第4章 线性分类器

Outline:

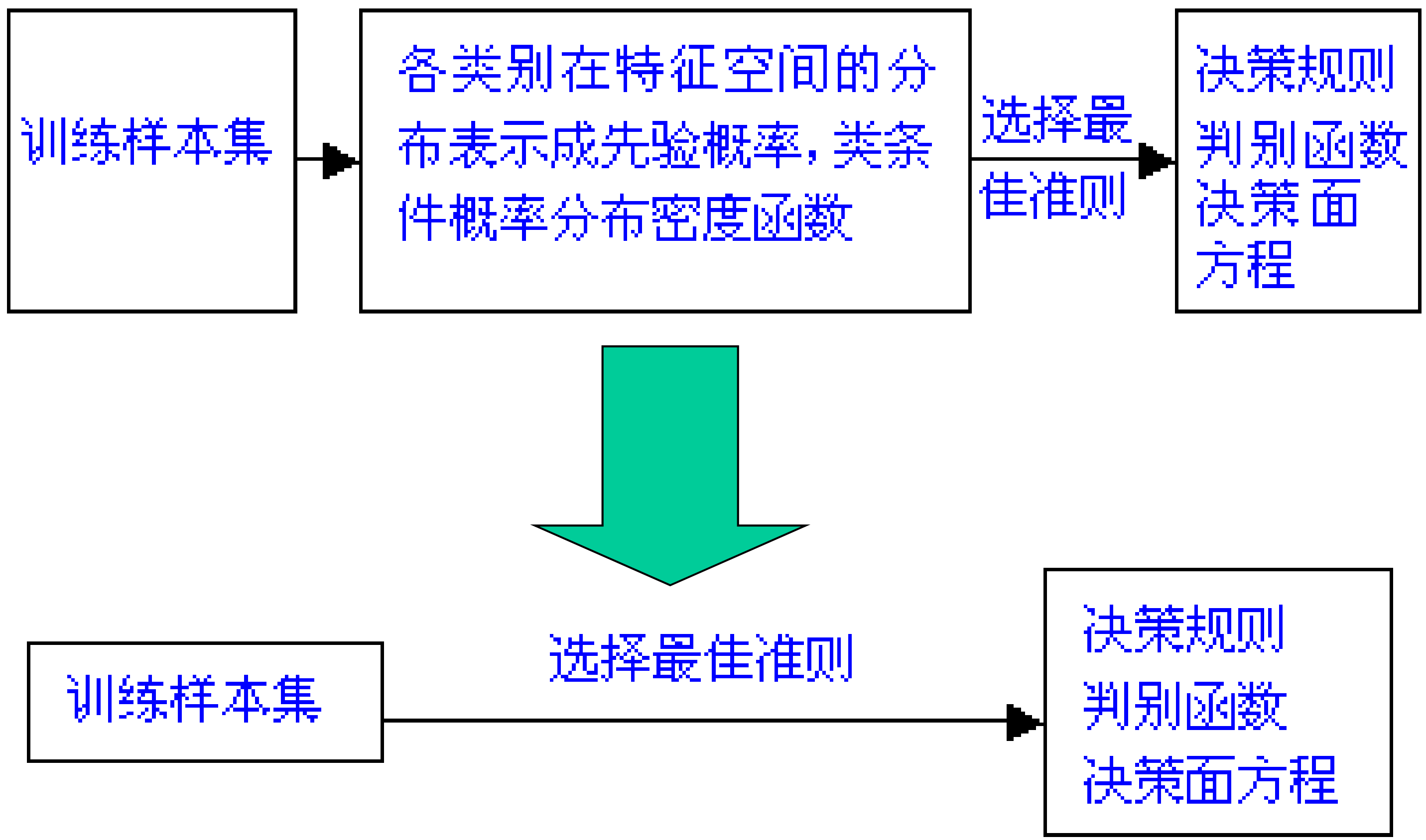
- 引言
- 线性判别函数基本概念
- 广义线性分类器
- 线性分类器设计步骤
- Fisher线性判别分析
- 感知器算法
- 最小平方误差判别
- 多类问题

引言

非参数判别分类方法的核心是由训练样本集提供的信息直接确定决策域的划分方法。这里最重要的概念是分类器设计用一种训练与学习的过程来实现。机器自动识别事物的能力通过训练学习过程来实现，其性能通过学习过程来提高，这是模式识别、人工神经元网络中最核心的内容。

引言

- 按Bayes决策理论设计分类器的步骤



引言

- 按照基于统计参数的决策分类方法，判别函数及决策面方程的类别确定是由样本分布规律决定的。但是在非参数判别方法的设计中，使用什么典型的分类决策方法却要预先由设计者确定，然后利用训练样本集提供的信息确定这些函数中的参数。非参数判别分类方法选择函数类型与确定参数是两个过程。

Outline:

- 引言
- 线性判别函数基本概念
- 广义线性分类器
- 线性分类器设计步骤
- Fisher线性判别分析
- 感知器算法
- 最小平方误差判别
- 多类问题

第四章 线性分类器

线性分类器

- 思路:
  - ✓首先选定判别函数类和一定的目标（准则/依据），利用样本集确定出函数类中的某些未知参数，使所选的准则最好

线性分类器

➤概念

$$g(x) = w^T x + w_0$$

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_d \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_d \end{bmatrix}$$

线性分类器

➢ 概念

两类： $g(x)=g_1(x)-g_2(x)$

决策规则：$\begin{cases} g(x)>0, & x\in\omega_1 \\ g(x)<0, & x\in\omega_2 \\ g(x)=0, & \textit{anyclass or refuse} \end{cases}$

$g(x)=0$ 定义了一个决策面，它把两个类别的样本点分开

线性分类器

➢ 概念

x_1, x_2 都在决策面H上

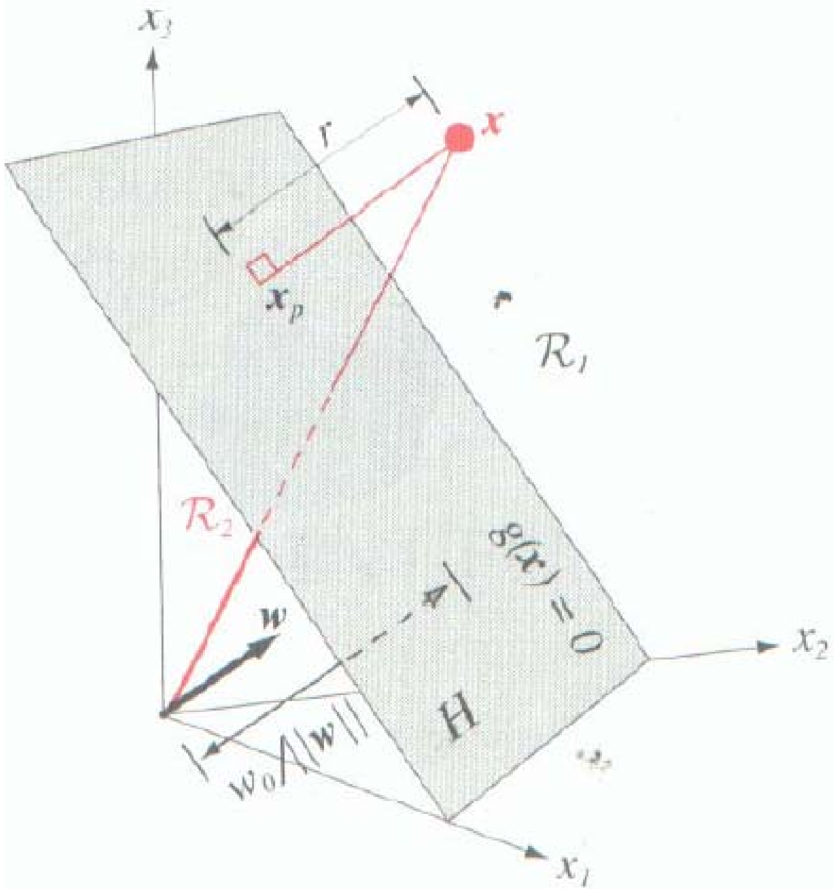
$$W^T x_1 + w_0 = W^T x_2 + w_0$$
$$\Rightarrow W^T(x_1 - x_2) = 0$$

