第五章 线性学习机器与线性分类器

苏智勇
可视计算研究组
南京理工大学
suzhiyong@njust.edu.cn
https://zhiyongsu.github.io

主要内容

5.1 引言

5.2 线性回归

5.3 线性判别函数的基本概念

5.4 Fisher线性判别分析

5.5 感知器

5.6 最小平方误差判别

5.7 罗杰斯特回归

5.8 最优分类超平面与线性支持向量机

5.9 多类线性分类器

5.1 引言

- **模式分类的三种主要途径**
  - **估计类条件概率密度** $P(x \mid \omega_i)$
    - 通过$P(\omega_i)$和$P(x \mid \omega_i)$，利用贝叶斯规则计算后验概率$P(\omega_i \mid x)$，然后通过最大后验概率做出决策。
    - 概率密度参数估计和非参数估计。
  - **直接估计后验概率** $P(\omega_i \mid x)$

    不需要先估计类条件概率密度$P(x \mid \omega_i)$。主要有K近邻方法等。
  - **直接计算判别函数**
    - 不需要估计$P(x \mid \omega_i)$或者$P(\omega_i \mid x)$。
    - 直接找到可用于分类的判别函数。常见的方法有神经网络等。

5.1 引言

- **基于概率密度（估计）的分类器设计：model-based method**
  - **已知**：类先验概率$P(\omega_i)$和类条件概率密度函数$P(x \mid \omega_i)$
  - **任务**：估计一个决策函数，借此进行分类
  - **方法**：参数估计、非参数估计
  - **特点**：需要大量的样本，需要知道某些概率及其形式

5.1 引言

- **基于样本的直接分类器设计：data-driven method**
  - **思路**：
    - 如果知道判别函数的形式，则可以直接从样本估计函数参数
    - 先选定判别函数类和一定的目标（准则），利用样本集确定函数类中的未定参数，使所选的准则最好
  - **形式化**：
    - 判别函数类：$\{g(\alpha), \alpha \in \Lambda\}$，$\alpha$为未定参数
    - 准则函数/目标函数：$L(\alpha)$
    - 求α^*：$L(\alpha^*) = \min_{\alpha} L(\alpha)$

5.1 引言

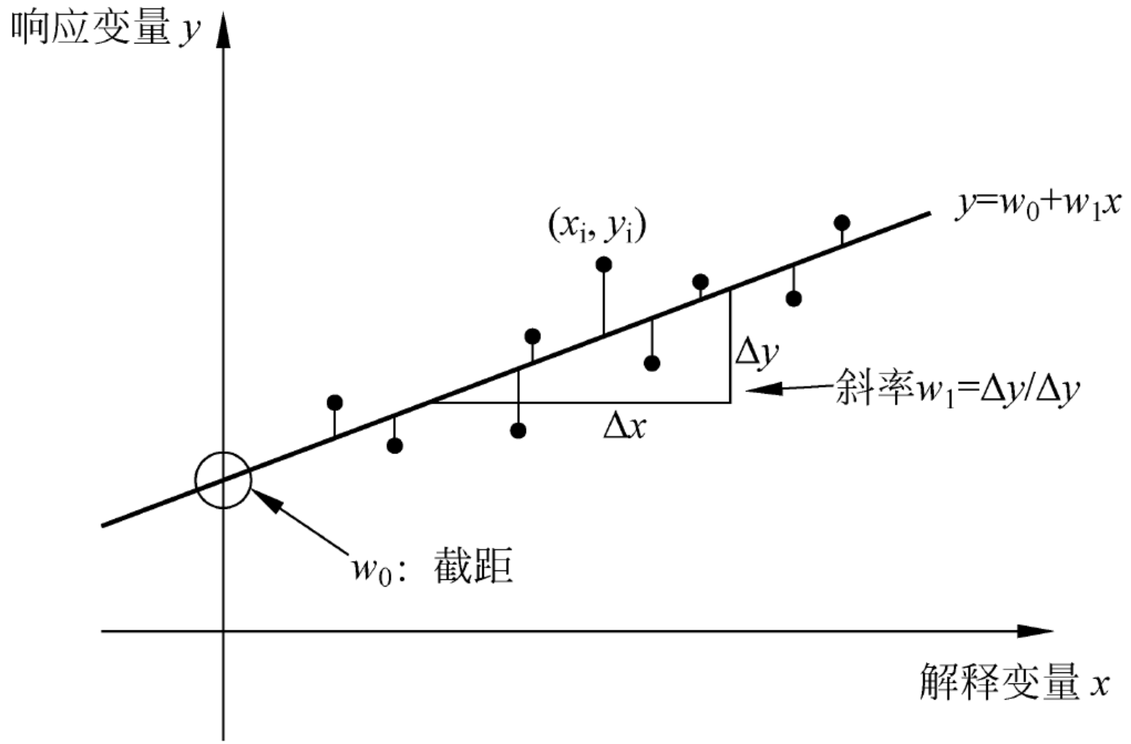
- **本章利用样本直接设计分类器的基本思想**
  - 给定一个判别函数，且已知该函数的参数形式
  - 采用样本来训练判别函数的参数
  - 对于新样本，采用判别函数对其进行觉得，并按照一些准则来完成分类
- **本章利用样本直接设计分类器的基本技术路线**
  - 假定有 n 个 d 维空间中的样本，每个样本的类别标签 已知，且一共有 c 个不同的类别。
  - 假定判别函数的形式已知，寻找一个判别函数。
  - 对于给定的新样本$\boldsymbol{x} \in \mathrm{R}^d$，判定它属于$\omega_1, \omega_2, \cdots, \omega_c$中的哪个类别。

5.2 线性回归

- **线性回归**
  - 通过数据发现或估计两个或多个变量之间可能存在的线性依赖关系的基本统计学方法
- **简单线性回归**
  - 解释变量 x、响应变量 y
  - 通过(x, y)的一系列观测样本，

    估计线性关系$y = w_0 + w_1 x$，

    即估计其中的系数w_0和w_1



5.2 线性回归

- **多元线性回归**

  _ 响应变量依赖于多个解释变量：$y = w_0 + w_1 x + \ldots + w_d x_d = \sum_{i=0}^{d} w_i x_i = \boldsymbol{w}^T x$

- **线性回归问题描述**

  _ 已知训练样本集$\left\{ \left(x_1, y_1\right), \ldots, \left(x_N, y_N\right) \right\}$，$x_j \in R^{d+1}$，$y_j \in R$；机器学习模型为

  $f(\boldsymbol{x}) = w_0 + w_1 x + \ldots + w_d x_d = \sum_{i=0}^{d} w_i x_i = \boldsymbol{w}^T \boldsymbol{x}$，其中，

  $\boldsymbol{w} = [w_0, w_1, ..., w_d]^T$ 是模型中的待定参数

5.2 线性回归

- **“最小二乘法” 求解线性回归问题**
  - 用训练样本集估计模型中的参数，使模型在**最小平方误差**意义下能够最好地拟合训练样本，即 $\min E = \frac{1}{N}\sum_{j=1}^{N}(f(x_j) - y_j)^2$
  - 目标函数可进一步写成矩阵形式 $E(\boldsymbol{w}) = \frac{1}{N} \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right\|^2 = \frac{1}{N}\left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right)^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$，其中，$\boldsymbol{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}$ 是全部训练样本的解释变量向量组成的矩阵，$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ 是全部训练样本的响应变量组成的向量

5.2 线性回归

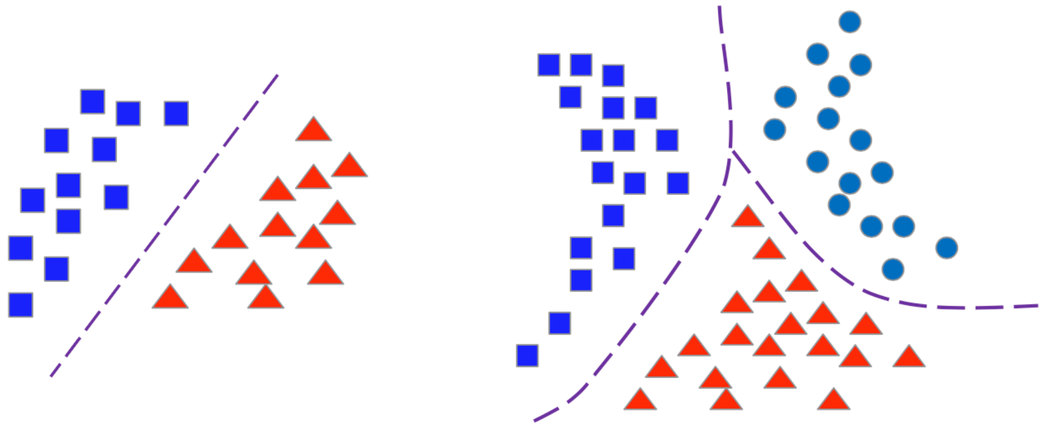
- **“最小二乘法” 求解线性回归问题**
  - 使目标函数$E(\boldsymbol{w})$最小化的参数$\boldsymbol{w}$应满足$\frac{\partial E(\boldsymbol{w})}{\partial(\boldsymbol{w})}=\frac{2}{N}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y})=0,$

    即 $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}=\boldsymbol{X}^T\boldsymbol{y}$
  - 当矩阵$(\boldsymbol{X}^T\boldsymbol{X})$可逆时，最优参数的解为$\boldsymbol{w}^*=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y},$

    $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$被称为$\boldsymbol{X}$的伪逆（pseudo-inverse）矩阵，记作$\boldsymbol{X}^+$
  - 线性回归给出了在最小平方误差意义下对解释变量与响应变量间线性关系的最好估计

