模式识别

中国科学技术大学 汪增福

- 第一章 绪论
- ✓ 第二章 统计模式识别中的几何方法
- 第三章 统计模式识别中的概率方法
- 第四章 分类器的错误率
- 第五章 统计模式识别中的聚类方法
- 第六章 结构模式识别中的句法方法
- 第七章 总结

第二章 统计模式识别中的几何方法

- 本章主要内容

主要讨论确定性简单模式的有监督分类问题，重点讨论统计模式识别中的几何分类法。

  - 统计分类的基本思想
  - 模式的相似性度量和最小距离分类器
  - 线性可分情况下的几何分类法
  - 非线性可分情况下的几何分类法
  - 线性可分问题的非迭代解法
  - 最优分类超平面

§ 2.1 统计分类的基本思想

基于特征向量表达的模式识别过程

特征抽取

→ 特征向量表示

→ 特征空间

→ 区域分割

求解

问题: 确定性简单模式的有监督分类

简单模式可以识别的前提条件

（1）具有表征其类别的类属性或（和）度量特征；

（2）同类样本在特征空间中组成若干个集群区域；

（3）不同类样本在特征空间中彼此分离。

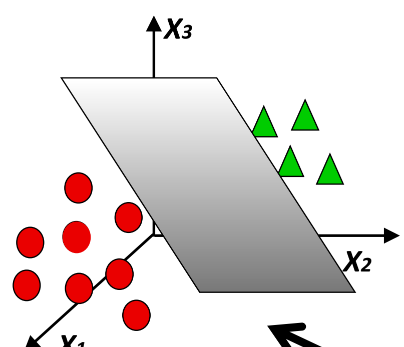
§ 2.1 统计分类的基本思想

如何完成给定的模式分类任务？

假设：抽取的模式特征满足可识别的三个前提条件。

观察与思考：观测条件不同，所得到的观测样本一般也不同，由此得到的模式特征自然也存在差异。但是，训练样本数足够多时，这些样本在特征空间的分布可反映出对应模式类别的总体分布规律。

结论：可根据足够多的训练样本，将特征空间划分成若干个区域，从而完成规定的模式分类任务。



分界面 → 分类器设计

§ 2.1 统计分类的基本思想

一些约定

- 用希腊字母作为模式类别的标示符

  用ω标记模式类

- 用花体英文字母表示样本集合

  用 $\mathscr{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n\}$ 表示由**n**个样本组成的样本集合

- 用粗体大写英文字母表示向量和矩阵

  用 $\boldsymbol{M}_{m\times n}$ 表示一个**m**行**n**列的矩阵

- 用普通体小写英文字母表示标量、向量分量和矩阵元素

  用 $\boldsymbol{X}=(x_1,x_2,\ldots x_n)^T$ 表示一个**n**维的列向量

§ 2.1 统计分类的基本思想

分类器设计 —类别数为2的情形

标记：● 相关的两个类别分别用○和△表示；

● 训练样本用 $X=(x_1,x_2,\ldots x_n)^T$ 表示。

任务：根据给定的训练样本找到一个分界面将整个特征空间分割成两个区域，使得每个区域内部仅包含来自同一类别的样本。

满足要求的分类器不止一个。这些分类器对于给定的训练样本而言可以做到正确分类，但对于训练集之外的样本而言，不同分类器给出的结果会存在差异。

为了对分类器性能做出评估，通常借助于训练集之外的样本集合（测试样本集）。

具体做法是：依次将测试样本集中的每一个样本输入分类器，并对判决结果进行统计，算出正确判决的样本数占样本总数的百分比，以此作为分类器识别率的估计。

§ 2.1 统计分类的基本思想

两个例子

- 高个与矮个

- 胖子与瘦子

§ 2.1 统计分类的基本思想

例1：一维三类的例子

问题： 根据身高特征，判断待识别个体的高矮程度。

● 矮个子　△ 中等个子　■ 高个子

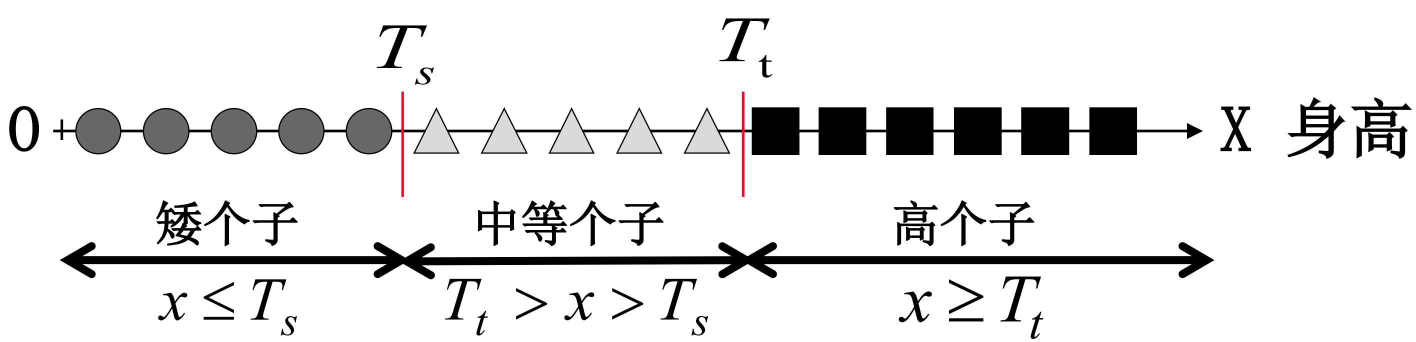
求解：Step1 - 获取类别已知的训练样本集合；

Step2 – 根据样本的 分布情况，确定两个边界，将特征空间划分成三个区间。

● 矮个子的样本

△ 中等个子的样本

■ 高个子的样本



§ 2.1 统计分类的基本思想

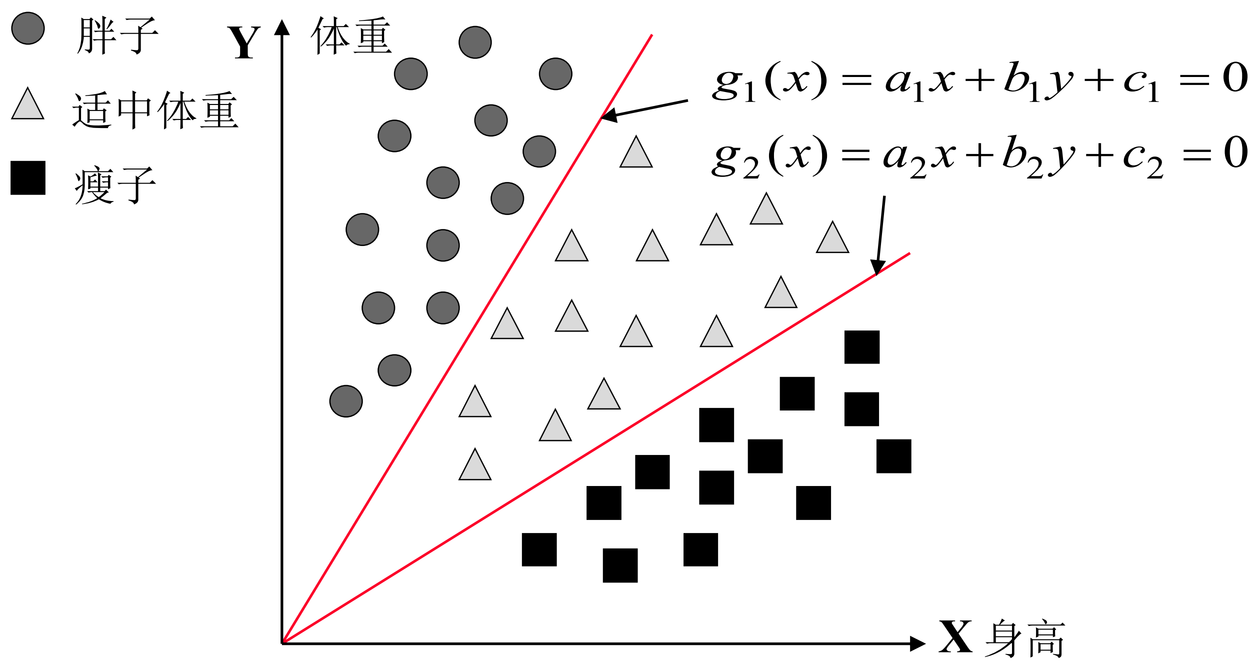
例2： 二维三类的例子

问题： 根据体重和身高特征，判断待识别个体的胖瘦程度。

● 胖子　　△ 不胖不瘦　　■ 瘦子

求解： **Step1** - 获取类别已知的训练样本集合；

Step2 – 根据样本的 分布情况，确定两个边界，将特征空间划分成三个区间。



§ 2.1 统计分类的基本思想

分类器设计关键步骤小结

- 选择合适的特征并据此构建特征空间；
- 获取已知类别的观测样本，将其映射到所构建的特征空间中；
- 根据观测样本在特征空间的分布情况，确定相应的分界面将整个特征空间划分为与待分类的类别数相等的子区域。

§ 2.2 相似性度量和最小距离分类器

分类器设计的进一步探讨

- 对分界面的要求

（1）将同一类别的已知训练样本归入相同的子区域；

（2）将不同类别的已知训练样本归入不同的子区域；

（3）对未知类别的测试样本有较好的分类能力。

需要研究相应的分类法则使据此设计的分类器有一定的预测能力。

§ 2.2 相似性度量和最小距离分类器

一种直观上易于理解的分类法则:相似性

- 相似性度量

✓ 两个模式在特征空间中的距离

在特征空间中两个模式距离越近，则越相似。

➡ 可以通过计算待分类样本到各个类别最近的样本之间的距离作为将它归入某个类别的依据。

✓ 作为相似性度量的距离函数应满足的性质

距离函数

（1） $d(\boldsymbol{X},\boldsymbol{Y})=d(\boldsymbol{Y},\boldsymbol{X})$

（2） $d(\boldsymbol{X},\boldsymbol{Y})\le d(\boldsymbol{X},\boldsymbol{Z})+d(\boldsymbol{Z},\boldsymbol{Y})$

（3） $d(\boldsymbol{X},\boldsymbol{Y})\ge 0$

（4） $d(\boldsymbol{X},\boldsymbol{Y})=0$ ， 当且仅当 $\boldsymbol{X}=\boldsymbol{Y}$

§ 2.2 相似性度量和最小距离分类器

经常使用的距离函数

- Minkowsky距离

$$d(X,Y)=\left[\sum_{i=1}^{n}\left|x_i-y_i\right|^{\lambda}\right]^{\frac{1}{\lambda}}$$

← 明氏距离

- Manhattan距离

$$d(X,Y)=\sum_{i=1}^{n}\left|x_i-y_i\right|$$

- Cityblock距离

$$d(X,Y)=\sum_{i=1}^{n}w_i\left|x_i-y_i\right|$$

- Euclidean距离

$$d(X,Y)=\left[\sum_{i=1}^{n}\left|x_i-y_i\right|^{2}\right]^{\frac{1}{2}}$$

← 欧氏距离

- Camberra距离

$$d(X,Y)=\sum_{i=1}^{n}\frac{\left|x_i-y_i\right|}{\left|x_i+y_i\right|}$$

- Mahalanobis距离

$$d^2(X,M)=(X-M)^T\Sigma^{-1}(X-M)$$

马氏距离

$$\begin{cases} M=E\{X\}=\int_{-\infty}^{\infty}\Lambda\int_{-\infty}^{\infty}Xp(X)dx_1dx_2\Lambda\, dx_n=\int_{-\infty}^{\infty}\Lambda\int_{-\infty}^{\infty}Xp(X)dX \\ \Sigma=E\{(X-M)(X-M)^T\}=\int_{-\infty}^{\infty}\Lambda\int_{-\infty}^{\infty}(X-M)(X-M)^T p(X)dX \end{cases}$$

