第4章 线性判别函数

- 线性判别函数基本概念
- Fisher线性判别
- 感知准则函数
- 最小平方误差准则函数
- 多类问题简介

4.1 引言

- 问题的提出

$\because P(\omega_i)$和$p(x|\omega_i)$估计困难

$\therefore$ 考虑直接方法，利用样本直接设计分类器

线性决策面是较简单的，易于实现，因此本章主要讨论线性分类器。

4.1 引言

- 判别函数形式

$$g(x) = \omega^T x + \omega_0$$

$$g_i(x) = \omega_i^T x + \omega_{i0} \qquad i = 1, 2, \cdots, c$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \qquad \omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_d \end{bmatrix}$$

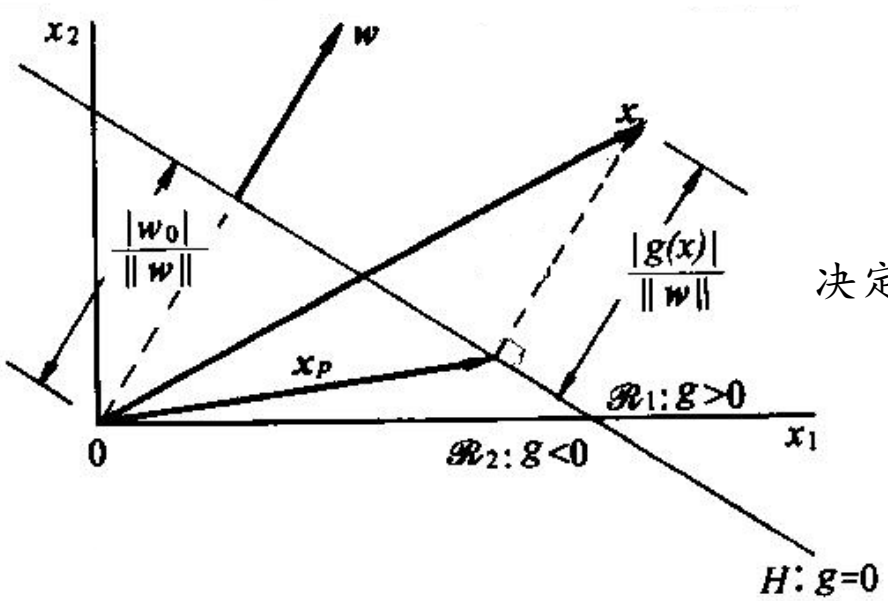
$$g_j(x) = \max g_i(x), i = 1, \cdots, c \quad x \in \omega_j$$

4.1 引言

- 两类情况

$$g(x)=g_1(x)-g_2(x)$$

$$\begin{cases} g(x)>0 & x\in\omega_1 \\ g(x)<0 & x\in\omega_2 \\ g(x)=0 & \text{任意分至某一类或拒绝} \end{cases}$$



决定分类界面的方向，ω_0决定分类界面的位置

4.1 引言

- 线性分类器的设计步骤

（1）有一已知类别的样本集H

（2）确定一准则函数 J

$$\begin{cases} J\text{是}H,\omega,\omega_0\text{的函数} \\ \text{极值解对应最好的“ 决策”} \quad -\omega^*,\omega_0^* \end{cases}$$

（3）利用最优化技术求出准则函数极值解 ω^*,ω_0^*

4.1 引言

- 与最优分类器的关系

贝叶斯分类器是在错误率或风险下为最优的分类器。线性分类器针对错误率或风险是“次优”的。但对于所采用的准则函数$J(\omega)$则是最优的。

- 线性可分性

已知一样本集，如果有一个线性分类器能把每个样本正确分类，则称这组样本集为线性可分的，否则称为线性不可分的。反之，如果样本集是线性可分的，则必然存在一个线性分类器能把每个样本正确分类。

4.2 Fisher线性判别

由1936年的经典论文开始最早研究线性判别函数，发明了最大似然估计方法。其著作有

《研究者用的统计方法》

《统计方法和科学推理》

《近交的理论》

《试验设计》

《自然选择的遗传理论》

《根据孟德尔遗传方式的亲属间的相关》

R. A. Fisher爵士,1890.2.17-1962.7.29,英国

现代统计学与现代演化论的奠基者之一

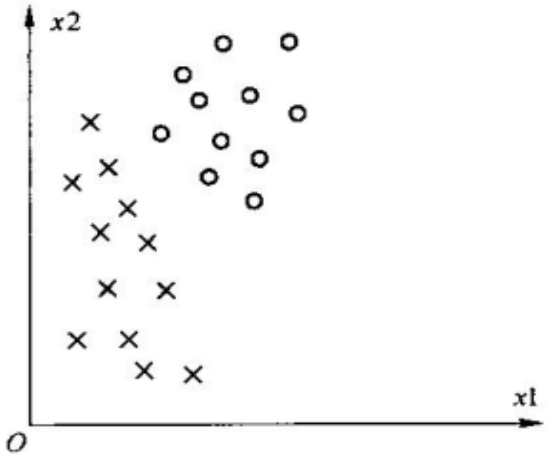
4.2 Fisher线性判别（两类）

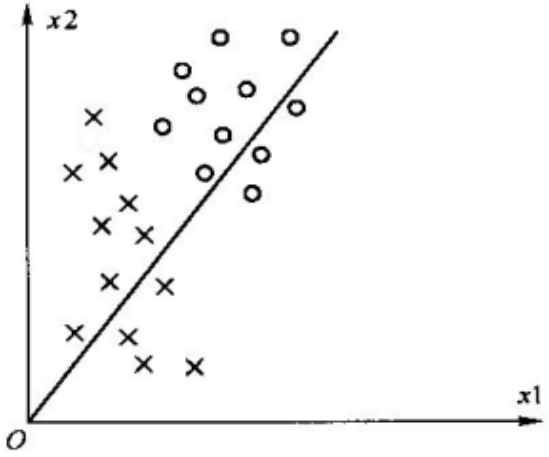
- 问题的提出

在低维空间中行得通的方法在高维空间里往往失效，因此需要降低空间维数。

- 基本原理

将高维空间数据向某一直线上投影，如何确定该直线在特征空间中的最佳方向？





Fisher法即寻找最佳的投影方向ω，使投影后样本最佳可分

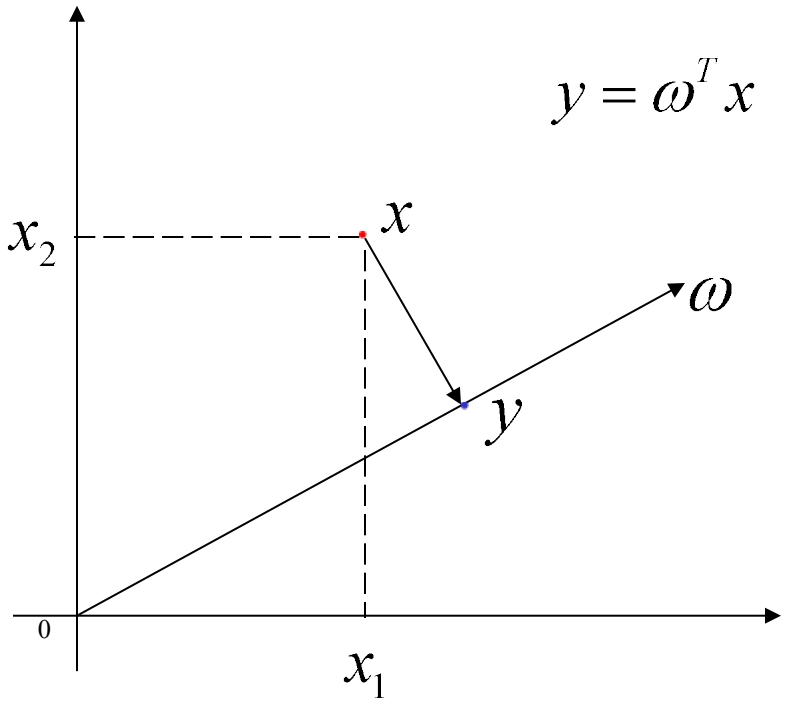
4.2 Fisher线性判别（两类）

- 投影的表示

$$y = \omega^T x$$

$$y = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_d x_d$$

$$\omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_d \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$



投影示意图

4.2 Fisher线性判别（两类）

- 基本参量定义（d维x空间）

（1）各类样本均值向量

$$m_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \qquad i = 1,2$$

（2）样本类内离散度矩阵S_i 和总类内离散度矩阵 S_ω

$$S_i = \sum_{x\in\omega_i}(x-m_i)(x-m_i)^T \qquad i = 1,2$$

$$S_\omega = S_1 + S_2$$

对称半正定阵，当N>d时通常非奇异

（3）样本类间离散度矩阵

$$S_b = (m_1 - m_2)(m_1 - m_2)^T$$

对称半正定阵

4.2 Fisher线性判别（两类）

- 补充定义—正定二次型

  设有实二次型 $f = X^T AX$，若对任意向量X都有f＞0，则称矩阵A为正定矩阵；若f＜0，则称A为负定矩阵；若f≧0，则称A为半正定矩阵。

- 补充定义—非奇异阵

  当|A|=0时，方阵A为奇异阵；|A| ≠0时，方阵A为非奇异阵。方阵A可逆的充要条件是|A| ≠0。

  可逆阵亦称非退化阵或非奇异阵或满秩阵

  不可逆的方阵也称退化阵或奇异阵或降秩阵

4.2 Fisher线性判别（两类）

- 基本参量定义（一维y空间）

（1）各类样本均值向量

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y \qquad i=1,2$$

（2）样本类内离散度矩阵$\tilde{S}_i^2$和总类内离散度矩阵 $\tilde{S}_\omega$

$$\tilde{S}_i^2 = \sum_{y\in\omega_i}(y-\tilde{m}_i)^2 \qquad i=1,2$$

$$\tilde{S}_\omega = \tilde{S}_1^2 + \tilde{S}_2^2$$

4.2 Fisher线性判别（两类）

- 定义Fisher准则函数

$$J_F(\omega)=\frac{(\tilde{m}_1-\tilde{m}_2)^2}{\tilde{S}_1^2+\tilde{S}_2^2} \qquad \frac{\text{越大越好}}{\text{越小越好}}$$

