线性判别函数

Author： 中山大学 17数据科学与计算机学院 YSY

https://github.com/ysyisyourbrother

简单线性判别函数

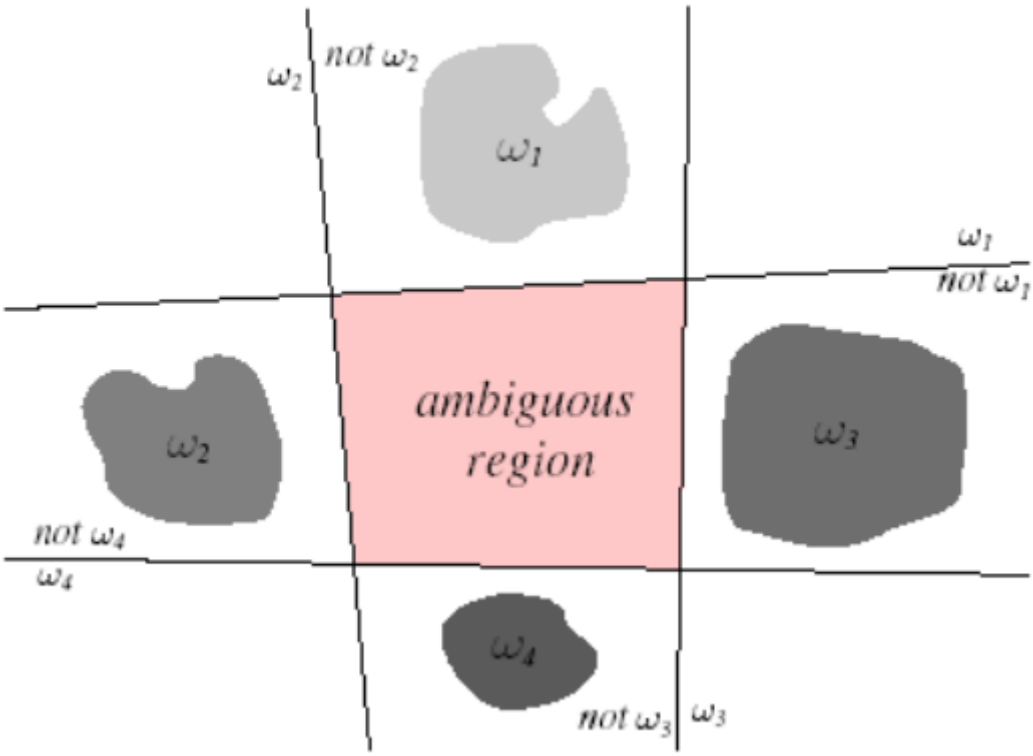
单分类

$$g(x) = w^T x + w_0 = r||w||$$
$$r = \frac{g(x)}{||w||}$$

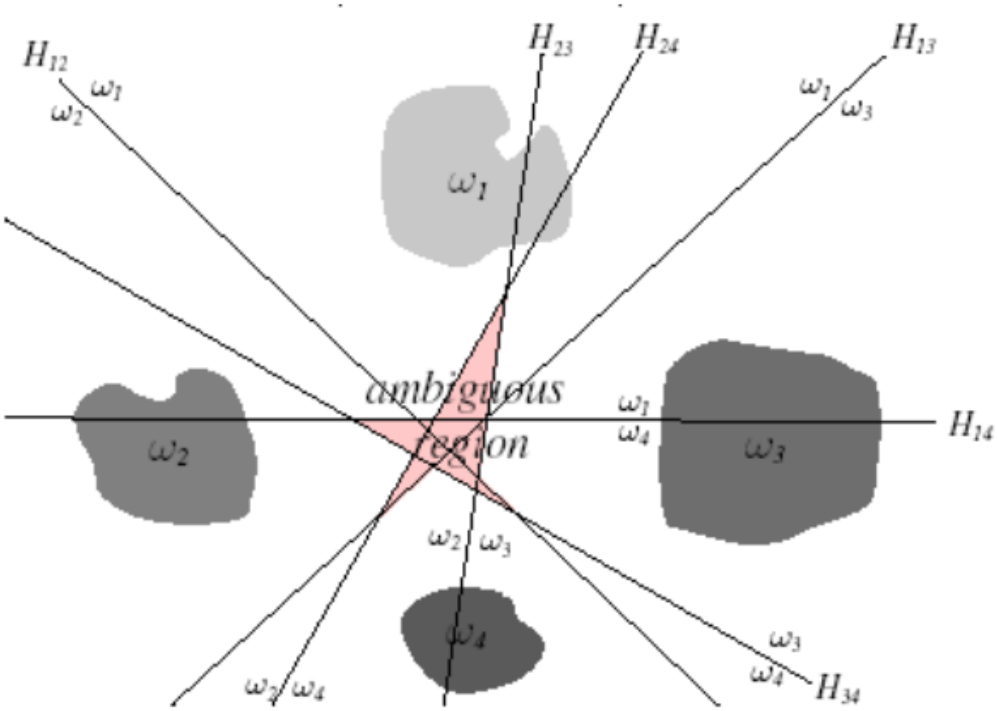
利用r可以进行分类。

多分类

- ovr



- ovo



判决面

$$(w_i - w_j)^T x + (w_{i0} - w_{j0}) = 0$$

算数距离

$$r = \frac{g_i(x) - g_j(x)}{||w_i - w_j||}$$

线性机的判决区域是凸的，限制了分类器的适应性和精确性。另外，判决区域是单连通的，使得对条件概率密度$p(x|w_i)$为单峰的问题设计线性机很适合。

广义线性判别函数

二次判别函数

$$g(x) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} x_i x_j$$

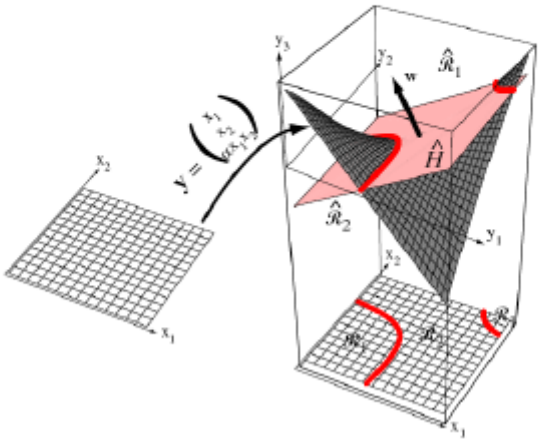
写成更一般的形式