W与超平面垂直

$$x = x_p + r\frac{W}{\|W\|}$$

x_p是x到H上的投影向量；

r是x到H的垂直距离；



线性分类器

- 概念

$$g(x)=W^T\left(x_p+r\frac{W}{\|W\|}\right)+w_0$$

$$=W^Tx_p+w_0+r\frac{W^TW}{\|W\|}=r\|W\|$$

$$\Rightarrow r=\frac{g(x)}{\|W\|}$$

若x为原点： $g(x)=w_0$

原点到超平面H的距离： $r_0=\frac{w_0}{\|W\|}$

Outline:

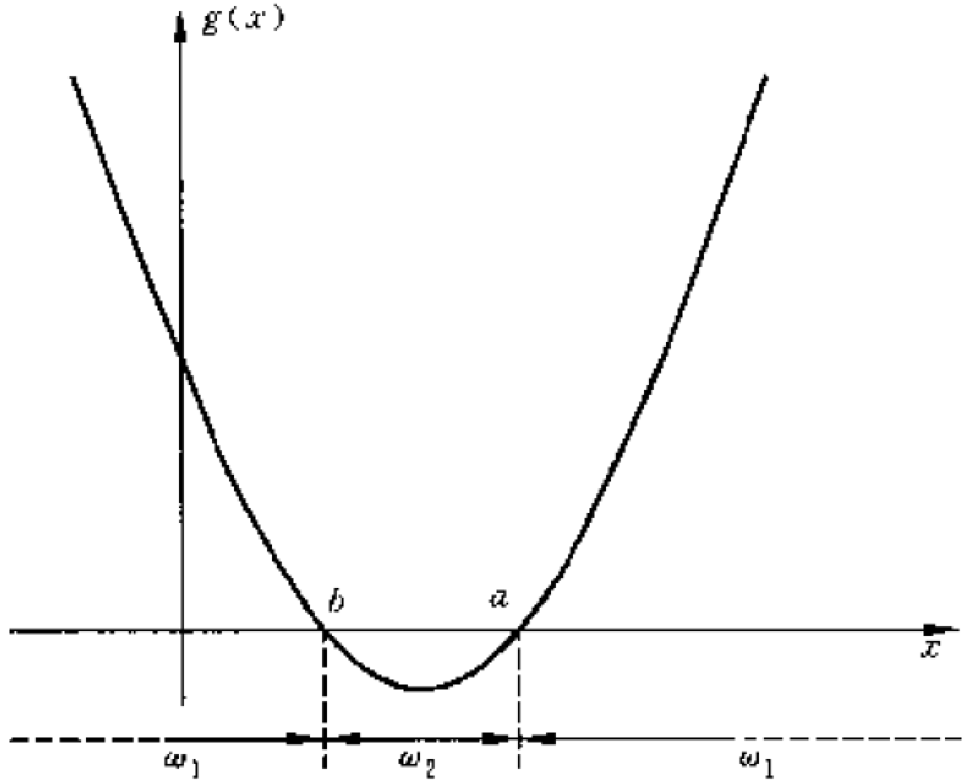
- 引言
- 线性判别函数基本概念
- 广义线性分类器
- 线性分类器设计步骤
- Fisher线性判别分析
- 感知器算法
- 最小平方误差判别
- 多类问题

广义线性判别函数
Generalized Linear Discriminant

第四章 线性分类器

➢欲设计这样一个一维样本的分类器，使其性能为：

$$\text{如果}\begin{cases} x < b \text{或} x > a & \text{决策} x \in \omega_1 \\ b \le x \le a & \text{决策} x \in \omega_2 \end{cases}$$



广义线性判别函数

判别函数：

$$g(x) = (x-a)(x-b)$$

决策规则：

$$\begin{cases} g(x) > 0 & x \in \omega_1 \\ g(x) \le 0 & x \in \omega_2 \end{cases}$$

映射:X→Y

$$g(x) = c_0 + c_1 x + c_2 x^2$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \Rightarrow g(x) = a^T Y = \sum_{i=1}^{3} a_i y_i$$

$$where: \quad a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}$$

广义线性判别函数

一种特殊的映射方法：

$$Y=\begin{bmatrix}1\\x_1\\x_2\\\vdots\\x_d\end{bmatrix}=\begin{bmatrix}1\\X\end{bmatrix}$$

增广样本向量

$$\overline{a}=\begin{bmatrix}w_0\\w_1\\w_2\\\vdots\\w_d\end{bmatrix}=\begin{bmatrix}w_0\\W\end{bmatrix}$$