原则：各类样本尽可能分开，类内样本尽量密集，需求令准则函数极大的ω^*。

4.2 Fisher线性判别（两类）

- 准则函数的显式化

$$(\tilde{m}_1-\tilde{m}_2)^2=(\omega^T m_1-\omega^T m_2)^2=\omega^T(m_1-m_2)(m_1-m_2)^T\omega=\omega^T S_b\omega$$

$$\because \tilde{S}_i^2=\sum_{x\in\omega_i}(y-\tilde{m}_i)^2=\sum_{x\in\omega_i}(\omega^T x-\omega^T m_i)^2=\omega^T\left[\sum_{x\in\omega_i}(x-m_i)(x-m_i)^T\right]\omega=\omega^T S_i\omega$$

$$\therefore \tilde{S}_1^2+\tilde{S}_2^2=\omega^T(S_1+S_2)\omega=\omega^T S_\omega\omega$$

$$\therefore J_F(\omega)=\frac{\omega^T S_b\omega}{\omega^T S_\omega\omega}$$

求令准则函数极大的 ω^*

4.2 Fisher线性判别（两类）

利用*Lagrange*乘子法

假设$\omega^T S_\omega \omega = C \neq 0$　　求令准则函数极大的ω^*

$$L(\omega,\lambda) = \omega^T S_b \omega - \lambda(\omega^T S_\omega \omega - C)$$

$$\frac{\partial L(\omega,\lambda)}{\partial \omega} = S_b\omega - \lambda S_\omega \omega = 0$$　　二次型对向量求导

$$S_b\omega^* = \lambda S_\omega \omega^*$$

$\because S_\omega$非奇异

$$\therefore S_\omega^{-1} S_b \omega^* = \lambda \omega^*$$

$$\because S_b = (m_1 - m_2)(m_1 - m_2)^T$$

$$\therefore \lambda\omega^* = S_\omega^{-1}(m_1 - m_2)\underbrace{(m_1 - m_2)^T \omega^*}$$　　内积为常量，以R替代

$$\therefore \omega^* = \frac{R}{\lambda} S_\omega^{-1}(m_1 - m_2)$$　　要求的是方向，常量系数可去除

4.2 Fisher线性判别（两类）

- 概念补充—二次型对向量求导

$$x^T Ax = [x_1 \quad \cdots \quad x_d] \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{dd} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} \sum_{\alpha=1}^{d} a_{\alpha 1} x_1 & \cdots & \sum_{\alpha=1}^{d} a_{\alpha d} x_d \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$= \sum_{\beta=1}^{d} \sum_{\alpha=1}^{d} a_{\alpha\beta} x_\alpha x_\beta$$

$$\frac{\partial}{\partial x_i}\left[x^T Ax\right] = \frac{\partial}{\partial x_i}\left|\sum_{\beta=1}^{d} \sum_{\alpha=1}^{d} a_{\alpha\beta} x_\alpha x_\beta\right| = 2a_{ii}x_i + \sum_{j\neq i}(a_{ij} + a_{ji})x_j$$

$$\therefore \frac{\partial}{\partial x}\left[x^T Ax\right] = (A + A^T)x \qquad \text{当} A = A^T, \frac{\partial}{\partial x}\left[x^T Ax\right] = 2Ax$$

4.2 Fisher线性判别（两类）

- 投影过程 $y = \omega^{*T} x$
- 一维阈值的选取

（1）当维数与样本数都很大时，可用贝叶斯分类器

（2）依据先验知识选择

$$y_0^{(1)} = \frac{\tilde{m}_1 + \tilde{m}_2}{2}$$

$$y_0^{(2)} = \frac{N_1 \tilde{m}_1 + N_2 \tilde{m}_2}{N_1 + N_2}$$

$$y_0^{(3)} = \frac{\tilde{m}_1 + \tilde{m}_2}{2} + \frac{\ln[P(\omega_1)/P(\omega_2)]}{N_1 + N_2 - 2}$$

$$y \begin{matrix} > \\ < \end{matrix} y_0 \rightarrow x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

4.2 Fisher线性判别（两类）

- Fisher线性判别步骤

（1）求两类样本均值向量 m_1和m_2

（2）求两类样本类内离散度矩阵 S_i

（3）求总类内离散度矩阵S_ω

（4）求向量 $\omega^* = S_\omega^{-1}(m_1 - m_2)$

（5）求出两类样本在ω^*上的投影点 $y = \omega^{*T} x$

（6）求各类在投影空间的均值 $\tilde{m}_i = \frac{1}{N_i} \sum_{y \in \omega_i} y$

（7）选取阈值 y_0

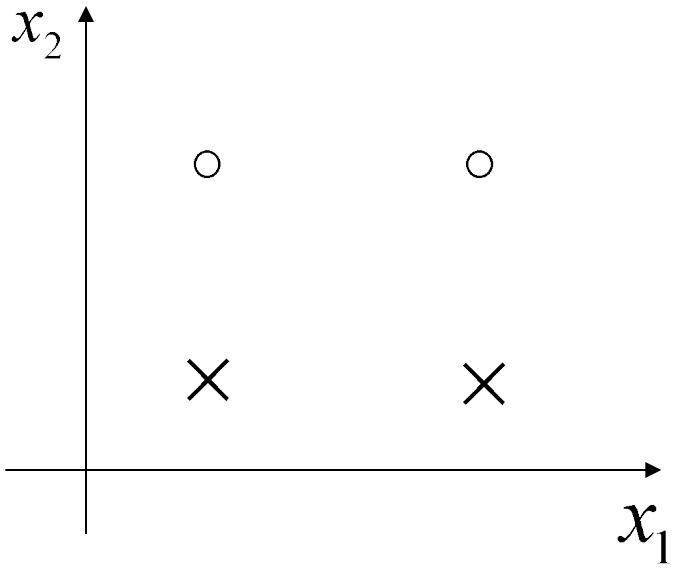
（8）对于未知样本，计算其在ω^*上的投影点y

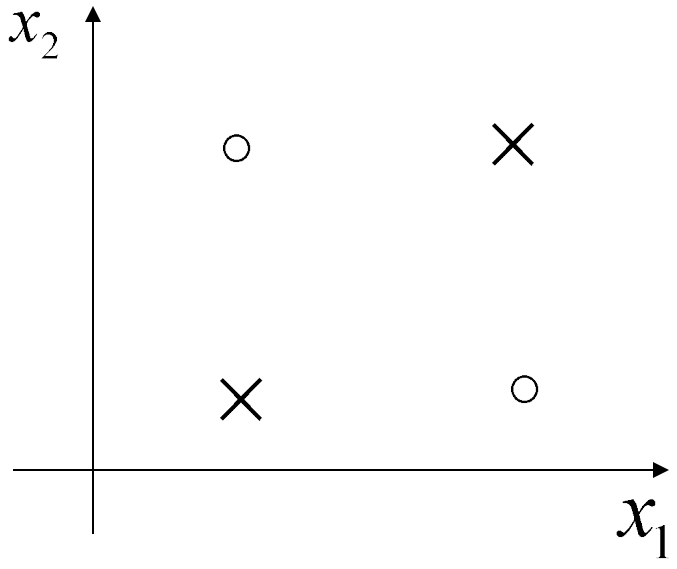
（9）根据决策规则分类

4.3感知准则函数（两类）

- 问题的提出

20世纪50年代由Rosenblatt提出，用于脑模型感知器，故称为感知准则函数。该模型未获成功，主要由于无法解决非线性问题，但其思想可沿用。





4.3感知准则函数（两类）

- 基本思想

分类器形式已定，只要估计出权值向量即完成分类器设计。感知器采用迭代的方法，是一种典型的赏罚过程，对正确分类的模式则“赏”，这里即“不罚”，权向量不变。对错误分类的模式则“罚”，即修正权向量。

4.3感知准则函数（两类）

- 样本及权向量的增广

$$g(x)=\sum_{i=1}^{d}\omega_i x_i+\omega_0=\sum_{i=1}^{d}a_i y_i=a^T y$$

$$y=\begin{bmatrix}x_1\\ \vdots\\ x_d\\ 1\end{bmatrix}=\begin{bmatrix}x\\ 1\end{bmatrix}\qquad a=\begin{bmatrix}\omega_1\\ \vdots\\ \omega_d\\ \omega_0\end{bmatrix}=\begin{bmatrix}\omega\\ \omega_0\end{bmatrix}$$

4.3感知准则函数（两类）

- 样本的规范化

如果样本集是线性可分的，必然存在某个权向量可使

$$\begin{cases} a^T y^i > 0 & y^i \in \omega_1 \\ a^T y^j < 0 & y^j \in \omega_2 \end{cases}$$

$$y = \begin{cases} y^i & y^i \in \omega_1 \\ -y^j & y^j \in \omega_2 \end{cases}$$

$$\therefore \begin{cases} a^T y > 0 & y^i \in \omega_1 \\ a^T y > 0 & y^j \in \omega_2 \end{cases}$$