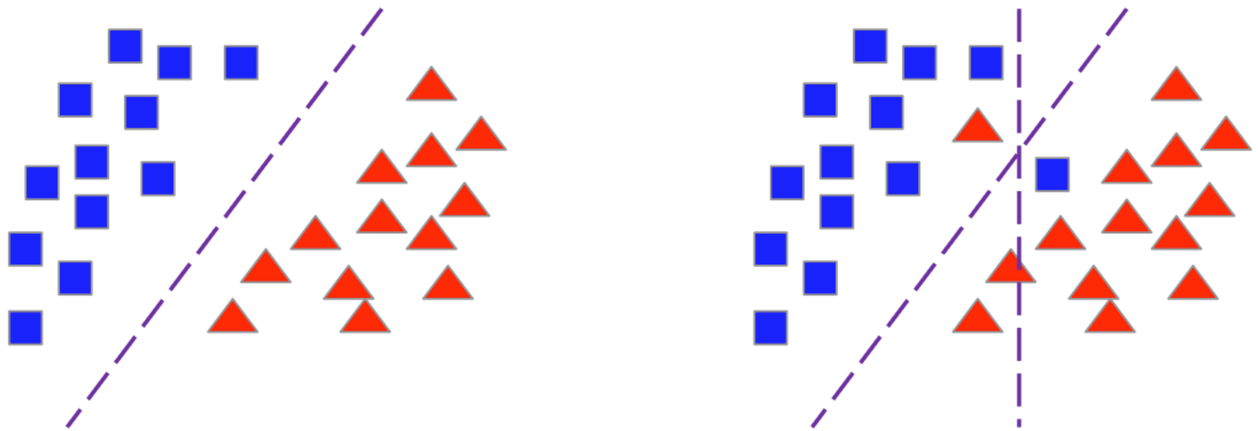
5.3 线性判别函数的基本概念

- **基于判别函数的分类器**
  - 采用已知类别标签的训练样本进行学习，获得若干个代数**界面**，这些界面将样本所在的空间分成若干个**相互不重叠的区域**。每个区域包含属于同一类的样本。
  - 表示界面的函数称为判别函数。
  - 判别函数是分类器最常用的表述形式。



5.3 线性判别函数的基本概念

- **线性可分**
  - 对于 n 个 d 维空间中的样本$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$，假定这些样本来着两个类别$\omega_1$和$\omega_2$。**如果存在一个线性判别函数能对这些样本正确地分类，则称这些样本是线性可分的**；否则是线性不可分的。



线性可分　　线性不可分

5.3 线性判别函数的基本概念

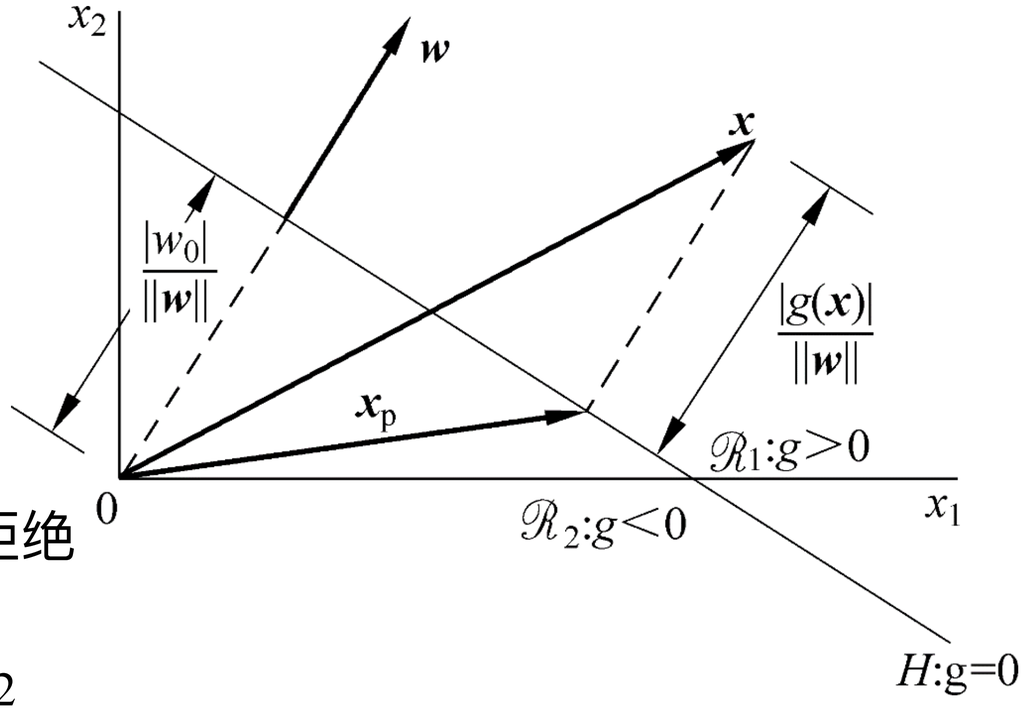
- **一般表达式**
  - $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$
    - $\boldsymbol{x}$：d维特征向量（样本向量），$\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T$
    - $\boldsymbol{w}$：权向量，$\boldsymbol{w} = [w_1, w_1, \cdots, w_d]^T$
  - 对于c类分类问题，$g_i(\boldsymbol{x})$，$i = 1,2,\cdots,c$，表示每个类别对应的判别函数
    - 决策规则：如果$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})$，$\forall j \neq i$，则$\boldsymbol{x} \in \omega_i$

5.3 线性判别函数的基本概念

- **决策规则**

  – 以二分类问题为例，令$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$

$$\begin{cases} \text{如果 } g(\boldsymbol{x}) > 0, & \text{则决策 } \boldsymbol{x} \in \omega_1 \\ \text{如果 } g(\boldsymbol{x}) < 0, & \text{则决策 } \boldsymbol{x} \in \omega_2 \\ \text{如果 } g(\boldsymbol{x}) = 0, & \text{可将 } \boldsymbol{x} \text{ 任意分到某一类，或拒绝} \end{cases}$$



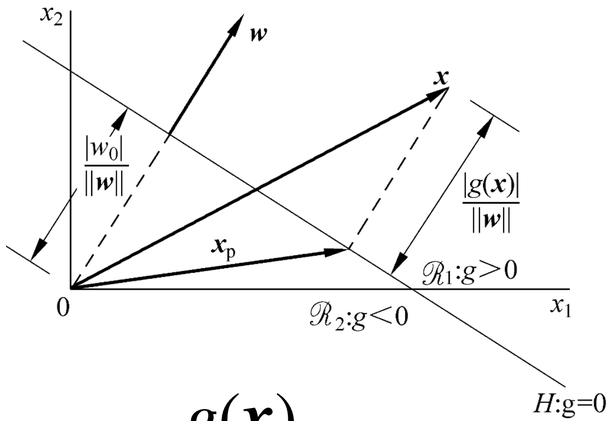
✓决策面/超平面H：$g(\boldsymbol{x}) = 0$，决策区域$\mathscr{R}_1$和$\mathscr{R}_2$

✓$\boldsymbol{w}$为超平面H的法向量

- 设$\boldsymbol{x}_1$，$\boldsymbol{x}_2$都在H上，则$\boldsymbol{w}^T\boldsymbol{x}_1 + w_0 = \boldsymbol{w}^T\boldsymbol{x}_2 + w_0$，即$\boldsymbol{w}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$
- $\boldsymbol{w}^T$跟决策面H上的任一向量正交，即$\boldsymbol{w}$为超平面H的法向量

5.3 线性判别函数的基本概念

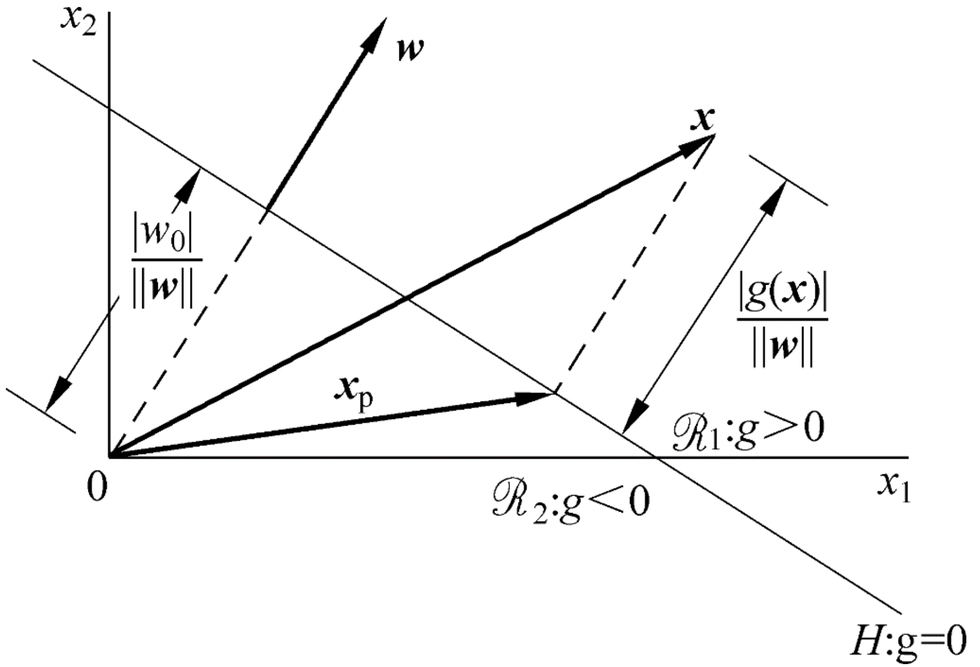
- **代数度量**
  - 判别函数$g(\boldsymbol{x})$正比于$\boldsymbol{x}$点到超平面H的代数距离（带正负号），$\boldsymbol{x}$在H正侧：$g(\boldsymbol{x})>0$；在负侧：$g(\boldsymbol{x})<0$
  - $g(\boldsymbol{x})$可以视为特征空间中某点$\boldsymbol{x}$到超平面H的距离的一种代数度量
    - $\boldsymbol{x}=\boldsymbol{x}_p+r\dfrac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$，$\boldsymbol{x}_p$为$\boldsymbol{x}$在$H$上的投影向量，$\dfrac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$为单位向量



    - $g(\boldsymbol{x})=\boldsymbol{w}^{\top}\left(\boldsymbol{x}_p+r\dfrac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)+w_0=\boldsymbol{w}^{\top}\boldsymbol{x}_p+w_0+r\dfrac{\boldsymbol{w}^{\top}\boldsymbol{w}}{\|\boldsymbol{w}\|}=r\|\boldsymbol{w}\|$ 或$r=\dfrac{g(\boldsymbol{x})}{\|\mathbf{w}\|}$
    - 若$\boldsymbol{x}$为原点，则$g(\boldsymbol{x})=w_0$，$r_0=\dfrac{w_0}{\|\boldsymbol{w}\|}$