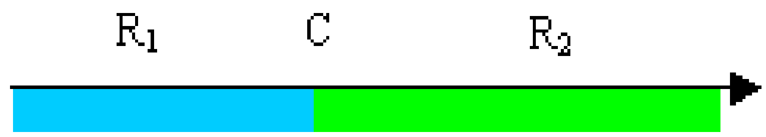
增广权向量

$$g(x)=w_0+\sum_{i=1}^{d}w_ix_i=\overline{a}^TY$$

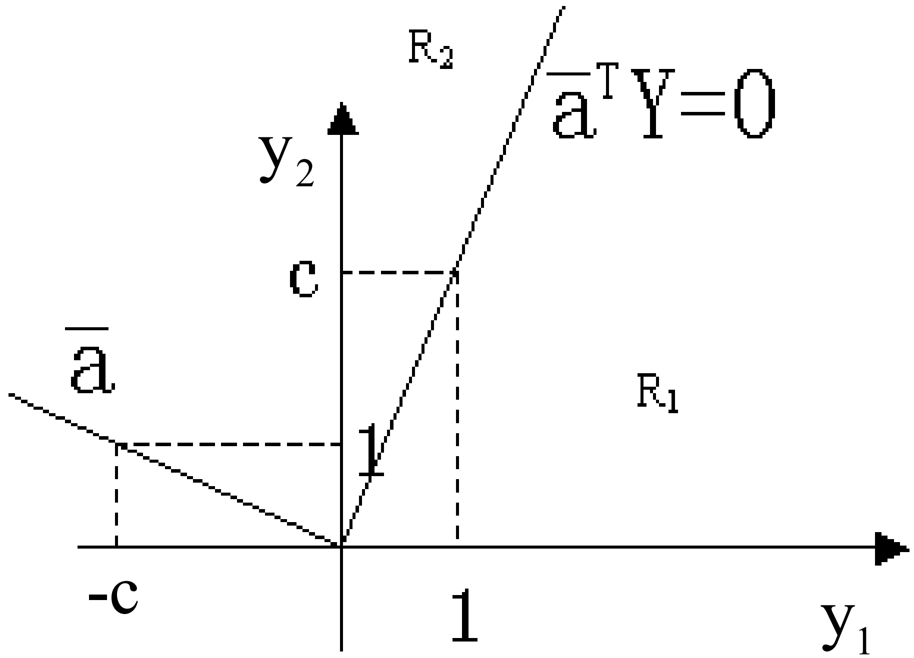
广义线性判别函数

例：

$$x - c = 0$$

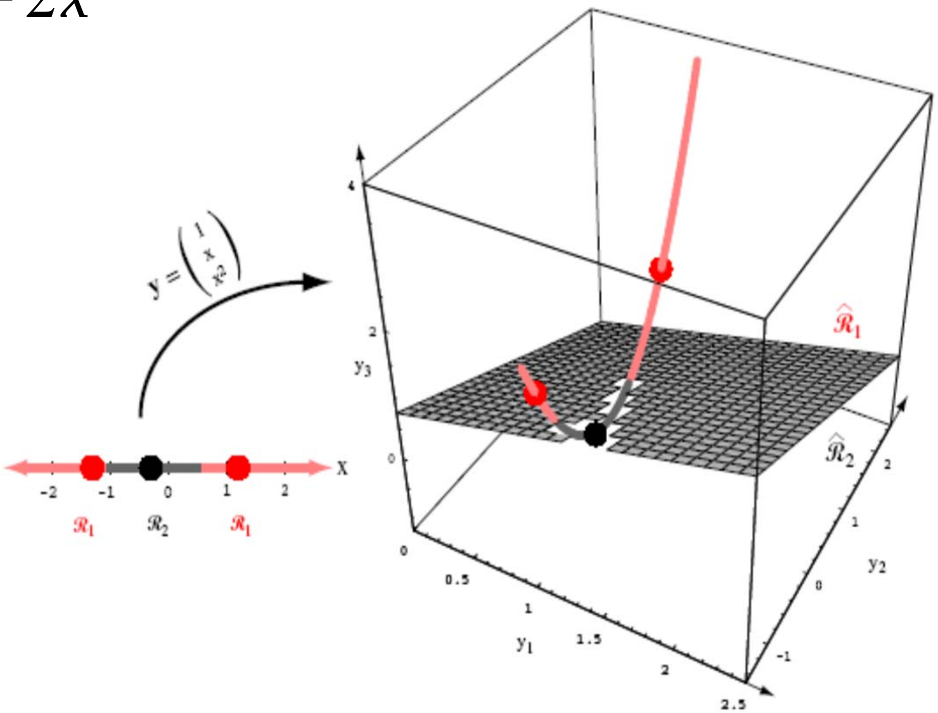


$$\begin{cases} Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ x \end{bmatrix} \\ \bar{a} = \begin{bmatrix} -c \\ 1 \end{bmatrix} \end{cases}$$



广义线性判别函数

$$g(x) = -1 + x + 2x^2$$



映射 $y=(1,x,x^2)^t$ 把一条直线映射为三维空间中的一条抛物线。由于两类问题，在三维空间中，一个平面就是一个分隔面。

第四章 线性分类器

Outline:

- 引言
- 线性判别函数基本概念
- 广义线性分类器
- 线性分类器设计步骤
- Fisher线性判别分析
- 感知器算法
- 最小平方误差判别
- 多类问题

线性分类器设计步骤

- 线性分类器设计任务是在给定样本集条件下，确定线性判别函数的各项系数，$w_0, w_1, \cdots, w_d$，以期对待测样本进行分类时，能满足相应的准则函数J为最优的要求。
  - ✓ **1、按需要确定一准则函数J**
  - ✓ **2、确定准则函数J达到极值时$w_0, w_1, \cdots, w_d$的具体数值，从而确定判别函数，完成分类器设计。**

线性分类器设计步骤

➢ 问题：

✓1. 实现这种转化的非线性变换可能非常复杂

✓2. 变换空间的维数可能非常高。——维数灾难

➢ 特例：线性判别函数的齐次简化

$$g(x) = w^T x + w_0 = \alpha^T y$$

$$y = \begin{bmatrix} 1 \\ x \end{bmatrix}, \qquad \alpha = \begin{bmatrix} w_0 \\ w \end{bmatrix}$$

增广样本向量　　　　增广权向量

Outline:

- 引言
- 线性判别函数基本概念
- 广义线性分类器
- 线性分类器设计步骤
- Fisher线性判别分析
- 感知器算法
- 最小平方误差判别
- 多类问题

Fisher 线性判别分析

➢ 问题：

把d维空间的样本投影到一条直线上，在这条直线上，样本能够较容易分开

N **samples:** $x_1, \cdots, x_N$

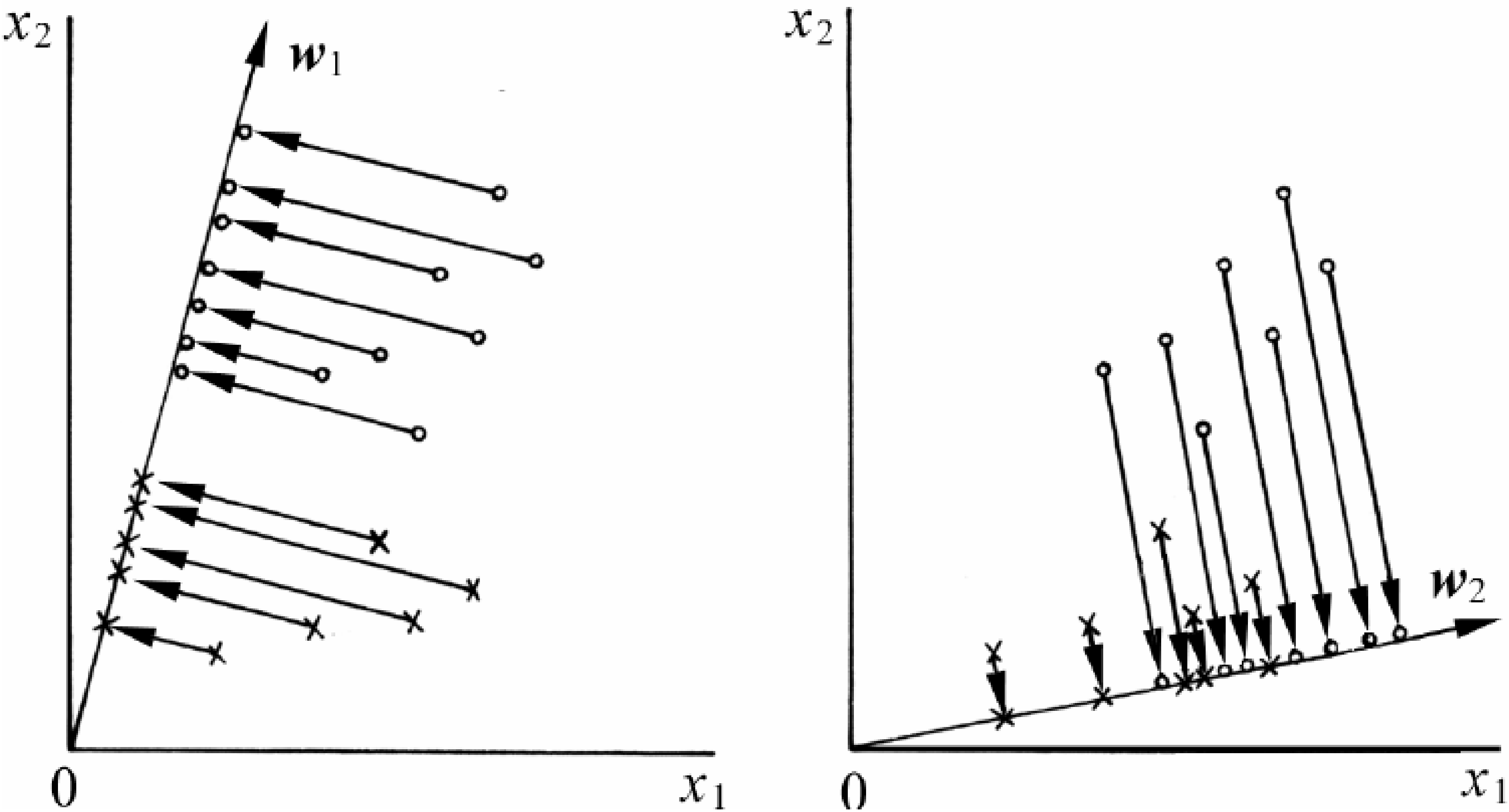
χ_1: N_1个样本构成的样本集

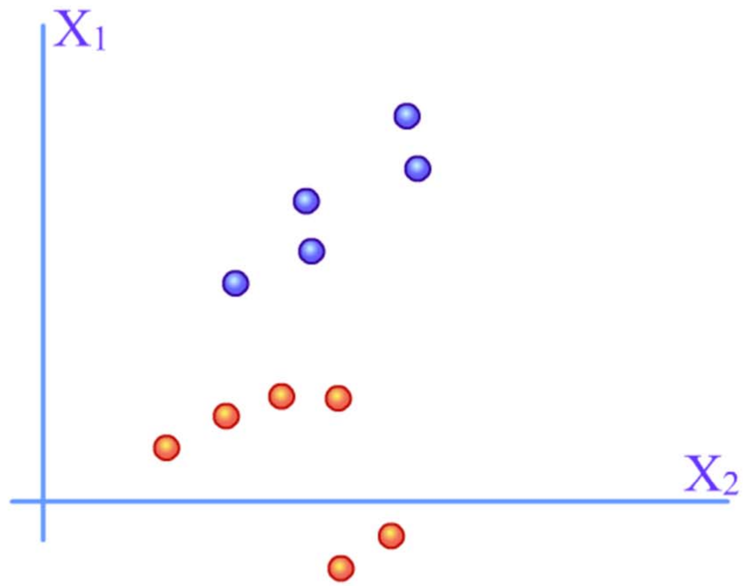
χ_2: N_2个样本构成的样本集

$N_1 + N_2 = N$

第四章 线性分类器

Fisher 线性判别分析





下一步

Fisher 线性判别分析

$$y_n = W^T x_n, \quad n = 1, 2, \cdots, N_i, i = 1, 2$$

$$m_i = \frac{1}{n} \sum_{x \in \chi_i} x \ , \quad i = 1, 2$$

Fisher 线性判别分析

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

S_i：类内离散度矩阵

$S_w = S_1 + S_2$：总的类内离散度矩阵

$$S_b = (m_1 - m_2)(m_1 - m_2)^T$$ 类间离散度矩阵

Fisher 线性判别分析

在Y空间（一维投影）：

类均值：$\tilde{m}_i = \frac{1}{N_i}\sum_{x_j \in \Upsilon_i} y_j,\ i=1,2$

类内离散度：$\tilde{S}_i^2 = \sum_{x_j \in \Upsilon_i} (y_j - \tilde{m}_i)^2,\ i=1,2$

总类内离散度：$\tilde{S}_w = \tilde{S}_1^2 + \tilde{S}_2^2$

类间离散度：$\tilde{S}_b^2 = (\tilde{m}_1 - \tilde{m}_2)^2$

Fisher准则函数(Fisher's Criterion):

$$\max \quad J_F(w) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$

即：使两类之间尽可能分开，各类内部尽可能聚集

Fisher 线性判别分析

带入$y=w^Tx$，可得

$$J_F(w)=\frac{w^T S_b w}{w^T S_w w}$$

$$w^*: \max_w J_F(w)$$

求解：令分母 $w^T S_w w = c \neq 0$ ，最大化分子。

定义Lagrange函数

$$L(w,\lambda)=w^T S_b w-\lambda(w^T S_w w-c)$$

$$\partial L/\partial w=0 \Rightarrow S_b w^* - \lambda S_w w^* = 0$$

即：$S_b w^* = \lambda S_w w^*$ w*为极值解。

Fisher 线性判别分析

第四章 线性分类器

（当N>d时，S_w通常是非奇异的），两边左乘 $\boldsymbol{S}_w^{-1}$

可得： $S_w^{-1}S_b w^* = \lambda w^*$

即w*为$S_w^{-1}S_b$ 矩阵的特征向量

将S_b带入，有： $\lambda w^* = S_w^{-1}(m_1 - m_2)(m_1 - m_2)^T w^*$

$$\overset{\Delta}{=} S_w^{-1}(m_1 - m_2)R$$

$$\boldsymbol{R} \overset{\Delta}{=} (\boldsymbol{m}_1 - \boldsymbol{m}_2)^T \boldsymbol{w}^* \text{ 为标量}$$