若$a^T y < 0$,则样本被错分

4.3感知准则函数（两类）

- 感知准则函数的构造
- 感知器（perceptron)

$$J_p(a) = \sum_{y \in Err} -a^T y$$

$-a^T y \geq 0$　　对被错分样本的处理

$J_p^*(a) = \min J_p(a) = 0$　准则函数为0时无错分

优化目标：令准则函数极小，采用梯度下降法

4.3感知准则函数（两类）

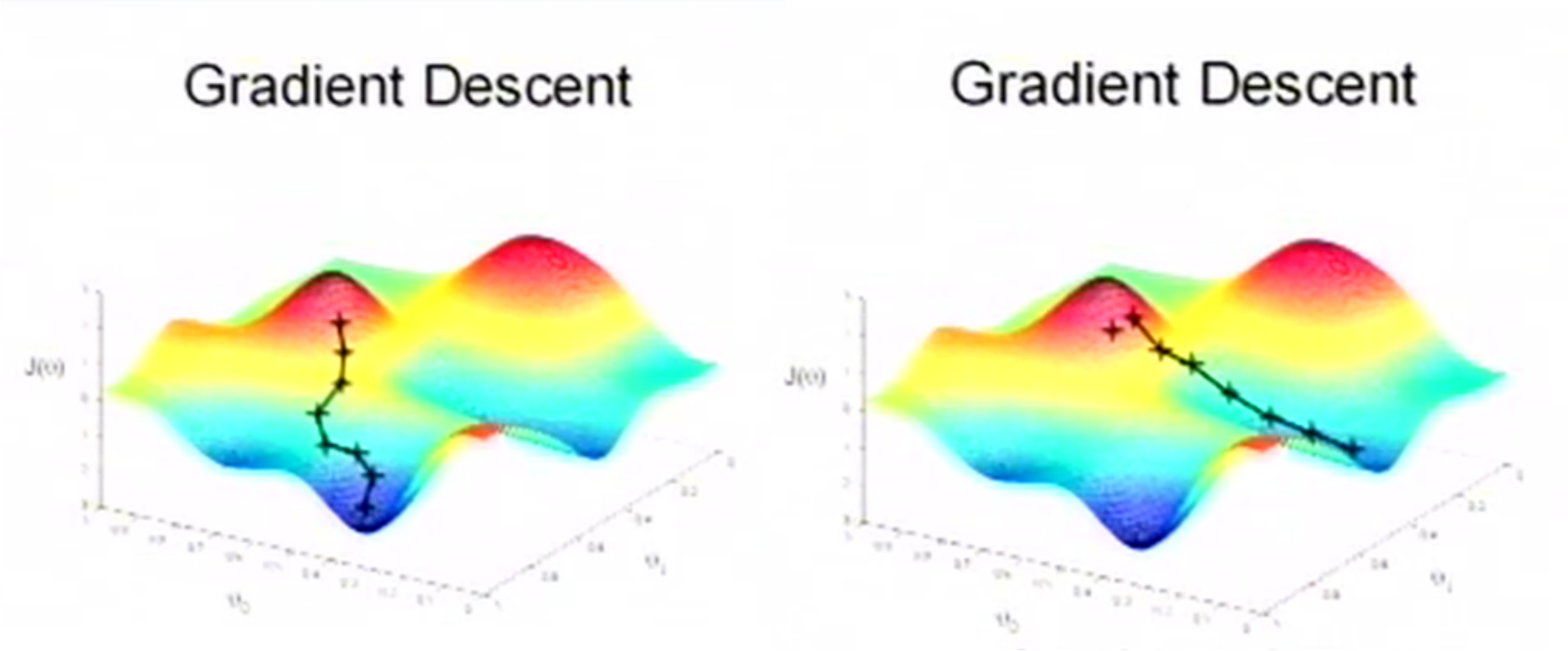
- 感知准则函数的梯度下降求解

$$J_p(a)=\sum_{y\in Err}-a^T y$$ 寻求最优的a令J极小

$$a(k+1)=a(k)-\rho(k)\nabla J$$ 迭代方式寻找极小值

思想：由某一初值开始，沿某一方向，按某一步长搜寻极小值，由于梯度是函数值增长最快的方向，沿负梯度即为函数下降最快的方向，沿此方向可最快到达极小点，故称梯度下降法。

梯度下降法演示



不同初始值的梯度下降效果

4.3感知准则函数（两类）

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

$$\because \nabla J_p(a)=\frac{\partial J_p(a)}{\partial a}=\frac{\partial}{\partial a}\sum_{y\in Err}(-a^T y)=\sum_{y\in Err}(-y)$$ 纯量对向量求导

$$\therefore a(k+1)=a(k)+\rho(k)\sum_{y\in Err} y$$ 负负为正

补充定义：纯量对向量求导

$$\frac{\partial(a^T y)}{\partial a_i}=\frac{\partial}{\partial a_i}\sum_{i=1}^{d+1}a_i y_i=y_i$$

$$\therefore \frac{\partial(a^T y)}{\partial a}=y$$

4.3感知准则函数（两类）

- 感知准则函数梯度下降法步骤

$$
\begin{cases}
a(1) = \text{初始随机数产生} \\
a(2) = a(1) - \rho(1)\nabla J(a(1)) \\
a(3) = a(2) - \rho(2)\nabla J(a(2)) \\
\vdots \\
a(k+1) = a(k) \qquad \text{收敛，结束}
\end{cases}
$$

4.3感知准则函数（两类）

- 步骤描述

（1）给定初始权向量a(1)和步长 $\rho(1)$

（2）找出被权向量a(k)错分的所有样本，转（3）；如无错分样本，算法结束。

（3）按迭代公式求新的权向量a(k+1)，转（2）

$$a(k+1)=a(k)+\rho(k)\sum_{y\in Err} y$$

4.3感知准则函数（两类）

例：已知两类样本，以感知器算法求解分类器（求判别函数）

$$\omega_1: \quad x^1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, x^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\omega_2: \quad x^3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, x^4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$a(1) = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}$$

4.3感知准则函数（两类）

解：写出规范化增广样本向量

$$x^1=\begin{pmatrix}0\\0\\1\end{pmatrix},x^2=\begin{pmatrix}0\\1\\1\end{pmatrix},x^3=\begin{pmatrix}-1\\0\\-1\end{pmatrix},x^4=\begin{pmatrix}-1\\-1\\-1\end{pmatrix}$$

$(1)\rho(k)=1$

$$(2)a^T(1)x^1=(-1\quad -1\quad 1)\begin{pmatrix}0\\0\\1\end{pmatrix}=1>0\text{正确分类}$$

$$a^T(1)x^2=(-1\quad -1\quad 1)\begin{pmatrix}0\\1\\1\end{pmatrix}=0\text{错分}$$

$$a^T(1)x^3=(-1\quad -1\quad 1)\begin{pmatrix}-1\\0\\-1\end{pmatrix}=0\text{错分}$$

$$a^T(1)x^4=(-1\quad -1\quad 1)\begin{pmatrix}-1\\-1\\-1\end{pmatrix}=1>0\text{正确分类}$$

4.3感知准则函数（两类）

$$(3)a(2)=a(1)+x^2+x^3=\begin{pmatrix}-1\\-1\\1\end{pmatrix}+\begin{pmatrix}0\\1\\1\end{pmatrix}+\begin{pmatrix}-1\\0\\-1\end{pmatrix}=\begin{pmatrix}-2\\0\\1\end{pmatrix}$$

$$\therefore g(x)=a^T(2)x=\begin{pmatrix}-2 & 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\1\end{pmatrix}=-2x_1+1$$

$$a^T(2)x^1=\begin{pmatrix}-2 & 0 & 1\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}=1>0\text{正确分类}$$

$$a^T(2)x^2=\begin{pmatrix}-2 & 0 & 1\end{pmatrix}\begin{pmatrix}0\\1\\1\end{pmatrix}=1>0\text{正确分类}$$

$$a^T(2)x^3=\begin{pmatrix}-2 & 0 & 1\end{pmatrix}\begin{pmatrix}-1\\0\\-1\end{pmatrix}=1>0\text{正确分类}$$

$$a^T(2)x^4=\begin{pmatrix}-2 & 0 & 1\end{pmatrix}\begin{pmatrix}-1\\-1\\-1\end{pmatrix}=1>0\text{正确分类}$$

4.3感知准则函数（两类）

- 批处理—把所有错分类样本一次性找出来修正
- 与人的学习方式不同，人采用“单样本修正”方式
- 单样本修正—找出一个错分样本即进行修正

4.3感知准则函数（两类）

- 按单样本修正法重作例题

$(1)\rho(k)=1$

$(2)a^T(1)x^1=(-1\quad -1\quad 1)\begin{pmatrix}0\\0\\1\end{pmatrix}=1>0$正确分类

$a^T(1)x^2=0$，错分，$a(2)=a(1)+x^2=\begin{pmatrix}-1\\0\\2\end{pmatrix}$

$a^T(2)x^3=(-1\quad 0\quad 2)\begin{pmatrix}-1\\0\\-1\end{pmatrix}=-1<0$，错分，$a(3)=a(2)+x^3=\begin{pmatrix}-2\\0\\1\end{pmatrix}$