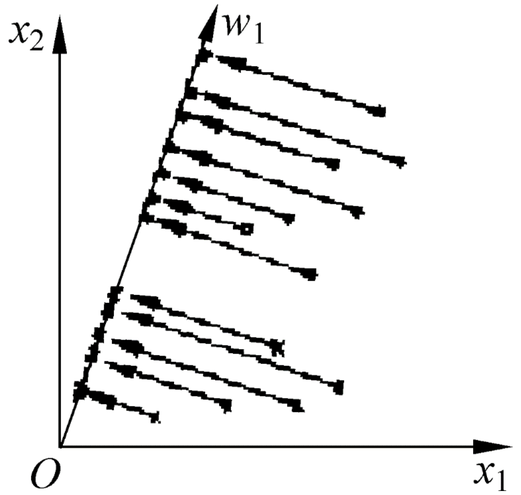
5.3 线性判别函数的基本概念

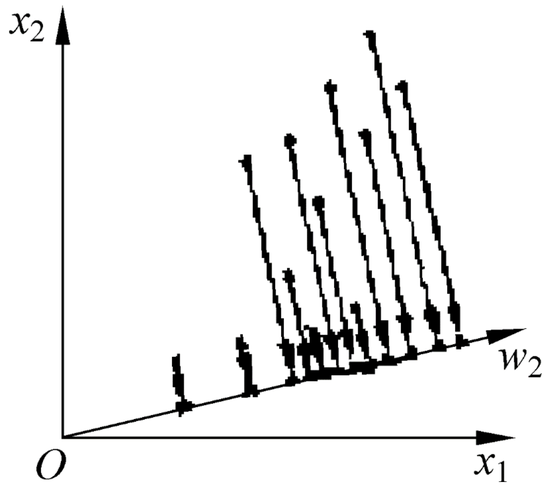
- **小结**
  - 用线性判别函数进行决策，就是用一个超平面H把特征空间分割成决策区域$\mathscr{R}_1$和$\mathscr{R}_2$
  - 超平面H的方向由权向量$\boldsymbol{w}$确定，它的位置由阈值权w_0确定
  - 判别函数$g(\boldsymbol{x})$正比于$\boldsymbol{x}$点到超平面的代数距离（带正负号）。当x在H正侧时，$g(\boldsymbol{x})>0$；在负侧时，$g(\boldsymbol{x})<0$。



5.4 Fisher线性判别分析

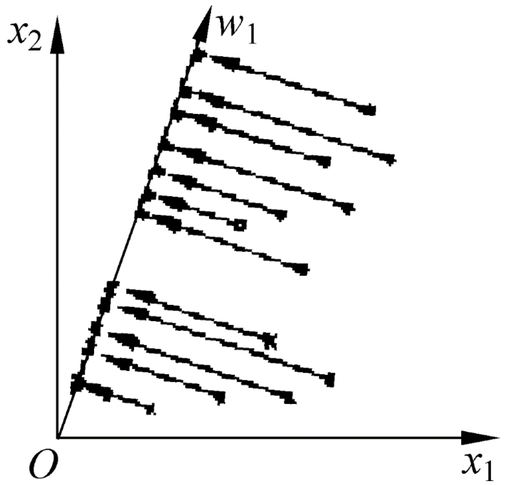
- **思路**
  - Fisher在1936年提出线下判别分析法(Linear Discriminant Analysis，LDA)
  - 把所有样本都投影到一个方向上，在**一维**空间中确定一个分类的阈值
  - **核心**：确定投影方向，使得**类间距离尽可能大，类内距离尽可能小**

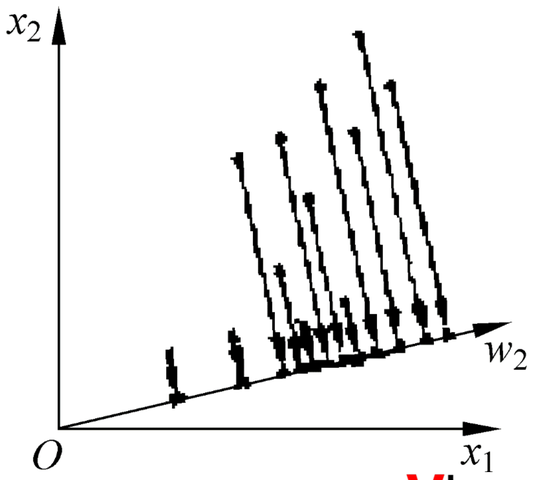




5.4 Fisher线性判别分析

- **基本定义**
  - 以两分类问题为例，训练样本集$X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$，每个样本为$d$维向量
  - ω_1类的样本$X_1 = \left\{\boldsymbol{x}_1^1, \ldots, \boldsymbol{x}_{N_1}^1\right\}$，$\omega_2$类的样本$X_2 = \left\{\boldsymbol{x}_1^2, \ldots, \boldsymbol{x}_{N_2}^2\right\}$
  - 寻找投影方向$\boldsymbol{w}$，使得投影后的样本为 $y_i = \boldsymbol{w}^T\boldsymbol{x}_i, i = 1,2,\ldots,N$





5.4 Fisher线性判别分析

- **投影前**
  - 类内离散度矩阵（within-class scatter matrix）

$$\boldsymbol{S}_i = \sum_{\boldsymbol{x}_j \in X_i} \left(\boldsymbol{x}_j - \boldsymbol{m}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{m}_i\right)^T, \quad i = 1,2,$$ 其中，类均值向量为 $$\boldsymbol{m}_i = \frac{1}{N_i}\sum_{\boldsymbol{x}_j \epsilon X_i} \boldsymbol{x}_j, \quad i = 1,2$$

  - 总类内离散度矩阵（pooled within-class scatter matrix）

$$\boldsymbol{S}_w = \boldsymbol{S}_1 + \boldsymbol{S}_2$$

  - 类间离散度矩阵（between-class scatter matrix）

$$\boldsymbol{S}_b = \left(\boldsymbol{m}_1 - \boldsymbol{m}_2\right)\left(\boldsymbol{m}_1 - \boldsymbol{m}_2\right)^T, \quad i = 1,2$$

5.4 Fisher线性判别分析

- **投影后一维空间**
  - 类内离散度

$$\tilde{S}_i^2 = \sum_{y_j \in Y_i} \left(y_j - \tilde{m}_i \right)^2, \quad \text{其中，两类均值为} \tilde{m}_i = \frac{1}{N_i} \sum_{y_j \epsilon Y_i} y_j = \frac{1}{N_i} \sum_{x_j \epsilon X_i} w^T x_j = w^T m_i, \quad i = 1,2$$

  - 总类内离散度矩阵

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

  - 类间离散度

$$\tilde{S}_b = (\tilde{m}_1 - \tilde{m}_2)^2, \quad i = 1,2$$

5.4 Fisher线性判别分析

- **Fisher准则函数（Fisher's Criterion）**

  - $\max J_F(\boldsymbol{w}) = \dfrac{\widetilde{S}_b}{\widetilde{S}_w} = \dfrac{(\widetilde{m}_1 - \widetilde{m}_2)^2}{\widetilde{S}_1^2 + \widetilde{S}_2^2}$，即，使投影后两类尽可能分开，而各类内部尽可能聚集
    - $\widetilde{S}_b = (\widetilde{m}_1 - \widetilde{m}_2)^2 = (\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_2)^2 = \boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}$
    - $\widetilde{S}_w = \widetilde{S}_1^2 + \widetilde{S}_2^2 = \boldsymbol{w}^T\boldsymbol{S}_1\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{S}_2\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}$
  - $\max J_F(\boldsymbol{w}) = \dfrac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$：广义的瑞利商（generalized Rayleigh quotient）

5.4 Fisher线性判别分析

- **求解**
  - 上述 Fisher 准则函数转化为：$\max \ \boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w} \quad \mathrm{s.t.} \ \ \boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w} = c \neq 0$
  - 定义拉格朗日函数 $L(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w} - \lambda\left(\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w} - c\right)$
  - 令$\dfrac{\partial L(\boldsymbol{w}, \lambda)}{\partial \boldsymbol{w}} = 0$，极值解$\boldsymbol{w}^*$应满足：$\boldsymbol{S}_b\boldsymbol{w}^* - \lambda\boldsymbol{S}_w\boldsymbol{w}^* = 0$
  - 假设$\boldsymbol{S}_w$是非奇异的（$N > d$时通常非奇异），可得$\mathrm{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{w}^* = \lambda\boldsymbol{w}^*$，即$\boldsymbol{w}^*$为$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的本征向量
  - 将$\boldsymbol{S}_b$带入可得 $\lambda\boldsymbol{w}^* = \boldsymbol{S}_w^{-1}\left(\boldsymbol{m}_1 - \boldsymbol{m}_2\right)\left(\boldsymbol{m}_1 - \boldsymbol{m}_2\right)^T\boldsymbol{w}^*$，只考虑$\boldsymbol{w}^*$的方向，得最优投影方向：$\boldsymbol{w}^* = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$：只需要求出原始样本的均值和方差！

5.4 Fisher线性判别分析

- **决策规则**
  - Fisher判别函数的最优解只给出了一个投影方向，不是分类面
  - 确定分类阈值w_0，得到分类面和决策规则

$$g(x) = \boldsymbol{w}^T\boldsymbol{x} + w_0 \lessgtr 0，\ 则x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

5.4 Fisher线性判别分析

- **确定分类阈值w_0**
  - **样本正态分布且两类协方差矩阵相同时**
    - 最优贝叶斯分类器为线性函数$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0,\ \boldsymbol{w} = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right),$

$$w_0 = -\frac{1}{2}\left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) - \ln\frac{P\left(\omega_2\right)}{P\left(\omega_1\right)}$$