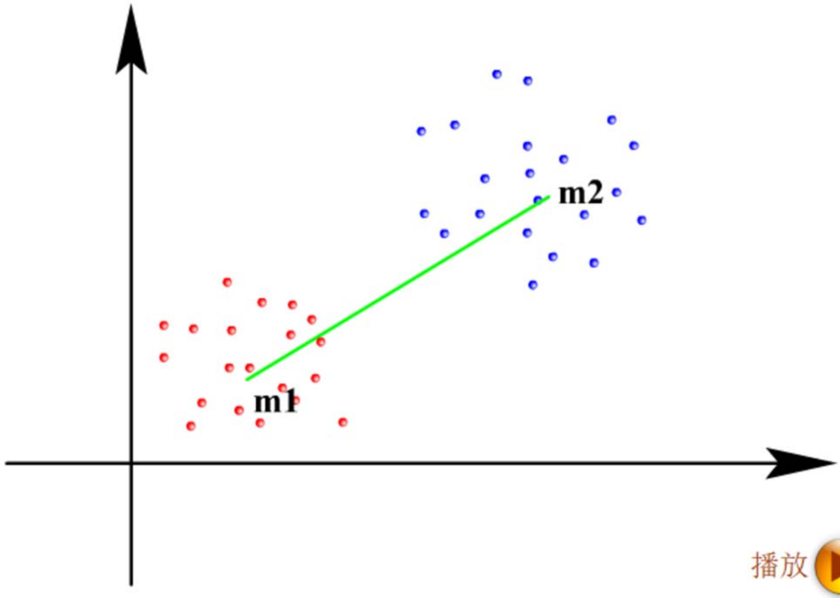
只考虑w*的方向，得 $\boldsymbol{w}^* = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$

Fisher 线性判别分析

- 由上式表示的最佳投影方向是容易理解的，因为其中一项$(m_1 - m_2)$是一向量，显然从两类均值在变换后距离最远这一点看，对与$(m_1 - m_2)$平行的向量投影可使两均值点的距离最远。但是如果同时又使类内密集程度较高这样一个指标来看，则需根据两类样本的分布离散程度对投影方向作相应的调整，这就体现在对$(m_1 - m_2)$向量按 S_W^{-1} 作一线性变换，从而使**Fisher**准则函数达到极值点。

Fisher 线性判别分析

- 下图表示了二维特征空间的两类样本分布及最佳投影方向。图中由于样本沿两类均值点连线向量分布较分散，因此由均值连线决定的向量作为线性加权向量并不能获得**Fisher**准则函数的极值，而图中向量w^*所体现的方向可预期有更好的分类效果。



播放

Fisher 线性判别分析

第四章 线性分类器

判别函数的确定（$\mathbf{w_0}$）：

$$w_0 = -\frac{\tilde{m}_1 + \tilde{m}_2}{2}$$

$$w_0 = -\frac{N_1 \tilde{m}_1 + N_2 \tilde{m}_2}{N_1 + N_2} = -\tilde{m}$$

这样，对于任意给定的未知样本**x**，计算其投影点 $W^{*T}x$

分类：

$$W^{*T}x + w_0 > 0 \rightarrow x \in \omega_1$$

$$W^{*T}x + w_0 < 0 \rightarrow x \in \omega_2$$

第四章 线性分类器

Outline:

- 引言
- 线性判别函数基本概念
- 广义线性分类器
- 线性分类器设计步骤
- Fisher线性判别分析
- 感知器算法
- 最小平方误差判别
- 多类问题

感知器算法

➢ 出发点

- ✓ 一旦判别函数的形式确定下来，不管它是线性的还是非线性的，剩下的问题就是如何确定它的系数。
- ✓ 在模式识别中，系数确定的一个主要方法就是通过对已知样本的训练和学习来得到。
- ✓ 感知器算法就是通过训练样本模式的迭代和学习，产生线性（或广义线性）可分的模式判别函数。

➢ 基本思想

- ✓ 采用感知器算法(**Perception Approach**)能通过对训练模式样本集的“学习”得到判别函数的系数。

➢ 背景

- ✓ “感知器”一词出自于**20**世纪**50**年代中期到**60**年代中期人们对一种分类学习机模型的称呼，它是属于有关动物和机器学习的仿生学领域中的问题。感知器的一些相关概念仍然沿用下来。

基本概念

✓感知器：**Perceptron，Rosenblatt**

✓线性可分性：训练样本集中的两类样本在特征空间可以用一个线性分界面正确无误地分开。在线性可分条件下，对合适的(广义)权向量a应有：

$$\text{如果 } \mathbf{y} \in \omega_1, \text{则} \mathbf{a}^T\mathbf{y} > 0$$

$$\text{如果 } \mathbf{y} \in \omega_2, \text{则} \mathbf{a}^T\mathbf{y} < 0$$

使用增广样本 向量和增广权向量，即用 **d+1**维向量表示样本的齐次化向量

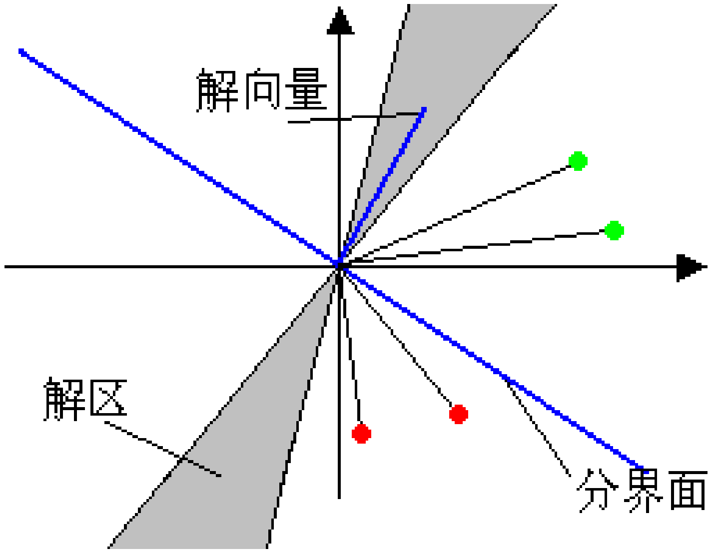
✓规范化样本向量：将第二类样本取其反向向量

- a叫做分离向量，或解向量

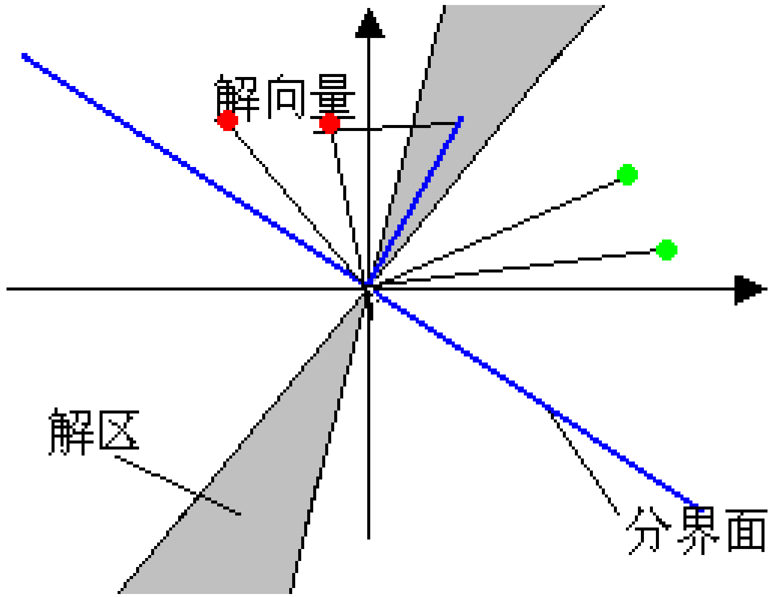
$$\mathbf{y}' = \begin{cases} \mathbf{y} & \text{如果 } \mathbf{y} \in \omega_1 \\ -\mathbf{y} & \text{如果 } \mathbf{y} \in \omega_2 \end{cases} \Rightarrow \mathbf{a}^T\mathbf{y}'_i > 0 \quad i = 1, \ldots, N$$