$a^T(3)x^4=(-2\quad 0\quad 1)\begin{pmatrix}-1\\-1\\-1\end{pmatrix}=1>0$

(3)回到(2)，$a^T(3)x^1=1>0$

$a^T(3)x^2=1>0$

$a^T(3)x^3=1>0$

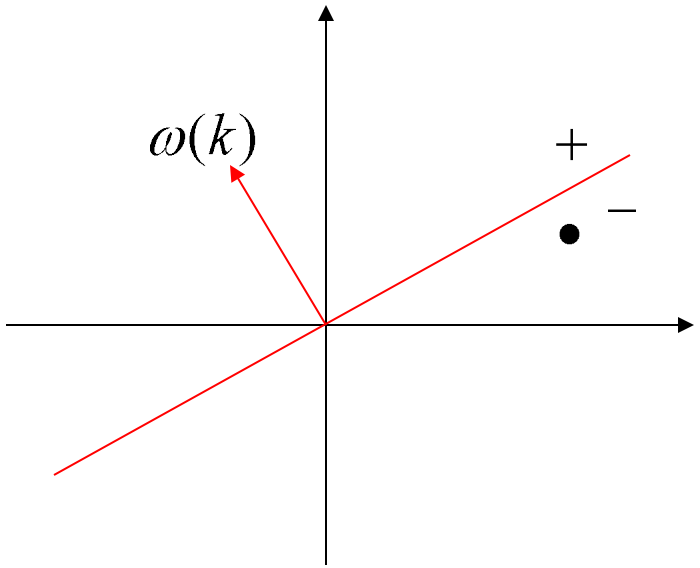
$a^T(3)x^4=1>0$

$\therefore g(x)=a^T(3)x=(-2\quad 0\quad 1)\begin{pmatrix}x_1\\x_2\\1\end{pmatrix}=-2x_1+1$

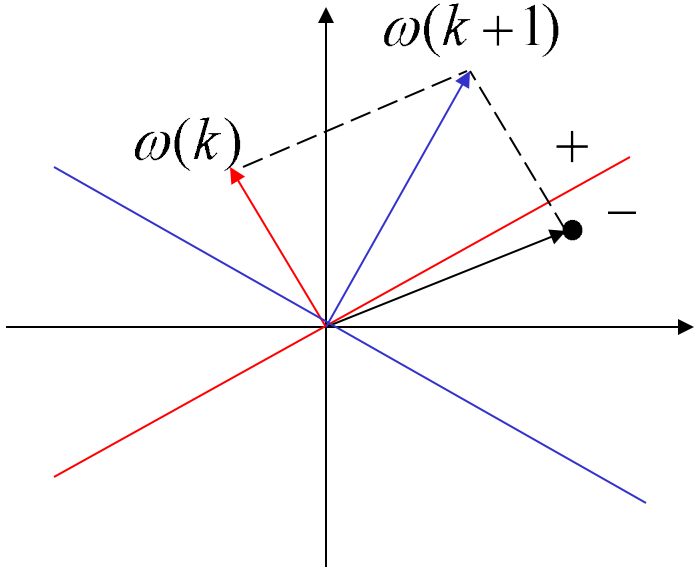
固定增量法

4.3感知准则函数（两类）

- 迭代方法要从理论上证明收敛，此处仅以图形方式说明



修正前错分类



修正后正确分类

4.3感知准则函数（两类）

- 结论：可以证明，采用梯度下降法对于线性可分的样本集，经过有限步修正，一定能找到一个使准则函数达到极小值的权向量a，即算法在有限步内收敛，其收敛速度取决于初始权向量和步长。
- 分类界面接近最终位置时，步长需选得较小，否则修正过头。但如各步步长都选得很小，收敛速度变慢。为此，采用变步长方案。

$$\rho(k) = \rho(1)/k$$

- 对于线性不可分情况，算法不收敛。

4.3感知准则函数（两类）

单样本修正与批处理的比较：

（1）学习开始时不能得到所有的训练样本，必须使用在线方法。

（2）训练样本数很大时，单样本方便，比批处理有效，因为批处理需要附加记忆来积累局部更新。

（3）单样本引入一些随机噪声，有助于逃离局部极小值。批处理引入一些平均滤波。

（4）单样本适合大规模分类问题，因为很多训练样本含有冗余信息，对梯度贡献类似，更新权值前计算所有样本浪费。

4.3感知准则函数（两类）

单样本修正与批处理的比较：

（5）批处理对梯度矢量估计较好，避免了权值变化相互影响，因此需要高精度映射时，选用批处理方法，但其难于通过提高计算速度来补偿增加的计算开销。

（6）批处理在复杂优化中有直接应用，单样本与批处理的相对有效性与求解的问题直接相关。

结论：许多情况下，单样本修正优于批处理，特别是对大的和冗余的训练集。

4.4最小平方误差准则函数

- 感知器对线性不可分情况不收敛

  ∵样本集是否线性可分无法确定

  ∴希望找到一种既适合线性可分又适合线性不可分情况的算法。

- 该算法具有如下特性：

  对于线性可分情况，一定能找到将样本全部正确分类的权向量。

  对于线性不可分情况，得到一个使误差平方和极小的权向量。

- 该准则函数称为平方误差准则函数。

4.4最小平方误差准则函数

- 定义平方误差准则函数

$a^{*T}y^i = b^i > 0$　　增广规范化样本，b^i为理想输出

$$Ya = b \qquad Y = \begin{bmatrix} y^{1T} \\ \vdots \\ y^{NT} \end{bmatrix} = \begin{bmatrix} y_1^1 & \cdots & y_{d+1}^1 \\ \vdots & \ddots & \vdots \\ y_1^N & \cdots & y_{d+1}^N \end{bmatrix}, b = \begin{bmatrix} b^1 & \cdots & b^N \end{bmatrix}^T$$

$e^i = b^i - a^T y^i$　　定义误差

$$J_s(a) = \frac{1}{2}\|Ya - b\|^2 = \frac{1}{2}\sum_{i=1}^{N}(a^T y^i - b^i)^2$$　　误差平方准则函数

4.4最小平方误差准则函数

- 最小二乘近似解

$Ya = b$　　通常有$N > d+1$

即方程个数大于未知数个数，属超定方程组，一般无解，但可求线性最小二乘解。

极小化$J_s(a) = \frac{1}{2}\|Ya - b\|^2 = \frac{1}{2}\sum_{i=1}^{N}(a^T y^i - b^i)^2$

可得a^*,采用何种算法?

4.4最小平方误差准则函数

一、伪逆法（解析方法）

令$\nabla J_s(a)=0$可得$J_s(a)$极小值

$$\nabla J_s(a)=\frac{\partial}{\partial a}\left(\frac{1}{2}\|Ya-b\|^2\right)=\frac{\partial}{\partial a}\frac{1}{2}\sum_{i=1}^{N}(a^T y^i-b^i)^2$$

$$=\sum_{i=1}^{N}(a^T y^i-b^i)y^i$$

$$=Y^T(Ya-b)$$

$$\therefore Y^T(Ya-b)=0 \qquad \therefore Y^TYa=Y^Tb$$

4.4最小平方误差准则函数

一、伪逆法（解析方法）

$$Y^TYa = Y^Tb$$

$\because Y^TY$为方阵，且一般为非奇异的，故可逆

$$\therefore a^* = (Y^TY)^{-1}Y^Tb$$

$Y^+ = (Y^TY)^{-1}Y^T$称为伪逆

为什么称为伪逆？

4.4最小平方误差准则函数

一、伪逆法（解析方法）

b的选取？

a^*依赖于b,当b取某些特殊值时，解有优良特性

$$b=\begin{bmatrix} N/N_1 \\ \vdots \\ N/N_1 \\ N/N_2 \\ \vdots \\ N/N_2 \end{bmatrix}\begin{matrix}\left.\begin{matrix}\\ \\ \\ \end{matrix}\right\}N_1\text{个} \\ \left.\begin{matrix}\\ \\ \\ \end{matrix}\right\}N_2\text{个}\end{matrix}$$

等价于Fisher线性判别

$$b=\begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \\ 1 \end{bmatrix}\left.\begin{matrix}\\ \\ \\ \\ \\ \\ \end{matrix}\right\}N\text{个, 样本数}N\to\infty$$

渐进逼近贝叶斯判别函数

4.4最小平方误差准则函数

例：已知样本为

$\omega_1 : x^1 = (0,0)^T, x^2 = (0,1)^T$

$\omega_2 : x^3 = (1,0)^T, x^4 = (1,1)^T$　　令$b = (1,1,1,1)^T$，利用伪逆法求判别函数

解：样本增广规范化

4.4最小平方误差准则函数

例：已知样本为

$\omega_1 : x^1 = (0,0)^T, x^2 = (0,1)^T$

$\omega_2 : x^3 = (1,0)^T, x^4 = (1,1)^T$　　令$b = (1,1,1,1)^T$，利用伪逆法求判别函数

解：样本增广规范化

$y^1 = (0,0,1)^T, y^2 = (0,1,1)^T, y^3 = (-1,0,-1)^T, y^4 = (-1,-1,-1)^T$

$$Y = (y^1, y^2, y^3, y^4)^T = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

$$Y^+ = (Y^T Y)^{-1} Y^T = \frac{1}{2} \begin{bmatrix} -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ 3/2 & 1/2 & -1/2 & 1/2 \end{bmatrix}$$

$$\therefore a = Y^+ b = (-2,0,1)^T$$

$$\therefore g(x) = a^T y = (-2,0,1)(x_1, x_2, 1)^T = -2x_1 + 1$$

4.4最小平方误差准则函数

- 上例中如何求逆？

$$\begin{cases} \dfrac{A^*}{|A|} = A^{-1} \qquad A^* \text{为伴随矩阵} \\ \begin{bmatrix} & 1 & & 0 \\ A & & \ddots & \\ & 0 & & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & & 0 & \\ & \ddots & & A^{-1} \\ 0 & & 1 & \end{bmatrix} \quad \text{初等变换} \end{cases}$$