    - 对于Fisher线性判别：

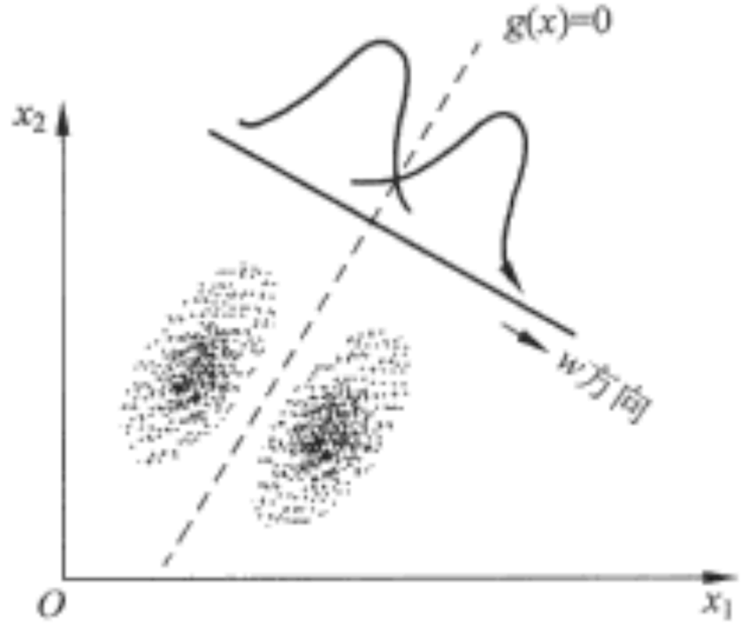
$$w_0 = -\frac{1}{2}\left(\boldsymbol{m}_1 - \boldsymbol{m}_2\right)^{T}\boldsymbol{S}_w^{-1}\left(\boldsymbol{m}_1 - \boldsymbol{m}_2\right) - \ln\frac{P\left(\omega_2\right)}{P\left(\omega_1\right)}$$

5.4 Fisher线性判别分析

- **确定分类阈值w_0**
  - **样本不服从正态分布时**
    - 可依据经验，如$w_0 = -\frac{1}{2}(\tilde{m}_1 + \tilde{m}_2)$或者$w_0 = -\tilde{m}$，（$\tilde{m}$是所有样本在投影后的均值）
    - 若$g(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{x} - \frac{1}{2}\left(\boldsymbol{m}_1 + \boldsymbol{m}_2\right)\right) \gtrless \log\frac{P\left(\omega_2\right)}{P\left(\omega_1\right)}$，则

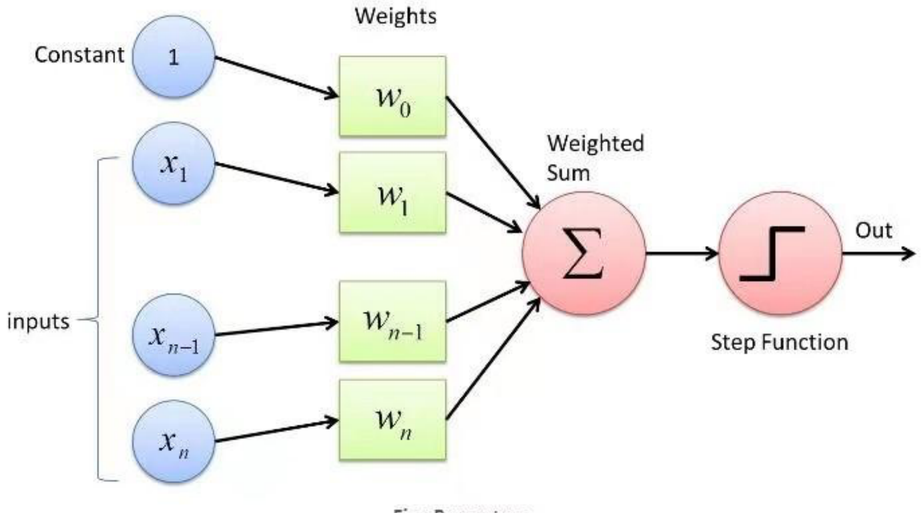
      $$\boldsymbol{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

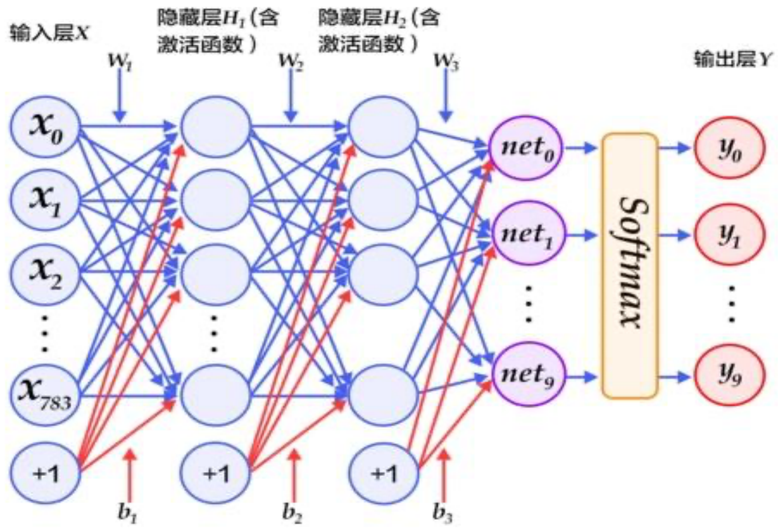
    - 通过与**两类均值投影的平分点**相比较做分类决策



5.5 感知器

- **Fisher 线性判断**：两步法
  - 确定最优的投影方向
  - 在投影方向上确定分类阈值
- **感知器**
  - 直接得到完整的线性判别函数
    $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$
  - 多层感知器神经网络和各种深度学习方法的基础





5.5 感知器

- **线性判别函数的齐次简化**

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 \quad \rightarrow \quad g(\boldsymbol{y}) = \boldsymbol{\alpha}^T\boldsymbol{y}$$

其中$\boldsymbol{y} = \left[1, x_1, x_2, \ldots, x_d\right]^T$, $\boldsymbol{\alpha} = \left[w_0, w_1, x_2, \ldots, w_d\right]^T$

– $\boldsymbol{y}$为增广的样本向量，$\boldsymbol{\alpha}$为增广的权向量

- **决策规则（两类ω_1、ω_2）**

– $g(\boldsymbol{y}) \lessgtr 0$，则$\boldsymbol{y} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

5.5 感知器

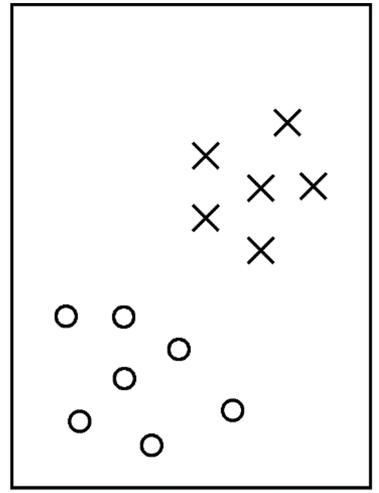
- **线性可分性**
  - 存在一个权向量$\boldsymbol{\alpha}$，使所有样本被正确分类，即

$$\exists \boldsymbol{\alpha},\ \forall i = 1,...,N,\ \begin{cases} \text{若 } \boldsymbol{y}_i \in \omega_1, & \text{则 } \boldsymbol{\alpha}^T \boldsymbol{y}_i > 0 \\ \text{若 } \boldsymbol{y}_i \in \omega_2, & \text{则 } \boldsymbol{\alpha}^T \boldsymbol{y}_i < 0 \end{cases}$$

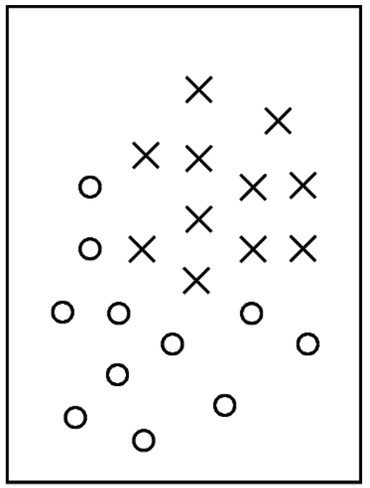
  - 令$\boldsymbol{y}_i' = \begin{cases} \boldsymbol{y}_i, & \text{若 } \boldsymbol{y}_i \in \omega_1 \\ -\boldsymbol{y}_i, & \text{若 } \boldsymbol{y}_i \in \omega_2 \end{cases}$，则线性可分性变为：

$$\exists \boldsymbol{\alpha},\ \forall i = 1,...,N,\ \boldsymbol{\alpha}^T \boldsymbol{y}_i > 0$$

  - $\boldsymbol{y}'$：规范化增广样本矩阵，仍然记作$\boldsymbol{y}$



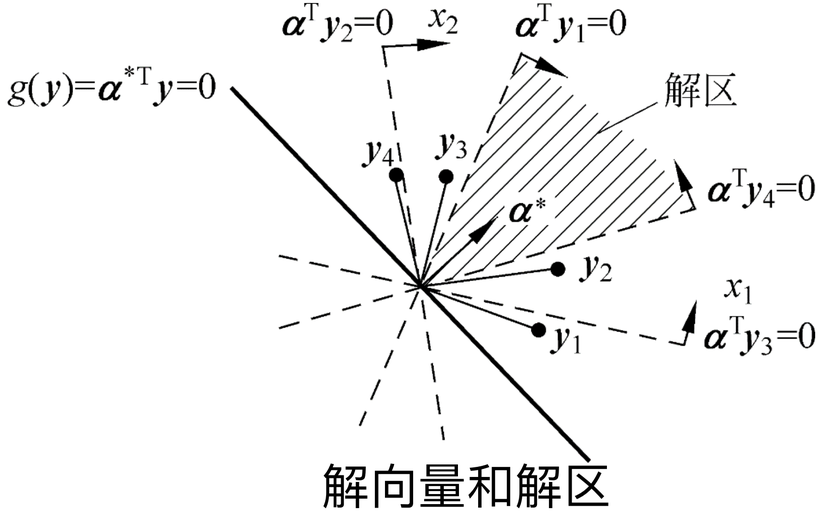
线性可分



线性不可分

5.5 感知器

- **求解**（线性可分情况）
  - **解向量**$\boldsymbol{\alpha}^*$：满足$\boldsymbol{\alpha}^T\boldsymbol{y}_i > 0,\ i = 1,2,...,N$
  - **解区**：权值空间中所有解向量组成的区域。
    - 对于样本$\boldsymbol{y}_i$，$\boldsymbol{\alpha}^T\boldsymbol{y}_i = 0$定义了权值空间中一个超平面$\hat{H}_i$。
    - 对于所有样本，解区就是每个样本对应超平面$\hat{H}_i$的正侧的**交集**。