感知器

解向量与解区



a: 第一类样本
(a) 未规范化



b: 第二类样本
(b) 规范化

感知器准则函数

- 对于一个合适的增广权向量a，
  - ✓ 对样本y(规范化增广样本向量)正确分类，则有：**$a^T y>0$**
  - ✓ 对样本y错误分类，则有：**$a^T y<0$**
- 准则函数

$$J_p(a)=\sum_{y\in Y^k}(-a^t y)$$

被错分类的规范化增广样本集

- ✓ **Y^k**表示被分离向量**a**错分的样本集合。
- ✓ 如果没有样本被错分，那么**Y(a)**为空，**$J_P(a)=0$;**
- ✓ 对于错分样本**$a^T y_i<0$**；所以**$J_P(a)\geq 0$**
- ✓ 当且仅当所有样本都被正确分类时，a为解向量，**Y^k**为空集（不存在错分样本）时，**$J_P(a)=0$**。此时 **$J_P(a)=0$**，即达到极小值。确定向量a的问题变为对**$J_P(a)$**求极小值的问题。

第四章 线性分类器

感知器准则函数

第四章 线性分类器

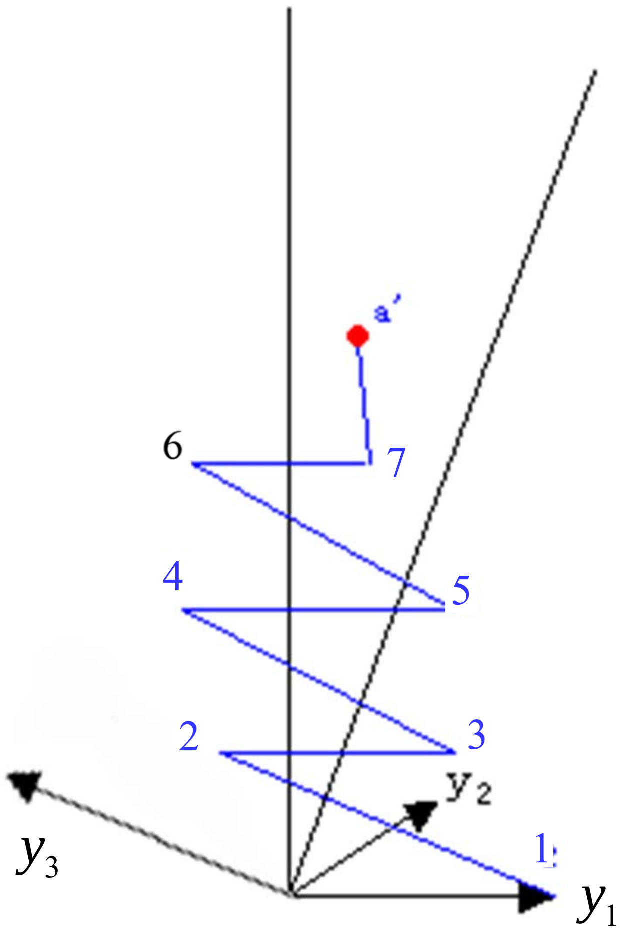
- 准则函数 $J_p(a)=\sum_{y\in Y^k}(-a^t y)$
  - ✓ 恒有$J_P(a)\geq 0$，且仅当a为解向量，Y^k为空集（不存在错分样本）时， $J_P(a)=0$，即达到极小值。确定向量a的问题变为对$J_P(a)$求极小值的问题。
  - ✓ 梯度下降算法：对(迭代)向量沿某函数的负梯度方向修正，可较快到达该函数极小值。
- 迭代算法寻找解向量a
  - ✓ 选择任意初始向量$a^{(1)}$
  - ✓ 选择任意初始向量$a^{(1)}$，计算$J_P(a)$对a的梯度 $\nabla J(a)=\sum_{y\in Y^k}(-y)$

    可见感知准则函数的梯度向量是所有被错分类的规范化增广样本向量之和，如果将迭代公式写成：

    $$a^{(k+1)}=a^{(k)}-\rho_k\nabla J(a^{(k)}),\rho_k \text{ ：步长系数} \qquad a^{(k+1)}=a^{(k)}+\rho_k\sum_{y\in Y^k} y,\rho_k \text{ ：步长系数}$$
  - ✓ 终止: 对所有样本正确分类

梯度下降算法图解

$$\mathbf{a}^{(k+1)} \leftarrow \mathbf{a}^{(k)} + \rho_k \sum_{y \in Y_k} \mathbf{y}$$



$a(1) = y_1 \qquad \rho_k = 1$

错分样本：

y_3
y_1
y_3
y_1

梯度下降算法寻找解向量

Begin

initialize

$$\mathbf{a}, \eta(\bullet), \text{threshold} \quad \theta, k=0$$

do

$$k \leftarrow k+1$$

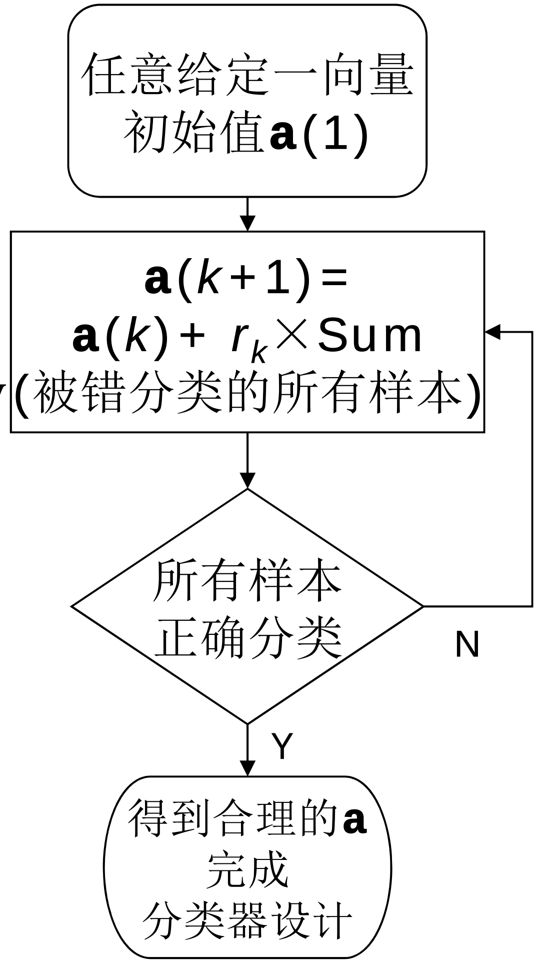
$$\mathbf{a} \leftarrow \mathbf{a} - \eta(k)\nabla J(\mathbf{a}) = \mathbf{a} + \eta(k)\sum_{y \in Y_k} \mathbf{y}$$

until

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

return a

End



- **How to set the learning rate η(k) ?**

$$\eta(k+1) = 0.9\eta(k), \eta(1) = 1$$

感知器算法是一种赏罚过程

- 感知器算法实质上是一种赏罚过程
  - ✓对正确分类的模式则“赏”，实际上是“不罚”，即权向量不变。
  - ✓对错误分类的模式则“罚”，使**a(k)**加上一个分量。
  - ✓当用全部模式样本训练过一轮以后，只要有一个模式是判别错误的，则需要进行下一轮迭代，即用全部模式样本再训练一次。
  - ✓如此不断反复直到全部模式样本进行训练都能得到正确的分类结果为止。