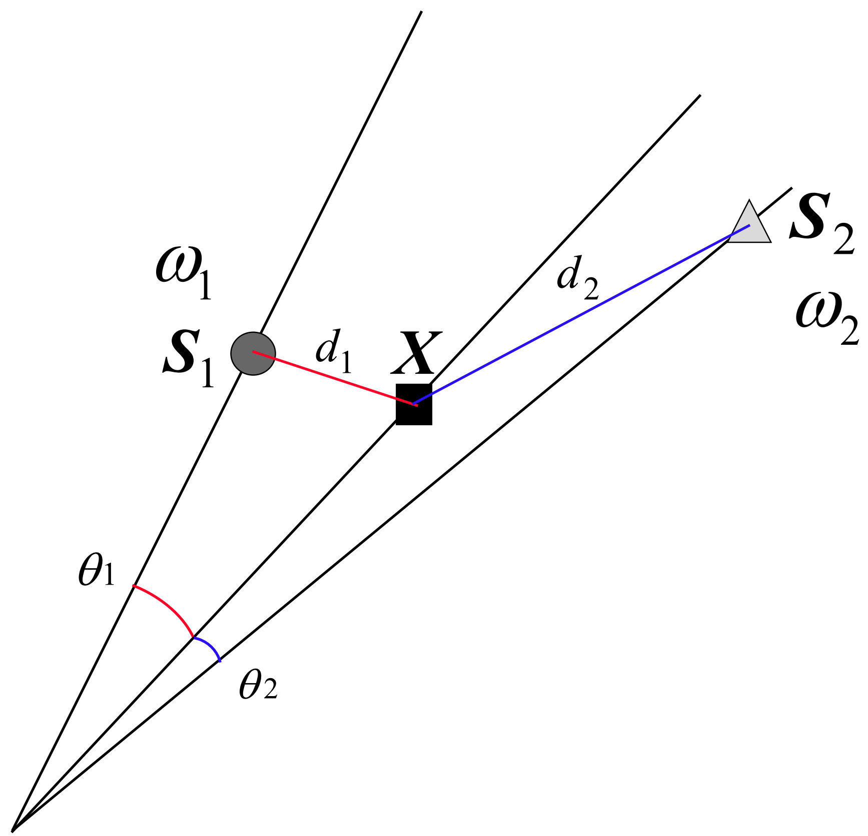
§ 2.2 相似性度量和最小距离分类器

向量夹角余弦

- 除距离函数外，适当定义的其它函数亦可作为相似性测度。

  例如，当样本呈扇状分布时，夹角余弦是很好的相似性测度。

$$S(X,Y)=\cos\theta=\frac{\boldsymbol{X}^T\boldsymbol{Y}}{\|\boldsymbol{X}\|\,\|\boldsymbol{Y}\|}$$



§ 2.2 相似性度量和最小距离分类器

最小距离分类器

- **基于标准样本的距离分类器**

类别 ω_1 ω_2 ω_3 …… ω_N

标准样本 $\boldsymbol{M}_1$ $\boldsymbol{M}_2$ $\boldsymbol{M}_3$ …… $\boldsymbol{M}_N$

分类规则：对于给定的未知样本$\boldsymbol{X}$，计算$\boldsymbol{X}$与$\boldsymbol{M}_1$、$\boldsymbol{M}_2$、……$\boldsymbol{M}_N$之间的距离，如果 $\boldsymbol{X}$与$\boldsymbol{M}_i$之间的距离为其中的最小者，则将$\boldsymbol{M}_i$的类别作为$\boldsymbol{X}$的类别。

➡ 若对 $\forall j \neq i$ ，有$d(\boldsymbol{X},\boldsymbol{M}_i) < d(\boldsymbol{X},\boldsymbol{M}_j), \forall j \neq i \Rightarrow \boldsymbol{X} \in \omega_i,\ \ j = 1,2,\ldots N$ 成立，则判决$\boldsymbol{X} \in \omega_i$ ，即将$\boldsymbol{X}$ 归入ω_i 类中。

➡ $$d(\boldsymbol{X},\boldsymbol{M}_i) < d(\boldsymbol{X},\boldsymbol{M}_j), \forall j \neq i \Rightarrow \boldsymbol{X} \in \omega_i,\quad j = 1,2,\ldots N$$

➡ $$d(\boldsymbol{X},\boldsymbol{M}_i) = \underset{j}{Min}\{d(\boldsymbol{X},\boldsymbol{M}_j)\} \Rightarrow \boldsymbol{X} \in \omega_i,\quad j = 1,2,\ldots N$$

局限性：当类内的观测样本散布较大时，分类性能下降较大。

最小距离分类器

● 最大值判决

对基于标准样本的距离分类器进行改写，可得到变形的分类器表示。

当采用欧氏距离作为距离测度时，有：

$$d^2(\boldsymbol{X},\boldsymbol{M}_j) = (\boldsymbol{X}-\boldsymbol{M}_j)^T(\boldsymbol{X}-\boldsymbol{M}_j)$$

$$= \boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{M}_j{}^T\boldsymbol{X} + \boldsymbol{M}_j{}^T\boldsymbol{M}_j$$

$$= \boldsymbol{X}^T\boldsymbol{X} - 2\underbrace{(\boldsymbol{M}_j{}^T\boldsymbol{X} - \frac{1}{2}\boldsymbol{M}_j{}^T\boldsymbol{M}_j)}$$

$$g_j(\boldsymbol{X}) = \boldsymbol{M}_j{}^T\boldsymbol{X} - \frac{1}{2}\boldsymbol{M}_j{}^T\boldsymbol{M}_j$$

由距离的非负性，有：

$$d(\boldsymbol{X},\boldsymbol{M}_i) < d(\boldsymbol{X},\boldsymbol{M}_j), \forall j \neq i \quad \Leftrightarrow \quad d^2(\boldsymbol{X},\boldsymbol{M}_i) < d^2(\boldsymbol{X},\boldsymbol{M}_j), \forall j \neq i$$

判决条件 ➡ $\boldsymbol{X}^T\boldsymbol{X} - 2g_i(\boldsymbol{X}) \;<\; \boldsymbol{X}^T\boldsymbol{X} - 2g_j(\boldsymbol{X})$

➡ $g_i(\boldsymbol{X}) \;>\; g_j(\boldsymbol{X})$

最终分类规则

$$g_i(\boldsymbol{X}) \;>\; g_j(\boldsymbol{X}),\; \forall j \neq i \;\Rightarrow\; \boldsymbol{X} \in \omega_i, \quad j = 1,2,\ldots N$$

最小距离分类器

- **最大值判决**

$$g_i(X) > g_j(X),\ \forall j \neq i \Rightarrow X \in \omega_i,\ \ j = 1,2,\ldots N$$

判别函数（当取欧氏距离作为距离测度时，具有线性形式）

记 $M_j = (m_{j,1}, m_{j,2}, \ldots, m_{j,n})^T$

$w_{j,k} = m_{j,k},\ k = 1,2,\ldots n$

$w_{j,n+1} = -\frac{1}{2} M_j^T M_j$

$$g_j(X) = M_j^T X - \frac{1}{2} M_j^T M_j$$

构造 $W'_j = (w_{j,1}, w_{j,2}, \ldots, w_{j,n},\ w_{j,n+1})^T$ 加长型向量

$X' = (x_1, x_2, \ldots x_n, 1)^T$ 增广型向量

权向量

则 $g_j(X) = M_j^T X - \frac{1}{2} M_j^T M_j$ 可写为

$$g_j(X') = W'^T_j X'$$ —— 线性判别函数

§ 2.2 相似性度量和最小距离分类器

最小距离分类器

- 基于分散样本的距离分类器

平均样本法

$$\boldsymbol{M}_j = \frac{1}{S_j}\sum_{k=1}^{S_j}\boldsymbol{Y}_{j,k}$$

局限性：没有考虑样本散布的影响，分类性能往往不理想。

平均距离法

特点：用待识别样本$\boldsymbol{X}$到各类别所有训练样本的平均距离作为分类依据。

$$\bar{d}(\boldsymbol{X},\omega_j) = \frac{1}{S_j}\sum_{k=1}^{S_j} d(\boldsymbol{X},\boldsymbol{Y}_{j,k}),\quad j=1,2,\ldots N$$

分类判决规则

$$\bar{d}(\boldsymbol{X},\omega_i) < \bar{d}(\boldsymbol{X},\omega_j), \forall j \neq i \ \Rightarrow\ \boldsymbol{X}\in\omega_i,\quad j=1,2,\ldots N$$

优缺点：具有一定的抗噪声能力，但计算和存储代价较大。

最小距离分类器

- 基于分散样本的距离分类器

最近邻法

方法要点：从整个训练样本集合中找出与测试样本最近邻的那个训练样本，并将其所属的类别作为测试样本的类别。

$$d_{\min}(\boldsymbol{X},\omega_j)=\underset{k=1,2,\ldots S_j}{Min}\{d(\boldsymbol{X},\boldsymbol{Y}_{j,k})\} \qquad j=1,2,\ldots N$$

分类判决规则

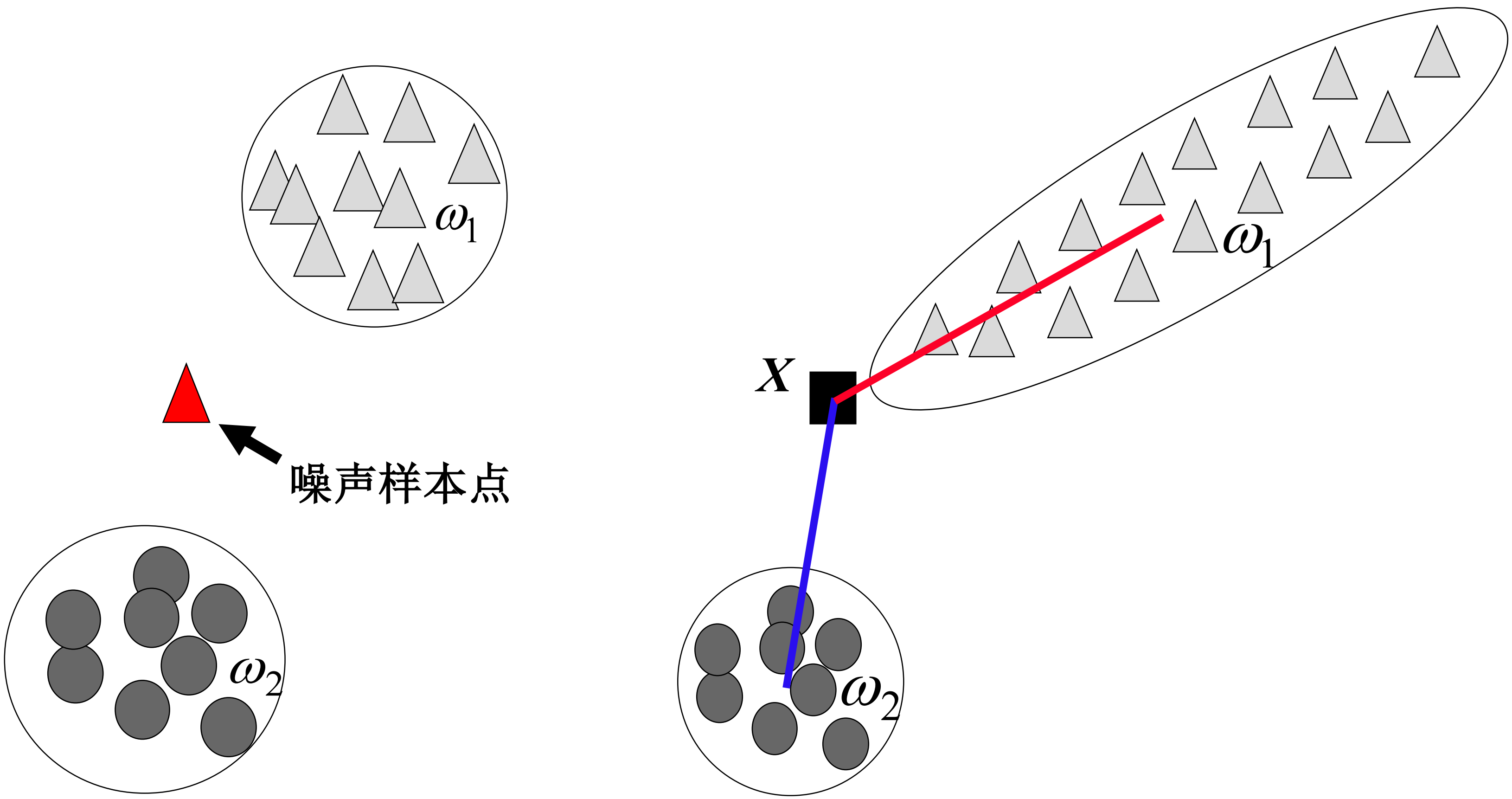
$$d_{\min}(\boldsymbol{X},\omega_i)<d_{\min}(\boldsymbol{X},\omega_j),\forall j\neq i \ \Rightarrow\ \boldsymbol{X}\in\omega_i,\quad j=1,2,\ldots N$$