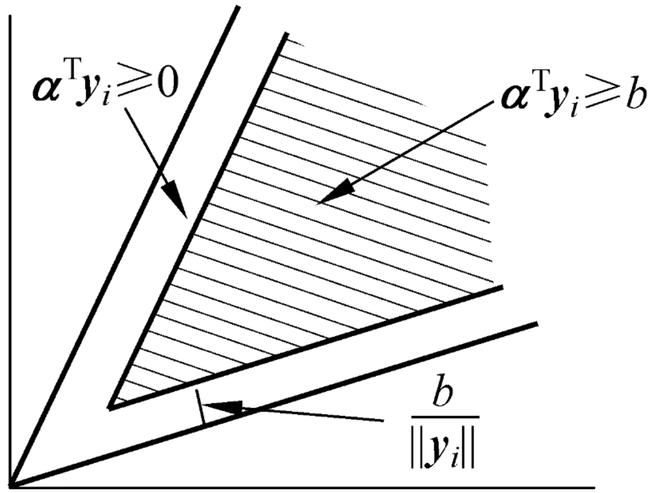


解向量和解区

余量

5.5 感知器

- **求解**（线性可分情况）
  - **余量**：把解区向中间缩小，不靠近边缘的解。
    即，引入余量$b > 0$，使得解向量满足$\boldsymbol{\alpha}^T \boldsymbol{y}_i > b$，
    $i = 1,2,...,N$
    - 靠近边缘的解容易受到噪声、数值计算误差的影响，解区中间更可靠。
    - B：线性判别函数的平移



余量

5.5 感知器

- **感知器准则函数**

$$J_p(\boldsymbol{\alpha}) = \sum_{\boldsymbol{\alpha}^T \mathbf{y}_k \le 0} \left(-\boldsymbol{\alpha}^T \mathbf{y}_k\right)$$

  - 对所有错分样本的求和表示对错分样本的惩罚
  - 当且仅当$J_p\left(\boldsymbol{\alpha}^*\right) = \min J_p(\boldsymbol{\alpha}) = 0$时，$\boldsymbol{\alpha}^*$是解向量

5.5 感知器

- **梯度下降法求解**

$$J_p(\boldsymbol{\alpha}) = \sum_{\boldsymbol{\alpha}^T \boldsymbol{y}_k \le 0} \left(-\boldsymbol{\alpha}^T \boldsymbol{y}_k\right)$$

- $\boldsymbol{\alpha}(t+1) = \boldsymbol{\alpha}(t) - \rho_t \nabla J_p(\boldsymbol{\alpha})$，$\rho_t$为调整步长
- $\nabla J_p(\boldsymbol{\alpha}) = \dfrac{\partial J_p(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}^T \boldsymbol{y}_k \le 0} \left(-\boldsymbol{y}_k\right)$
- $\therefore \boldsymbol{\alpha}(t+1) = \boldsymbol{\alpha}(t) + \rho_t \sum_{\boldsymbol{\alpha}^T \boldsymbol{y}_k \le 0} \boldsymbol{y}_k$，即每步迭代时把错分的样本按照某个系数加到权向量上。

5.5 感知器

- **梯度下降迭代算法步骤**
  - (1) 任意选择初始的权向量$\boldsymbol{\alpha}(0)$，置$t=0$
  - (2) 每次考察一个样本$\boldsymbol{y}_j$，若$\boldsymbol{\alpha}(t)^T\boldsymbol{y}_j \le 0$，则$\boldsymbol{\alpha}(t+1)=\boldsymbol{\alpha}(t)+\boldsymbol{y}_j$，否则继续
  - (3) 考察另一个样本，重复(2)，直至对所有样本都有$\boldsymbol{\alpha}(t)^T\boldsymbol{y}_j > 0$，即$J_p(\boldsymbol{\alpha})=0$

5.5 感知器

- **说明**
  - 若考虑余量b，将错分判断条件变成$\boldsymbol{\alpha}(t)^T\mathbf{y}_j \leq b$即可
  - 收敛性：对于线性可分的样本集，梯度下降的迭代算法经过有限次修正后一定会收敛到一个解向量$\boldsymbol{\alpha}^*$
  - 可使用可变步长减少迭代步数：$\rho_i = \dfrac{\left|\boldsymbol{\alpha}(k)^T\mathbf{y}_j\right|}{\left\|\mathbf{y}_j\right\|^2}$

5.6 最小平方误差判别

- **考虑线性不可分情况**
  - $\boldsymbol{\alpha}^T\boldsymbol{y}_i > 0,\ i = 1,2,...,N$，不等式组无法同时满足：最小化错分样本数
- **形式化描述**
  - 把$\boldsymbol{\alpha}^T\boldsymbol{y}_i > 0$变成$\boldsymbol{\alpha}^T\boldsymbol{y}_i = b_i > 0,\ i = 1,2,\cdots,N$
  - 或矩阵形式$\boldsymbol{Y}\boldsymbol{\alpha} = \boldsymbol{b}$，其中$\boldsymbol{Y} = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & \cdots & y_1\hat{d} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_N\hat{d} \end{bmatrix}$,

    $\boldsymbol{b} = [b_1, b_2, \cdots, b_N]^T$，　$\hat{d}$是增广的样本向量的维数，$\hat{d} = d + 1$

5.6 最小平方误差判别

- **求解**

  –通常$N > \hat{d}$，为矛盾方程组，无法求得精确解

  –误差为$\boldsymbol{e} = \boldsymbol{Y\alpha} - \boldsymbol{b}$，可求得方程组的最小平方误差解，即$\boldsymbol{\alpha}^*$：$\min J_s(\boldsymbol{\alpha})$

  –$J_s(\boldsymbol{\alpha})$为最小平方误差（MSE）准则函数

$$J_s(\boldsymbol{\alpha}) = \| \boldsymbol{Y\alpha} - \boldsymbol{b} \|^2 = \sum_{i=1}^{N} \left(\boldsymbol{\alpha}^T \boldsymbol{y}_i - \boldsymbol{b}_i \right)^2$$

5.6 最小平方误差判别

- **求解**

  – **伪逆法求解**

  ✓令 $\nabla J_s(\boldsymbol{\alpha}) = 2\boldsymbol{Y}^T(\boldsymbol{Y}\boldsymbol{\alpha} - \boldsymbol{b}) = 0$，可得 $\boldsymbol{\alpha}^* = \left(\boldsymbol{Y}^T\boldsymbol{Y}\right)^{-1}\boldsymbol{Y}^T\boldsymbol{b} = \boldsymbol{Y}^+\boldsymbol{b}$,

  其中$\boldsymbol{Y}^+ = \left(\boldsymbol{Y}^T\boldsymbol{Y}\right)^{-1}\boldsymbol{Y}^T$是长方矩阵$\boldsymbol{Y}$的伪逆

5.6 最小平方误差判别

- **求解**
  - **梯度下降法迭代求解**

✓任意选择初始的权向量$\boldsymbol{\alpha}(0)$，置$t=0$

✓按照梯度下降的方式迭代更新权向量$\boldsymbol{\alpha}(t+1)=\boldsymbol{\alpha}(t)-\rho_t\boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{\alpha}-\boldsymbol{b}\right)$，
直到$\nabla J_s(\alpha)\le\xi$或$\left\|\boldsymbol{\alpha}(t+1)-\boldsymbol{\alpha}(t)\right\|\le\xi$（$\xi$：误差灵敏度）

✓也可通过单样本修正法调整权向量
$\boldsymbol{\alpha}(t+1)=\boldsymbol{\alpha}(t)+\rho_t\left(\boldsymbol{b}_k-\boldsymbol{\alpha}(t)^T\boldsymbol{y}_k\right)\boldsymbol{y}_k$，其中，$\boldsymbol{y}_k$是使得$\boldsymbol{\alpha}(t)^T\boldsymbol{y}_k\neq\boldsymbol{b}_k$
的样本：最小均方根算法（LMS 算法）、Widrow-Hoff算法

5.6 最小平方误差判别

- **求解**
  - **梯度下降法迭代求解**

✓$\boldsymbol{b}$的选择

✦若$\boldsymbol{b}_i = \begin{cases} \frac{N}{N_1} & \boldsymbol{y}_i \in \omega_1 \\ \frac{N}{N_2} & \boldsymbol{y}_i \in \omega_2 \end{cases}$，同类样本的$\boldsymbol{b}_i$相同，则MSE的解等价于Fisher线性判别，且

$w_0^* = -\boldsymbol{m}^T \boldsymbol{w}^*$，其中，$N_1$、$N_2$分别是第一和第二类样本数，$N$为样本总数，

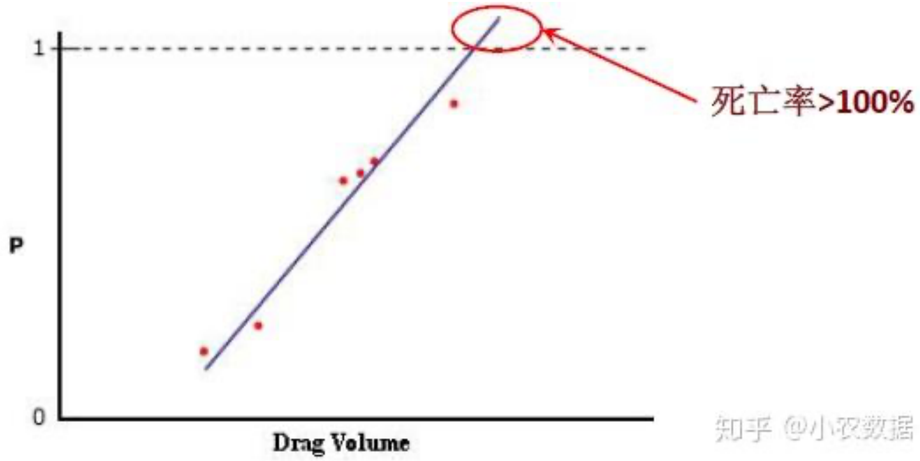
$$\boldsymbol{m} = \frac{1}{N}\left(N_1 \boldsymbol{m}_1 + N_2 \boldsymbol{m}_2\right)$$