$$g(x) = \sum_{i=1}^{\hat{d}} a_i y_i(x)$$

其中，a是权向量，$y_i(x)$可以是任意的函数，用于将在x空间上的d维的点映射到y空间上的$\hat{d}$维的点。

如下面图。原本的数据点可能只有x_1和x_2两个属性，在二维平面上分布，很难区分，如果加入了$x_1 x_2$这个属性，原本的下图中二维平面上的点就投影成了一个马鞍型，然后取一个超平面分割，就可以划分出线性判别函数无法划分的结果。

$$g(x) = x_1 + x_2 + \alpha x_1 x_2, \quad y = \begin{pmatrix} x_1 \\ x_2 \\ \alpha x_1 x_2 \end{pmatrix}$$



优点

- 能解决在低维线性不可分的问题

缺点

- 维数灾难
- 要求大量的训练样本
- 基于假设：**映射到高位空间并不给数据附加错误的结构及相关性。**

两类的线性可分情况：

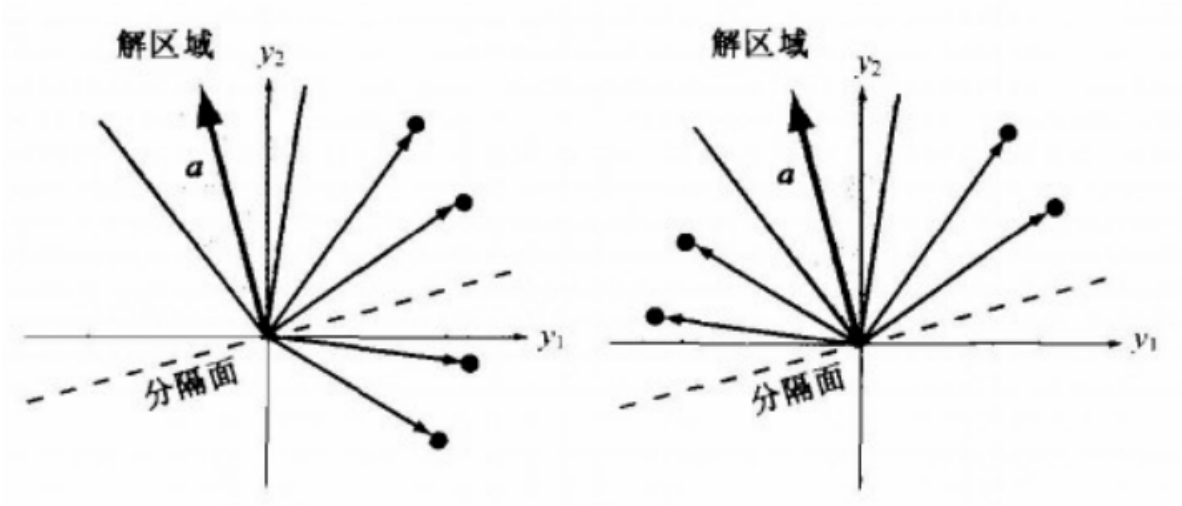
假设一个包含n个样本的集合，数据都有标签，我们希望找到一个判别函数$g(x) = a^t y$的权向量a。假设这个解的错误概率很小，我们就把它当成一个正确分类的权向量，如果这个向量存在，那这些样本就可以被称为**“线性可分”**