优缺点：简单、实用，但计算和存储代价较大。另外，在样本点受到噪声污染时容易造成误分类。

§ 2.2 相似性度量和最小距离分类器

最小距离分类器

- 基于分散样本的距离分类器
  - 平均样本法
  - 平均距离法
  - 最近邻法



§ 2.2 相似性度量和最小距离分类器

最小距离分类器

- 基于分散样本的距离分类器
  - 平均样本法
  - 平均距离法
  - 最近邻法

进一步讨论

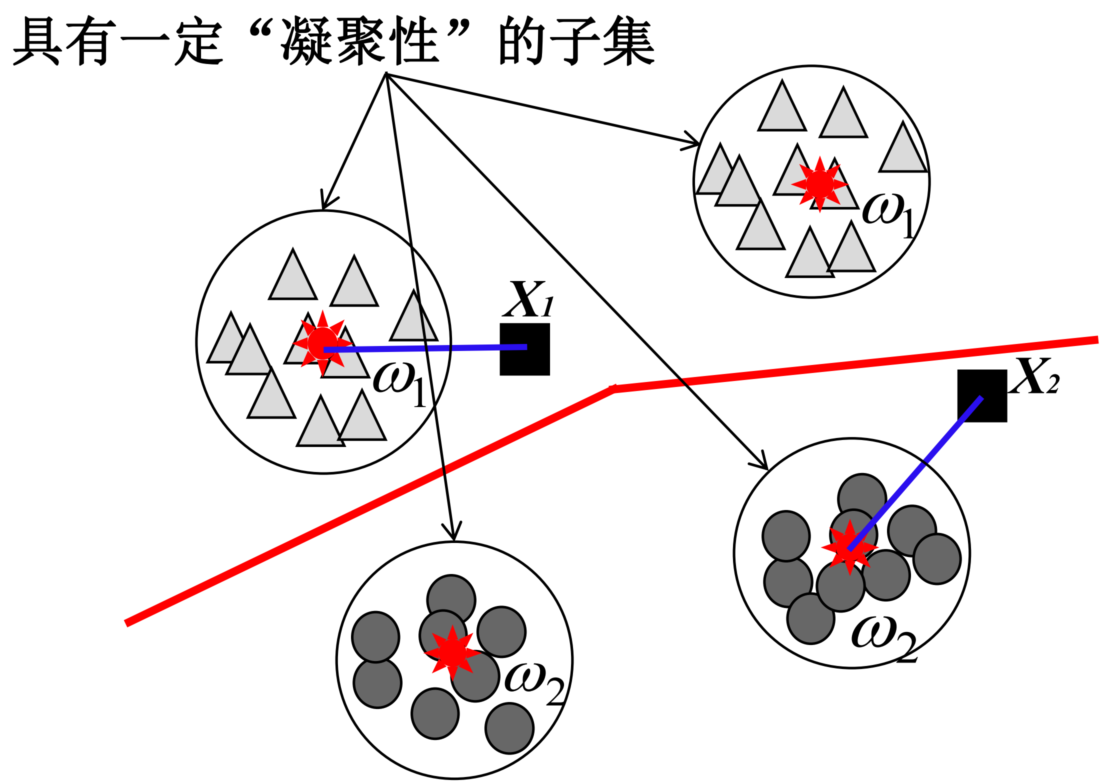
平均样本法和最近邻法代表了两种极端的情形。

平均样本法用一个点代表一个类别（过分集中），当某些类别的样本点在特征空间中散布程度较大时，将导致对有些样本点的误分类。与此不同，最近邻法将每一个样本点都视为所属类别的一个代表（过分分散），当样本点受到噪声污染时容易造成误分类。

§ 2.2 相似性度量和最小距离分类器

● 提高最小距离分类器性能的若干考虑

方法1 引入集群操作
- 用各个子集的平均样本作为其标准样本
- 每个子集的样本个数不小于设定的阈值
- 用所得标准样本集，借助于最近邻法完成分类



§ 2.2 相似性度量和最小距离分类器

- 提高最小距离分类器性能的若干考虑

方法2　K近邻操作

设待识别的测试样本为X

Step1. 计算X到训练样本集中每一个样本的距离，并按大小排序；

Step2. 提取与X最近邻的前K个样本点作为进一步判决的依据；

Step3. 统计X 与最近邻的前K个样本点中属于类别ω_j的个数k_j，$j=1,2,\ldots N$。

$$\sum_{j=1}^{N} k_j = K$$

判决规则

$$k_i > k_j, \forall j \neq i \;\; \Rightarrow \;\; X \in \omega_i, \quad j = 1,2,\ldots N$$

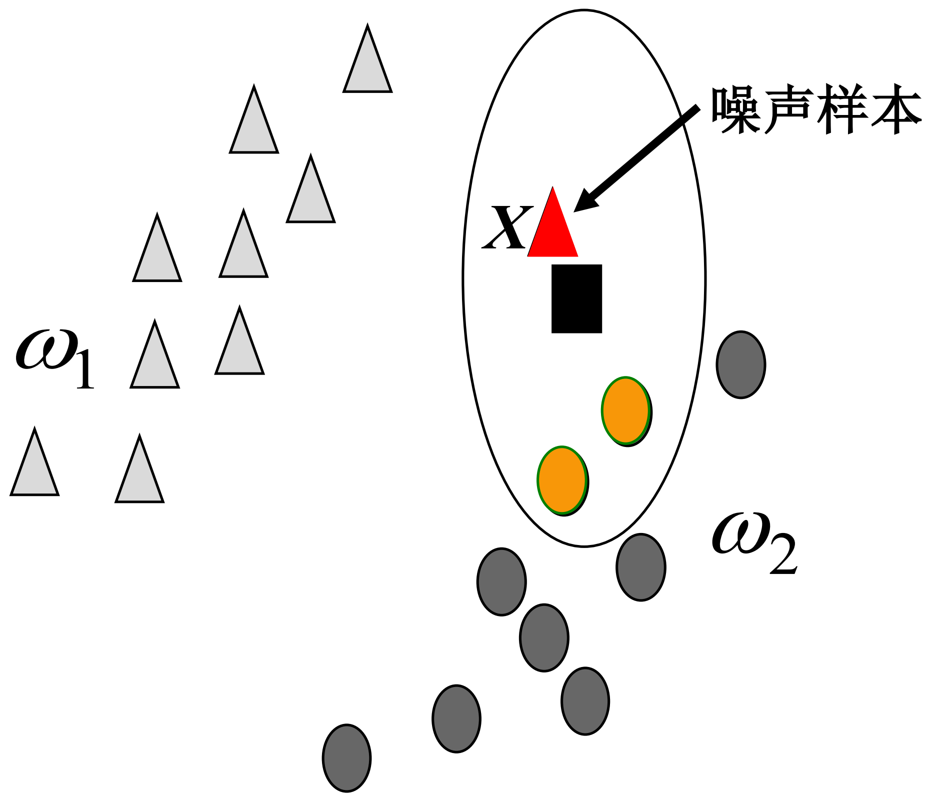
K值的选取

- 希望大些以减小噪声
- 希望选到的近邻样本都很近

§ 2.2 相似性度量和最小距离分类器

● 提高最小距离分类器性能的若干考虑

K近邻方法图示



§ 2.3 线性可分情况下的几何分类法

把利用几何方法将特征空间分解成不同类别子区域的方法称为几何分类法，相应的分类器称为几何分类器。

线性判别函数和线性分类器

思路：首先解决两类问题的几何分类器设计问题，然后将结果拓展到多类情况。

在两类情况下，基于标准样本的最小距离分类器的分类规则可写成：

$$g_i(\boldsymbol{X}) > g_j(\boldsymbol{X}),\ \forall j \neq i \Rightarrow \boldsymbol{X} \in \omega_i,\quad j = 1,2$$

其中，$g_j(X), j = 1, 2$分别为两个类别的判别函数。若采用欧氏距离作为距离测度，则它们具有线性函数的形式。

§ 2.3 线性可分情况下的几何分类法

两类情况下的线性分类器

如前所述，在设计分类器时为每个类别都定义了属于自己的判别函数。

与前不同，现希望找到一个“不直接与类别相关联的” 函数 $G(X)$，希望通过对其的操作来解决相应的分类问题。
为叙述方便起见，将$G(X)$称为决策面函数。

$G(\boldsymbol{X})=0$定义了相应特征空间中的一个曲面。

分界面或决策面，将特征空间划分为两部分 $\begin{cases} \boldsymbol{G}(\boldsymbol{X})>0 \\ \boldsymbol{G}(\boldsymbol{X})<0 \end{cases}$

希望以此分界面为界，同类别的样本位于分界面的同一侧，不同类别的样本位于分界面的不同一侧。

规定使类别 ω_1 的所有样本都位于 $\boldsymbol{G}(\boldsymbol{X})>0$ 一侧，使类别 ω_2 的所有样本都位于 $\boldsymbol{G}(\boldsymbol{X})<0$ 一侧。

§ 2.3 线性可分情况下的几何分类法

两类情况下的线性分类器

一旦找到使所有训练样本都能被正确分类的$G(X)$，则分类器设计问题即告解决。

分类规则

$$G(X)\begin{cases} >0 \quad \Rightarrow \quad X\in\omega_1 \\ =0 \quad \Rightarrow \quad \text{任意判决或拒绝判决} \\ <0 \quad \Rightarrow \quad X\in\omega_2 \end{cases}$$

确定$G(X)$
- 函数形式 ← 试探法确定（原则：先简单后复杂）
- 函数中所含参数

根据训练样本，用待定系数法确定

§ 2.3 线性可分情况下的几何分类法

两类情况下的线性分类器

设特征空间的维数为n，则待求决策面函数为

$$G(\boldsymbol{X}) = \boldsymbol{W}^T \boldsymbol{X} + w_{n+1} = \sum_{k=1}^{n} w_k x_k + w_{n+1}$$