4.4最小平方误差准则函数

二、梯度下降法

伪逆法可得$a^* = Y^+ b$，但需计算$Y^+ = (Y^T Y)^{-1} Y^T$

问题{ 要求$Y^T Y$非奇异
求Y^+计算量大，同时可能引入较大的计算误差

$\therefore$ 实际往往不采用此解析方法

4.4最小平方误差准则函数

- MSE的梯度下降法（批处理）

$$\nabla J_s(a) = Y^T(Ya - b)$$

梯度下降算法

$$\begin{cases} a(1), \text{任意} \\ a(k+1) = a(k) - \rho_k Y^T(Ya - b) \end{cases}$$

可以证明, 若$\rho_k = \dfrac{\rho_1}{k}$, ρ_1为正常数

则所得权向量收敛于使$\nabla J_s(a) = 2Y^T(Ya - b) = 0$的$a^*$,即伪逆解

算法优点:

$$\begin{cases} \text{无论} Y^T Y \text{是否奇异, 都能找到解} \\ \text{只计算} Y^T Y, \text{比计算伪逆计算量小} \end{cases}$$

4.4最小平方误差准则函数

- MSE梯度下降法的单样本改进（单样本）

$$\begin{cases} a(1), \text{任意} \\ a(k+1) = a(k) + \rho_k (b_k - a(k)^T y^k) y^k \end{cases}$$

