第五章

线性判别函数

线性判别函数

中山大學

- 已知判别函数的参数形式，用训练的方法来估计判别函数的参数值；
- 不要求知道有关的概率密度函数的确切的参数形式，注意在高斯模型且协方差相等情形下判别函数形式为线性；
- 判别函数形式为线性（即样本分量的某种线性函数）；
- 特点是简单，然而判别结果未必为最优。
- 寻找线性判别函数的问题将被形式化为极小化准则问题，通常采用梯度下降法来求解。

本章目录

中山大學

3

3

本章目录(*Cont.*)

中山大學

4

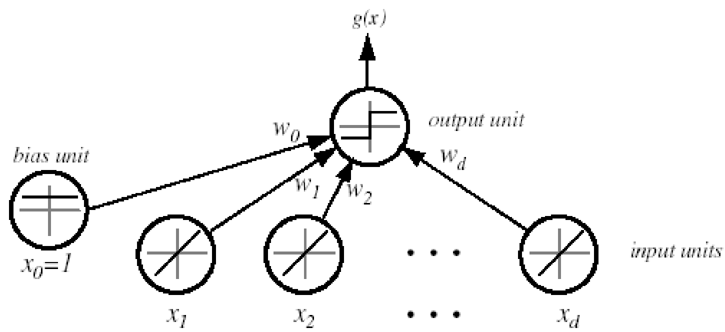
4

5.2 线性判别函数和判定面

两类情况的线性判别函数

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + \mathbf{w}_0$$

w称为权向量，$\mathbf{w}_0$为偏置



而c类问题将有c个线性判别函数。

5

判定面:

如果 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 都在判定面 $g(\mathbf{x})=0$上, 则

$$\mathbf{w}^t(\mathbf{x}_1-\mathbf{x}_2)=0 \quad (\text{正交})$$

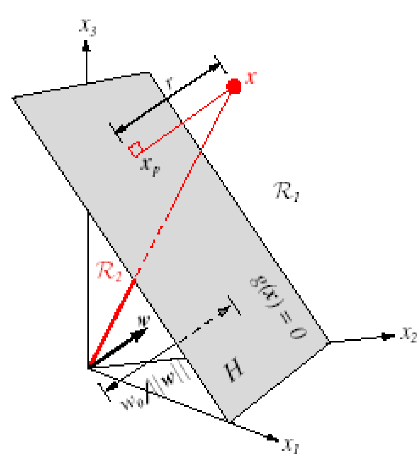
$$\mathbf{x}=\mathbf{x}_p+r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

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

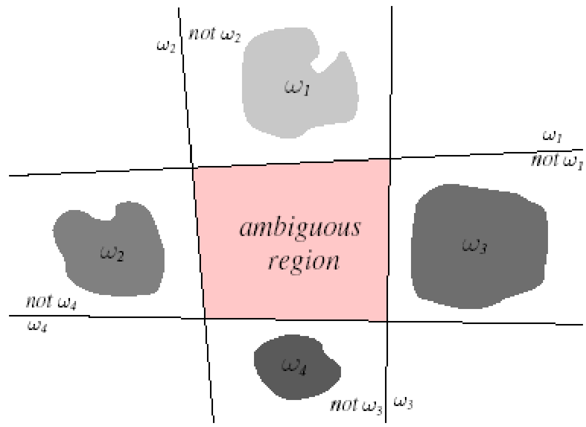
$$=r\|\mathbf{w}\|$$

$$r=\frac{g(\mathbf{x})}{\|\mathbf{w}\|}$$

$r>0$, 正侧； $r<0$为负侧。



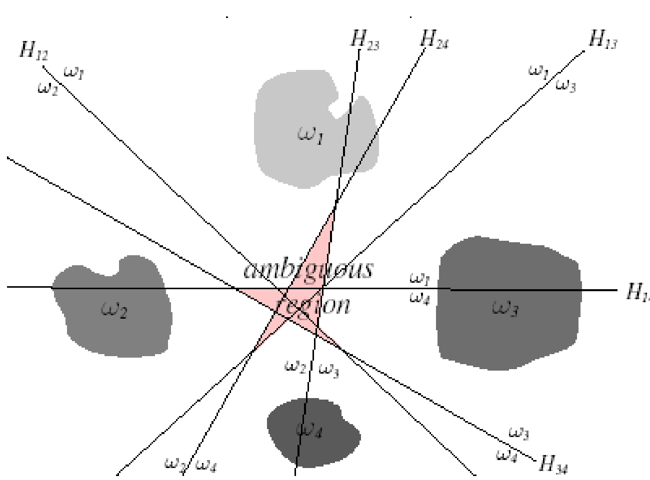
多类情况(1)



7

7

多类情况(2)

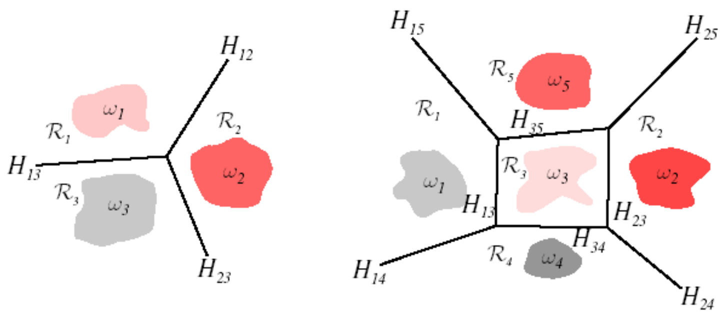


8

8

线性机 (Linear Machine)

中山大學



$g_i(x) = w_i^t x + w_{i0}, \quad i = 1, \cdots, c$

如果对所有的 $j \neq i$ 有 $g_i(\mathbf{x}) > g_j(\mathbf{x})$，则判定

$x \in \omega_i$

相邻区域的分界是超平 面 H_{ij} 的一部分：

$g_i(\mathbf{x}) = g_j(\mathbf{x}) \quad \text{or} \quad (\mathbf{w}_i - \mathbf{w}_j)^t x + (w_{i0} - w_{j0}) = 0$

9

9

线性机 (Linear Machine)

中山大學

- **线性机的判别区域是凸的**
  - **限制了分类器的适应性和精确性**
- **每个判别区域是单连通**
  - **适应条件概率密度 $p(x|w_i)$ 为单峰的问题**

10

10

5.3 广义线性判别函数

二次判别函数

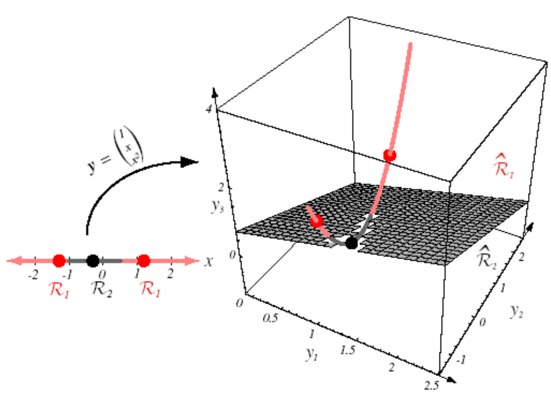
$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij} x_i x_j$$

广义线性判别函数

$$g(\mathbf{x}) = \sum_{i=1}^{\hat{d}} a_i y_i(\mathbf{x}) \text{ or } g(\mathbf{x}) = \mathbf{a}^t \mathbf{y}$$

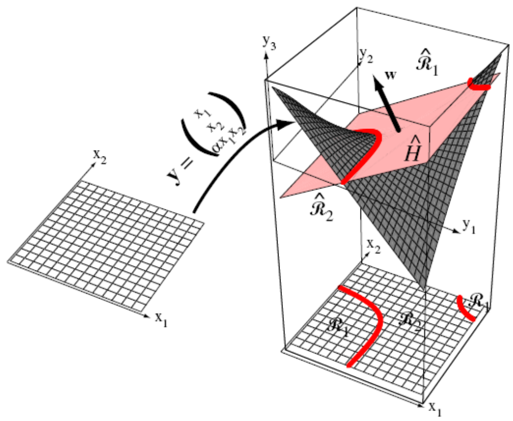
例子

$$g(x) = a_1 + a_2 x + a_3 x^2, \quad y = \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix}$$



中山大學

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



13

13

中山大學

广义判别函数的优缺点

- **优点**
  - 在高维空间，能得到简单的判定面
- **缺点**
  - 维数灾难
  - 例如，一个完整的二次型判别函数包含项的个数是 $(d+1)(d+2)/2$
  - 要求大量训练样本，