加长向量表示

$$G(\boldsymbol{X}) = \boldsymbol{W}'^T \boldsymbol{X}' = \sum_{k=1}^{n+1} w_k x_k$$

$$\boldsymbol{W}' = (w_1, w_2, \ldots, w_n, w_{n+1})^T$$

$$\boldsymbol{X}' = (x_1, x_{2,\ldots} x_n, 1)^T$$

任务：根据训练样本确定所求权向量。

§ 2.3 线性可分情况下的几何分类法

两类情况下的线性分类器

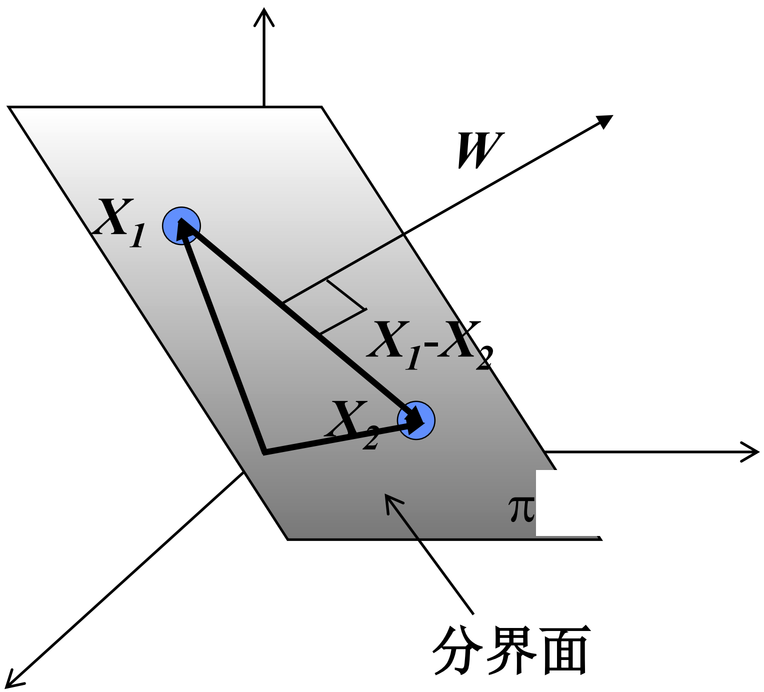
选用线性判别函数时，决策面方程的几何意义：

当特征空间是一维空间时，它是一个点；
当特征空间是二维空间时，它是一条直线；
当特征空间是三维空间时，它是一个平面；
当特征空间是多维空间时，它是一个超平面。

§ 2.3 线性可分情况下的几何分类法

两类情况下的线性分类器

权向量的重要性质



设X_1和X_2是决策面上任意两点，

显然有：

$$W^T X_1 + w_{n+1} = W^T X_2 + w_{n+1} = 0$$

$$\Rightarrow \quad W^T (X_1 - X_2) = 0$$

因$X_1 - X_2$是决策面上任意一个向量，

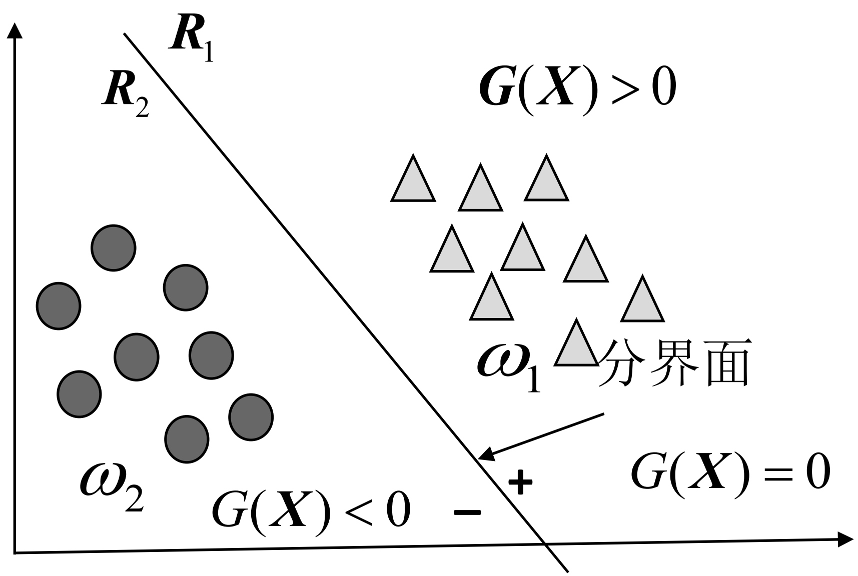
故，权向量W和决策面π正交。

即，权向量和决策面法线方向一致。

§ 2.3 线性可分情况下的几何分类法

两类情况下的线性分类器

和决策面相关的一些表述



- 在两类情况下决策面把整个特征空间分成两个半空间。

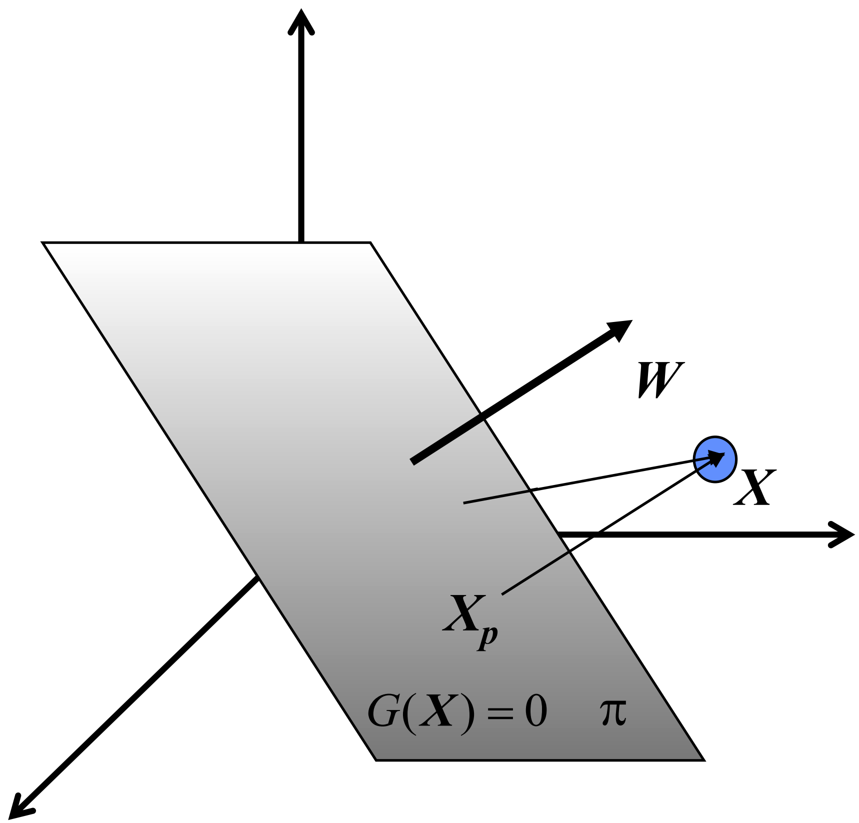
$$\begin{cases} \omega_1\text{的决策域 } \boldsymbol{R}_1 \text{ 构成的半空间} \\ \omega_2\text{的决策域 } \boldsymbol{R}_2 \text{ 构成的半空间} \end{cases}$$

- 决策面法向量的正方向指向 $\boldsymbol{R}_1$
- $\boldsymbol{R}_1$位于决策面的正面一侧
- $\boldsymbol{R}_2$位于决策面的反面一侧

§ 2.3 线性可分情况下的几何分类法

两类情况下的线性分类器

决策面函数$G(X)$在X处取值的几何意义



特征空间中一点X到决策面的距离

自X引垂线交决策面于X_p

则所求距离由$X-X_p$的模给出

根据定义$X-X_p$与决策面正交，故它和权向量方向或者一致或者相反。

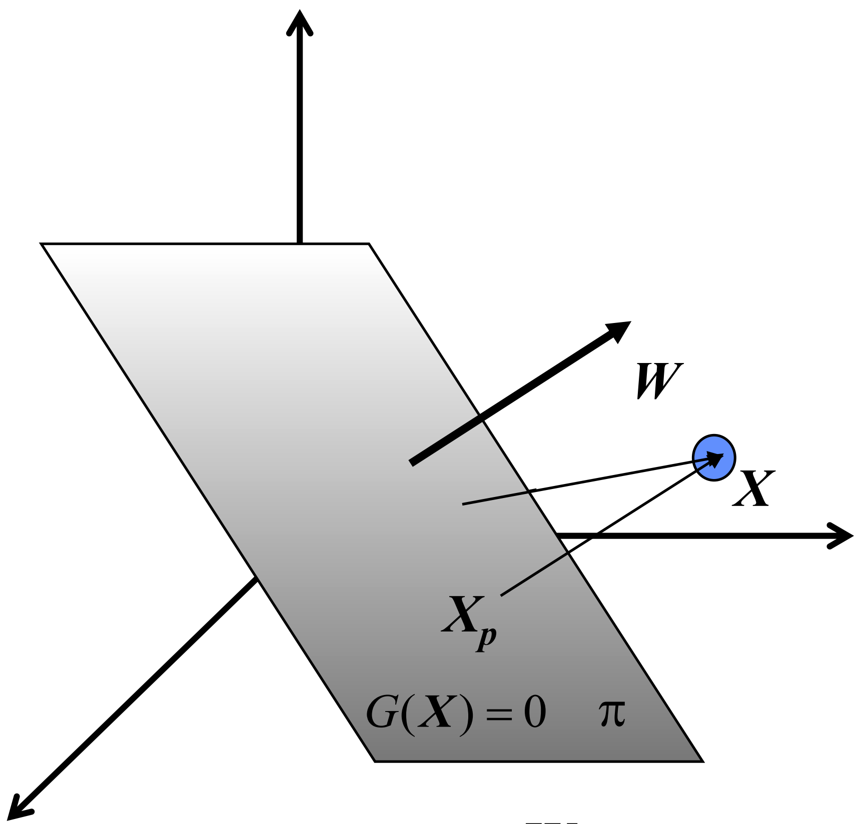
$$X - X_p = r\frac{W}{\|W\|}$$

r是一个代数量，其大小等于所求距离，符号反映了其所处的位置。

§ 2.3 线性可分情况下的几何分类法

两类情况下的线性分类器

决策面函数*G(X)*在*X*处取值的几何意义



$$\boldsymbol{X}-\boldsymbol{X}_p = r\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|}$$

$$\boldsymbol{W}^T\boldsymbol{X}-\boldsymbol{W}^T\boldsymbol{X}_p = r\frac{\boldsymbol{W}^T\boldsymbol{W}}{\|\boldsymbol{W}\|} = r\|\boldsymbol{W}\|$$

$$(\boldsymbol{W}^T\boldsymbol{X}+w_{n+1})-(\boldsymbol{W}^T\boldsymbol{X}_p+w_{n+1}) = r\|\boldsymbol{W}\|$$

$$G(\boldsymbol{X}) = \boldsymbol{W}^T\boldsymbol{X}+w_{n+1}$$

$$G(\boldsymbol{X}_p) = \boldsymbol{W}^T\boldsymbol{X}_p+w_{n+1} = 0$$

$$G(\boldsymbol{X}) = r\|\boldsymbol{W}\|$$

$$r = \frac{G(\boldsymbol{X})}{\|\boldsymbol{W}\|}$$

结论：*G(X)*给出了特征空间中一个点*X*到决策面距离的度量

§ 2.3 线性可分情况下的几何分类法

多类情况下的线性分类器

- 总体线性可分
  若能用一个超平面将任意一个类别的样本同其它所有类别的样本分开
- 成对线性可分
  若能用一个超平面将所有类别中任意两个类别的样本分开

⬇

线性可分

如何得到多类情况下的线性分类器？

➡ 由求解多个两类问题的线性分类器得到。

- 分解 ⟶ 三种情况
- 综合

§ 2.3 线性可分情况下的几何分类法

第一种情况：输入样本集合总体线性可分

类别 $\omega_1 \ \omega_2 \ \omega_3 \quad \cdots\cdots \quad \omega_{\mathrm{N}}$

用 $\omega_i\,/\,\overline{\omega}_i$ 分法把N类判决问题转化为N个两类问题求解。

$\omega_i\,/\,\overline{\omega}_i \qquad i=1,2,.N$

第i个类别之外的所有类别

第i个类别

判决规则

$$G_i(X) = W_i^{'T} X' \begin{cases} > 0 \Rightarrow X \in \omega_i \\ = 0 \Rightarrow \text{任意判决或拒绝判决} \\ < 0 \Rightarrow X \in \overline{\omega}_i \end{cases}$$