$\because b_k$任意给定

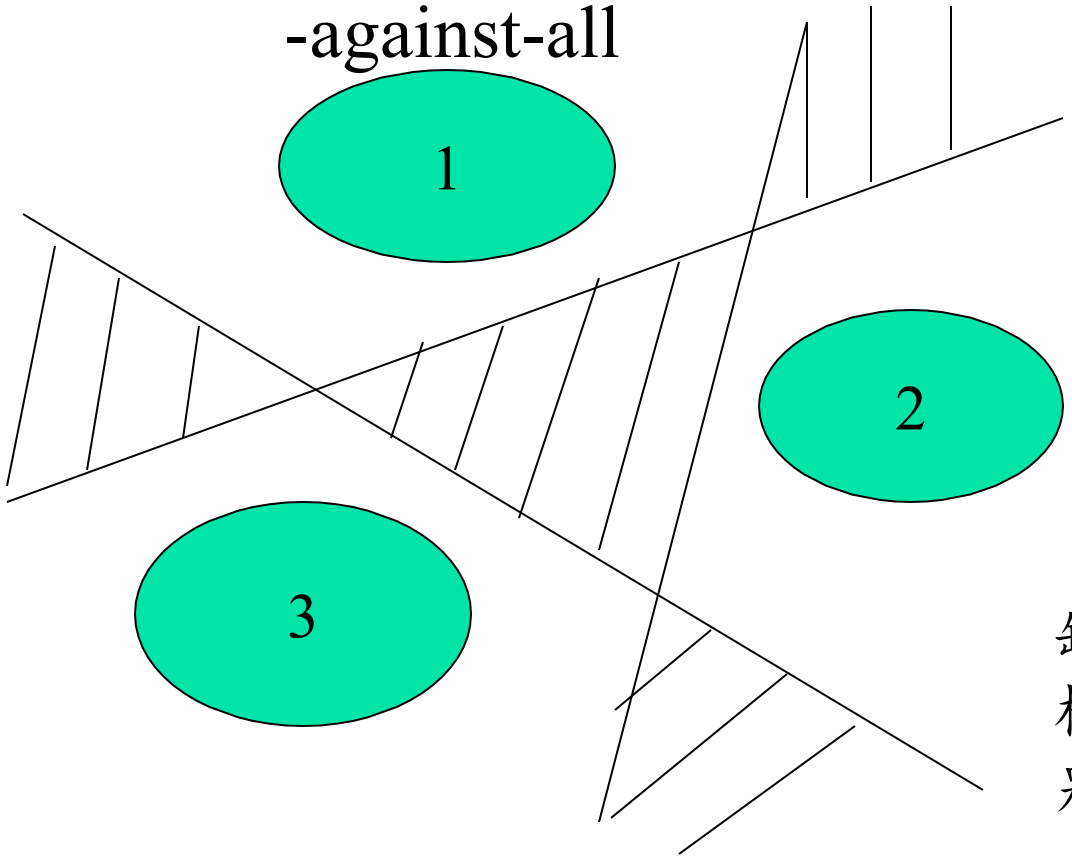
$\therefore b_k = a(k)^T y^k$几乎不可能成立

$\therefore$ 应设$\rho_k = \dfrac{\rho_1}{k}$,以保证算法收敛

此算法称为$Widrow-Hoff$算法

4.5线性分类器的多类推广

1、按两类问题处理 one -against-all



C类C个判别函数，若

$$g_i(x) > 0 \qquad x \in \omega_i$$

若两个判别函数同时大于0，则样本落入不确定区域

缺点：存在不确定区域，样本落入其中无法确定类别，错误率大。

4.5线性分类器的多类推广

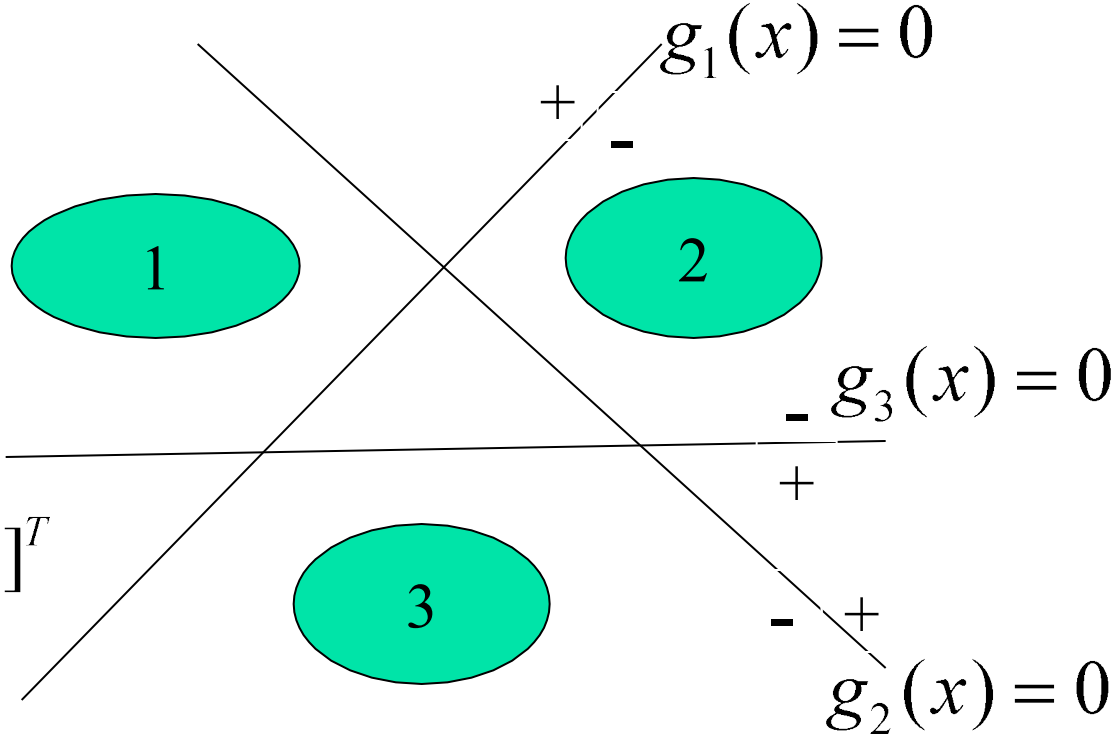
- 例：设有一个三类问题，其判别函数为

$$g_1(x) = -x_1 + x_2$$

$$g_2(x) = x_1 + x_2 - 5$$

$$g_3(x) = -x_2 + 1$$

待分类样本为$\mathbf{x} = [6 \quad 5]^T$



4.5线性分类器的多类推广

- 例：设有一个三类问题，其判别函数为

$$g_1(x) = -x_1 + x_2$$

$$g_2(x) = x_1 + x_2 - 5$$

$$g_3(x) = -x_2 + 1$$

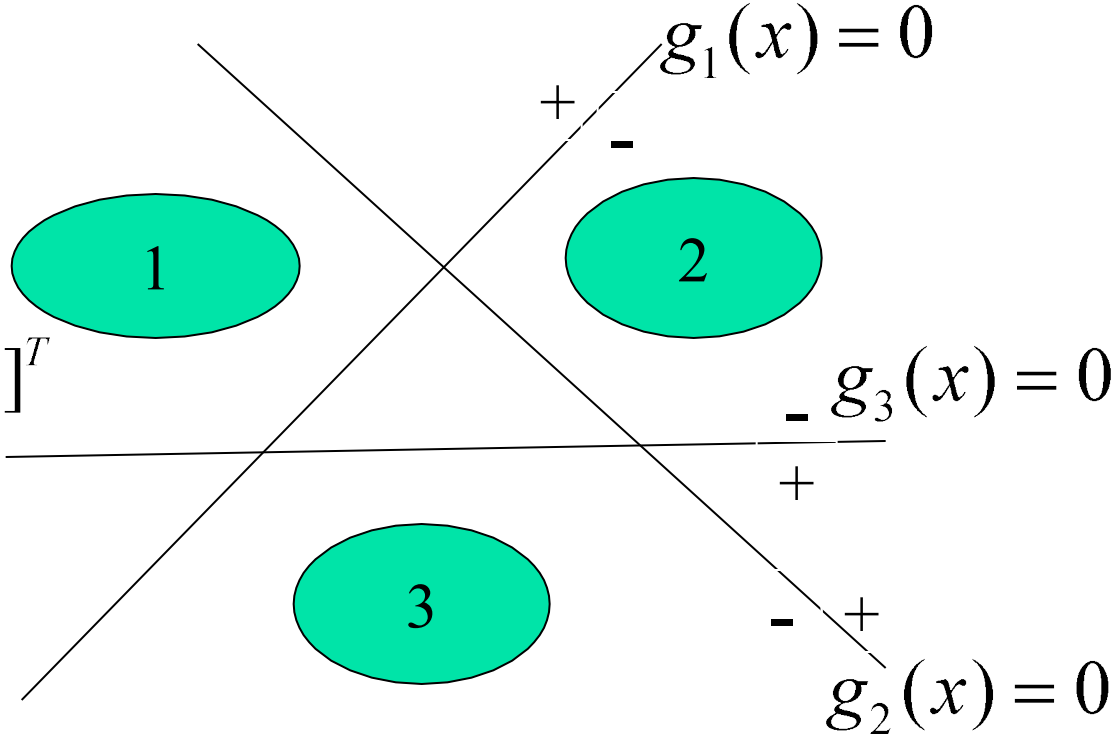
待分类样本为$\mathbf{x} = [6 \quad 5]^T$

$$g_1(x) = -1 < 0$$

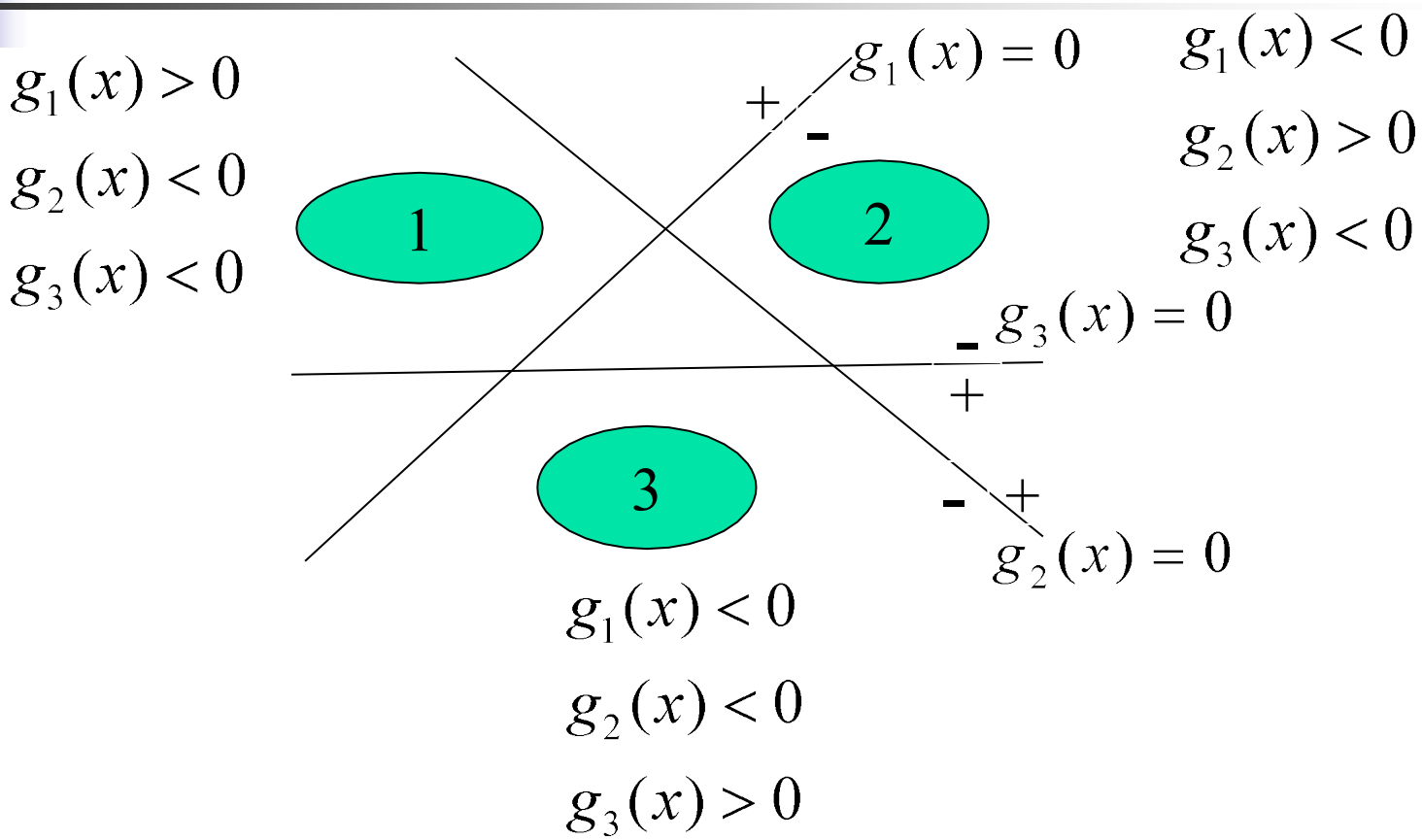
$$g_2(x) = 6 > 0$$

$$g_3(x) = -4 < 0$$

$$\because g_2(x) > 0 \therefore \in \omega_2$$



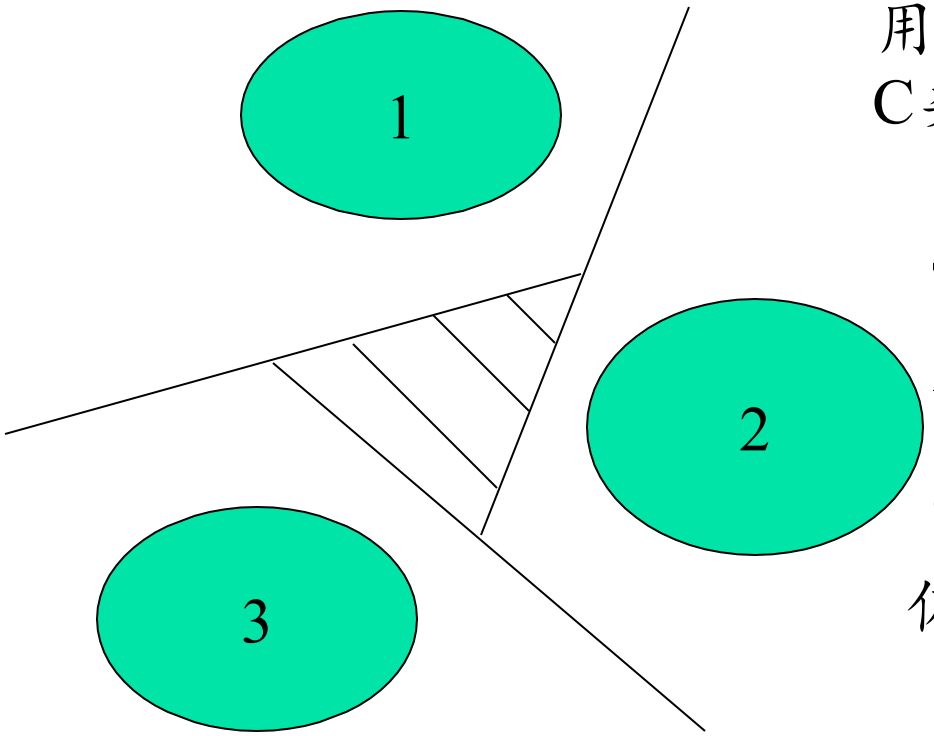
4.5线性分类器的多类推广



两个判别函数同时大于0则落入不确定区

4.5线性分类器的多类推广

2、多类化两类 one-against-one



用线性分类器将类别两两分开，
C类问题需要C(C-1)/2个判别函数

$$C_c^2 = \frac{C(C-1)}{2!} = \frac{C(C-1)}{2}$$

决策规则:

$$g_{ij}(x) > 0 \qquad i, j = 1, \cdots, c; i \neq j \qquad x \in \omega_i$$

仍然存在不确定区域

4.5线性分类器的多类推广

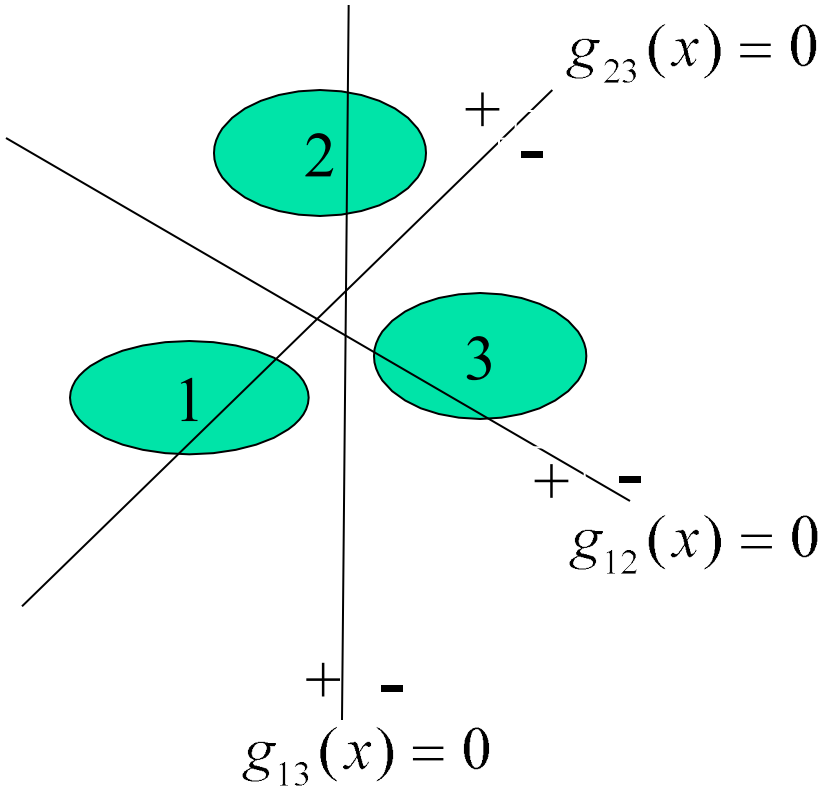
- 例：设有一个三类问题，其判别函数为

$$g_{12}(x) = -x_1 - x_2 + 5$$

$$g_{13}(x) = -x_1 + 3$$

$$g_{23}(x) = -x_1 + x_2$$

待分类样本为$\mathbf{x} = [4 \quad 3]^T$



4.5线性分类器的多类推广

- 例：设有一个三类问题，其判别函数为

$$g_{12}(x) = -x_1 - x_2 + 5$$

$$g_{13}(x) = -x_1 + 3$$

$$g_{23}(x) = -x_1 + x_2$$

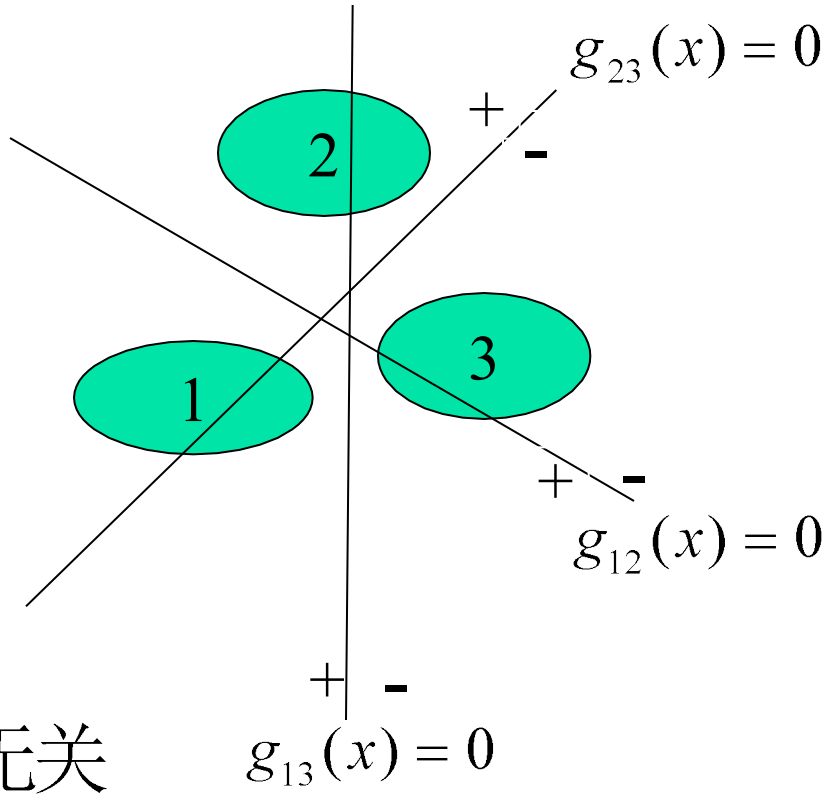
待分类样本为$\mathbf{x} = [4 \quad 3]^T$

$$g_{12}(x) = -2 < 0 \Rightarrow g_{21} = 2 > 0$$

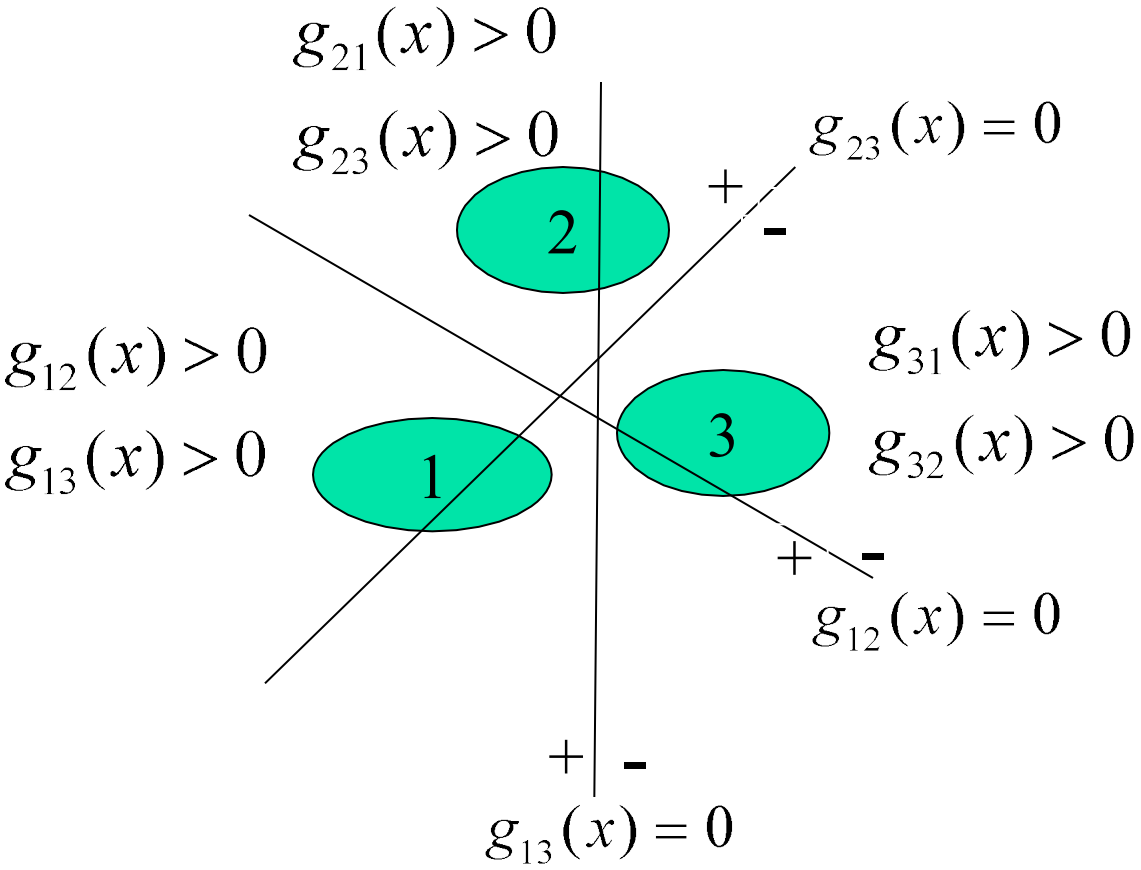
$$g_{13}(x) = -1 < 0 \Rightarrow g_{31} = 1 > 0$$