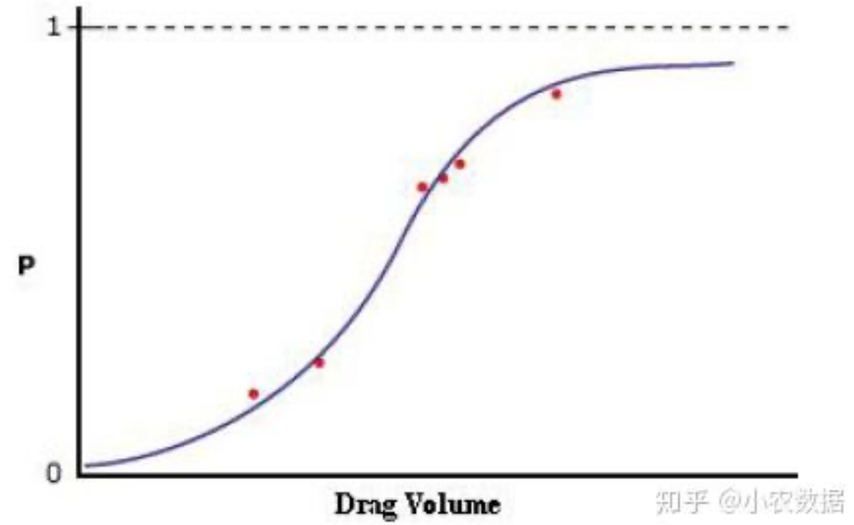
✦若$\boldsymbol{b}_i = 1(\forall i = 1, \cdots, N)$，则当$N \to \infty$时，MSE的解以最小平方误差逼近贝叶斯判别函数

$g_0(\boldsymbol{x}) = P(\omega_1 | \boldsymbol{x}) - P(\omega_2 | \boldsymbol{x})$，即$\boldsymbol{\alpha}^*$使得$\varepsilon^2 = \int \left[\boldsymbol{\alpha}^T \boldsymbol{y} - g_0(\boldsymbol{x})\right]^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$极小。

5.7 罗杰斯特回归（logistic regression）

- **线性回归**
  - $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$
  - 线性回归是**发散型方程式**，在许多应用上会有不适合的情形发生。
  - 将线性回归的值压制在0~1之间，才能产生出合理的值。





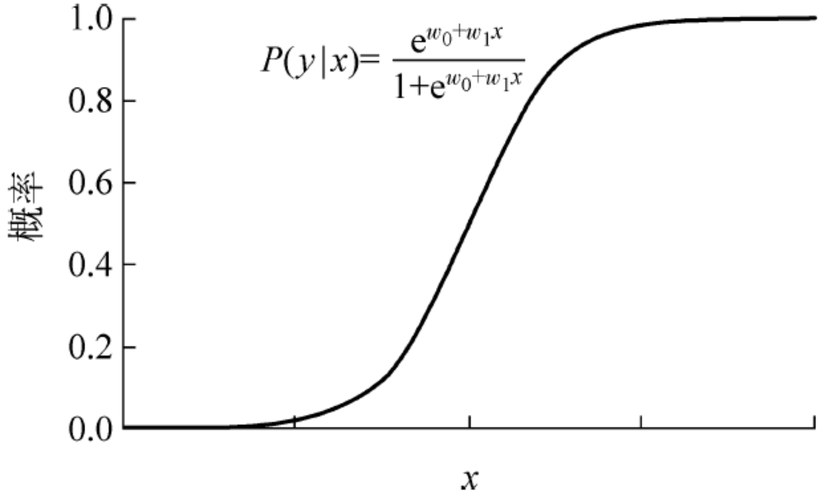
5.7 罗杰斯特回归（logistic regression）

- **罗杰斯特函数（logistic function）**

  _ 形式一：$P(y=1\,|\,x)=P(y\,|\,x)=\dfrac{e^{w_0+w_1x}}{1+e^{w_0+w_1x}}$

  _ 形式二：$\theta(s)=\dfrac{e^s}{1+e^s}=\dfrac{1}{1+e^{-s}}$，常用形式

  – 在神经网络中称为Sigmoid函数



5.7 罗杰斯特回归（logistic regression）

- **罗杰斯特回归（logistic regression）**

$$\text{logit}(\boldsymbol{x}) = \ln\left(\frac{P(y\,|\,\boldsymbol{x})}{1-P(y\,|\,\boldsymbol{x})}\right) = \omega_0 + \omega_1 x_1 + \ldots + \omega_m x_m$$

  - $P(y\,|\,\boldsymbol{x}) = \dfrac{e^{\omega_0+\omega_1 x_1+\cdots+\omega_m x_m}}{1+e^{\omega_0+\omega_1 x_1+\cdots+\omega_m x_m}}$
  - 几率：$\dfrac{P(y\,|\,x)}{1-P(y\,|\,x)} = e^{\omega_0+\omega_1 x}$，医学中表示患病的可能性和不患病可能性之比
  - 对数几率：$\ln\left(\dfrac{P(y\,|\,x)}{1-P(y\,|\,x)}\right) = \omega_0 + \omega_1 x$

5.7 罗杰斯特回归（logistic regression）

- **决策规则**

  _ 若logit$(\boldsymbol{x}) \lessgtr 0$，则$\boldsymbol{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

- **决策规则学习算法：最大似然法**

  _ 设共有N个独立的训练样本$\left\{ \left(\boldsymbol{x}_1, \boldsymbol{y}_1\right), \ldots, \left(\boldsymbol{x}_N, \boldsymbol{y}_N\right) \right\}$，$\boldsymbol{x}_j \in R^{d+1}$，

  $y_j \in \{-1,1\}$，假设样本类别从某未知概率$f(\boldsymbol{x})$中产生，以概率$f(\boldsymbol{x})$属于所关心类别，即$P(y \mid \boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}), & y = +1 \\ 1 - f(\boldsymbol{x}), & y = -1 \end{cases}$，用罗杰斯特函数$h(\boldsymbol{x}) = \theta(\boldsymbol{\omega}^T \boldsymbol{x})$估计$f(\boldsymbol{x})$，其中$\boldsymbol{\omega}$为罗杰斯特函数中待求参数组成的向量

5.7 罗杰斯特回归（logistic regression）

- **决策规则学习算法：最大似然法**
  - 模型在样本上的似然函数：

$$P(y_j|\boldsymbol{x}_j)=\begin{cases}h\left(\boldsymbol{x}_j\right), & y_j=+1\\ 1-h\left(\boldsymbol{x}_j\right), & y_j=-1\end{cases}$$

或者

$$l\left(h\mid\left(\boldsymbol{x}_j,y_j\right)\right)\triangleq P\left(y_j\mid\boldsymbol{x}_j,h\right)=\theta\left(y_j\boldsymbol{w}^T\boldsymbol{x}_j\right)$$

  - 对于所有样本，模型的似然函数是

$$L(\boldsymbol{w})=\prod_{j=1}^{N}P\left(y_j|\boldsymbol{x}_j\right)=\prod_{j=1}^{N}\theta\left(y_j\boldsymbol{w}^T\boldsymbol{x}_j\right)$$

5.7 罗杰斯特回归（logistic regression）

– 梯度下降法最优化目标函数

优化问题：

$$\min E(\boldsymbol{w}) = -\frac{1}{N}\ln\left(L(\boldsymbol{w})\right) = -\frac{1}{N}\ln\left(\prod_{j=1}^{N}\theta\left(y_j\boldsymbol{w}^T\boldsymbol{x}_j\right)\right)$$

$$= \frac{1}{N}\sum_{j=1}^{N}\ln\left(\frac{1}{\theta\left(y_j\boldsymbol{w}^T\boldsymbol{x}_j\right)}\right) = \frac{1}{N}\sum_{j=1}^{N}\ln\left(1+e^{-y_j\boldsymbol{w}^T\boldsymbol{x}_j}\right)$$

5.7 罗杰斯特回归（logistic regression）

– 梯度下降法最优化目标函数

算法步骤：

✓（1）记时刻为$k=0$，初始化参数$\omega(0)$

✓（2）计算目标函数的负梯度方向 $\nabla E=-\frac{1}{N}\sum_{j=1}^{N}\frac{y_j\boldsymbol{x}_j}{1+e^{y_j\boldsymbol{w}(\boldsymbol{k})^T\boldsymbol{x}_j}}$

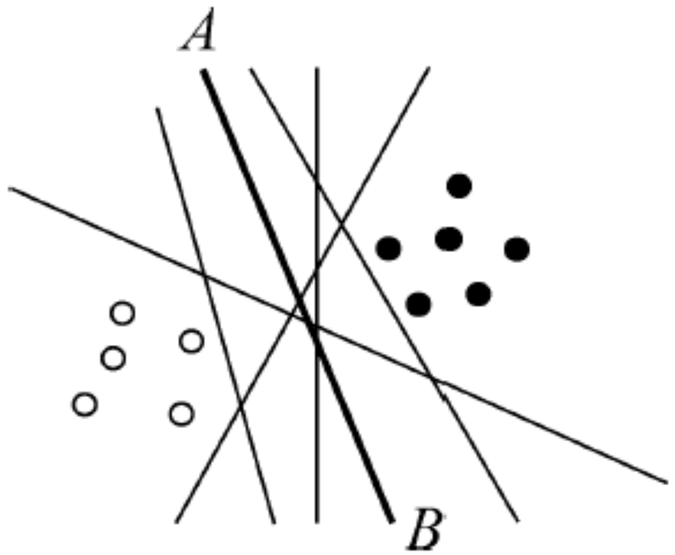
✓按步长（学习率）η更新下一时刻参数$\boldsymbol{w}(k+1)=\boldsymbol{w}(k)-\eta\nabla E$，检查是否达到终止条件，如未达到，令$k=k+1$，重新进行（2）

✓（3）算法停止，输出得到的参数ω

终止条件可以是似然函数的梯度小于某个预设值，训练过程不再有显著更新，或迭代达到预设的上限等

5.8 最优分类超平面与线性支持向量机

- **问题提出**
  - 只要一个样本集线性可分，解区中的任何向量都是一个解向量
  - 感知器算法采用不同的初值和迭代参数会得到不同的解
  - 哪个解更好?