待求权向量

§ 2.3 线性可分情况下的几何分类法

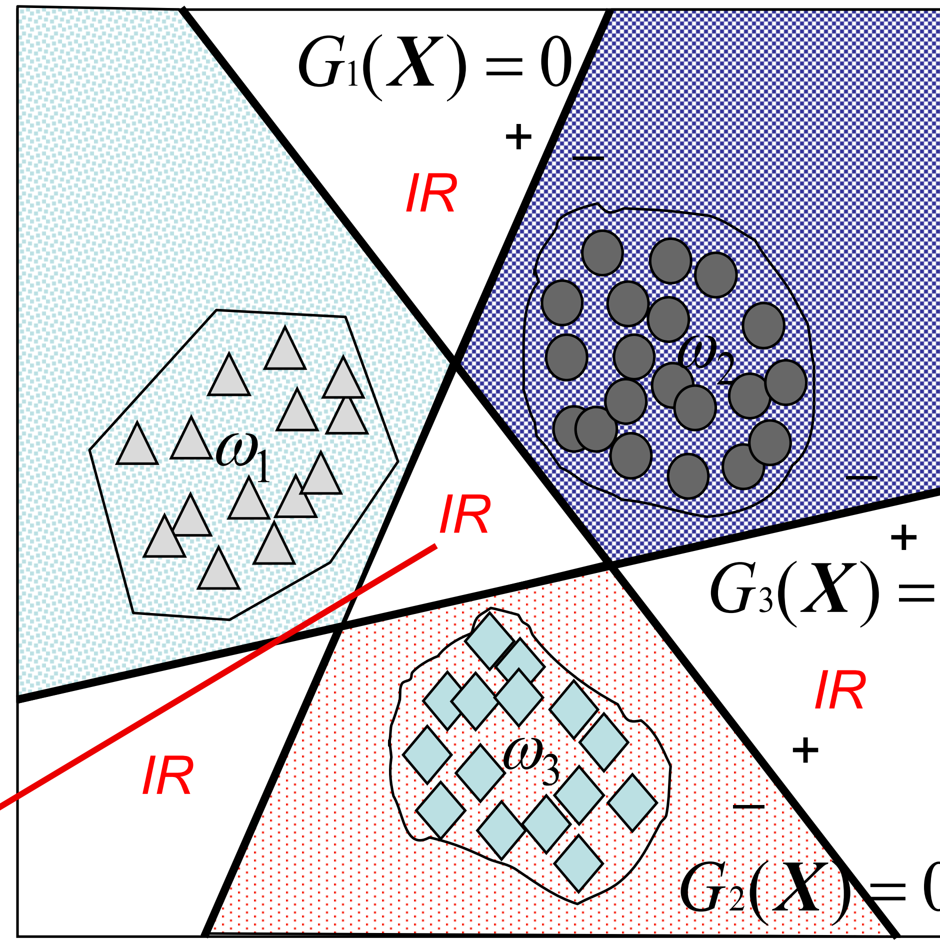
第一种情况： 输入样本集合总体线性可分

示例：两维三类

$$\begin{cases} \omega_1 / \overline{\omega}_1 \\ \omega_2 / \overline{\omega}_2 \\ \omega_3 / \overline{\omega}_3 \end{cases}$$

决策域ω_i ,i=1,2,3的确定

$$G_i(X) > 0, G_j(X) < 0, \forall j \neq i$$



不确定区域

§ 2.3 线性可分情况下的几何分类法

第二种情况： 输入样本集合成对线性可分

类别 $\omega_1 \ \omega_2 \ \omega_3 \quad \cdots\cdots \quad \omega_\mathrm{N}$

用 ω_i / ω_j 两分法把N类判决问题转化为 C_N^2 个两类问题求解。

$\omega_i / \omega_j \quad i \neq j \qquad C_N^2$

第 j 个类别

第 i 个类别

判决规则

$$G_{ij}(X) = W_{ij}^{'T} X' \begin{cases} > 0 & \Rightarrow \quad X \in \omega_i \\ = 0 & \Rightarrow \quad \text{任意判决或拒绝判决} \\ < 0 & \Rightarrow \quad X \in \omega_j \end{cases}$$

待求权向量

§ 2.3 线性可分情况下的几何分类法

第二种情况：输入样本集合成对线性可分

示例：两维三类

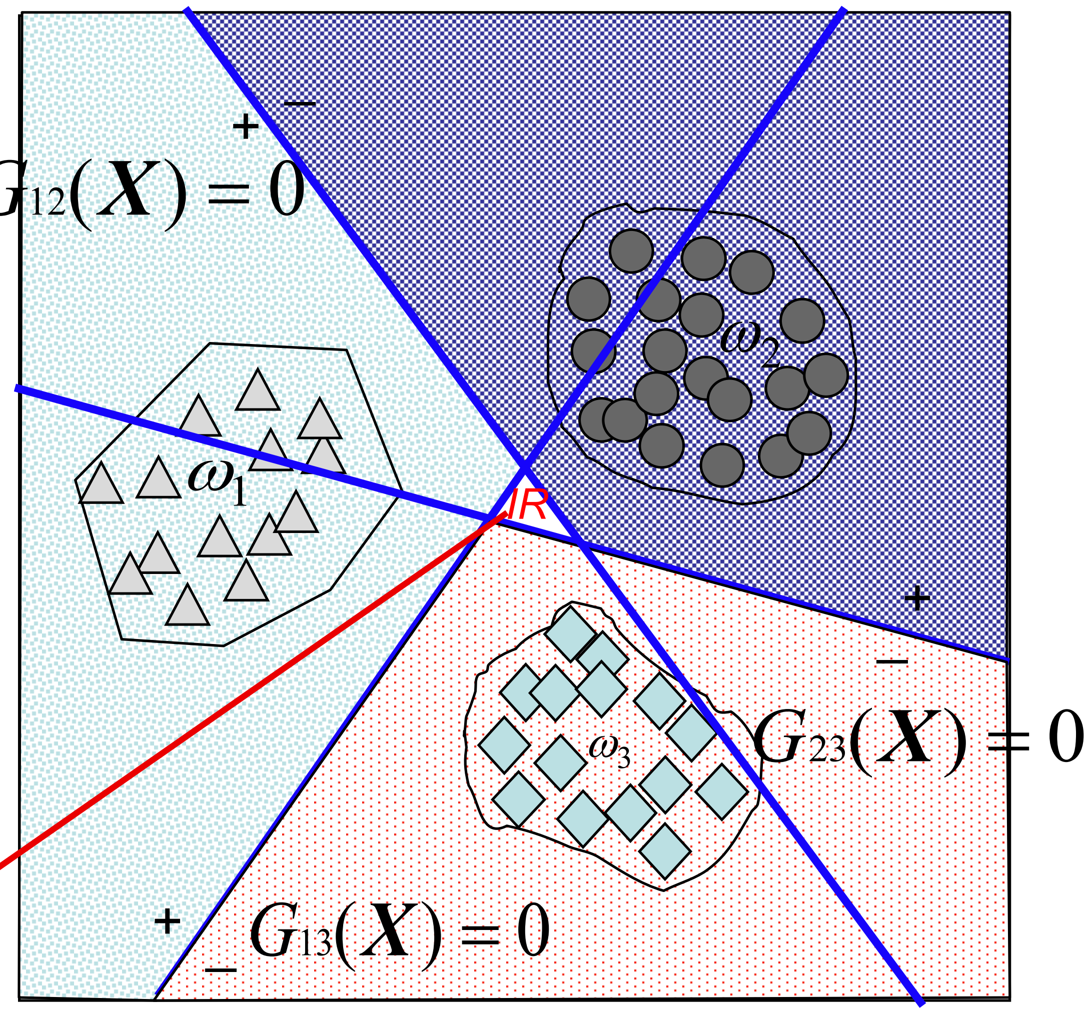
$$\begin{cases} \omega_1 / \omega_2 \\ \omega_1 / \omega_3 \\ \omega_2 / \omega_3 \end{cases}$$

$$G_{ij}(\boldsymbol{X}) = -G_{ji}(\boldsymbol{X})$$

ω_i ,i=1,2,3的决策域

$$G_{ij}(\boldsymbol{X}) > 0, \forall j \neq i$$

不确定区域



§ 2.3 线性可分情况下的几何分类法

第三种情况： 输入样本集合成对线性可分，且存在没有不确定区域的解。

分类策略： 每个类别定义一个判别函数，并采用最大值判决。

N个判别函数 $$g_j(\boldsymbol{X}) = \boldsymbol{W}_j^{'T}\boldsymbol{X}', j = 1,2,\ldots N$$

分类判决规则 $$g_i(\boldsymbol{X}) > g_j(\boldsymbol{X}),\ \forall j \neq i \Rightarrow \boldsymbol{X} \in \omega_i,\quad j = 1,2,\ldots N$$

➡ ω_i / ω_j 两分问题

i和j固定时，是一个两分问题。

定义 $$G_{ij}(\boldsymbol{X}) = g_i(\boldsymbol{X}) - g_j(\boldsymbol{X})$$

属于ω_i类别的样本的判决规则

$$G_{ij}(\boldsymbol{X}) = \boldsymbol{W}_{ij}^{'T}\boldsymbol{X}' > 0, \forall j \neq i \Rightarrow \boldsymbol{X} \in \omega_i,\quad j = 1,2,\ldots N$$

➡ ω_i ,i=1,2, …,N的决策域

第三种情况和第二种情况的判决规则形式上相似

§ 2.3 线性可分情况下的几何分类法

第三种情况和和第二种情况在内涵上的区别

第二种情况

N类判决问题转化为 C_N^2 个彼此独立的 ω_i / ω_j 两分问题

第三种情况

N类判决问题形式上亦可转化为 C_N^2 个ω_i / ω_j 两分问题

但彼此之间不是完全独立的!!!