$\hat{d} > d$, 代表自由度

14

14

增广向量（Augmented Vectors）

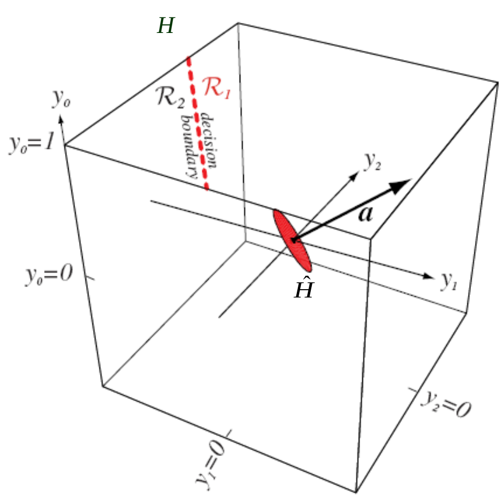
$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i, \quad x_0 = 1$$

$$\mathbf{y} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix}$$

15

15

增广空间的判别面



$\mathbf{y}$ 到超平面 $\hat{H}$的距离:

$$|\mathbf{a}^t\mathbf{y}| / \|\mathbf{a}\| = |g(\mathbf{x})| / \|\mathbf{a}\|$$

$\mathbf{x}$ 到 H的距离:

$$|g(\mathbf{x})| / \|\mathbf{w}\| \geq |g(\mathbf{x})| / \|\mathbf{a}\|$$

16

16

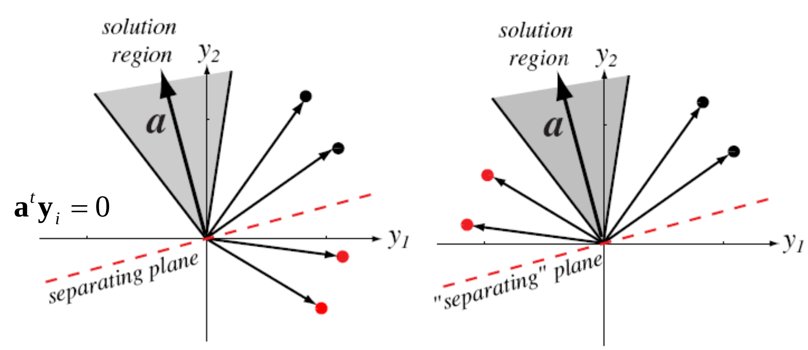
5.4 两类线性可分的情况

- N个样本: $\mathbf{y}_1, \ldots, \mathbf{y}_n$
- 类标: ω_1, ω_2
- 目标：确定判别函数 $g(\mathbf{x})=\mathbf{a}^t\mathbf{y}$ 的权向量a
- 线性可分
  - 存在能将所有样本正确分类的权向量
- 规范化
- 分离向量 (解向量)

17

17

权空间以及解区域



$$\omega_1 : \mathbf{a}^t\mathbf{y}_i > 0$$

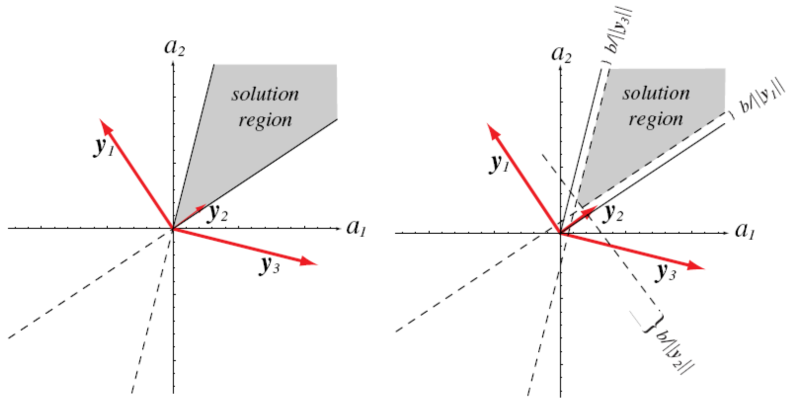
$$\omega_2 : \mathbf{a}^t\mathbf{y}_i < 0$$

for all $\mathbf{a}^t\mathbf{y}_i > 0$

18

18

边沿裕量的作用



$$\mathbf{a}_i^t \mathbf{y} \geq b > 0$$

19

19

5.4.2 基本梯度下降算法

定义准则函数 $J(\mathbf{a})$

如果$\mathbf{a}$是解向量，则$J(\mathbf{a})$达到最小化

a(k+1)=a(k)-η(k)∇J(a(k))

初始化 $\mathbf{a}$ 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 $|\eta(k)\nabla J(\mathbf{a})| < \theta$

return $\mathbf{a}$

end

20

20

学习率的选择：最小二阶逼近

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

$$\mathbf{H}: \text{Hessian 矩阵}, H_{ij} = \left.\frac{\partial^2 J}{\partial a_i \partial a_j}\right|_{\mathbf{a}=\mathbf{a}(k)}$$

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

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

$$\text{选择 } \eta(k) = \frac{\|\nabla J\|^2}{\nabla J^t \mathbf{H} \nabla J} \text{ 使得} J(\mathbf{a}(k+1)) \text{ 最小化}$$

21

21

牛顿下降法

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

$$\mathbf{H}: \text{Hessian 矩阵}, H_{ij} = \left.\frac{\partial^2 J}{\partial a_i \partial a_j}\right|_{\mathbf{a}=\mathbf{a}(k)}$$

$$\text{选择 } \mathbf{a} = \mathbf{a}(k+1) = \mathbf{a}(k) - \mathbf{H}^{-1}\nabla J \text{ 使得 } J(\mathbf{a}) \text{ 最小化}$$

初始化 $\mathbf{a}$ 和阈值 θ

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

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

  return $\mathbf{a}$

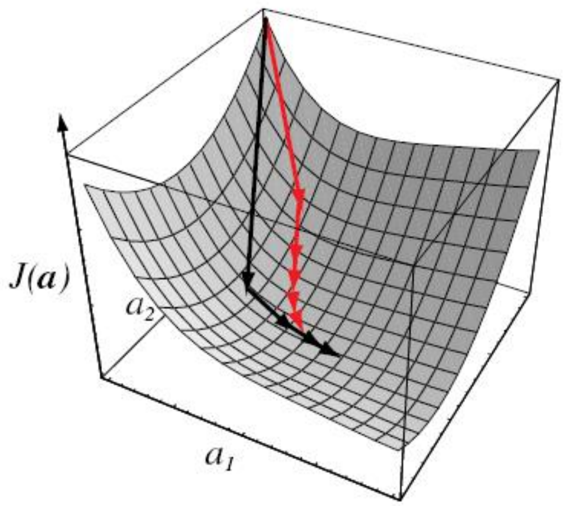
end

22

22

梯度下降 vs. 牛顿算法

中山大學



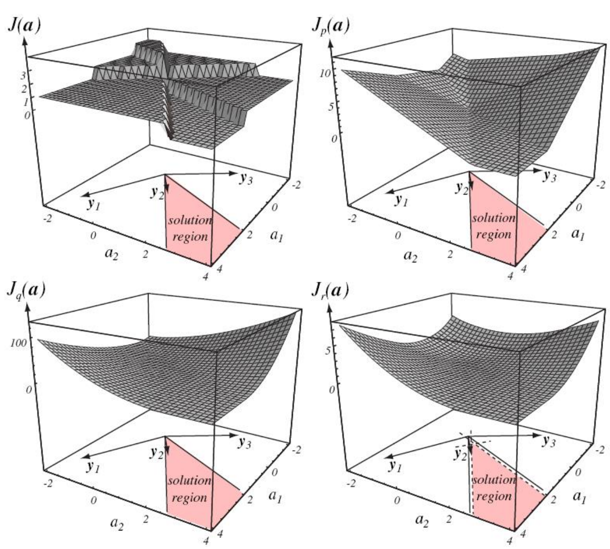
牛顿算法计算H逆矩阵的时间开销大

23

23

四种学习准则

中山大學



- 错分模式的总数
- 感知器准则
- 平方误差
- 具有裕量的平方误差

24

24

5.5 批处理感知器算法

中山大學

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

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

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

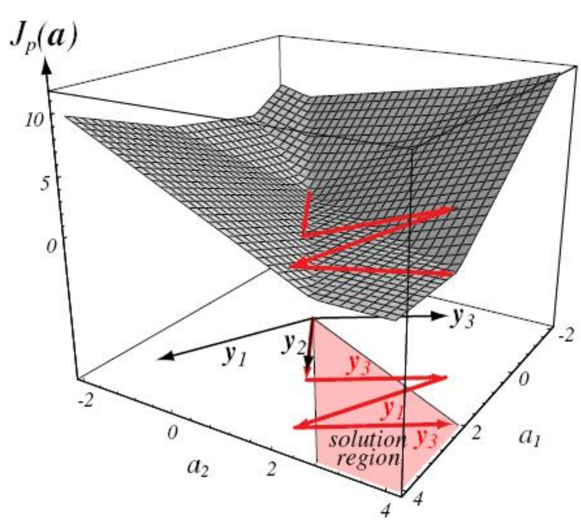
return $\mathbf{a}$

end

25

批处理感知器

中山大學



26

中山大學

固定增量单样本感知器

begin

  初始化 $\mathbf{a}, k \leftarrow 0$

    do $k \leftarrow (k+1) \bmod n$

      if $\mathbf{y}_k$ 被 $\mathbf{a}$ 错分， then $\mathbf{a} \leftarrow \mathbf{a} + \mathbf{y}_k$

    until 所有模式被正确分类

    return $\mathbf{a}$

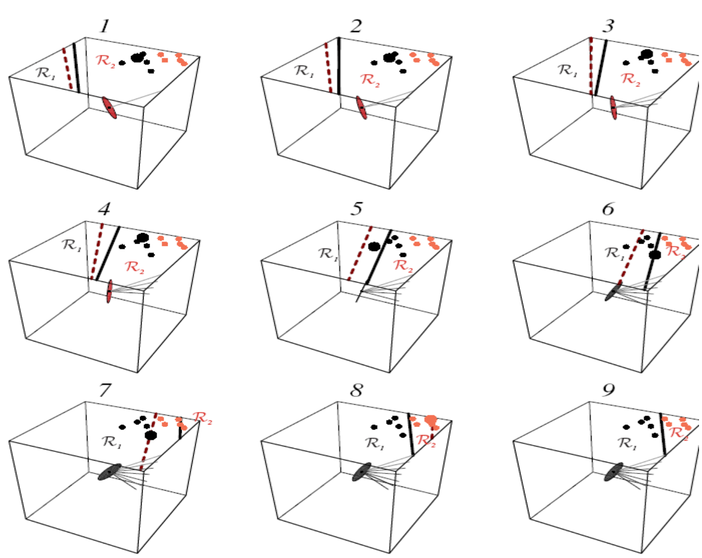
end

27

27

中山大學

固定增量单样本感知器



28

28

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

中 山 大 學

设$\hat{\mathbf{a}}$为任意的解向量，则对任何的i有$\hat{\mathbf{a}}^t\mathbf{y}_i > 0$

设α为一正的比例因子，则有

$$\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}} = \left(\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\right) + \mathbf{y}^k$$

$$\left\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\right\|^2 = \left\|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\right\|^2 + 2\left(\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\right)^t\mathbf{y}^k + \left\|\mathbf{y}^k\right\|^2$$