$a^t y_i > 0$就标记为第一类，小于0则为第二类。**如果我们把负样本都乘一个负数，让我们对所有样本找一个$a^t y > 0$的权向量a，这样就是一种规范化操作。**这样的解a称为**分离向量**或者**解向量**

求解分离向量：

对每一个样本，等式$a^t y_i = 0$确定了一个穿过权空间原点的超平面（维数由a的维数决定），y_i为法向量（此时a是未知，y是已知）。

解向量必须在每个超平面的正侧也就是$a^t y > 0$的那一侧（**在二维上测试就可以发现，a向量指向的那一侧，就是$a^t y > 0$因的那一侧**），然后同方向形成的交集区域就是**解区域**



更新权向量算法

解向量存在通常不是唯一的，它们构成解区域。

1. 基本梯度下降算法

初始化 **a** 以及阈值 θ, $\eta(\bullet)$,$k \leftarrow 0$
  do $k \leftarrow k+1$
    $a \leftarrow a-\eta(k)\nabla J(\mathbf{a})$
  until $\left|\eta(k)\nabla J(\mathbf{a})\right| < \theta$
  return **a**
end

$$a(k+1) = a(k) - \eta \nabla J(a(k))$$

设定学习率的原则：

将要优化的损失函数使用泰勒展开到二阶。注意如果是J是二阶的，那么海塞矩阵H不变，因为二阶求两次导后和变量无关。

$$J(a) \approx J(a(k)) + \nabla J^T(a - a(k)) + \frac{1}{2}(a - a(k))^T H(a - a(k))$$

$$J(a(k+1)) \approx J(a(k)) - \eta(k)||\nabla J||^2 + \frac{1}{2}\eta^2(k)\nabla J^T H \nabla J$$

为让梯度∇J取得极大值（可证明二阶偏导数小于零），对$\eta(k)$求偏导数，得

$$\eta(k) = \frac{||\nabla J||^2}{\nabla^T J H \nabla J}$$

2. 牛顿下降法

初始化 $\mathbf{a}$和阈值 θ

do $\mathbf{a} \leftarrow \mathbf{a} - \mathbf{H}^{-1}\nabla J(\mathbf{a})$

until $\left|\mathbf{H}^{-1}\nabla J(\mathbf{a})\right| < \theta$

return $\mathbf{a}$

end

当H为奇异矩阵时，不能使用牛顿算法。此外，计算H的逆的时间复杂度为$O(n^3)$。

Q：为什么直接求了最优的步长，还要迭代更新?

因为越靠近最低点，斜率越小。而上面损失函数只展开到了二阶，如果要计算的点a(k+1)和a(k)相对远一点，a(k+1)点在上面的展开到二阶的式子中更近就会到达最低点，而真实的损失函数中可能还没到最低点，因此不能一步到位，要一步步学习。

3. 批处理感知器算法

只考虑$a^t y_i > 0$也就是规范后的样本点，**为了统一正负标签的样本，把负标签的样本点属性都取相反数**，这样我们就可以保证对所有正确样本点有$a^t y_i > 0$，于是对错误样本点，我们就会得到$a^t y_i < 0$，我们要最大化这个值，就相当于最小化取反的值