第四章 线性分类器

- **Example:**

Training Dataset:

$$\omega_1: \ x_1 = (3,3)^T, x_2 = (4,3)^T$$

$$\omega_2: \ x_3 = (1,1)^T$$

增本向量：

$$\omega_1: \ x_1 = (1,3,3)^T, x_2 = (1,4,3)^T$$

$$\omega_2: \ x_3 = (1,1,1)^T$$

Normalization:

$$\omega_1: \ x_1 = (1,3,3)^T, x_2 = (1,4,3)^T$$

$$\omega_2: \ x_3 = (-1,-1,-1)^T$$

➢ Example:

$$\textbf{Assume: } \mathbf{w}_1 = \mathbf{x}_1 = \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} \to \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix} \to \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \to \begin{pmatrix} -2 \\ 0 \\ 0 \end{pmatrix} \to \cdots \to \begin{pmatrix} -5 \\ 2 \\ 1 \end{pmatrix}$$

Outline:

- 引言
- 线性判别函数基本概念
- 广义线性分类器
- 线性分类器设计步骤
- Fisher线性判别分析
- 感知器算法
- 最小平方误差判别
- 多类问题

最小平方误差准则

- 规范化增广样本向量y_i，增广权向量a，正确分类要求： $a^T y_i > 0,\ i=1,\dots,N$
- 线性分类器设计⇔求一组N个线性不等式的解
- 样本集增广矩阵Y及一组N个线性不等式的的矩阵表示：

$$Y = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_1^T \\ \cdots \\ \mathbf{y}_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1\hat{d}} \\ y_{21} & y_{22} & \cdots & y_{2\hat{d}} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{N\hat{d}} \end{bmatrix}, \boldsymbol{Ya} > 0$$

- 引入余量(目标向量) $b=[b_1, b_2, \dots, b_N]^T$,
  - ✓ b_i任意给定正常数， $a^T y_i = b_i > 0$
- N个线性方程的的矩阵表示：$Y\mathbf{a} = \mathbf{b}$

最小平方误差准则

平方误差准则函数

- 定义误差向量 e=**Y**a-b：
- 定义平方误差准则函数J_s(a):

$$J_s(\mathbf{a}) = \|\mathbf{e}\|^2 = \|Y\mathbf{a} - \mathbf{b}\|^2 = \sum_{i=1}^{N} (\mathbf{a}^T \mathbf{y}_i - b_i)^2$$

- 最小二乘近似解（**MSE**解）：

$$\mathbf{a}^* = \underset{\mathbf{a}}{\arg\min}\, J_s(\mathbf{a})$$

MSE方法的思想：对每个样本，设定一个“理想”的判别函数输出值，以最小平方误差为准则求最优权向量

第四章 线性分类器

MSE准则函数的伪逆解

$$J_s(a) = \|e\|^2 = \|Ya - b\|^2 = \sum_{i=1}^{N} (a^T y_i - b_i)^2$$

$$\nabla J_s(a) = \sum_{i=1}^{N} 2(a^T y_i - b_i) y_i = 2Y^T (Ya - b)$$

$$\nabla J_s(a^*) = 0 \Leftrightarrow Y^T Y a^* = Y^T b$$

$$a^* = (Y^T Y)^{-1} Y^T b = Y^+ b$$

$$Y^+ Y = I, \quad YY^+ \neq I$$

Y^TY通常是非奇异的，存在逆矩阵。如果Y非奇异则$a=Y^{-1}b$

Y的伪逆矩阵

第四章 线性分类器

MSE方法的迭代解

- $a^* = Y^+ b$, $Y^+ = (Y^T Y)^{-1} Y^T$
  - ✓计算量大
  - ✓如果$(Y^T Y)$是奇异无法计算
- 实际中常用梯度下降法：

$$\nabla J_s(\mathbf{a}) = \sum_{i=1}^{N} 2(\mathbf{a}^T \mathbf{y}_i - b_i)\mathbf{y}_i = 2Y^T(Y\mathbf{a} - \mathbf{b})$$

$$\begin{cases} \mathbf{a}(1), \text{任意初始化} \\ \mathbf{a}(k+1) = \mathbf{a}(k) - r_k Y^T (Y\mathbf{a} - \mathbf{b}) \end{cases}$$