由于$\mathbf{y}^k$为被错分，固$\mathbf{a}(k)^t\mathbf{y}^k < 0$，所以

$$\left\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\right\|^2 \le \left\|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\right\|^2 - 2\alpha\hat{\mathbf{a}}^t\mathbf{y}^k + \left\|\mathbf{y}^k\right\|^2$$

设β为模式向量的最大长度，即

$\beta^2 = \max_i \left\|\mathbf{y}_i\right\|^2$，并令$\gamma$为解向量与所有模式向量最小的内积，即

$$\gamma = \min_i \left[\hat{\mathbf{a}}^t\mathbf{y}_i\right] > 0$$

29

29

定理 5.1 *(Cont.)*

中 山 大 學

得到不等式

$$\left\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\right\|^2 \le \left\|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\right\|^2 - 2\alpha\gamma + \beta^2$$

选择 $\alpha = \dfrac{\beta^2}{\gamma}$，就有

$\left\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\right\|^2 \le \left\|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\right\|^2 - \beta^2$，过了k步矫正后

$$\left\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\right\|^2 \le \left\|\mathbf{a}(1) - \alpha\hat{\mathbf{a}}\right\|^2 - k\beta^2$$

平方距离非负，经过不 超k_0次矫正后矫正将终止， 其中

$$k_0 = \frac{\left\|\mathbf{a}(1) - \alpha\hat{\mathbf{a}}\right\|^2}{\beta^2}$$

30

30

难点:

- 取决于与解向量接近正交的样本

$$\mathbf{a}(1)=0$$

$$k_0=\frac{\alpha^2\|\hat{\mathbf{a}}\|^2}{\beta^2}=\frac{\beta^2\|\hat{\mathbf{a}}\|^2}{\gamma^2}=\frac{\max_i\|\mathbf{y}_i\|^2\|\hat{\mathbf{a}}\|^2}{\min_i\left[\mathbf{y}_i^t\hat{\mathbf{a}}\right]^2}$$

- 样本几乎共面

带裕量的变增量感知器

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

可以证明当样本线性可分，如果

$$\eta(k)\geq 0,\quad \lim_{m\to\infty}\sum_{k=1}^{m}\eta(k)=\infty,\quad \lim_{m\to\infty}\frac{\sum_{k=1}^{m}\eta^2(k)}{\left(\sum_{k=1}^{m}\eta(k)\right)^2}=0$$

$e.g.,\quad \eta(k)\sim 1/k$

则$a(k)$收敛于一个解向量。

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

批处理变增量感知器

initialize $\mathbf{a}$, $\eta(\bullet)$, $k \leftarrow 0$

do $k \leftarrow (k+1) \bmod n$

$Y_k = \{\ \}$

$j \leftarrow 0$

do $j \leftarrow j+1$

if $\mathbf{y}_j$ 被错分类 then 把 $\mathbf{y}_j$ 加进 Y_k

until $j = n$

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

until $Y_k = \{\ \}$

return $\mathbf{a}$

end

33

33

理论与实践

- **理论**
  - 对任何有限的可分样本集，对任意的初始权向量，对任意非负的裕度，对任意符合条件的比例因子，都能得到解。
- **实践**
  - 边沿裕度 b 最好选择接近 $\eta(\mathrm{k})\|\mathbf{y}^k\|^2$
  - $\mathbf{y}^k$ 分量的比例因子对算法会产生很大的影响

34

34

中山大學

平衡 Winnow 算法

$$
\begin{aligned}
&\text{initialize } \mathbf{a}^+, \mathbf{a}^-, \eta(\bullet), k \leftarrow 0, \alpha > 1 \\
&\quad z_k = \mathrm{Sgn}[\mathbf{a}^{+t}\mathbf{y}_k - \mathbf{a}^{-t}\mathbf{y}_k] \text{---（判别模式是否被错分）} \\
&\quad \text{if } z_k = 1 \text{ then } a_i^+ \leftarrow \alpha^{+y_i} a_i^+,\ a_i^- \leftarrow \alpha^{-y_i} a_i^- \ \text{ for all } i \\
&\quad \text{if } z_k = -1 \text{ then } a_i^+ \leftarrow \alpha^{-y_i} a_i^+,\ a_i^- \leftarrow \alpha^{+y_i} a_i^- \ \text{ for all } i \\
&\quad \text{return } \mathbf{a}^+, \mathbf{a}^- \\
&\text{end}
\end{aligned}
$$

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

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

35

35

中山大學

平衡 Winnow 算法的优点

- **在训练过程中，两个候选权向量分别朝各自的恒定方向运动**
  - **两个向量的“间隔” 始终不会变大**
  - **收敛性比感知器收敛性定理还要更加一般化**
- **通常比感知器算法收敛得更快**
  - **在有大量不相关或冗余特征的情况下尤其明显**

36

36

5.6 松弛算法

中山大學

$$J_q(\mathbf{a}) = \sum_{\mathbf{y}\in Y} \left(\mathbf{a}^t\mathbf{y}\right)^2 \qquad J_p(\mathbf{a}) = \sum_{\mathbf{y}\in Y} \left(-\mathbf{a}^t\mathbf{y}\right)$$

$$J_r(\mathbf{a}) = \frac{1}{2}\sum_{\mathbf{y}\in Y} \frac{\left(\mathbf{a}^t\mathbf{y}-b\right)^2}{\|\mathbf{y}\|^2} \qquad \nabla J_r = \sum_{\mathbf{y}\in Y} \frac{\mathbf{a}^t\mathbf{y}-b}{\|\mathbf{y}\|^2}\mathbf{y}$$

- 主要的区别在于：
  - Jq的梯度是连续的，而Jp的梯度是不连续的；
  - Jp的另一个问题，可能依赖于模值最大的样本向量。

37

37

批处理裕量松弛算法

中山大學

a(1)

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \eta(k)\sum_{\mathbf{y}\in Y} \frac{b-\mathbf{a}^t\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}$$

initialize $\mathbf{a}, \eta(\bullet), b, k \leftarrow 0$

do $k \leftarrow (k+1) \bmod n$

$Y_k = \{\ \}, \quad j \leftarrow 0$

do $j \leftarrow j+1$

if $\mathbf{a}^t\mathbf{y}^j \le b$ then 把 $\mathbf{y}^j$ 加进 Y_k

until $j = n$

$\mathbf{a} \leftarrow \mathbf{a} + \eta(k)\sum_{\mathbf{y}\in Y_k} \frac{b-\mathbf{a}^t\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}$

until $Y_k = \{\ \}$

return $\mathbf{a}$

end

38

38

单样本裕量松弛算法

initialize $\mathbf{a}, \eta(\bullet), k \leftarrow 0$

do $k \leftarrow (k+1) \bmod n$

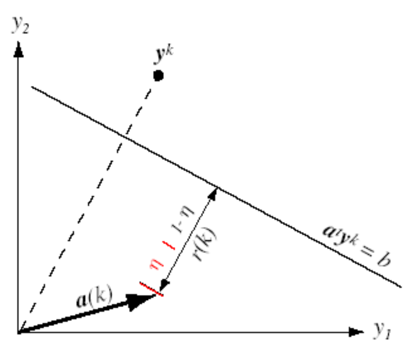
if $\mathbf{a}^t \mathbf{y}^k \le b$ then $\mathbf{a} \leftarrow \mathbf{a} + \eta(k) \dfrac{b - \mathbf{a}^t \mathbf{y}^k}{\|\mathbf{y}\|^2} \mathbf{y}^k$

until $\mathbf{a}^t \mathbf{y}^k > b$ for all $\mathbf{y}^k$

return $\mathbf{a}$

end

几何解释



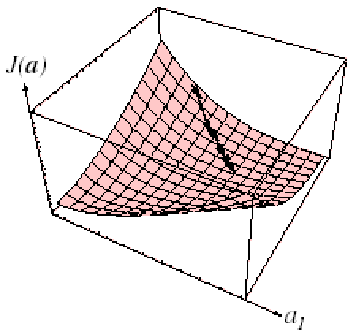
$$r(k) = \frac{b - \mathbf{a}^t(k)\mathbf{y}^k}{\|\mathbf{y}\|^2}$$

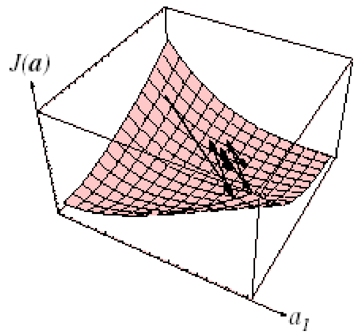
$$\mathbf{a}^t(k+1)\mathbf{y}^k - b = (1-\eta)\left[\mathbf{a}^t(k)\mathbf{y}^k - b\right]$$

其中，$\eta < 1$称为欠松弛，$\eta > 1$称为过松弛.