我们可以统计错分的样本数来作为损失函数J:

$$J_P(a) = \sum_{y \in \gamma}(-a^t y)$$

这个式子的值和错分样本到判决边界的距离之和成正比，我们目标是最小化这个函数。

$J_p(\mathbf{a}) = \sum_{\mathbf{y}\in Y}\left(-\mathbf{a}^t\mathbf{y}\right)$, Y：被$\mathbf{a}$错分的样本集

$\nabla J_p = \sum_{\mathbf{y}\in Y}\left(-\mathbf{y}\right)$, $\mathbf{a}(k+1) = \mathbf{a}(k) + \eta(k)\sum_{\mathbf{y}\in Y}\mathbf{y}$

初始化 $\mathbf{a}$, $\eta(\bullet)$, 准则 θ, $k \leftarrow 0$

do $k \leftarrow k+1$

$\mathbf{a} \leftarrow \mathbf{a} + \eta(k)\sum_{\mathbf{y}\in Y_k}\mathbf{y}$

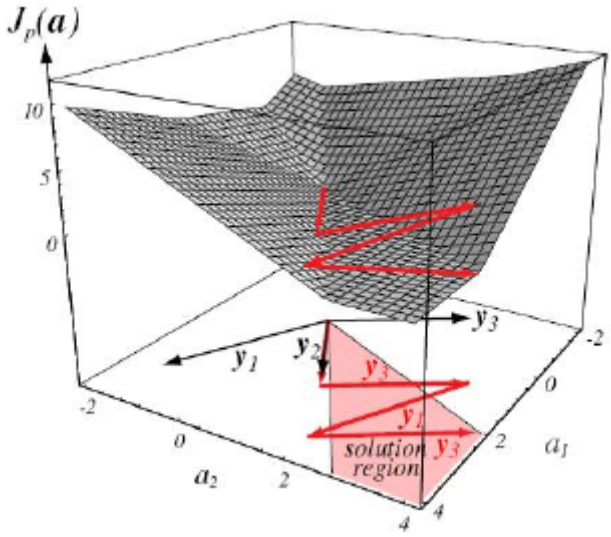
until $\left|\eta(k)\sum_{\mathbf{y}\in Y_k}\mathbf{y}\right| < \theta$

return $\mathbf{a}$

end

下面是一个例子:

从a(1)=0和学习率为1开始求解向量。



4. 固定增量单样本感知器

```
begin
初始化 a, k ← 0
    do k ← (k+1) mod n
        if y_k 被 a 错分， then a ← a + y_k
    until 所有模式被正确分类
    return a
end
```

定理5.1 如果训练样本是线性可分的，则固定增量单样本感知器给出的权向量序列必定终止于某个解向量。

证明

设$\hat{a}$为解向量，则有$\hat{a}^T y > 0, \forall i$ 令α为一个正的比例因子

$$a(k+1) - \alpha\hat{a} = a(k) - \alpha\hat{a} + y^k$$
$$||a(k+1) - \alpha\hat{a}||^2 = ||a(k) - \alpha\hat{a}||^2 + 2(a(k) - \alpha\hat{a})^T y^k + ||y^k||^2$$

由于$a(k)^T y^k$是错分类的样本，有$a(k)^T y^k < 0$。

$$||a(k+1) - \alpha\hat{a}||^2 \leq ||a(k) - \alpha\hat{a}|| - 2\alpha\hat{a}^T y^k + ||y^k||^2$$

令

$$\beta^2 = \max_i ||y_i||^2$$
$$\gamma = \min_i \hat{a}^T y_i > 0$$

则

$$||a(k+1) - \alpha\hat{a}||^2 \leq ||a(k) - \alpha\hat{a}|| - 2\alpha\gamma + \beta^2$$

选择$\alpha = \frac{\beta^2}{\gamma}$，则有

$$||a(k+1)-\alpha\hat{a}||^2 \leq ||a(k)-\alpha\hat{a}||^2 - \beta^2 \leq ||a(1)-\alpha\hat{a}||^2 - k\beta^2$$