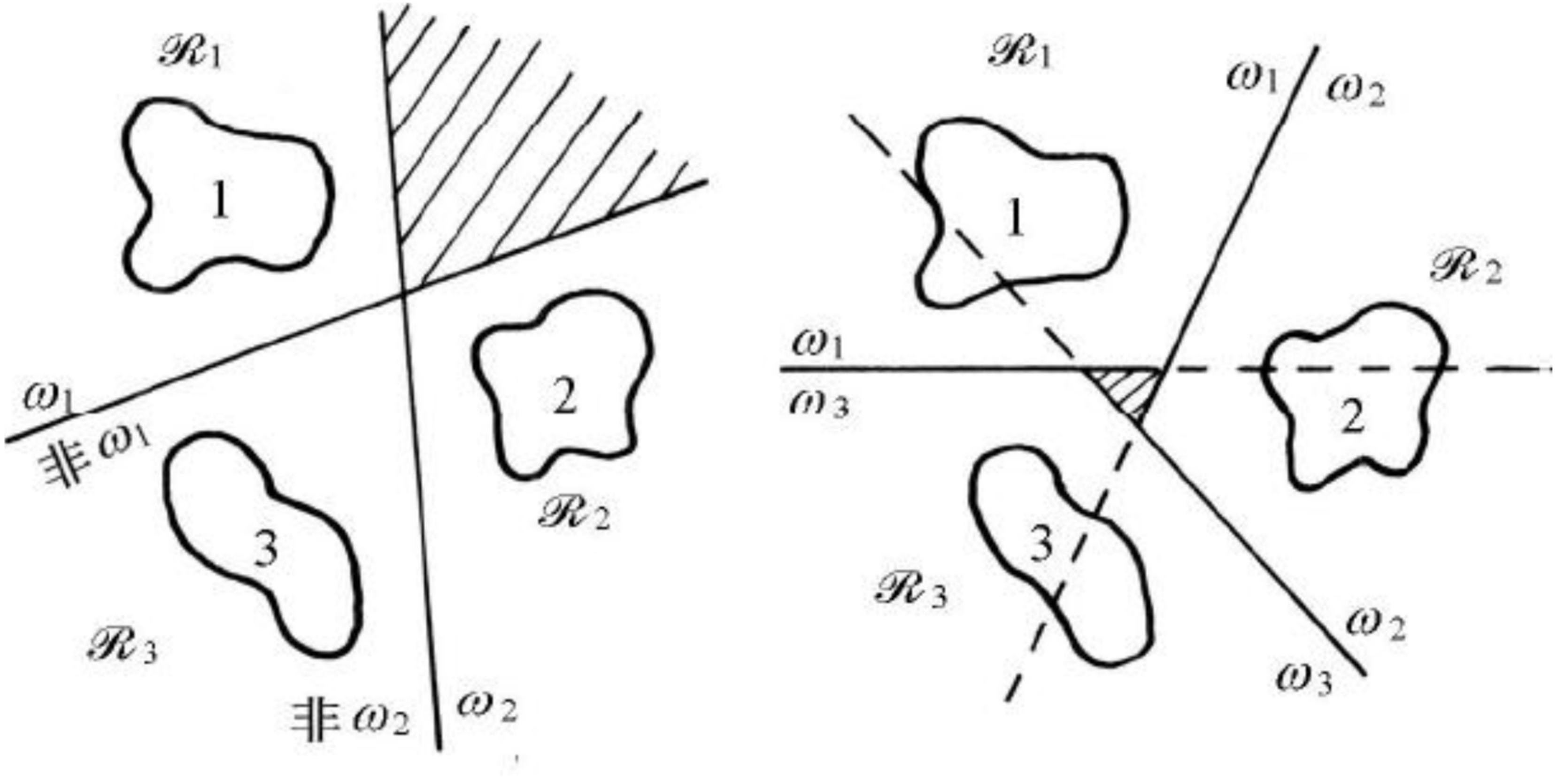
Outline:

- 引言
- 线性判别函数基本概念
- 广义线性分类器
- 线性分类器设计步骤
- Fisher线性判别分析
- 感知器算法
- 最小平方误差判别
- 多类问题

多类问题 Multicategory Problems

- 基本考虑
  - ✓思路一：转化为多个两类问题
    - A. c 类转化为c－1 个两类问题。ω_i与非ω_i
    - B. 每类转化为c－1 个两类问题，ω_i与 ω_j，　$j \neq i$
      c 类则c(c −1) 个两类问题，但其中半数相同，
      故c 类转化为c(c −1)/2个两类问题

多类问题



(a) ω_i/ 非 ω_i (b) ω_i/ ω_j

第四章 线性分类器

多类问题

思路二：线性机器(Linear Machine)

c 个线性判别函数

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

$$if \;\; g_i(x) > g_j(x), \quad \forall_j \neq i \Rightarrow x \in \omega_i$$

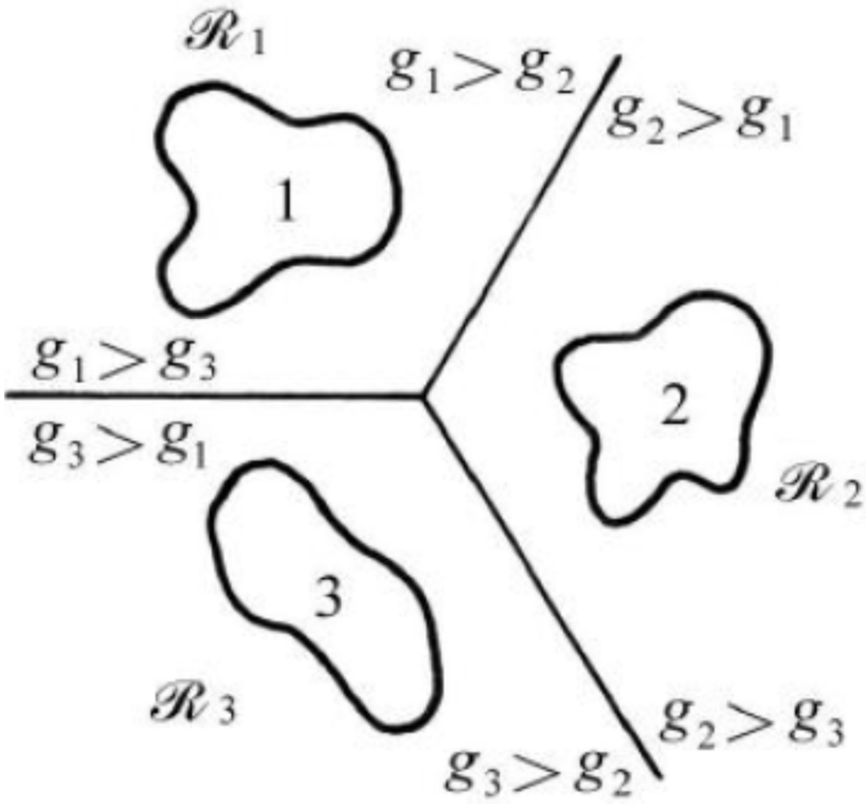
$$if \;\; g_i(x) = g_j(x), \quad j \neq i \Rightarrow refuse$$

增广权向量 $\alpha_i = \begin{bmatrix} w_i \\ w_{i0} \end{bmatrix}$, 线性机器可表示为：

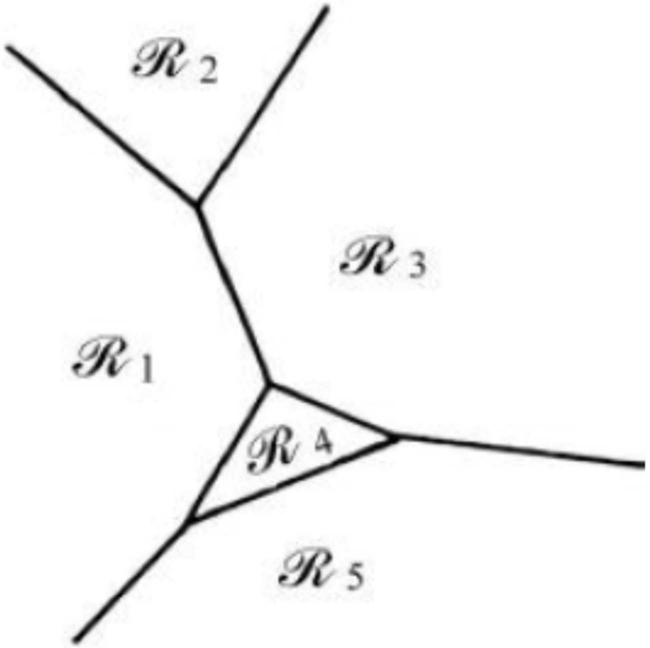
$$L(\alpha_1, \alpha_2, \cdots, \alpha_c)$$

多类问题

第四章 线性分类器



(a) 三类



(b) 五类