中山大學

欠松弛 和 过松弛





41

41

中山大學

收敛性证明

$$\mathbf{a}(k+1)=\mathbf{a}(k)+\eta\frac{[b-\mathbf{a}^t(k)\mathbf{y}^k]}{\|\mathbf{y}^k\|^2}\mathbf{y}^k$$

因为

$$\|\mathbf{a}(k+1)-\hat{\mathbf{a}}\|^2$$

$$=\|\mathbf{a}(k)-\hat{\mathbf{a}}\|^2-2\eta\frac{[b-\mathbf{a}^t(k)\mathbf{y}^k]}{\|\mathbf{y}^k\|^2}[\hat{\mathbf{a}}-\mathbf{a}(k)]^t\mathbf{y}^k$$

$$+\eta^2\frac{[b-\mathbf{a}^t(k)\mathbf{y}^k]^2}{\|\mathbf{y}^k\|^2}$$

又$[\hat{\mathbf{a}}-\mathbf{a}(k)]^t\mathbf{y}^k>b-\mathbf{a}^t(k)\mathbf{y}^k\geq 0\quad(\because \hat{\mathbf{a}}^t\mathbf{y}^k>b)$

所以$\|\mathbf{a}(k+1)-\hat{\mathbf{a}}\|^2\leq\|\mathbf{a}(k)-\hat{\mathbf{a}}\|^2-\eta(2-\eta)\frac{[b-\mathbf{a}^t(k)\mathbf{y}^k]^2}{\|\mathbf{y}^k\|^2}$

42

42

收敛性证明

限制$0<\eta<2$

当$k\to\infty$, $\|\mathbf{a}(k)-\hat{\mathbf{a}}\|$ 到达一个有限的距离 $r(\hat{\mathbf{a}})$

对解区域内的所有 $\hat{\mathbf{a}}$

假设 $\mathbf{a}'$ 和 $\mathbf{a}''$ 为公共交集上的点

则对解区域上的所有 $\hat{\mathbf{a}}$都有$\|\mathbf{a}'-\hat{\mathbf{a}}\|=\|\mathbf{a}''-\hat{\mathbf{a}}\|$

i.e., $\hat{\mathbf{a}}$ 位于一个 d-1维的超球面上（解区域为 d维）

如果对所有i都有 $\hat{\mathbf{a}}^t\mathbf{y}_i>0$，那么对都有 d维向量 $\mathbf{v}$,

当ε 足够小时

对所有$i=1,\cdots,n,$都有$(\hat{\mathbf{a}}+\varepsilon\mathbf{v})^t\mathbf{y}_i>0$

5.7 不可分的情况

- 数目少于$2\hat{d}$的样本集很可能是线性可分的（第九章会再次讨论）
- 对足够多数据的情况，往往线性不可分
- 每个矫正过程产生一个无穷权向量序列
- 很多启发式规则被用于修改误差矫正算法，修改的目的是在不可分的问题中得到令人接受的结果，同时保持它对可分问题的处理能力

5.8 最小平方误差方法

中山大學

$$\begin{pmatrix} y_{10} & y_{11} & \cdots & y_{1d} \\ y_{20} & y_{21} & \cdots & y_{2d} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ y_{n0} & y_{n1} & \cdots & y_{nd} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_n \end{pmatrix}$$

$$J_s(\mathbf{a}) = \sum_{i=1}^{n} \left(\mathbf{a}^t \mathbf{y}_i - b_i\right)^2 = \left\|\mathbf{Y}\mathbf{a} - \mathbf{b}\right\|^2$$

45

45

最小平方误差方法

中山大學

$$\nabla J_s = \sum_{i=1}^{n} 2\left(\mathbf{a}^t \mathbf{y}_i - b_i\right)\mathbf{y}_i = 2\mathbf{Y}^t\left(\mathbf{Y}\mathbf{a} - \mathbf{b}\right) = 0$$

$$\mathbf{Y}^t\mathbf{Y}\mathbf{a} = \mathbf{Y}^t\mathbf{b}$$

$$\mathbf{a} = \left(\mathbf{Y}^t\mathbf{Y}\right)^{-1}\mathbf{Y}^t\mathbf{b} = \mathbf{Y}^+\mathbf{b}$$

伪逆 $\mathbf{Y}^+ = \left(\mathbf{Y}^t\mathbf{Y}\right)^{-1}\mathbf{Y}^t$

46

46

例子

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

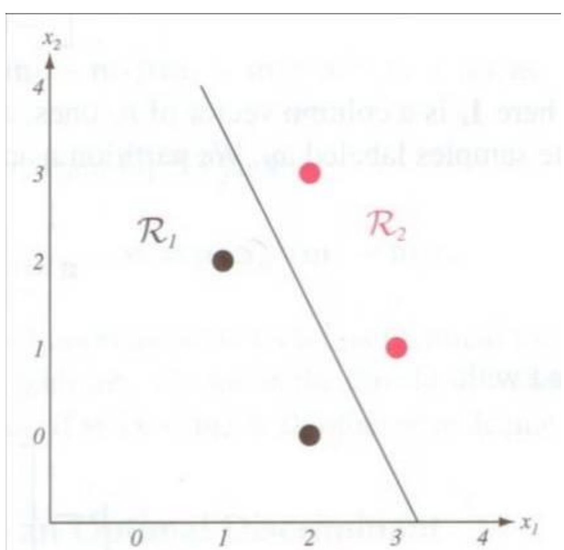
$$\text{判别边界}:\mathbf{a}^t\begin{pmatrix}1\\x_1\\x_2\end{pmatrix}=0$$

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

47

47

例子



48

48

5.8.2 MSE与 Fisher 线性判别的关系

$$D_1=\left\{\mathbf{x}_1,\cdots,\mathbf{x}_{n_1}\right\},\ D_2=\left\{\mathbf{x}_{n_1+1},\cdots,\mathbf{x}_{n_1+n_2}\right\}$$

增量模式: $\mathbf{y}_i=\begin{bmatrix}1\\ \mathbf{x}_i\end{bmatrix}$

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

49

49

与 Fisher 线性判别的关系

$$\mathbf{Y}^t\mathbf{Y}\mathbf{a}=\mathbf{Y}^t\mathbf{b}$$

$$\begin{pmatrix}\mathbf{1}_1^t & -\mathbf{1}_2^t\\ \mathbf{X}_1^t & -\mathbf{X}_2^t\end{pmatrix}\begin{pmatrix}\mathbf{1}_1 & \mathbf{X}_1\\ -\mathbf{1}_2 & -\mathbf{X}_2\end{pmatrix}\begin{pmatrix}w_0\\ \mathbf{w}\end{pmatrix}=\begin{pmatrix}\mathbf{1}_1^t & -\mathbf{1}_2^t\\ \mathbf{X}_1^t & -\mathbf{X}_2^t\end{pmatrix}\begin{pmatrix}\frac{n}{n_1}\mathbf{1}_1\\ \frac{n}{n_2}\mathbf{1}_2\end{pmatrix}$$

$$\mathbf{m}_i=\frac{1}{n_i}\sum_{\mathbf{x}\in D_i}\mathbf{x},\quad \mathbf{S}_W=\sum_{i=1}^{2}\sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$$

$$\begin{pmatrix}n & (n_1\mathbf{m}_1+n_2\mathbf{m}_2)^t\\ (n_1\mathbf{m}_1+n_2\mathbf{m}_2) & S_W+n_1\mathbf{m}_1\mathbf{m}_1^t+n_2\mathbf{m}_2\mathbf{m}_2^t\end{pmatrix}\begin{pmatrix}w_0\\ \mathbf{w}\end{pmatrix}=\begin{pmatrix}0\\ n(\mathbf{m}_1-\mathbf{m}_2)\end{pmatrix}$$

50

50

中山大學

与 Fisher 线性判别的关系

$$w_0 = -\mathbf{m}^t\mathbf{w}$$ m是所有样本均值

$$\left[\frac{1}{n}\mathbf{S}_W + \frac{n_1 n_2}{n^2}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\right]\mathbf{w} = \mathbf{m}_1 - \mathbf{m}_2$$

$$\frac{n_1 n_2}{n^2}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} = (1-\alpha)(\mathbf{m}_1 - \mathbf{m}_2)$$

$$\mathbf{w} = \alpha n \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

51

51

中山大學

5.8.3 最优判别的渐近

$b=1_n$时，MSE的解等同于以最小均方误差近似Bayes 判别函数：

$$g_0(\mathbf{x}) = P(\omega_1 | \mathbf{x}) - P(\omega_2 | \mathbf{x})$$

证明：按照概率定律

$$p(\mathbf{x}) = p(\mathbf{x}|\omega_1)P(\omega_1) + p(\mathbf{x}|\omega_2)P(\omega_2)$$