由于平方距离不能是负的，经过$k_0 = \frac{||a(1)-\alpha\hat{a}||^2}{\beta^2}$次后迭代终止。

5. 带裕量的变增量感知器

为了让递归过程不收敛到边界点上，可以引入边沿裕量，就相当于样本点和超平面的距离没超过多少，即使分对了，也要改进。边沿裕量：

$$a^t(k)y^k \leq b$$

初始化 $\mathbf{a}$ 、阈值 θ, 裕量 $b, \eta(\bullet), k \leftarrow 0$
do $k \leftarrow (k+1) \bmod n$
if $\mathbf{a}^t\mathbf{y}^k \leq b$ then $\mathbf{a} \leftarrow \mathbf{a} + \eta(k)\mathbf{y}^k$
until $\mathbf{a}^t\mathbf{y}^k > b$ for all $k = 1, \cdots, n$
return $\mathbf{a}$
end

在二维平面上绘图可以看到，对正确的点再进行这个算法，确实可以让超平面和样本点的距离增大。

也可以同时考虑多个错分的点，用批处理的方法：

$$y^k = \sum_{y \in \gamma_k} y$$

理论上，对于任何有限的可分样本集，对任意的初始权向量，对任意非负的裕量，对任意比例因子$\eta(k)$都能得到解。但是在实践中，边界裕量最好接近$\eta(k)||y^k||^2$，而且y^k分量的比例因子对算法会产生很大的影响。

6. 松弛算法

将 $J_p(a) = \sum_{y \in Y} -a^T y$ 替换成 $J_q(a) = \sum_{y \in Y} (a^T y)^2$

然而，J_q在解区域过于光滑，容易收敛到边界上，除此之外，它得到的值依赖模值最大的样本向量。故改用**归一化后再加上裕量**的准则函数：

$$J_r = \frac{1}{2} \sum_{y \in Y} \frac{(a^T y - b)^2}{||y||^2}$$

$$\nabla J_r = \sum_{y \in Y} \frac{a^T y - b}{||y||^2} y$$

这里的$\gamma(a)$是满足$a^t y \leq b$的错误样本集。当错误样本集为空的时候，梯度为0，停止更新。这里为减b是因为当损失函数不断梯度下降，最终都会收敛到裕量为b的情况。

因此得到的改进批处理裕量松弛算法为：

7. 平衡Winnow

initialize $\mathbf{a}^+, \mathbf{a}^-, \eta(\bullet), k \leftarrow 0, \alpha > 1$

$z_k = \mathrm{Sgn}[\mathbf{a}^{+t}\mathbf{y}_k - \mathbf{a}^{-t}\mathbf{y}_k]$---（判别模式是否被错分）

if $z_k = 1$ then $a_i^+ \leftarrow \alpha^{+y_i} a_i^+, a_i^- \leftarrow \alpha^{-y_i} a_i^-$ for all i

if $z_k = -1$ then $a_i^+ \leftarrow \alpha^{-y_i} a_i^+, a_i^- \leftarrow \alpha^{+y_i} a_i^-$ for all i

return $\mathbf{a}^+, \mathbf{a}^-$

end

其中，α^{+y_i}表示增加因子，$\alpha^{+y_i} > 1$;

α^{-y_i}表示减少因子，$1 > \alpha^{+y_i} > 0$.