$$g_{23}(x) = -1 < 0 \Rightarrow g_{32} = 1 > 0$$

$\because g_{31} > 0, g_{32} > 0 \therefore \mathbf{x} \in \omega_3$，与$g_{21}$无关



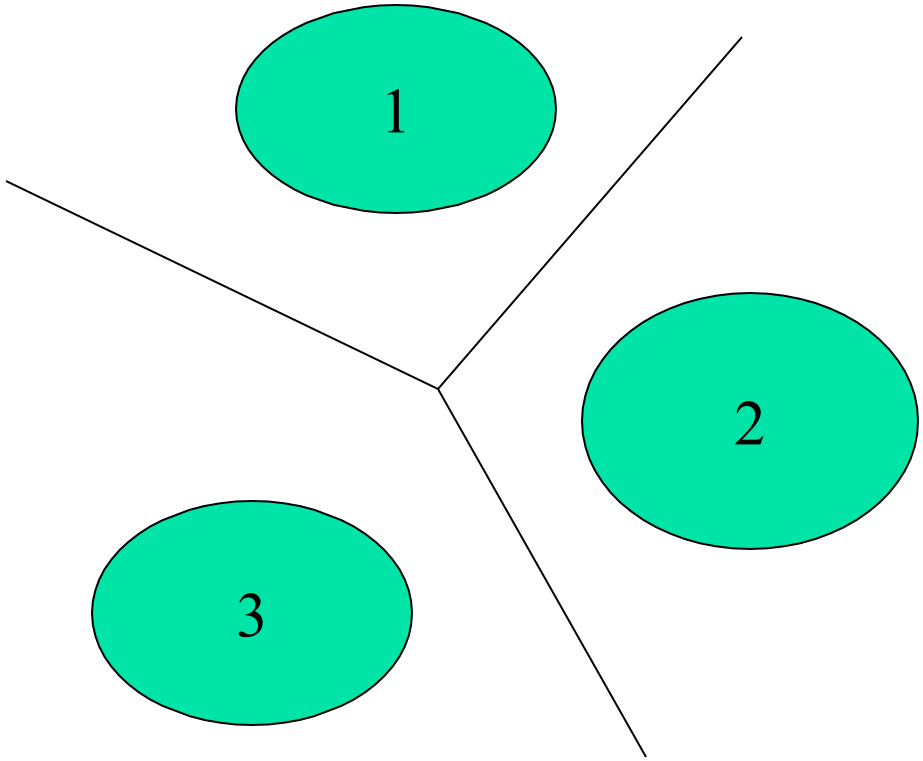
4.5线性分类器的多类推广



仍然存在无效区域

4.5线性分类器的多类推广

3、直接按多类问题处理



C类C个判别函数

$$g_i(x) = \max g_j(x) \qquad x \in \omega_i$$

不存在不确定区域，多类问题多采用此方案。

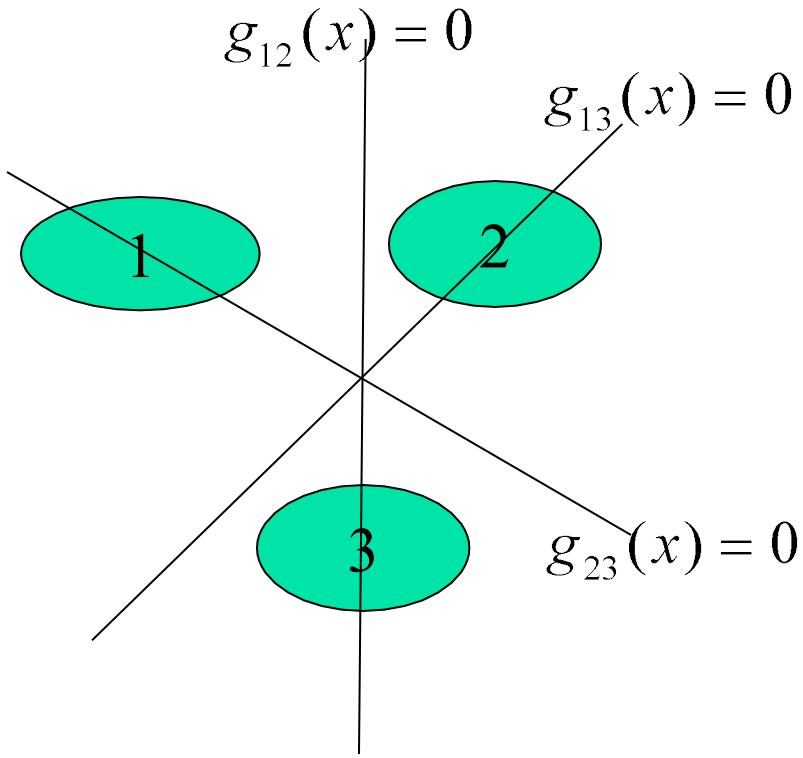
4.5线性分类器的多类推广

- 例：设有一个三类问题，其判别函数为

$$g_1(x) = -x_1 + x_2$$

$$g_2(x) = x_1 + x_2 - 1$$

$$g_3(x) = -x_2$$



4.5线性分类器的多类推广

- 例：设有一个三类问题，其判别函数为

$$g_1(x) = -x_1 + x_2$$

$$g_2(x) = x_1 + x_2 - 1$$

$$g_3(x) = -x_2$$

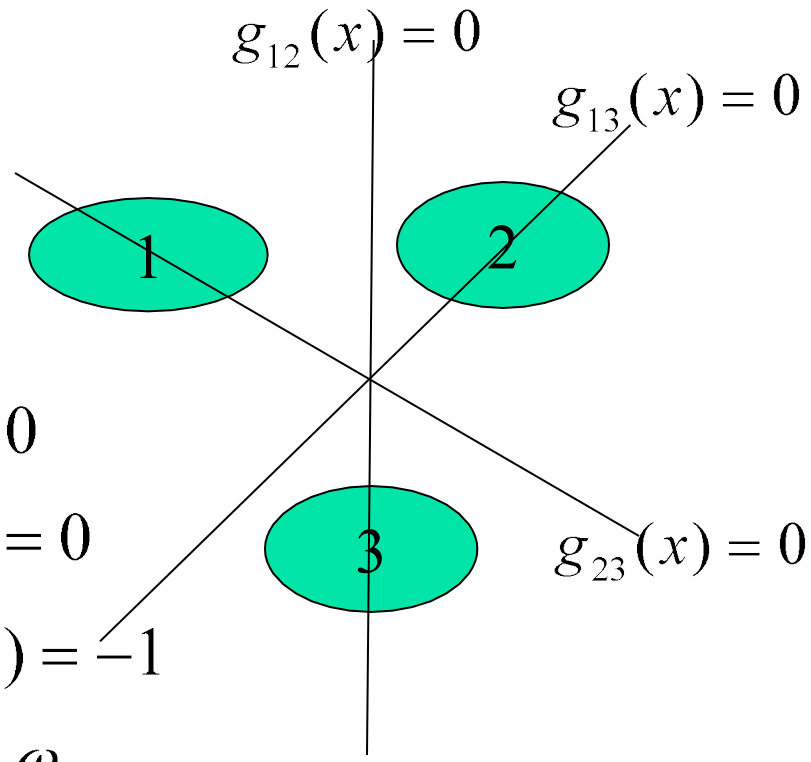
$$g_{12}(x) = g_1(x) - g_2(x) = -2x_1 + 1 = 0$$

$$g_{13}(x) = g_1(x) - g_3(x) = -x_1 + 2x_2 = 0$$

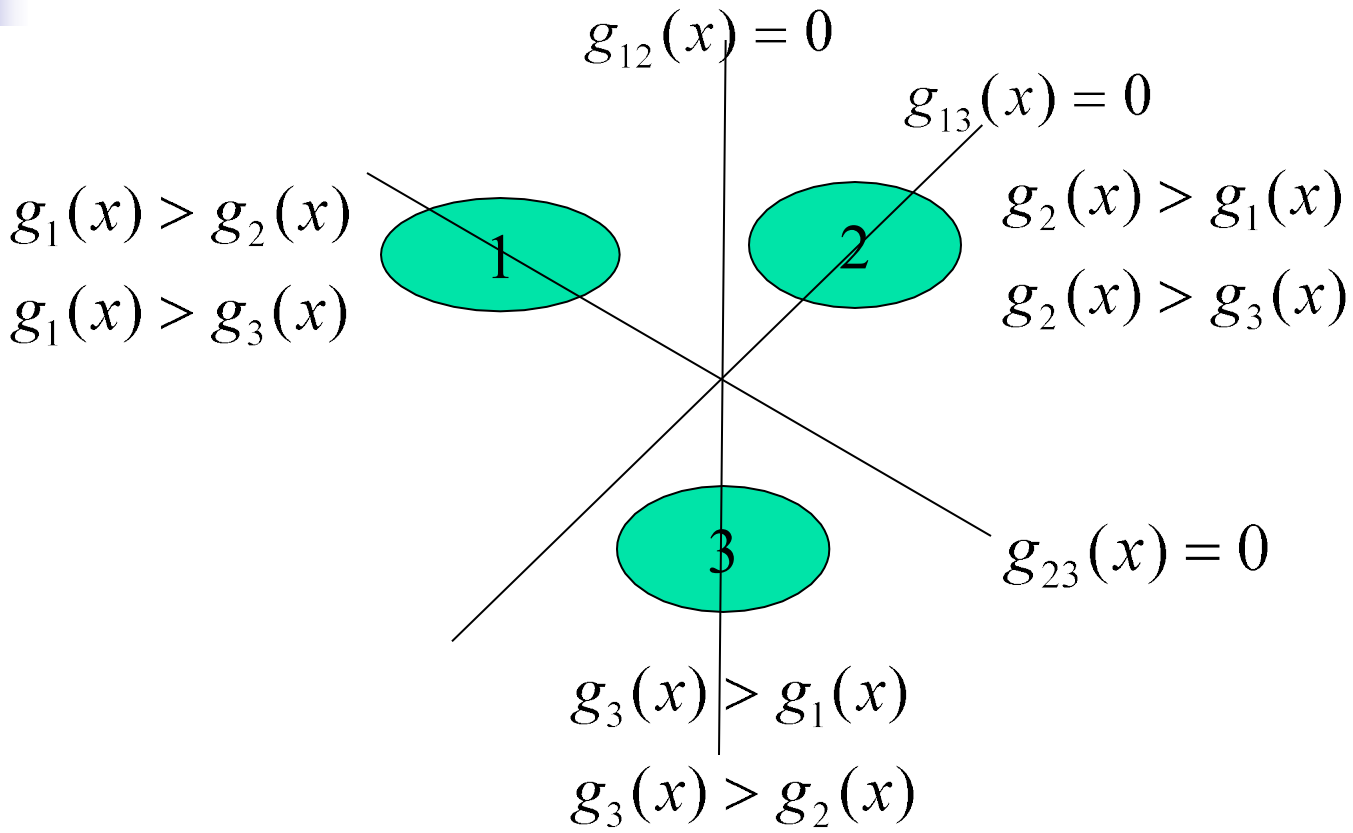
$$g_{23}(x) = g_2(x) - g_3(x) = x_1 + 2x_2 - 1 = 0$$

$$\mathbf{x} = [1 \quad 1]^T, g_1(x) = 0, g_2(x) = 1, g_3(x) = -1$$

$$\because g_2(x) > g_1(x), g_2(x) > g_3(x) \therefore \mathbf{x} \in \omega_2$$



4.5线性分类器的多类推广



无不确定区域，经常采用

4.5线性分类器的多类推广

- 多类感知器例题

给出三类训练样本如下：

$\omega_1 : \mathbf{x}^1 = (0,0)^T, \omega_2 : \mathbf{x}^2 = (1,1)^T, \omega_3 : \mathbf{x}^3 = (-1,1)^T$

增广形式为$\omega_1 : \mathbf{x}^1 = (0,0,1)^T, \omega_2 : \mathbf{x}^2 = (1,1,1)^T, \omega_3 : \mathbf{x}^3 = (-1,1,1)^T$

三类初始权值$\mathbf{a}_1(1) = \mathbf{a}_2(1) = \mathbf{a}_3(1) = (0,0,0)^T$

取步长$\rho = 1$，设计感知器，由于此处是多类问题，

没有任何一类的样本应乘以-1。

4.5线性分类器的多类推广

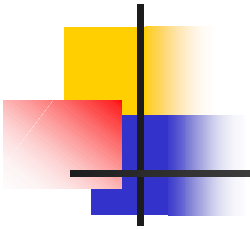
- 训练样本数的选择问题

为保证线性分类器的分类效果，训练样本数的选择应遵循如下公式

$$N_k = 2(d+1)$$

d为模式维数。

训练样本数一般为N_k的10到20倍。



本章小结

- 线性判别概念
- Fisher线性判别
- 感知准则函数
- 最小平方误差准则函数
- 线性分类的多类推广