独立同分布抽取样本，得到

$$g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}, \quad \mathbf{y} = \mathbf{y}(\mathbf{x})$$

定义均方近似误差为

$$\varepsilon^2 = \int \left[\mathbf{a}^t\mathbf{y} - g_0(\mathbf{x})\right]^2 p(\mathbf{x})d\mathbf{x}$$

52

52

最优判别的渐近

当$\mathbf{b}=\mathbf{1}_n$时，最小均方误差准则函数为：

$$J_s(\mathbf{a})=\sum_{\mathbf{y}\in Y_1}\left(\mathbf{a}^t\mathbf{y}-1\right)^2+\sum_{\mathbf{y}\in Y_2}\left(\mathbf{a}^t\mathbf{y}+1\right)^2$$

$$=n\left[\frac{n_1}{n}\frac{1}{n_1}\sum_{\mathbf{y}\in Y_1}\left(\mathbf{a}^t\mathbf{y}-1\right)^2+\frac{n_2}{n}\frac{1}{n_2}\sum_{\mathbf{y}\in Y_2}\left(\mathbf{a}^t\mathbf{y}+1\right)^2\right]$$

利用大数定理，n趋向无穷大时

$$J(\mathbf{a})=\frac{1}{n}J_s(\mathbf{a})=P(\omega_1)E_1\left[\left(\mathbf{a}^t\mathbf{y}-1\right)^2\right]+P(\omega_2)E_2\left[\left(\mathbf{a}^t\mathbf{y}+1\right)^2\right]$$

其中，$E_1\left[\left(\mathbf{a}^t\mathbf{y}-1\right)^2\right]=\int\left(\mathbf{a}^t\mathbf{y}-1\right)^2 p(\mathbf{x}\mid\omega_1)d\mathbf{x}$

$$E_2\left[\left(\mathbf{a}^t\mathbf{y}+1\right)^2\right]=\int\left(\mathbf{a}^t\mathbf{y}+1\right)^2 p(\mathbf{x}\mid\omega_2)d\mathbf{x}$$

53

最优判别的渐近

$$g_0(\mathbf{x})=\frac{p(\mathbf{x},\omega_1)-p(\mathbf{x},\omega_2)}{p(\mathbf{x})}$$

$$\bar{J}(\mathbf{a})=\int\left(\mathbf{a}^t\mathbf{y}-1\right)^2 p(\mathbf{x},\omega_1)d\mathbf{x}+\int\left(\mathbf{a}^t\mathbf{y}+1\right)^2 p(\mathbf{x},\omega_2)d\mathbf{x}$$

$$=\int\left(\mathbf{a}^t\mathbf{y}\right)^2 p(\mathbf{x})d\mathbf{x}-2\int\mathbf{a}^t\mathbf{y}g_0(\mathbf{x})p(\mathbf{x})d\mathbf{x}+1$$

$$=\int\left[\mathbf{a}^t\mathbf{y}-g_0(\mathbf{x})\right]^2 p(\mathbf{x})d\mathbf{x}+\left[1-\int g_0^2(\mathbf{x})p(\mathbf{x})d\mathbf{x}\right]$$

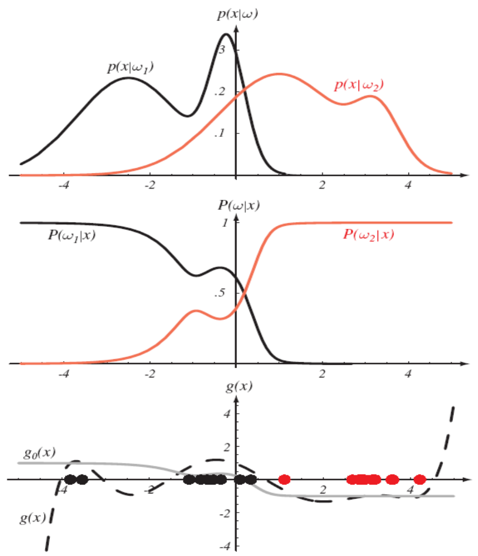
$$=\varepsilon^2+\text{独立于}\mathbf{a}\text{的成分}$$

于是，MSE的解等同于以最小均方误差近似Bayes 判别函数：

$$g_0(\mathbf{x})=P(\omega_1\mid\mathbf{x})-P(\omega_2\mid\mathbf{x})$$

54

MSE 解和 Bayes 判别函数



55

55

Widrow-Hoff 或最小均方算法

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

$$\mathbf{a}(k+1) = \mathbf{a}(k) - \eta(k)\mathbf{Y}^t(\mathbf{Ya}-\mathbf{b})$$

LMS 算法

initialize $\mathbf{a},\ \mathbf{b}$, 阈值 $\theta, \eta(\bullet), k \leftarrow 0$

do $k \leftarrow (k+1) \bmod n$

$\mathbf{a} \leftarrow \mathbf{a} + \eta(k)\left(b_k - \mathbf{a}^t\mathbf{y}^k\right)\mathbf{y}^k$

until $\left|\eta(k)\left(b_k - \mathbf{a}^t\mathbf{y}^k\right)\mathbf{y}^k\right| < \theta$

return $\mathbf{a}$

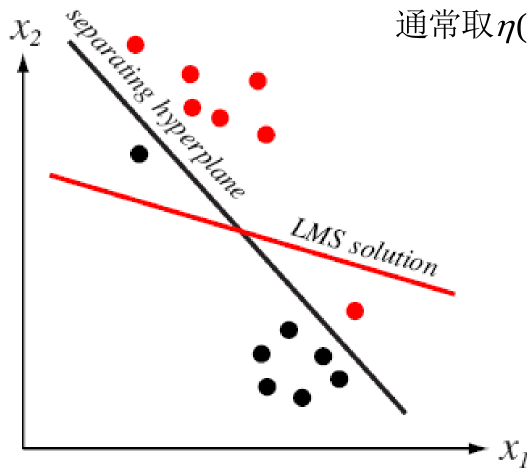
end

56

56

LMS算法未必收敛于分类超平面

通常取$\eta(k)=\eta(1)/k$



5.8.5 随机近似法

中山大學

类标可变 -

Bayes判别函数的带噪声版本

根据 $p(\mathbf{x}\,|\,\omega_i), i=1,2$选择x

标记 $\theta=+1$ ，如果 $\mathbf{x}$ 属于 ω_1

标记 $\theta=-1$，如果 $\mathbf{x}$ 属于 ω_2

$$P(\theta=1\,|\,\mathbf{x})=P(\omega_1\,|\,\mathbf{x}),\quad P(\theta=-1\,|\,\mathbf{x})=P(\omega_2\,|\,\mathbf{x})$$

$$E_{\theta|\mathbf{x}}[\theta]=\sum_{\theta}\theta P(\theta\,|\,\mathbf{x})=P(\omega_1\,|\,\mathbf{x})-P(\omega_2\,|\,\mathbf{x})=g_0(\mathbf{x})$$

用有限级数展开来近似 $g_0(\mathbf{x})$

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

选择一个权向量 $\hat{\mathbf{a}}$ 来最小化均方近似误差

$$\varepsilon^2 = E\left[\left(\mathbf{a}^t\mathbf{y} - g_0(\mathbf{x})\right)^2\right]$$

$$J_m(\mathbf{a}) = E\left[\left(\mathbf{a}^t\mathbf{y} - \theta\right)^2\right]$$

$$\nabla J_m = 2E\left[\left(\mathbf{a}^t\mathbf{y} - \theta\right)\mathbf{y}\right]$$

•因为θ本质上来说是$g_0(x)$的一个含噪版本

MSE 解: $\hat{\mathbf{a}} = E\left[\mathbf{y}\mathbf{y}^t\right]^{-1} E\left[\theta\mathbf{y}\right]$

59

59

Widrow-Hoff 算法和收敛条件

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \eta(k)\left[\theta_k - \mathbf{a}^t\mathbf{y}_k\right]\mathbf{y}_k$$

$-\nabla J_m$

$E\left[\mathbf{y}\mathbf{y}^t\right]$ 非奇异和 对 $\eta(k)$ 满足

$$\lim_{m\to\infty} \sum_{k=1}^{\infty} \eta(k) = \infty$$

$$\lim_{m\to\infty} \sum_{k=1}^{\infty} \eta^2(k) < \infty$$

① 阻止权向量收敛得太快以至于系统误差永远存在
② 保证随机波动最终会被抑制

$$\Rightarrow \lim_{k\to\infty} E\left[\left\|\mathbf{a}(k) - \hat{\mathbf{a}}\right\|^2\right] = 0$$

60

60

牛顿算法:

取J_m的二阶偏导矩阵

$$D = 2E\left[\mathbf{y}\mathbf{y}^t\right]$$

得到J_m极小化的牛顿算法:

$$\mathbf{a}(k+1) = \mathbf{a}(k) + E\left[\mathbf{y}\mathbf{y}^t\right]^{-1} E\left[\left(\theta - \mathbf{a}^t\mathbf{y}\right)\mathbf{y}\right]$$

用样本估计代替期望，得到迭代算法:

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \mathbf{R}_{k+1}\left(\theta - \mathbf{a}^t(k)\mathbf{y}_k\right)\mathbf{y}_k$$