两个向量之间的间隔始终不会变大。收敛性比感知器的收敛性定理更加一般化。通常也比感知器计算法收敛的更快。

最小平方误差方法

要求解

$$Ya = b$$

如果Y是非奇异矩阵，可以用矩阵的逆求解，但假如Y是一个长方形矩阵，a通常没有精确的解，但我们可以寻找一个权向量a，它使得Ya和b最接近，定义MSE误差如下：

$$J = |Ya - b|^2 = \sum_{i=1}^{n} (a^t y_i - b_i)^2$$

直接求导等于0就可以得到

$$Y^t Ya = Y^t b$$

因为Y的转置和Y的乘积是方阵而且一般是可逆的于是我们可以定义出伪逆来求解原来的方程

伪逆 $Y^+ = (Y^T Y)^{-1} Y^T$，当Y是方针且非奇异，那么伪逆矩阵就是Y的逆矩阵。

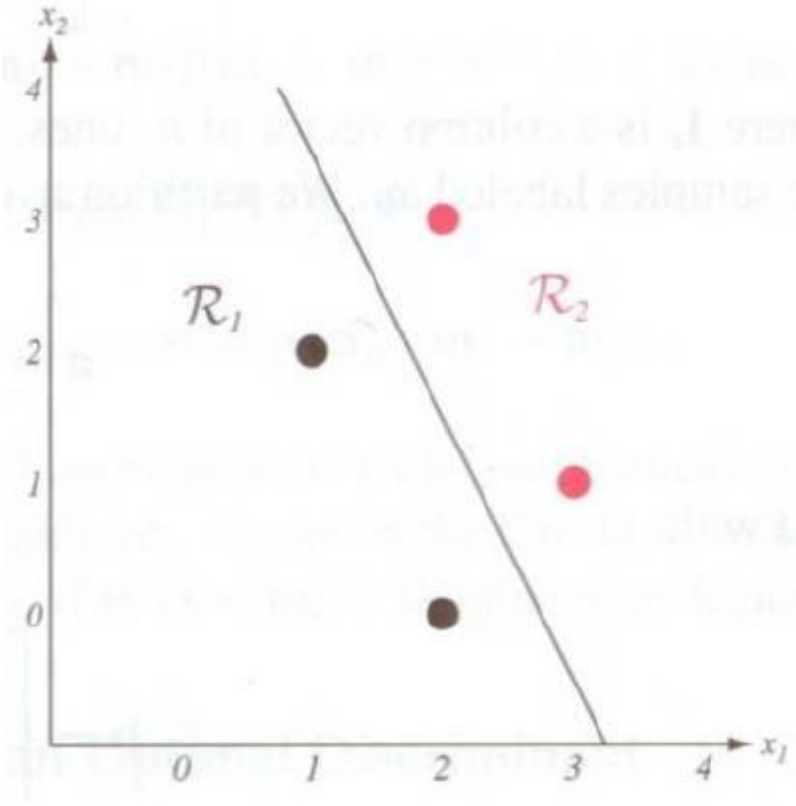
对于最小平方误差 $Y^T Ya = Y^T b$，由于b的不同，解的性质也不同，我们希望通过最小化平方误差准则函数，得到一个在可分和不可分情况下都有用的判别函数。这里的b代表了裕量。

$$\omega_1:(1,2)^t,(2,0)^t;\quad \omega_2:(3,1)^t,(2,3)^t$$

$$\text{判别边界}:\mathbf{a}^t\begin{pmatrix}1\\x_1\\x_2\end{pmatrix}=0$$

$$\mathbf{Y}=\begin{pmatrix}1&1&2\\1&2&0\\-1&-3&-1\\-1&-2&-3\end{pmatrix},\mathbf{b}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix},\mathbf{a}=\mathbf{Y}^{+}\mathbf{b}=\begin{pmatrix}11/3\\-4/3\\-2/3\end{pmatrix}$$

先对不同类的进样本点行规范化，然后根据属性(1,X1,X2)组成Y矩阵，随意设置裕量都为1，解方程即可



MSE 与 Fisher 线性判别的联系

$$\mathbf{Y}=\begin{pmatrix}\mathbf{1}_1&\mathbf{X}_1\\-\mathbf{1}_2&-\mathbf{X}_2\end{pmatrix},\mathbf{a}=\begin{pmatrix}w_0\\\mathbf{w}\end{pmatrix},\mathbf{b}=\begin{pmatrix}\frac{n}{n_1}\mathbf{1}_1\\\frac{n}{n_2}\mathbf{1}_2\end{pmatrix}$$

利用$Y^TYa=Y^Tb$，可得$w=\alpha nS_W^{-1}(m_1-m_2)$，其中$S_W=\sum_{i=1}^{2}\sum_{x\in D_i}(x-m_i)(x-m_i)^T$

MSE 与 Bayes 判别函数的联系

当 $b = 1_n$的时候，MSE的解以最小均方误差逼近贝叶斯判别函数。

Ho-Kashyap 算法

感知器和松弛算法对于不可分的情况不收敛，MSE方法不管样本是否可分，都能得到一个权向量，但是不能保证在可分的情况下，这个向量一定是分类向量。所以对b也进行更新。为了保证$b > 0$，更新的时候对负方向的梯度截断。

$$\min_{\mathbf{a},\mathbf{b}} J_s(\mathbf{a},\mathbf{b}) = \|\mathbf{Ya}-\mathbf{b}\|^2 \quad \text{subject to } \mathbf{b}>0$$

$$\nabla_{\mathbf{a}} J_s = 2\mathbf{Y}^t(\mathbf{Ya}-\mathbf{b}), \quad \nabla_{\mathbf{b}} J_s = -2(\mathbf{Ya}-\mathbf{b})$$

$$\mathbf{a}(k) = \mathbf{Y}^+\mathbf{b}(k)$$

---

start with $\mathbf{b} > 0$ and let

$$\mathbf{b}(k+1) = \mathbf{b}(k) - \eta(k)\left[\nabla_{\mathbf{b}} J_s - \left|\nabla_{\mathbf{b}} J_s\right|\right]$$

$$= \mathbf{b}(k) + 2\eta(k)[(\mathbf{Ya}-\mathbf{b}) + |(\mathbf{Ya}-\mathbf{b})|]$$