权向量$W_{ij}^{'}$，不仅和ω_i和ω_j有关，亦和其他存在关联关系的类别有关。

原因:在第三种情况下每一个模式类别都分别建立了一个判别函数。

$$G_{ij}(X) = g_i(X) - g_j(X)$$
$$G_{ik}(X) = g_i(X) - g_k(X)$$
$$G_{jk}(X) = g_j(X) - g_k(X)$$

不同的决策面函数之间可能存在关联。其中一些决策面函数可由其它一些决策面函数经线性组合得到。

§ 2.3 线性可分情况下的几何分类法

第三种情况和和第二种情况在内涵上的区别

例：三类问题

$$\left\{\begin{aligned} &\omega_1 : g_1(\boldsymbol{X}) \\ &\omega_2 : g_2(\boldsymbol{X}) \\ &\omega_3 : g_3(\boldsymbol{X}) \end{aligned}\right. \Rightarrow \omega_i / \omega_j, i = 1,2,3$$

$$\left\{\begin{aligned} &G_{12}(\boldsymbol{X}) = g_1(\boldsymbol{X}) - g_2(\boldsymbol{X}) \\ &G_{13}(\boldsymbol{X}) = g_1(\boldsymbol{X}) - g_3(\boldsymbol{X}) \\ &G_{23}(\boldsymbol{X}) = g_2(\boldsymbol{X}) - g_3(\boldsymbol{X}) \end{aligned}\right. \Rightarrow$$

$$\begin{aligned} &\boldsymbol{G}_{23}(\boldsymbol{X}) \\ &= \boldsymbol{g}_2(\boldsymbol{X}) - \boldsymbol{g}_3(\boldsymbol{X}) \\ &= (\boldsymbol{g}_1(\boldsymbol{X}) - \boldsymbol{g}_3(\boldsymbol{X})) - (\boldsymbol{g}_1(\boldsymbol{X}) - \boldsymbol{g}_2(\boldsymbol{X})) \\ &= \boldsymbol{G}_{13}(\boldsymbol{X}) - \boldsymbol{G}_{12}(\boldsymbol{X}) \end{aligned}$$

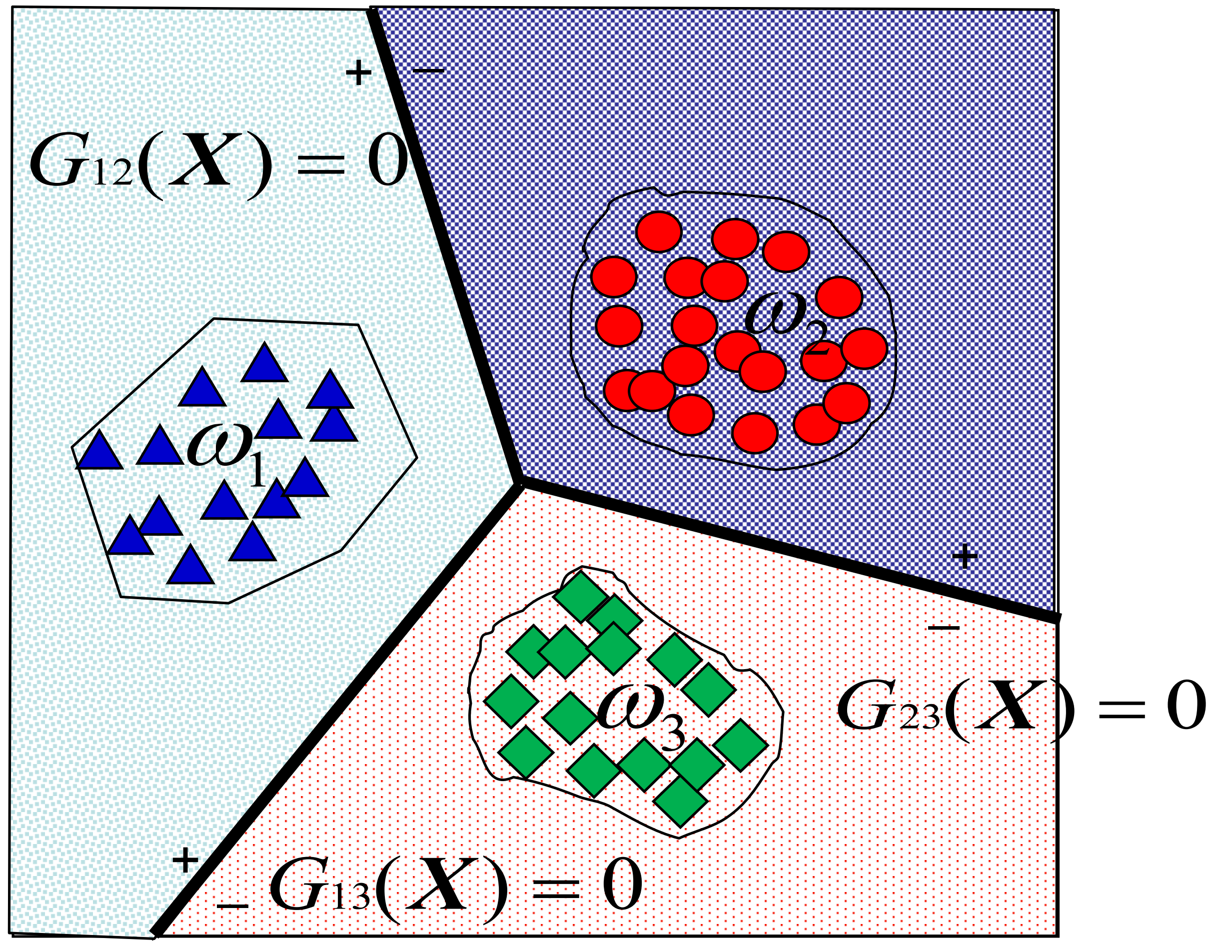
$\Downarrow$

三者之间不是相互独立的!

§ 2.3 线性可分情况下的几何分类法

在第三种情况下

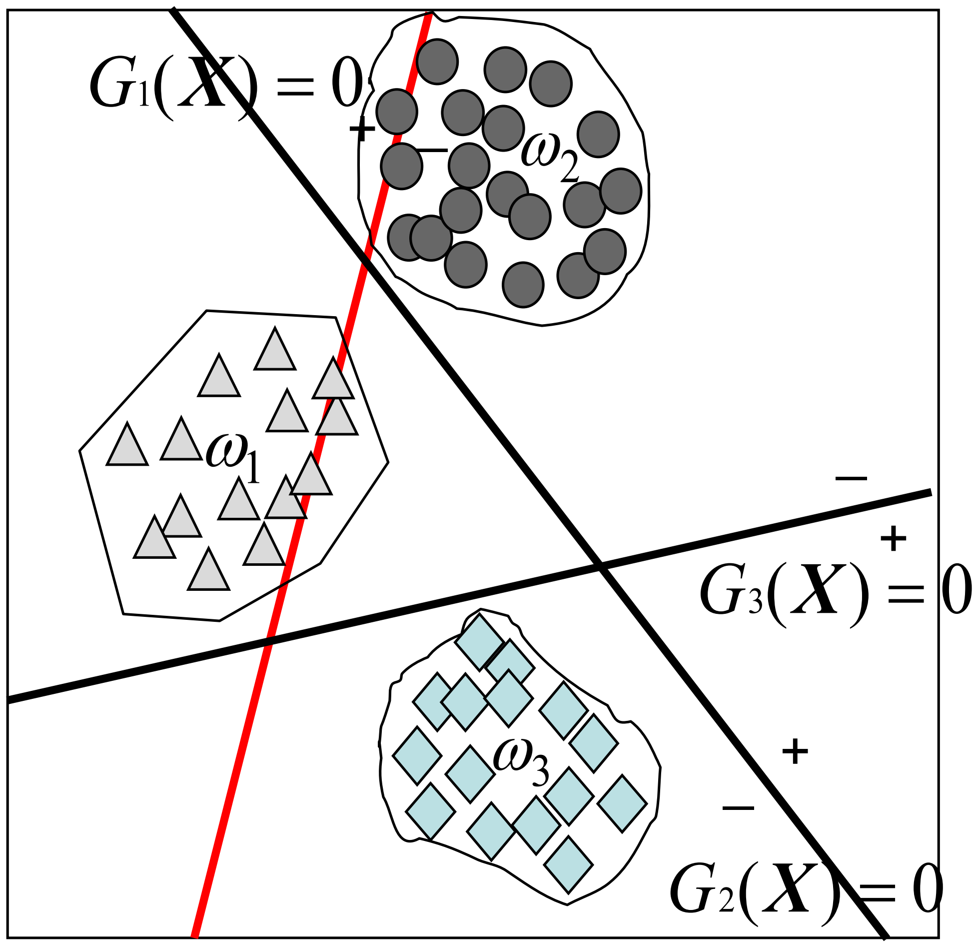
- N类分类问题中，独立决策面函数的个数为（N -1）。
- 只要给定的分类问题是线性可分的，则其决策域不存在不确定区域。



§ 2.3 线性可分情况下的几何分类法

三种情况之间的关系

- 在第二种情况下线性可分的样本集在第一种情况下未必是线性可分的。



§ 2.3 线性可分情况下的几何分类法

三种情况之间的关系

- 在第二种情况下线性可分的样本集在第一种情况下未必是线性可分的。
- 第一种情况下线性可分的样本集在第二种情况下也一定是线性可分的。
- 若在第三种情况下是线性可分的，则在第二种情况下也是线性可分的。
- 若在第二种情况下是线性可分的，则在第三种情况下不一定有解。

分类器的评价

分类错误率　**VS**　不确定区域

§ 2.3 线性可分情况下的几何分类法

线性判别函数的参数确定

线性分类器的数学表示

$$G(X) = W'^T X' \qquad \text{解权向量} \quad W^*$$

问题：确定权向量

如何确定权向量?

利用已预分类的训练样本集 $\mathscr{X} = \{X_1, X_2, \ldots, X_N\}$

ω_i / ω_j 两分问题

$$G(X) = W^T X \begin{cases} > 0 & \Rightarrow \quad X \in \omega_i \\ < 0 & \Rightarrow \quad X \in \omega_j \end{cases}$$

$$\mathscr{X} = \{X_1, X_2, \ldots, X_N\} \begin{cases} X_1, X_2, \ldots, X_{n_i} \in \omega_i \\ X_{n_i+1}, X_{n_i+2}, \ldots, X_N \in \omega_j \end{cases}$$

$$n_i + n_j = N$$

§ 2.3 线性可分情况下的几何分类法

线性判别函数的参数确定

$$\mathscr{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\dots,\boldsymbol{X}_N\}\begin{cases}\boldsymbol{X}_1,\boldsymbol{X}_2,\dots,\boldsymbol{X}_{n_i}\in\omega_i\\ \boldsymbol{X}_{n_i+1},\boldsymbol{X}_{n_i+2},\dots,\boldsymbol{X}_N\in\omega_j\end{cases}$$

根据预分类结果，$\boldsymbol{W}^*$ 满足

$$\boldsymbol{W}^T\boldsymbol{X}_k>0,\quad k=1,2,3,\dots,n_i$$

$$\boldsymbol{W}^T\boldsymbol{X}_k<0,\quad k=n_i+1,\dots,N \quad\Rightarrow\quad \boldsymbol{W}^T(-\boldsymbol{X}_k)>0,\quad k=n_i+1,\dots,N$$