其中 $\mathbf{R}_{k+1}^{-1} = \mathbf{R}_k^{-1} + \mathbf{y}_k\mathbf{y}_k^t$

或得到等价的结果$\mathbf{R}_{k+1} = \mathbf{R}_k - \dfrac{\mathbf{R}_k\mathbf{y}_k}{1+\mathbf{y}_k^t\mathbf{R}_k\mathbf{y}_k}$

61

Stochastic Approximation Procedure

在统计学文献中，像J_m 和∇J_m一样

具有 E[f(α, x)] 形式的函数

统称为“回归函数”(Regression Function),

这类的迭代算法就叫“随机近似算法”

(Stochastic Approximation Prodedure)

62

5.9 Ho-Kashyap 算法

中山大學

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

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

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

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

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

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

Ho-Kashap rule:

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

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

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

63

63

收敛性证明

中山大學

$$\mathbf{e}(k) = (\mathbf{YY}^+ - \mathbf{I})\mathbf{b}(k) \quad \text{将(82)式代入(80)式}$$

$$\mathbf{e}(k+1) = (\mathbf{YY}^+ - \mathbf{I})\left(\mathbf{b}(k) + 2\eta\mathbf{e}^+(k)\right)$$

$$= \mathbf{e}(k) + 2\eta(\mathbf{YY}^+ - \mathbf{I})\mathbf{e}^+(k)$$

$$\frac{1}{4}\|\mathbf{e}(k+1)\|^2 = \frac{1}{4}\|\mathbf{e}(k)\|^2 + \eta\mathbf{e}^t(k)(\mathbf{YY}^+ - \mathbf{I})\mathbf{e}^+(k)$$

$$+ \left\|\eta(\mathbf{YY}^+ - \mathbf{I})\mathbf{e}^+(k)\right\|^2$$

$$\mathbf{Y}^t\mathbf{Ya}(k) = \mathbf{Y}^t\mathbf{b}(k) \Rightarrow \mathbf{Y}^t\mathbf{e}(k) = 0$$

$$\eta\mathbf{e}^t(k)(\mathbf{YY}^+ - \mathbf{I})\mathbf{e}^+(k) = -\eta\mathbf{e}^t(k)\mathbf{e}^+(k) = -\eta\left\|\mathbf{e}^+(k)\right\|^2$$

其中， $\mathbf{Y}^+ = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t$

64

64

收敛性证明

$$\mathbf{YY}^{+}=\mathbf{Y}\left(\mathbf{Y}^{t}\mathbf{Y}\right)^{-1}\mathbf{Y}^{t}$$

$$\left(\mathbf{YY}^{+}\right)^{t}\left(\mathbf{YY}^{+}\right)=[\mathbf{Y}\left(\mathbf{Y}^{t}\mathbf{Y}\right)^{-1}\mathbf{Y}^{t}][\mathbf{Y}\left(\mathbf{Y}^{t}\mathbf{Y}\right)^{-1}\mathbf{Y}^{t}]=\mathbf{YY}^{+}$$

$$\left\|\eta\left(\mathbf{YY}^{+}-\mathbf{I}\right)\mathbf{e}^{+}(k)\right\|^{2}=\eta^{2}\mathbf{e}^{+t}(k)\left(\mathbf{YY}^{+}-\mathbf{I}\right)^{t}\left(\mathbf{YY}^{+}-\mathbf{I}\right)\mathbf{e}^{+}(k)$$

$$=\eta^{2}\left\|\mathbf{e}^{+t}(k)\right\|^{2}-\eta^{2}\mathbf{e}^{+t}(k)\mathbf{YY}^{+}\mathbf{e}^{+}(k)$$

$$\frac{1}{4}\left(\|\mathbf{e}(k)\|^{2}-\|\mathbf{e}(k+1)\|^{2}\right)$$

$$=\eta(1-\eta)\left\|\mathbf{e}^{+t}(k)\right\|^{2}+\eta^{2}\mathbf{e}^{+t}(k)\mathbf{YY}^{+}\mathbf{e}^{+}(k)>0$$

$$0<\eta<1$$

收敛性证明

$\|\mathbf{e}(k)\|^{2}$ 单调递减且收敛到一个有限的值 $\|\mathbf{e}\|^{2}$

$\Rightarrow \mathbf{e}^{+}(k)$ 收敛到0

$[\because \mathbf{b}(k+1)=\mathbf{b}(k)+2\eta(k)\mathbf{e}^{+}(k), \mathbf{e}(k)=(\mathbf{YY}^{+}-\mathbf{I})\mathbf{b}(k)]$

$\Rightarrow$ $\mathbf{e}(k)$ 的正分量收敛到0

对线性可分的样本,

$\mathbf{Y}\hat{\mathbf{a}}=\hat{\mathbf{b}},\ \hat{\mathbf{b}}>0$

$\mathbf{e}^{t}(k)\mathbf{Y}\hat{\mathbf{a}}=\left(\mathbf{Y}^{t}\mathbf{e}(k)\right)^{t}\hat{\mathbf{a}}=0\Rightarrow$ 对所有的k，有$\mathbf{e}^{t}(k)\hat{\mathbf{b}}=0$

$\Rightarrow$ $\mathbf{e}(k)$的负分量收敛到0

$\because (\mathbf{e}^{+}(k)-\mathbf{e}^{-}(k))^{t}\hat{\mathbf{b}}=0\Rightarrow \mathbf{e}^{-t}(k)\hat{\mathbf{b}}=\mathbf{e}^{+t}(k)\hat{\mathbf{b}}=0$

不可分情况

$$\mathbf{e}(k+1)=\mathbf{e}(k)+2\eta(\mathbf{YY}^{+}-\mathbf{I})\mathbf{e}^{+}(k)$$

$$\frac{1}{4}\left(\|\mathbf{e}(k)\|^{2}-\|\mathbf{e}(k+1)\|^{2}\right)$$

$$=\eta(1-\eta)\|\mathbf{e}^{+}(k)\|^{2}+\eta^{2}\mathbf{e}^{+t}(k)\mathbf{YY}^{+}\mathbf{e}^{+}(k)$$

$\|\mathbf{e}(k)\|^{2}$ 仍然收敛到 $\|\mathbf{e}\|^{2}$

$\mathbf{e}^{+}(k)$ 收敛到0

但是如果 $\mathbf{e}^{t}(k)\hat{\mathbf{b}}=0, \hat{\mathbf{b}}>0$ 不成立，

就不能证明 $\mathbf{e}(k)$ 的负成分收敛到0

亦即，当一个非零误差 向量没有正分量，问题 为不可分情况

67

67

一个改进算法

$\mathbf{b}(1)>0,\quad \mathbf{a}(1)$ 任意

$$\mathbf{b}(k+1)=\mathbf{b}(k)+\left(\mathbf{e}(k)+|\mathbf{e}(k)|\right)$$

$$\mathbf{a}(k+1)=\mathbf{a}(k)+\eta\mathbf{Y}^{+}|\mathbf{e}(k)| \quad (\because \mathbf{a}(k)=\mathbf{Y}^{+}\mathbf{b}(k), \text{将上式代入})$$

$$\mathbf{e}(k)=\mathbf{Ya}(k)-\mathbf{b}(k)$$

68

68

中山大學

改进算法变形:

$$\mathbf{b}(1) > 0, \quad \mathbf{a}(1)\ \text{任意}$$

$$\mathbf{b}(k+1) = \mathbf{b}(k) + \left(\mathbf{e}(k) + \left|\mathbf{e}(k)\right|\right)$$

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \eta \mathbf{R}\mathbf{Y}^t \left|\mathbf{e}(k)\right|$$

$\mathbf{R}$: 任意, 常量, 正定, 对称, $\hat{d} \times \hat{d}$ 矩阵

$$\because \mathbf{Y}^+ = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t$$

69

中山大學

收敛性证明

$$\mathbf{e}(k+1) = \mathbf{Y}\mathbf{a}(k+1) - \mathbf{b}(k+1)$$

$$= \left(\eta \mathbf{Y}\mathbf{R}\mathbf{Y}^t - \mathbf{I}\right)\left|\mathbf{e}(k)\right| \quad (\because \text{由(89)式,(90)式})$$

$$\left\|\mathbf{e}(k+1)\right\|^2 = \left|\mathbf{e}(k)\right|^t \left(\eta^2 \mathbf{Y}\mathbf{R}\mathbf{Y}^t\mathbf{Y}\mathbf{R}\mathbf{Y}^t - 2\eta \mathbf{Y}\mathbf{R}\mathbf{Y}^t + \mathbf{I}\right)\left|\mathbf{e}(k)\right|$$

$$\left\|\mathbf{e}(k)\right\|^2 - \left\|\mathbf{e}(k+1)\right\|^2 = \left(\mathbf{Y}^t\left|\mathbf{e}(k)\right|\right)^t \mathbf{A} \left(\mathbf{Y}^t\left|\mathbf{e}(k)\right|\right)$$

$$\mathbf{A} = 2\eta\mathbf{R} - \eta^2 \mathbf{R}\mathbf{Y}^t\mathbf{Y}\mathbf{R}$$