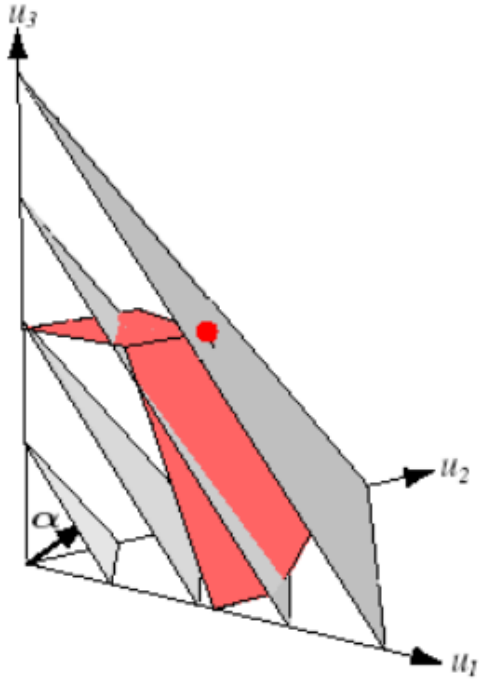
Ho-Kashap rule:

$$\mathbf{b}(1) > 0, \quad \mathbf{b}(k+1) = \mathbf{b}(k) + 2\eta(k)\mathbf{e}^+(k)$$

$$\mathbf{e}^+(k) = \frac{1}{2}(\mathbf{e}(k) + |\mathbf{e}(k)|), \mathbf{e}(k) = \mathbf{Ya}(k) - \mathbf{b}(k)$$

$$\mathbf{a}(k) = \mathbf{Y}^+\mathbf{b}(k)$$

线性规划



$$\min_{\mathbf{u}} \boldsymbol{\alpha}^t \mathbf{u}$$

$$\text{s.t.} \mathbf{Au} \geq \boldsymbol{\beta}$$

常用求解方法：

单纯形法

其中额外要求$\mathbf{u} \geq 0$

为了将线性判别与线性规划联系起来，我们可以发现把u看成权向量a是不对的，因为大多数情况下，权向量有负的分量。但是我们可以通过如下技巧构造权向量a

$$a = a^+ - a^-$$
$$a^+ = \frac{1}{2}(|a| + a)$$
$$a^- = \frac{1}{2}(|a| - a)$$

引入人工变量$\tau \geq 0$，在线性可分的情况下，有

$$a^T y_i + \tau \geq b_i$$

我们希望得到满足上式的极小化的τ，如果$\tau = 0$，那么样本线性可分，且我们可以得到解。如果$\tau > 0$，那么就没有分裂向量，此时可以证明样本是不可分的。

$$\min J_p^{'}(\mathbf{a}) = \sum_{\mathbf{y}_i \in Y'} \left(b_i - \mathbf{a}^t \mathbf{y}_i \right), \quad Y' = \left\{ \mathbf{y}_i \,|\, \mathbf{a}^t \mathbf{y}_i \le b_i \right\}$$

等价问题:

$$\min_{\tau} \; z = \sum_{i=1}^{n} \tau_i \text{ subject to } \tau_i \ge 0, \mathbf{a}^t \mathbf{y}_i + \tau_i \ge b_i$$

等价问题:

$$\min_{\mathbf{u}} \; \boldsymbol{\alpha}^t \mathbf{u} \text{ subject to } \mathbf{A}\mathbf{u} \ge \boldsymbol{\beta}, \mathbf{u} \ge 0$$

其中

$$\mathbf{u} = \begin{bmatrix} \mathbf{a}^+ \\ \mathbf{a}^- \\ \boldsymbol{\tau} \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \mathbf{y}_1^t & -\mathbf{y}_1^t & 1 & 0 & \cdots & 0 \\ \mathbf{y}_2^t & -\mathbf{y}_2^t & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_n^t & -\mathbf{y}_n^t & 0 & 0 & \cdots & 1 \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{1}_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$