5.8.1 最优分类超平面

- **分类超平面**
  - 假设有训练样本集$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)$，$\boldsymbol{x}_i \in R^d$，$y_i \in \{+1, -1\}$，其中每个样本是$d$维向量，$y$是类别标号，$\omega_1$类用$+1$表示，$\omega_2$类用$-1$表示。这些样本线性可分，即存在超平面

$$g(\boldsymbol{x}) = (\boldsymbol{w} \bullet \boldsymbol{x}) + b = 0$$

    把所有N个样本都没有错误地分开。

- **最优分类超平面**
  - 一个分类超平面，如果它能将训练样本没有错误地分开，且两类训练样本中离超平面最近的样本与超平面之间的距离是最大的，则这个超平面称作**最优分类超平面**（optimal seperating hyperplane），简称最优超平面（optimal hyperplane）。

    ✓两类样本中离分类面最近的样本到分类面的距离称作**分类间隔**（margin）

    ✓最优超平面也称作**最大间隔超平面**

5.8.1 最优分类超平面

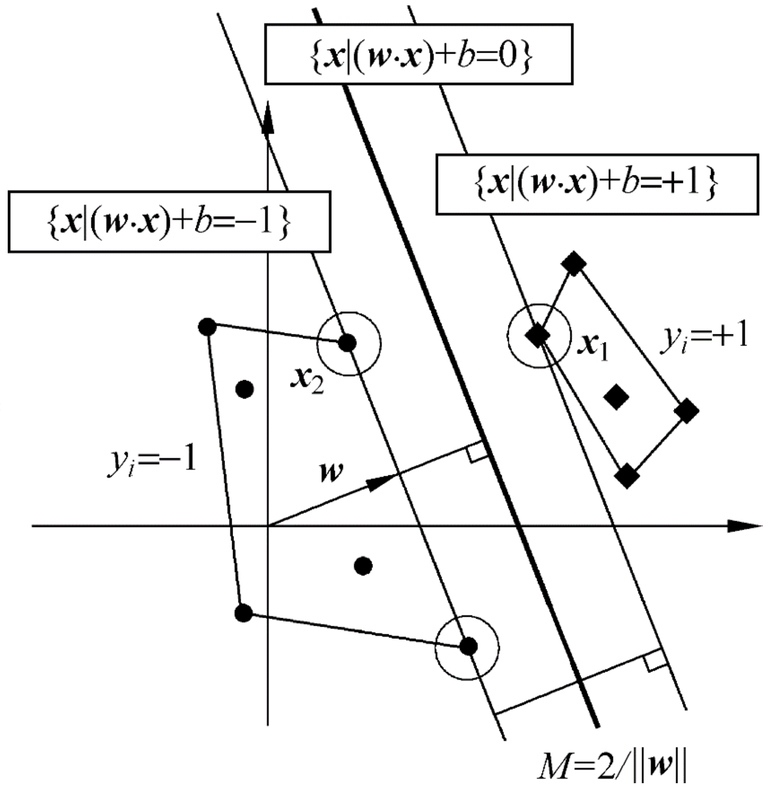
- **最优分类超平面**
  - 最优超平面定义的分类决策函数

$$f(\boldsymbol{x}) = \mathrm{sgn}\big(g(\boldsymbol{x})\big) = \mathrm{sgn}\big((\boldsymbol{w} \bullet \boldsymbol{x}) + b\big)$$

  - 规范化的分类超平面

$$\begin{cases} (\boldsymbol{w} \bullet \boldsymbol{x}_i) + b \geq 1, \ y_i = +1 \\ (\boldsymbol{w} \bullet \boldsymbol{x}_i) + b \leq -1, \ y_i = -1 \end{cases}$$

  即$y_i[(\boldsymbol{w} \bullet \boldsymbol{x}_i) + b] \geq 1, \ \ i = 1,2,\cdots,N$



5.8.1 最优分类超平面

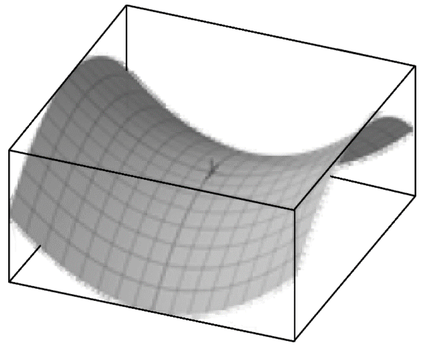
- **最优分类超平面**

  – 求解最优超平面（$\boldsymbol{w}^*$）

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad y_i\left[(\boldsymbol{w}\bullet\boldsymbol{x}_i)+b\right]-1\geq 0, \quad i=1,2,\cdots,N$$

引入拉格朗日系数

$$\boldsymbol{\alpha}_i \geq 0, \quad i=1,\cdots,N$$



鞍点

等价问题

$$\min_{\boldsymbol{w},b}\max_{\boldsymbol{\alpha}} L(\boldsymbol{w},b,\boldsymbol{\alpha}) = \frac{1}{2}(\boldsymbol{w}\bullet\boldsymbol{w}) - \sum_{i=1}^{N}\boldsymbol{\alpha}_i\left\{y_i\left[(\boldsymbol{w}\bullet\boldsymbol{x}_i)+b\right]-1\right\}$$

最优解在$L(\boldsymbol{w},b,\boldsymbol{\alpha})$的鞍点上取得，即$L(\boldsymbol{w},b,\boldsymbol{\alpha})$对$\boldsymbol{w}$和$b$的偏导数均为0

5.8.1 最优分类超平面

- **最优分类超平面**
  - 求解最优超平面（$\boldsymbol{w}^*$）

    最优解处：

$$\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \boldsymbol{x}_i, \quad \text{且} \quad \sum_{i=1}^{N} y_i \alpha_i^* = 0$$

    最优超平面的对偶问题（the dual problem）：

$$\max_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \left(\boldsymbol{x}_i \bullet \boldsymbol{x}_j \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{N} y_i \alpha_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \cdots, N$$

5.8.1 最优分类超平面

- **最优分类超平面**
  - 求解最优超平面（$\boldsymbol{w}^*$）

通过对偶问题的解，可求出原问题的解：

$$\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \boldsymbol{x}_i$$

$$f(\boldsymbol{x}) = \operatorname{sgn}\{g(\boldsymbol{x})\} = \operatorname{sgn}\{(\boldsymbol{w}^* \bullet \boldsymbol{x}) + b\} = \operatorname{sgn}\left\{ \sum_{i=1}^{N} \alpha_i^* y_i (\boldsymbol{x}_i \bullet \boldsymbol{x}) + b^* \right\}$$

5.8.1 最优分类超平面

- **最优分类超平面**

  – 求解最优超平面（b^*）

- 库恩-塔克（Kuhn-Tucker）条件：拉格朗日泛函的鞍点处满足

$$\alpha_i\left\{y_i\left[(\boldsymbol{w}\bullet\boldsymbol{x}_i)+b\right]-1\right\}=0,\quad i=1,2,\cdots,N$$

  等号成立的样本所对应的α_i才会大于0，求解b：

$$y_i\left[\left(\boldsymbol{w}^*\bullet\boldsymbol{x}_i\right)+b^*\right]-1=0$$

5.8.2 线性不可分情况

- **样本集非线性可分**

  – 定义

  对样本集$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2)\cdots, (\boldsymbol{x}_N, y_N)$，$\boldsymbol{x}_i \in R^d$，$y_i \in \{+1, -1\}$，不等式

$$y_i[(\boldsymbol{w} \bullet \boldsymbol{x}_i) + b] - 1 \geq 0, \quad i = 1,2,\cdots, N$$

  不可能对所有样本同时满足

  – 策略

  每个样本引入一个非负的松弛变量ξ_i，约束条件变为

$$y_i[(\boldsymbol{w} \bullet \boldsymbol{x}_i) + b] - 1 + \xi_i \geq 0, \quad i = 1,2,\cdots, N$$

5.8.2 线性不可分情况

- 广义最优分类面
  - 定义

$$\min_{w,b,\xi_i} \frac{1}{2}(\boldsymbol{w} \bullet \boldsymbol{w}) + C\sum_{i=1}^{N} \xi_i$$

$$\text{s.t.} \quad y_i\left[(\omega \bullet x_i) + b\right] - 1 + \xi_i \geq 0, \quad i = 1,2,\cdots,N$$

$$\xi_i \geq 0, \quad i = 1,2,\cdots,N$$

5.8.2 线性不可分情况

- **广义最优分类面**

  – 求解

  转化为拉格朗日泛函的鞍点问题

$$\min_{w,b,\xi_i} \max_{\alpha} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$$
$$= \frac{1}{2}(\boldsymbol{w} \cdot \boldsymbol{w}) + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\{\gamma_i[(\boldsymbol{w} \cdot \boldsymbol{x}_i) + b] - 1 + \xi_i\} - \sum_{i=1}^{N}\gamma_i\xi_i$$

5.8.2 线性不可分情况

- **广义最优分类面**

  – 求解

  广义最优分类面的对偶优化问题

$$\max_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \left(\boldsymbol{x}_i \bullet \boldsymbol{x}_j \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{N} y_i a_i = 0, \ \text{且} \ 0 \le \alpha_i \le C, \ i = 1,2,\cdots,N$$

原问题的解向量满足：$\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \boldsymbol{x}_i$

广义最优分类面的判别函数：$f(\boldsymbol{x}) = \text{sgn}\{g(\boldsymbol{x})\} = \text{sgn}\{(\boldsymbol{w}^* \bullet \boldsymbol{x}) + b\} = \text{sgn}\left\{ \sum_{i=1}^{N} a_i^* y_i (\boldsymbol{x}_i \bullet \boldsymbol{x}) + b^* \right\}$