其中 η 足够小, $\mathbf{A}$ 正定

$$\left\|\mathbf{e}(k)\right\|^2 > \left\|\mathbf{e}(k+1)\right\|^2$$

70

线性情况

中山大學

$\|\mathbf{e}(k)\|$ 收敛到 $\|\mathbf{e}\|$

$\mathbf{Y}^t\mathbf{e}(k)$ 必收敛到0

假设存在 $\hat{\mathbf{a}}$ 和 $\hat{\mathbf{b}}$, 使得

$$\mathbf{Y}\hat{\mathbf{a}} = \hat{\mathbf{b}}$$

$$|\mathbf{e}(k)|^t \mathbf{Y}\hat{\mathbf{a}} = |\mathbf{e}(k)|^t \hat{\mathbf{b}} \geq 0$$

$\therefore |\mathbf{e}(k)|^t \mathbf{Y}$ 收敛到0

$\Rightarrow |\mathbf{e}(k)|$ 收敛到0

71

71

不可分情况

中山大學

$\mathbf{Y}^t|\mathbf{e}(k)|$ 收敛到0

但 $\mathbf{e}(k)$ 既不为0， 也不收敛到0

72

72

对 R 和 h 的最简单选择

$$\mathbf{R}=\mathbf{I}$$

$$\mathbf{A}=2\eta\mathbf{I}-\eta^2\mathbf{Y}^t\mathbf{Y}$$

如果$0<\eta<2/\lambda_{\max}$，则$\mathbf{A}$正定

$\lambda_{\max}:\mathbf{Y}^t\mathbf{Y}$的最大特征值

$\lambda_{\max}$的最差边界：$\lambda_{\max}\le\sum_i\|\mathbf{y}_i\|^2/\hat{d}$

h 的最优选择

$$\|\mathbf{e}(k)\|^2-\|\mathbf{e}(k+1)\|^2=|\mathbf{e}(k)|^t\,\mathbf{Y}(2\eta\mathbf{R}-\eta^2\mathbf{R}\mathbf{Y}^t\mathbf{Y}\mathbf{R})\mathbf{Y}^t\,|\mathbf{e}(k)|$$

对上式求关于η的微分,得到最优 η:

$$\eta(k)=\frac{|\mathbf{e}(k)|^t\,\mathbf{YRY}^t\,|\mathbf{e}(k)|}{|\mathbf{e}(k)|^t\,\mathbf{YRY}^t\mathbf{YRY}^t\,|\mathbf{e}(k)|}$$

当 $\mathbf{R}=\mathbf{I}$,

$$\eta(k)=\frac{\|\mathbf{Y}^t\,|\mathbf{e}(k)|\|^2}{\|\mathbf{YY}^t\,|\mathbf{e}(k)|\|^2}$$

R 的最优选择

用对称的$\mathbf{R}+\delta\mathbf{R}$,来代替$\mathbf{R}$，并忽略第二项，得到

$$\delta\left(\|\mathbf{e}(k)\|^2-\|\mathbf{e}(k+1)\|^2\right)$$

$$=|\mathbf{e}(k)|\mathbf{Y}\left[\delta\mathbf{R}^t(\mathbf{I}-\eta\mathbf{Y}^t\mathbf{Y}\mathbf{R})+(\mathbf{I}-\eta\mathbf{R}\mathbf{Y}^t\mathbf{Y})\delta\mathbf{R}\right]\mathbf{Y}^t|\mathbf{e}(k)|$$

可通过选择

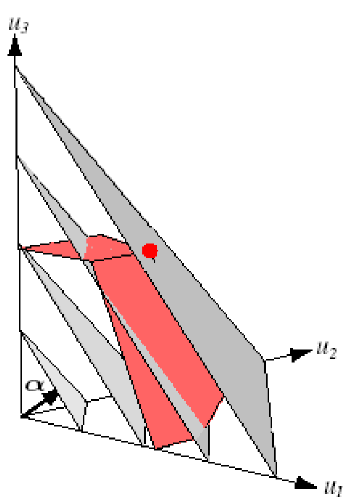
$$\mathbf{R}=\frac{1}{\eta}\left(\mathbf{Y}^t\mathbf{Y}\right)^{-1}$$

使得平方误差向量下降达到最大

（因为$\eta\mathbf{R}\mathbf{Y}^t=\mathbf{Y}^+$：实际上就是Ho-Kashap algorithm）

75

5.10 线性规划



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

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

常用求解方法：

单纯形法

其中额外要求$\mathbf{u}\geq 0$

76

分离向量分解

$$\mathbf{a} = \mathbf{a}^{+} - \mathbf{a}^{-}$$

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

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

$$\mathbf{a}^{+} \geq 0, \quad \mathbf{a}^{-} \geq 0$$

77

线性可分情形:

假设有n个样本$y_1, y_2, ..., y_n$,

我们希望a对所有的i都满足:

$$a^t y_i \geq b_i > 0$$

在线性规划中表达的表达方法为:

引入$\tau \geq 0$，满足:

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

导出问题为:

求解a和极小化的τ，满足

$\tau \geq 0$和$a^t y_i \geq b_i$

如果所求得$\tau = 0$，样本就是线性可分的且可得到一解;

如果所求得τ为正数，可以证明样本不可分;

78

公式表达

中山大學

$$\min_{\mathbf{u}} \ z = \boldsymbol{\alpha}^t \mathbf{u} \quad \text{subject to } \ \mathbf{Au} \ge \boldsymbol{\beta}, \mathbf{u} \ge 0$$

$$\mathbf{u} = \begin{bmatrix} \mathbf{a}^+ \\ \mathbf{a}^- \\ \tau \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{y}_1^t & -\mathbf{y}_1^t & 1 \\ \mathbf{y}_2^t & -\mathbf{y}_2^t & 1 \\ \vdots & \vdots & \vdots \\ \mathbf{y}_n^t & -\mathbf{y}_n^t & 1 \end{bmatrix}$$

用有限步可以求得z=τ的极小值。

如果所求得$\tau = 0$，样本就是线性可分的且可得到一解；

如果所求得τ为正数，解没有用，但可以证明样本非线性可分；

79

极小化感知器准则函数

中山大學

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

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:

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

其中

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

80

5.11 支持向量机

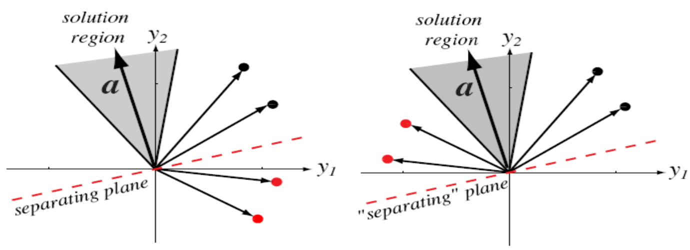
Support Vector Machines （SVM）

$$\mathbf{y}_k = \varphi(\mathbf{x}_k), \quad z_k = \pm 1$$

$g(\mathbf{y}) = \mathbf{a}^t\mathbf{y}$, $\mathbf{y}$ 属于增量空间

分隔超平面保证

$$z_k g(\mathbf{y}_k) \geq 0, \quad k = 1, \cdots, n$$



81

81

SVM

$\mathbf{y}$到超平面的距离: $\dfrac{|g(\mathbf{y})|}{\|\mathbf{a}\|}$

正值边沿裕度 b,

$$\frac{z_k g(\mathbf{y}_k)}{\|\mathbf{a}\|} \geq b, k = 1, \cdots, n$$

等价形式:

$$\frac{z_k (g(\mathbf{y}_k) - b)}{\|\mathbf{a}\|} \geq 0$$

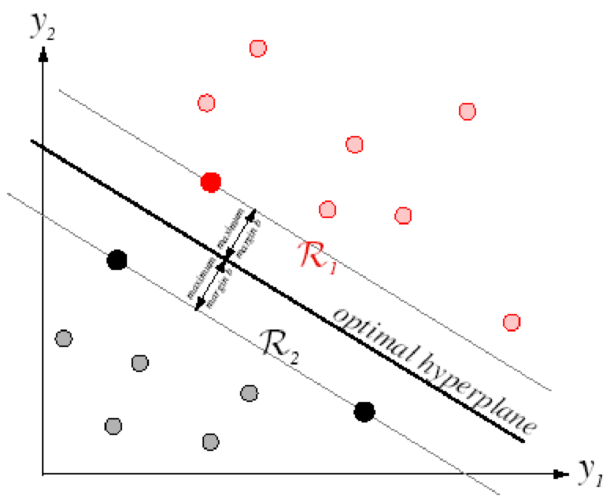
目标: 找到一个 $\mathbf{a}$ 使得 ：

$$\min_{\mathbf{a}} \|\mathbf{a}\|^2 \text{ subject to } z_k (g(\mathbf{y}_k) - b) \geq 0, k = 1, \cdots, n$$

82

82

SVM



83

83

SVM 的训练方法

- 对感知器训练方法进行改进。
- 通过选择当前最坏分类的模式来更新权向量达到训练目的。
- 在训练的结束阶段，该模式将成为一个支持向量。
- 寻找最坏分类的模式，计算量很大。