规范化处理：各分量分别乘以-1，并将所得新向量标记为$\boldsymbol{X}_k$。

$$\Downarrow$$

$$\boldsymbol{W}^T\boldsymbol{X}_k>0,\quad k=1,2,\dots,N$$

决策面方程组 $\boldsymbol{W}^T\boldsymbol{X}_k=0,\quad k=1,2,\dots,N$

$$\boldsymbol{X}_k^T\boldsymbol{W}=0,\quad k=1,2,\dots,N \quad\Rightarrow$$

以$\boldsymbol{X}_k$为法向量的一组超平面

过 $\boldsymbol{W}^{n+1}$ 的原点

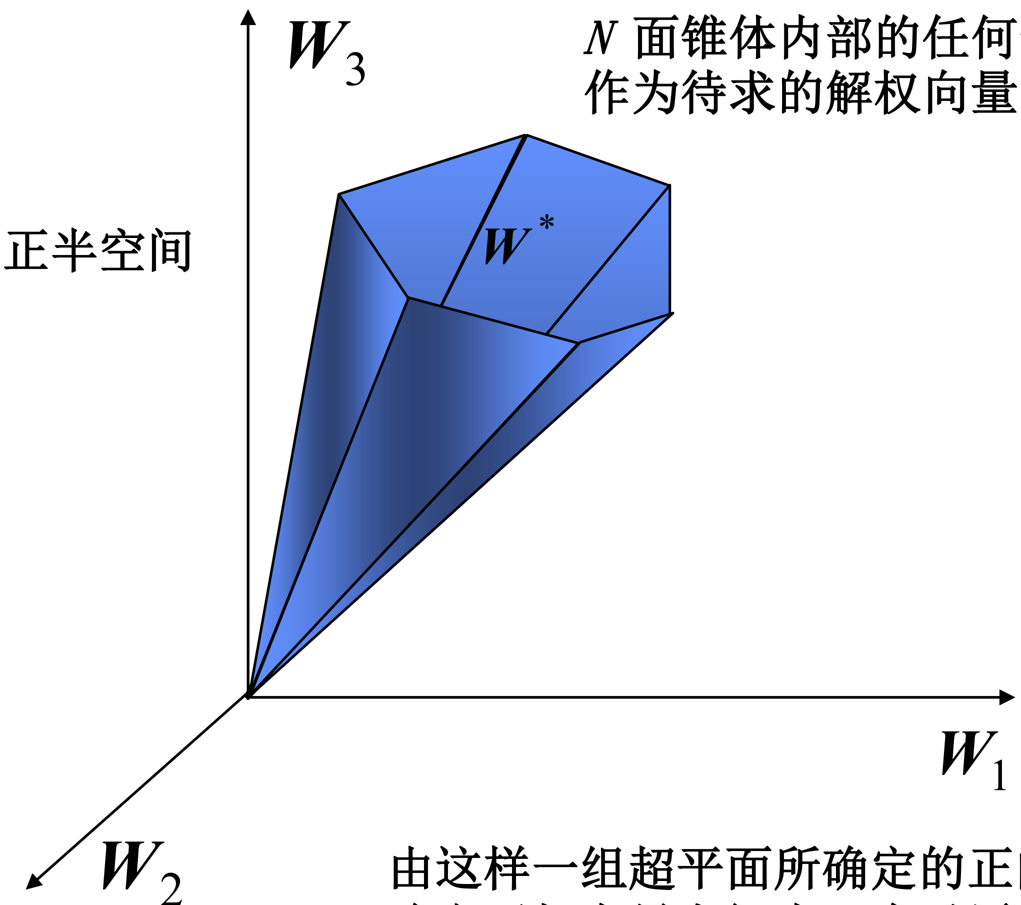
权向量全体构成一个(*n+1*)维的权向量空间 $\boldsymbol{W}^{n+1}$。

§ 2.3 线性可分情况下的几何分类法

线性判别函数的参数确定

N 面锥体内部的任何一点均有资格作为待求的解权向量。

X_k 指向的一侧为正半空间



由这样一组超平面所确定的正的半空间的交集确定了权向量空间中一个以原点为顶点的凸的 N 面超锥体区域。

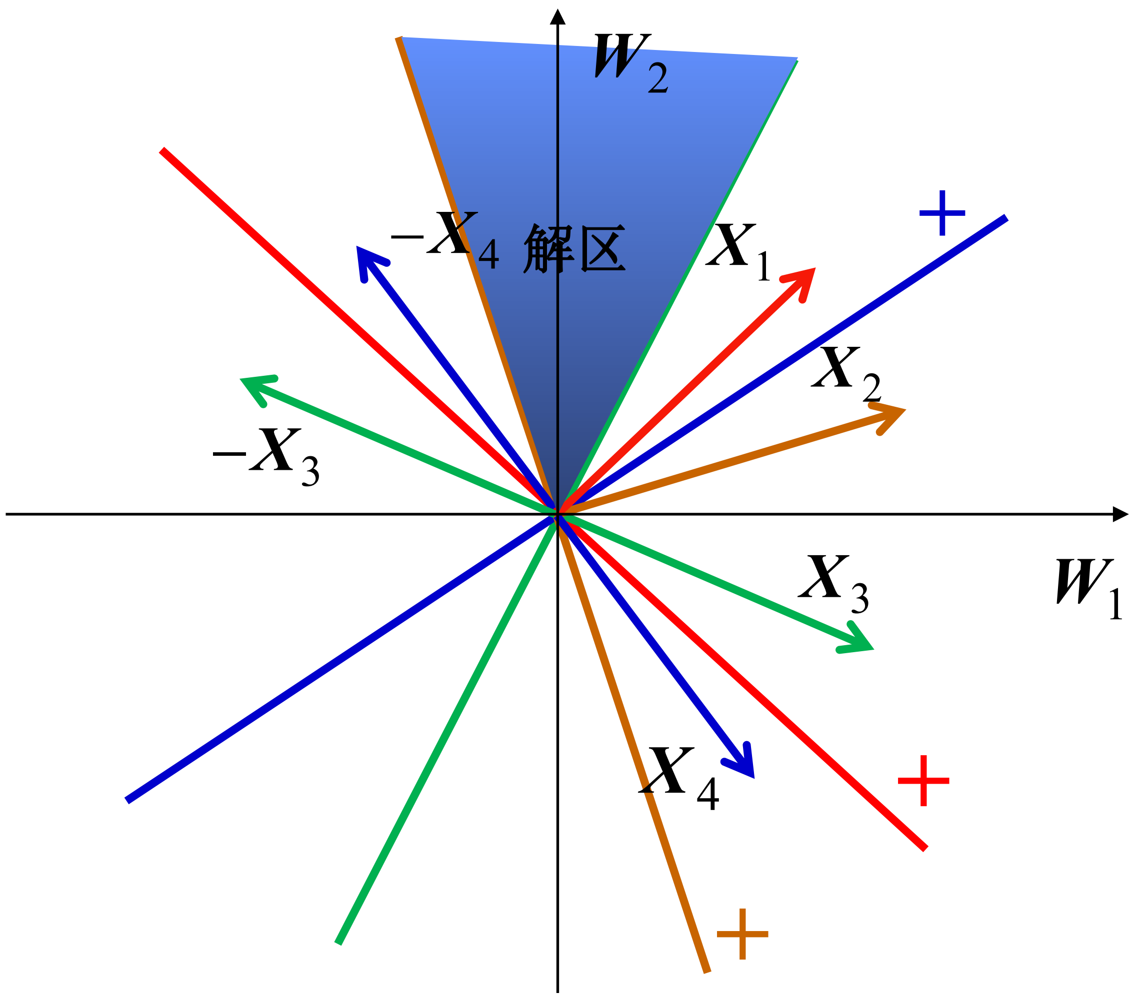
线性判别函数的参数确定

小结

- W^*一般不唯一。落在解区内的任何一点均有资格作为解权向量。
- 每一个训练样本都对解区提供了一个限制:训练样本越多，对解区的限制越严格，解区相对也越小。
- 离解区边界越远的解权向量,可靠性越高。

§ 2.3 线性可分情况下的几何分类法

线性判别函数的参数确定



$X_1, X_2 \in \omega_i$

$X_3, X_4 \in \omega_j$

§ 2.3 线性可分情况下的几何分类法

线性判别函数的参数确定

求解解权向量的步骤

Step1. 采集训练样本，构成预分类的样本集合。

$$\mathscr{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$$

Step2. 选用（或确定）一个准则函数J。

$$J = J(\boldsymbol{W}, \boldsymbol{X})$$

准则函数的极值解和最优分类判决相对应

Step3. 求准则函数的极值解得到待求的解权向量。

§ 2.3 线性可分情况下的几何分类法

感知器算法

准则函数的选取

其取值正比于所有被错分样本到决策面的距离之和

$$J_p(\boldsymbol{W},\boldsymbol{X}) = \sum_{\boldsymbol{X}\in\mathscr{X}_{\boldsymbol{W}}} -\boldsymbol{W}^T\boldsymbol{X}, \quad \boldsymbol{W} \neq \boldsymbol{0}$$

感知准则函数

$\mathscr{X}_{\boldsymbol{W}}$ …… 被误分的样本集合

采用样本规范化表示时，被误分是指 $\boldsymbol{W}^T\boldsymbol{X} \leq 0$

当且仅当 $\mathscr{X}_{\boldsymbol{W}}$为空集时，准则函数达到最小值

$$J_p(\boldsymbol{W},\boldsymbol{X}) = 0$$

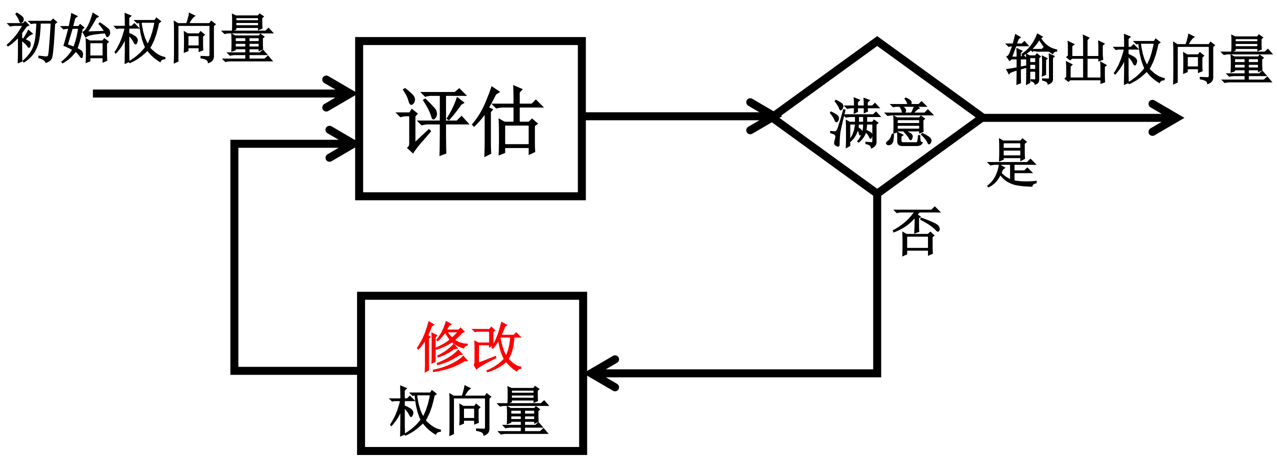
使准则函数达到最小值的权向量***W****即为所求解权向量

§ 2.3 线性可分情况下的几何分类法

感知器算法

如何求感知器准则函数的极值解W^*？

借助于优化技术 —— 固定增量的感知器算法



1. 如果所选权向量不能做到对所有输入训练样本正确分类，那么应该如何对其进行修改？
2. 迭代算法收敛吗？它在什么情况下收敛？如何判断其收敛性？

§ 2.3 线性可分情况下的几何分类法

感知器算法　ω_i / ω_j 两分问题

Step1. 赋初值：迭代步数 k=0，固定比例因子 $0 \le \rho \le 1$；

$\boldsymbol{W}(0)$，连续正确分类计数器 $N_c = 0$；

Step2. 读入训练样本集合 $\mathscr{X} = \{\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{N-1}\}$；

Step3. 取样本 $\boldsymbol{X} = \boldsymbol{X}_{[k]_N}$　$[k]_N = k \mod(N)$；

计算 $G(\boldsymbol{X}) = \boldsymbol{W}(k)^T \boldsymbol{X}$；

Step4. 修正权向量：

当 $\boldsymbol{X} \in \omega_i$ 时，

若 $G(\boldsymbol{X}) \le 0$，则 $\boldsymbol{W}(k+1) = \boldsymbol{W}(k) + \rho \boldsymbol{X}$，$N_c = 0$；