5.8.2 线性不可分情况

- **广义最优分类面**

  – 求解

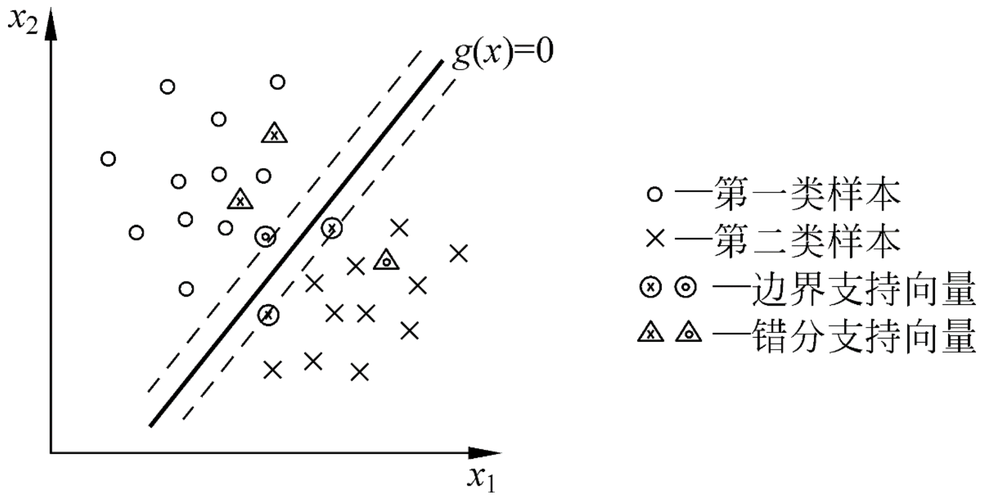
  库恩-塔克条件：拉格朗日泛函的鞍点处满足

$$\alpha_i\{y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]-1+\xi_i\}=0,\ \ i=1,2,\cdots,N$$

$$\gamma_i\xi_i=(C-\alpha_i)\xi_i=0,\ \ i=1,2,\cdots,N$$

  边界支持向量：$0<\alpha_i<C,\ \ \xi_i=0$

  错分支持向量：$\alpha_i=C,\ \ \xi_i>0$



5.9 多类线性分类器

- **多分类问题**

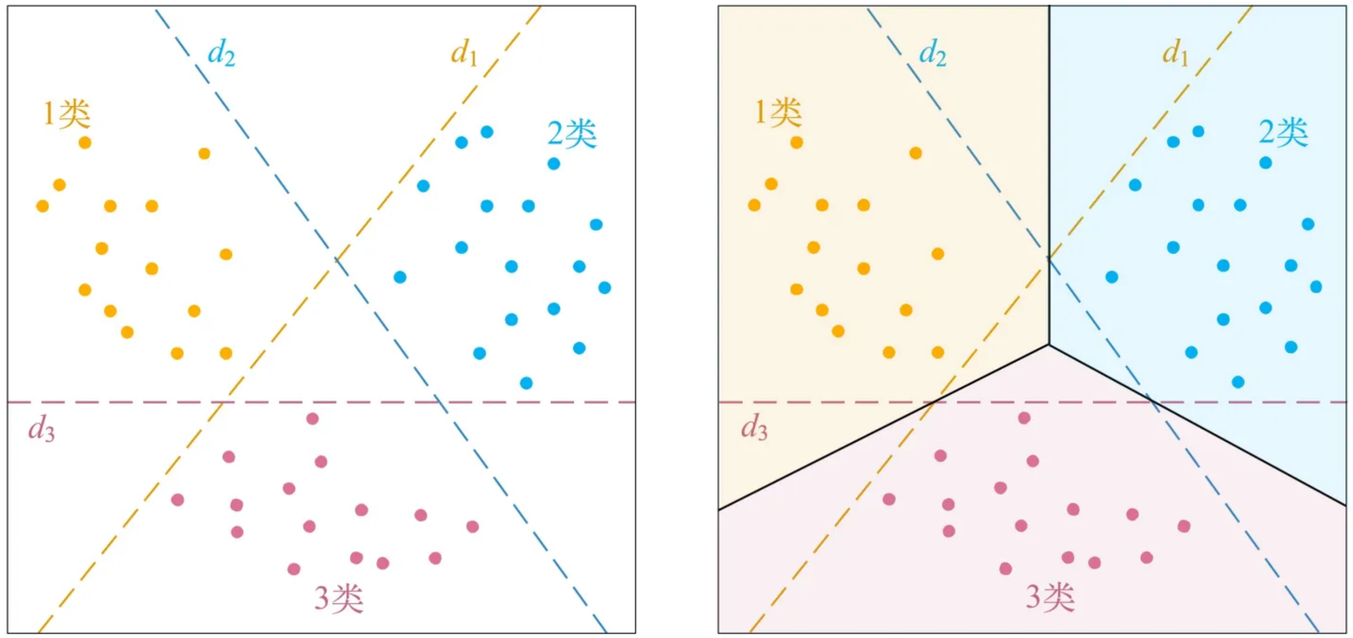
  给定含N个样本的训练集$X = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$，其中$K$维特征向量$\boldsymbol{x}_n \in \mathrm{R}^K$，类标签$y_n \in \{1,2,\ldots,M\}, n = 1,\ldots,N$。训练集数据共$M$个类。任务是找到决策函数$y = f(\boldsymbol{x})$ (或者说一个规则)用于预测新数据的类别。

  – 间接：将多分类问题分解为多个两类问题

  – 直接：设计多类分类器

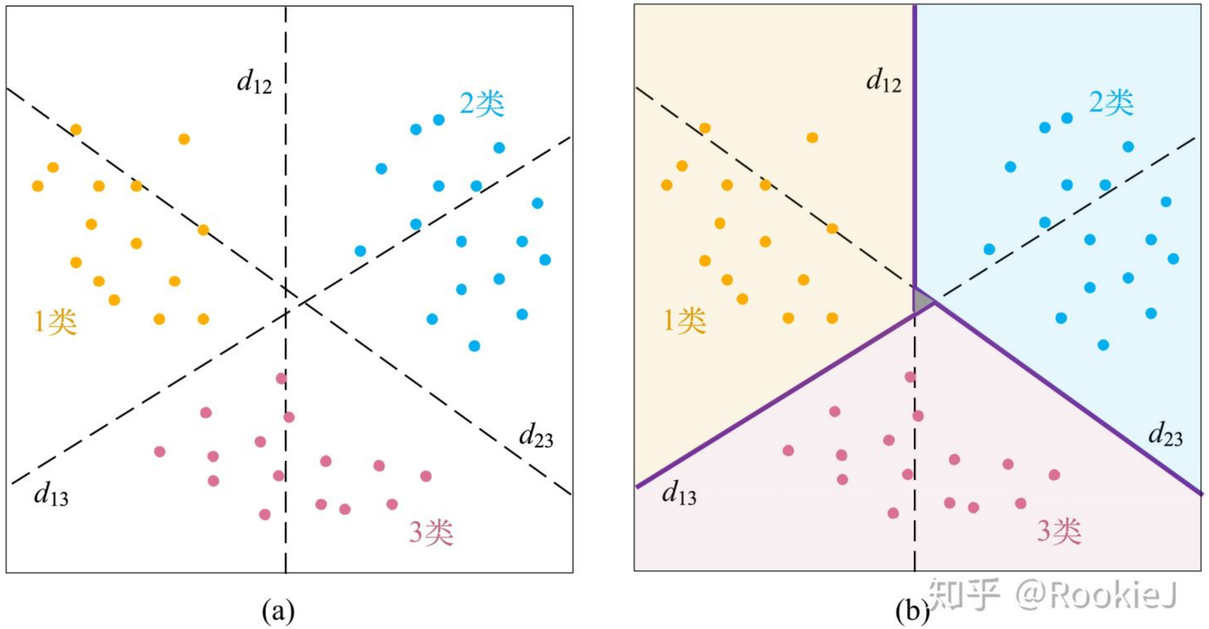
5.9.1 多个两类分类器的组合

- **方案一：一类对余类(one-against-all，one-against-the-rest)**
  - c类转化为$c-1$个两类问题
  - 问题：训练样本不均衡；歧义区



5.9.1 多个两类分类器的组合

- **方案二：成对分类(one-against-one，pairwise classification)**
  - 问题：每两类构造一个分类器，则c类需要$\frac{c(c-1)}{2}$个两类分类器
  - 问题：分类器多



5.9.2 多类线性判别函数

- **定义**

对c类设计c个判别函数

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}, \quad i = 1,2,\cdots,c$$

取判别函数最大的类，即

$$若g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}),\ \forall j \neq i,\ 则\boldsymbol{x} \in \omega_i$$

表示为增广向量形式：$g_i(\boldsymbol{x}) = \boldsymbol{\alpha}_i^T \boldsymbol{y}, \quad i = 1,2,\cdots,c$，其中，$\boldsymbol{\alpha}_i = \begin{bmatrix} \boldsymbol{w}_i \\ w_{i0} \end{bmatrix}$为增广权向量。

多类线性判别函数也称作多类线性机器，可记作$L\left(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_c\right)$。

5.9.2 多类线性判别函数

- **求解：逐步修正法求解线性机器（多类线性可分）**

(1) 任意选择初始的权向量$\alpha_i(0)$，$i = 1,2,...,c$，置$t = 0$

(2)考察某个样本$\boldsymbol{y}^k \in w_i$，若$\boldsymbol{\alpha}_i(t)^T\boldsymbol{y}^k > \boldsymbol{\alpha}_j(t)^T\boldsymbol{y}^k$，则所有权向量不变；若存在某个类$j$，使$\boldsymbol{\alpha}_i(t)^T\boldsymbol{y}^k \le \boldsymbol{\alpha}_j(t)^T\boldsymbol{y}^k$，则选择$\boldsymbol{\alpha}_j(t)^T\boldsymbol{y}^k$最大的类别$j$，对各类权值进行如下修正

$$\begin{cases} \boldsymbol{\alpha}_i(t+1) = \boldsymbol{\alpha}_i(t) + \rho_t\boldsymbol{y}^k \\ \boldsymbol{\alpha}_j(t+1) = \boldsymbol{\alpha}_j(t) - \rho_t\boldsymbol{y}^k \\ \boldsymbol{\alpha}_l(t+1) = \boldsymbol{\alpha}_l(t), \quad l \neq i,j \end{cases}$$

(3)如果所有样本都分类正确，则停止；否则考查另一个样本，重复（2）