84

84

最优化学习

$$\min_{\mathbf{a}} L(\mathbf{a}, b, \boldsymbol{\alpha})$$

$$L(\mathbf{a}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{a}\|^2 - \sum_{k=1}^{n} \alpha_k \left[z_k (\mathbf{a}^t \mathbf{y}_k - b) \right],$$

$$\boldsymbol{\alpha} \geq 0$$

85

85

Kuhn-Tucker 定理

$$\min_{\mathbf{a}} L(\mathbf{a}, b, \boldsymbol{\alpha})$$

$$L(\mathbf{a}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{a}\|^2 - \sum_{k=1}^{n} \alpha_k \left[z_k (\mathbf{a}^t \mathbf{y}_k - b) \right] \quad (*)$$

$$\frac{\partial L(\mathbf{a}, b, \boldsymbol{\alpha})}{\partial \mathbf{a}} = 0, \quad \frac{\partial L(\mathbf{a}, b, \boldsymbol{\alpha})}{\partial b} = 0$$

$$\mathbf{a} = \sum_{k=1}^{n} \alpha_k z_k \mathbf{y}_k, \quad \sum_{k=1}^{n} \alpha_k z_k = 0$$

代入公式（*），且b=1，得

$$L(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{k,j}^{n} \alpha_k \alpha_j z_k z_j \mathbf{y}_j^t \mathbf{y}_k$$

Maximizing L，满足：$\sum_{k=1}^{n} \alpha_k z_k = 0$

86

86

中山大學

由Kuhn-Tucker 定理

等价于二次规划问题

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) \text{ subject to } \sum_{k=1}^{n} z_k \alpha_k = 0, \boldsymbol{\alpha} \geq 0$$

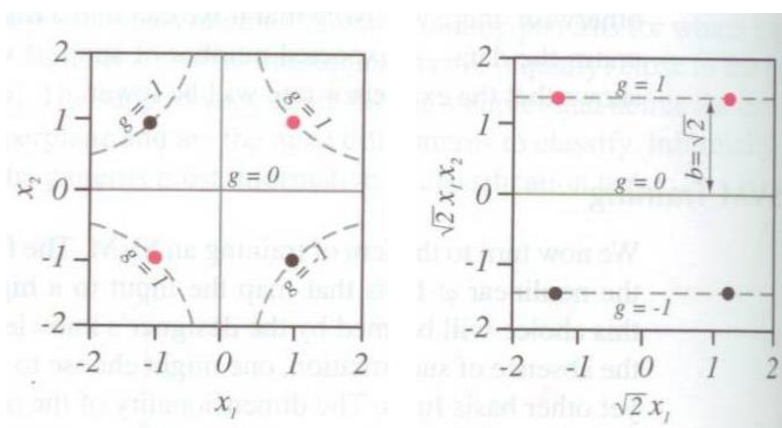
$$L(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{k,j}^{n} \alpha_k \alpha_j z_k z_j \mathbf{y}_j^t \mathbf{y}_k$$

87

87

中山大學

例子: XOR 问题的 SVM



88

88

例子: XOR 问题的 SVM

$$\omega_1 : \mathbf{x}_1 = (1 \quad 1)^t, \mathbf{x}_3 = (-1 \quad -1)^t$$

$$\omega_2 : \mathbf{x}_2 = (1 \quad -1)^t, \mathbf{x}_4 = (-1 \quad 1)^t$$

$$\varphi \text{ functions } :1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2$$

$$\max_{\boldsymbol{\alpha}} \sum_{k=1}^{4} \alpha_k - \frac{1}{2}\sum_{k,j} \alpha_k \alpha_j z_k z_j \mathbf{y}_j^t \mathbf{y}_k$$

$$\text{subject to } \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0, \text{ and } \alpha_k \geq 0$$

例子: XOR 问题的 SVM

$$解: \alpha_k^* = 1/8, k = 1, \cdots, 4$$

这四个训练样本都是支 持向量

$$\mathbf{a}^* = \sum_{k=1}^{4} \alpha_k^* z_k \mathbf{y}_k$$

$$最终判别函数: g(\mathbf{x}) = x_1 x_2$$

$$判定超平面: g = 0$$

$$裕度: b^* = 1/\|\mathbf{a}^*\| = \sqrt{2}$$

核方法和非线性SVM

中山大學

上述例子具有典型性:

1. 需要非线性映射将Φ, 将x映射到

高维空间的特征$y=\Phi(x)$, 使得样本线性可分;

2. 支持向量机求解方法仅涉及内积

$y_i \cdot y_j = \Phi(x_i) \cdot \Phi(x_j)$

如果能找到函数: $K(x_i, x_j)=\Phi(x_i) \cdot \Phi(x_j)$,

就可以简化计算,避免维数问题.

$K(x_i, x_j)$通常要求满足Mercer条件:

对$\int g^2(u)du < \infty, g \neq 0$,有

$\iint k(u,v)g(u)g(v)dudv>0$

91

核方法和非线性SVM

中山大學

这时,由泛函分析相关理论,有

$$k(u,v)=\sum_{k=1}^{\infty} a_k \varphi_k(u)\varphi_k(v)$$

其中, $a_k > 0$

通常称$k(u,v)$为Mercer核。

于是,由Mercer核可以构造非线性SVM

92

SVM 的优点

- 所获得的分类器的复杂度可以采用支持向量的个数，而不是变换空间的维数来刻画
- 不像一些别的方法一样容易发生过拟合（overfitting）现象

93

5.12 多类问题的推广

广义线性判别函数

$$g_i(\mathbf{x}) = \mathbf{a}_i^t \mathbf{y}(\mathbf{x}),\ i = 1, \cdots, c$$

如果存在一组 $\hat{\mathbf{a}}_1, \cdots, \hat{\mathbf{a}}_c$ 使得

如果 $\mathbf{y}_k \in Y_i$, 则对$j \neq i$，有$\hat{\mathbf{a}}_i^t \mathbf{y}_k > \hat{\mathbf{a}}_j^t \mathbf{y}_k$

那么称样本线性可分

94

5.12.1 Kesler 构造法

让 $\mathbf{y}_k \in Y_1$，则

$$\hat{\mathbf{a}}_1^t \mathbf{y}_k - \hat{\mathbf{a}}_j^t \mathbf{y}_k > 0, \quad j = 2, \cdots, c$$

$\Rightarrow$

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_c \end{bmatrix}, \boldsymbol{\eta}_{12} = \begin{bmatrix} \mathbf{y} \\ -\mathbf{y} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \cdots, \boldsymbol{\eta}_{1c} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ -\mathbf{y} \end{bmatrix}$$

要求 $\hat{\boldsymbol{\alpha}}^t \boldsymbol{\eta}_{1j} > 0, \ j = 2, \cdots, c$

Kesler 构造法

- 将数据的维数乘以 c,且将样本数目乘以 c-1；
- 为了证明收敛性，可通过将多类误差矫正法转化为两类问题来实现；

5.12.2 固定增量规则的收敛性

令 L_k表示权向量为 $\mathbf{a}_1(k), \cdots, \mathbf{a}_c(k)$的线性机

令 $\mathbf{y}^k$ 为需要矫正的第 k个样本，$\mathbf{y}^k \in Y_i$

则至少存在一个 $j \neq i$，使得 $\mathbf{a}_i^t(k)\mathbf{y}^k \leq \mathbf{a}_j^t(k)\mathbf{y}^k$

固定增量矫正：

$$\mathbf{a}_i(k+1) = \mathbf{a}_i(k) + \mathbf{y}^k$$

$$\mathbf{a}_j(k+1) = \mathbf{a}_j(k) - \mathbf{y}^k$$

$$\mathbf{a}_l(k+1) = \mathbf{a}_l(k), \quad l \neq i \text{ 且 } l \neq j$$

97

97

证明： 有限步后必收敛

$$\boldsymbol{\alpha}(k) = \begin{bmatrix} \mathbf{a}_1(k) \\ \vdots \\ \mathbf{a}_c(k) \end{bmatrix}, \quad \boldsymbol{\eta}_{ij}^k = \begin{array}{c} \\ \left[\begin{array}{c} \vdots \\ \mathbf{y}^k \\ \vdots \\ -\mathbf{y}^k \\ \vdots \end{array}\right] \end{array}\begin{array}{l} \\ i \\ \\ j \\ \\ \end{array}, \quad \boldsymbol{\alpha}^t(k)\boldsymbol{\eta}_{ij}^k \leq 0$$

$$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \boldsymbol{\eta}_{ij}^k$$

此时，多类情况与两类的情况对应。对两类问题, $\boldsymbol{\alpha}(k)$ 在有限步矫正后必终止在一个解向量上

98

98

MSE 算法推广

对于多类问题:

$a_i^t y = 1,\quad$ 对所有$y \in Y_i$

$a_i^t y = 0,\quad$ 对所有$y \notin Y_i$

的最小均方误差解确保$a_i^t y$以

最小均方误差逼近$\mathrm{P}(\omega_{\mathrm{i}} | \mathrm{x})$

99

MSE 算法推广

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_c \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_c \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_c \end{bmatrix},$$

其中: B_i的第i列为1, 其余的列为0.

平方误差矩阵 $(\mathbf{YA}-\mathbf{B})^t(\mathbf{YA}-\mathbf{B}), \mathbf{A} = \mathbf{Y}^+\mathbf{B}$

100

本章小结

- 本章给出了一些判别函数，它们都是某个参数集的线性函数，而这些参数一般被称为权系数。
- 通过构造准则函数，采用梯度下降法迭代求解。也可以通过线性代数运算直接求权参数。
- SVM，被映射到一个高维空间，具有最大间隔的最优超平面
- 如何找到一个适当的非线性映射，下一章涉及

101