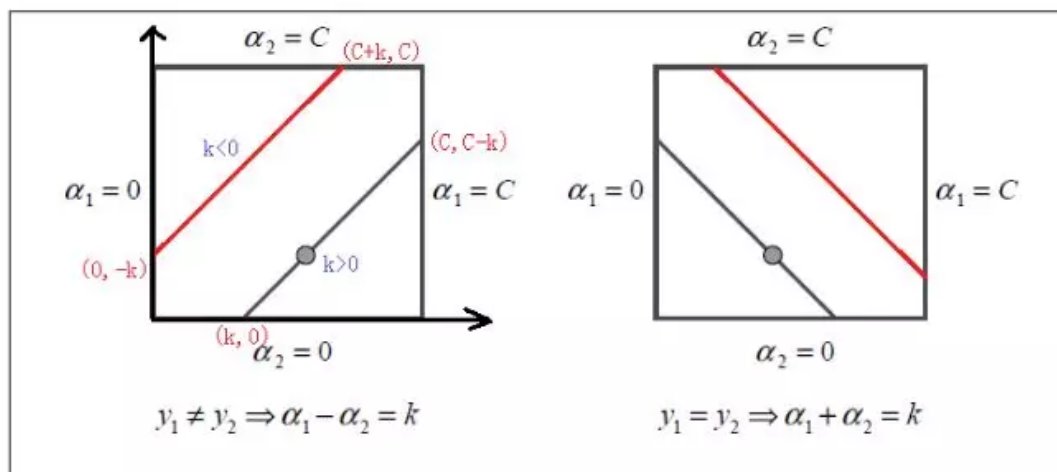


Figure 1. The two Lagrange multipliers must fulfill all of the constraints of the full problem. The inequality constraints cause the Lagrange multipliers to lie in the box. The linear equality constraint causes them to lie on a diagonal line. Therefore, one step of SMO must find an optimum of the objective 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， $L = \max(0, \alpha_1^{old} + \alpha_2^{old} - C)$; $H = \min(C, \alpha_2^{old} + \alpha_1^{old})$

过上述约束的修剪，最优解就可以记为 α_2^{new} 了。

$$\alpha_2^{\text{new}} = \begin{cases} H, & \alpha_2^{\text{new,unc}} > H \\ \alpha_2^{\text{new,unc}}, & L \leq \alpha_2^{\text{new,unc}} \leq H \\ L, & \alpha_2^{\text{new,unc}} < L \end{cases}$$

2.5 求解 α_1^{new}

由于其他N-2个变量固定，因此

$$\alpha_1^{\text{old}} y_1 + \alpha_2^{\text{old}} y_2 = \alpha_1^{\text{new}} y_1 + \alpha_2^{\text{new}} y_2$$

所以可求得

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

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， α_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x1与x2输入特征相同；

也可以如下理解，对（3）式求二阶导数得：

$$\eta = K_{11} + K_{22} - 2K_{12}$$

计算方法：

即当 $\alpha_2^{\text{new}} = L$ 和 $\alpha_2^{\text{new}} = H$ 分别代入（9）式中，

令 $\alpha_1^{\text{new}} = L_1$ 和 $\alpha_1^{\text{new}} = H_1$ ，其中 $s = y_1 y_2$ ，得：

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

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

代入目标函数（1）内，比较 $\Psi(\alpha_1 = L_1, \alpha_2 = L_2)$ 与 $\Psi(\alpha_1 = H_1, \alpha_2 = H_2)$ 的大小， α_2 取较小的函数值对应的边界点。

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

其中,

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

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。

3.1 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

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

3.2 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 有较大的变换，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，则在整体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。

写在结束的话

SMO算法是求解拉格朗日乘子 α_i 的一种快速实现算法，SMO算法比较抽象，若花了较长时间还没有弄懂的话就跳过吧，会用就行了。

推荐阅读文章

[支持向量机（三）：图解KKT条件和拉格朗日乘子法](#)

[支持向量机（二）：算法详细解析](#)

[支持向量机（一）：支持向量机的分类思想](#)



长按二维码关注

机器学习算法那些事

微信：beautifulife244

砥砺前行 不忘初心