否则 $\boldsymbol{W}(k+1) = \boldsymbol{W}(k)$，$N_c += 1$；

当 $\boldsymbol{X} \in \omega_j$ 时，

若 $G(\boldsymbol{X}) \ge 0$，则 $\boldsymbol{W}(k+1) = \boldsymbol{W}(k) - \rho \boldsymbol{X}$，$N_c = 0$；

否则 $\boldsymbol{W}(k+1) = \boldsymbol{W}(k)$，$N_c += 1$；

Step5. 若 $N_c \ge N$，算法结束；否则 $k = k+1$，返回step3。

§ 2.3 线性可分情况下的几何分类法

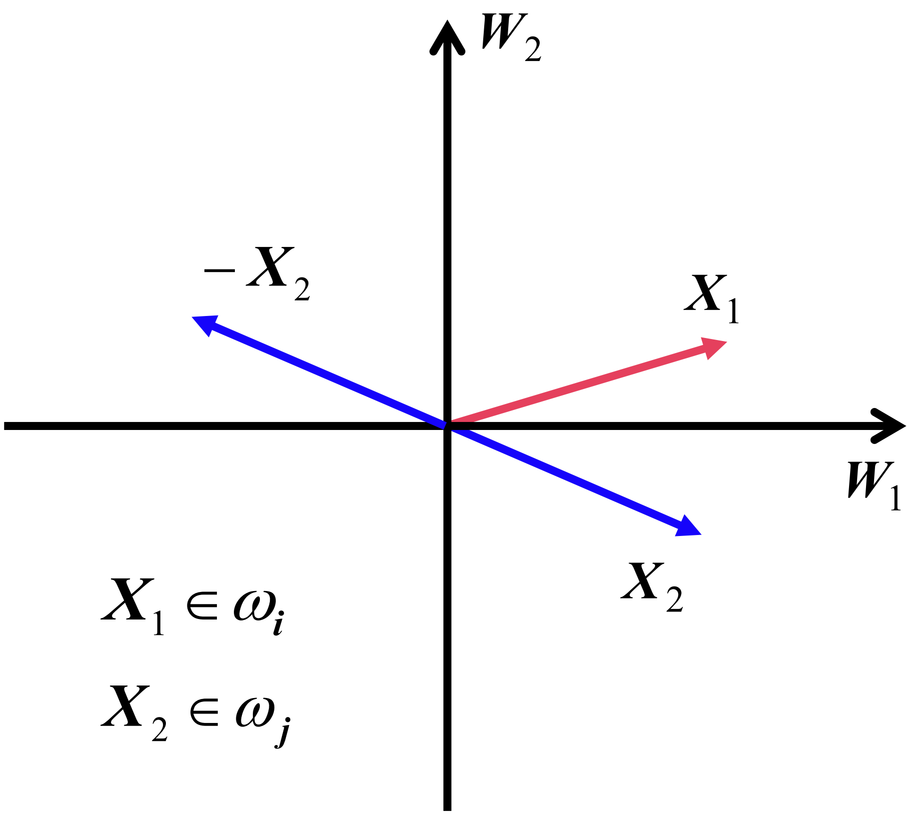
感知器算法

权向量更新规则

当 $X \in \omega_i$，$G(X) \le 0$ 时，
$W(k+1) = W(k) + \rho X$ → $W(1) = X_1$

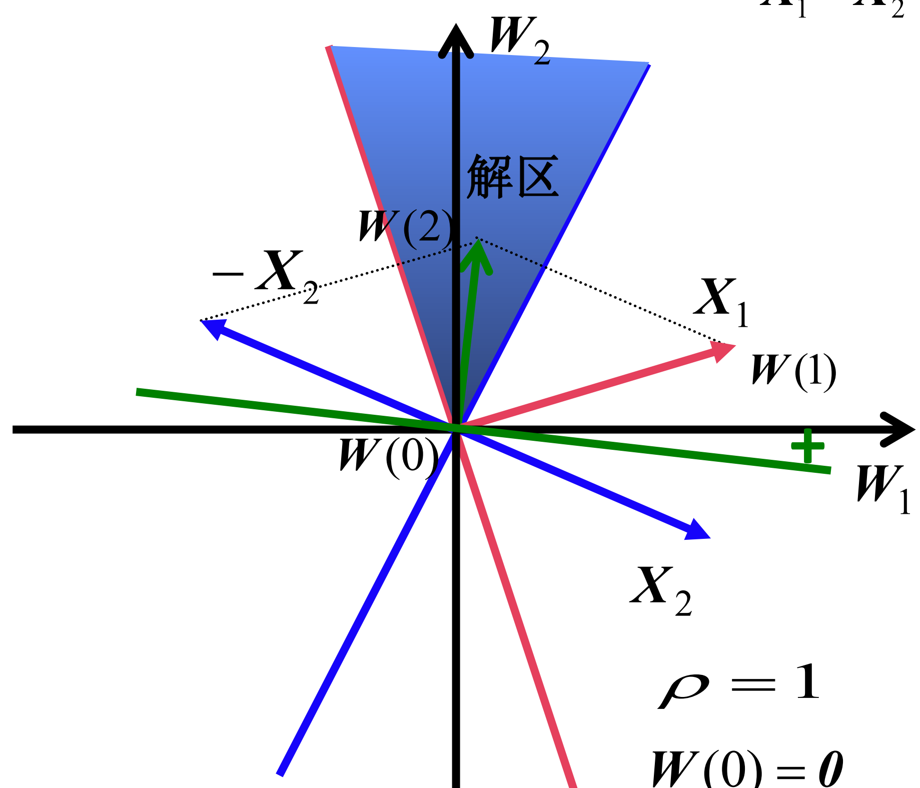
当 $X \in \omega_j$，$G(X) \ge 0$ 时，
$W(k+1) = W(k) - \rho X$ → $W(2) = W(1) - X_2 = X_1 - X_2$

举例



$X_1 \in \omega_i$

$X_2 \in \omega_j$



$\rho = 1$

$W(0) = 0$

§ 2.3 线性可分情况下的几何分类法

感知器算法

权向量更新规则

当 $X \in \omega_i$，$G(X) \le 0$ 时，
$$W(k+1) = W(k) + \rho X$$

当 $X \in \omega_j$，$G(X) \ge 0$ 时，
$$W(k+1) = W(k) - \rho X$$

修改权向量会影响下一次迭代的结果！

$$\begin{aligned} G^{k+1}(X) &= W(k+1)^T X \\ &= (W(k) + \rho X)^T X \\ &= W(k)^T X + \rho X^T X \end{aligned}$$

$$\rho X^T X \ge 0$$

$$G^k(X) = W(k)^T X \le 0$$

$$G^{k+1}(X) = G^k(X) + \rho X^T X$$

$$G^{k+1}(X) = G^k(X) + \text{正数}$$

§ 2.3 线性可分情况下的几何分类法

感知器算法

权向量更新规则

当 $X \in \omega_i$，$G(X) \le 0$ 时，
$W(k+1) = W(k) + \rho X$

当 $X \in \omega_j$，$G(X) \ge 0$ 时，
$W(k+1) = W(k) - \rho X$

$$G^{k+1}(X) = W(k+1)^T X$$
$$= (W(k) - \rho X)^T X$$
$$= W(k)^T X - \rho X^T X$$

$\rho X^T X \ge 0$

$W(k)^T X \ge 0$

$G^{k+1}(X) = G^k(X) - \rho X^T X$ ⟹ $G^{k+1}(X) = G^k(X) -$ 正数

§ 2.3 线性可分情况下的几何分类法

问题：什么问题是感知器算法无法胜任的？

- 不可分类的问题
- 非线性可分，但线性不可分的问题
- 多类分类问题直接解决不了

可以处理的问题为线性可分的两类问题

问题：所有线性可分两类问题都可完美解决？

§ 2.3 线性可分情况下的几何分类法

收敛性定理

如果训练样本集 $\mathscr{X}$ 是线性可分的，则基于固定增量法则的感知器算法经过有限次迭代后将收敛到正确的解权向量。

几个约定

- 采用加长向量表示
- 执行规范化处理：若 $X_m \in \omega_j$，则 $-X_m \to X_m$
- 选择步长因子ρ = 1

感知器算法权向量的更新规则

对 $\boldsymbol{X} \in \mathscr{X}$，若 $\boldsymbol{G}(\boldsymbol{X}) = \boldsymbol{W}^T \boldsymbol{X} \le 0$，则 $\boldsymbol{W} + \boldsymbol{X} \to \boldsymbol{W}$，否则 $\boldsymbol{W} \to \boldsymbol{W}$。

§ 2.3 线性可分情况下的几何分类法

收敛性定理的证明

经规范化处理后，$\mathscr{X}$线性可分是指：

$\exists W^*$，使对$\forall X \in \mathscr{X}$ 有，$G(X) = W^{*T} X > 0$ 换个说法

对预先任意给定的正常数 C_p,总可找到大正数C,使

$$(CW^*)^T X > C_p, \forall X \in \mathscr{X}$$

特别地，若令

$$M = Max(\|X\|^2)$$

则总可找到大的正数C，使

$$(CW^*)^T X = W_s^T X > M, \quad \forall X \in \mathscr{X}$$

根据定义，为解权向量。

$$W(k+1) = W(k)$$

故:只要证明每进行一次权向量的更新，都使$W(k+1)$接近W_s一个正的非无穷小量就可以了。

$$W(k+1) = W(k) + X$$

§ 2.3 线性可分情况下的几何分类法

收敛性定理的证明

考察：更新前后权向量与W_s之间欧氏距离的变化

$W(k+1)$与W_s之间的欧氏距离 $\|W_s - W(k+1)\|$

$W(k)$与W_s之间的欧氏距离 $\|W_s - W(k)\|$

$$\|W_s - W(k)\|^2 - \|W_s - W(k+1)\|^2$$

$$= (W_s - W(k))^T (W_s - W(k)) - (W_s - W(k+1))^T (W_s - W(k+1))$$

$$= -2W_s^T W(k) + W(k)^T W(k) + 2W_s^T W(k+1) - W(k+1)^T W(k+1)$$

$$= -2W(k)^T X + 2W_s^T X - X^T X$$

$$W(k+1) = W(k) + X$$

$$\geq M$$

$$W(k)^T X < 0$$

$$W_s^T X > M$$

$$X^T X = \|X\|^2 \leq M$$

$$W(k)^T X < 0$$

$$W_s^T X > M$$

$$X^T X = \|X\|^2 \leq M$$

∴每一次更新都使权向量接近W_s一个正的非无穷小量

即经有限次迭代后W将收敛于W_s！

中国科学技术大学 University of Science and Technology of China

§ 2.3 线性可分情况下的几何分类法

基于固定增量的感知器算法小结

算法简单，且当训练样本集线性可分时，算法经有限次迭代后可收敛到解权向量，但算法的收敛速度不理想。

有办法改善吗？

回顾：求解解权向量的步骤

Step1. 采集训练样本，构成预分类的样本集合。

$$\mathscr{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$$

Step2. 选用或确定一个准则函数J。

$$J = J(\boldsymbol{W}, \boldsymbol{X})$$

准则函数的极值解和最优分类判决相对应

Step3. 求准则函数的极值解得到待求的解权向量。

§ 2.3 线性可分情况下的几何分类法

梯度下降法

若干数学预备知识

$\boldsymbol{X}=(x_1,x_2,\dots x_n)^T$ **n维列向量**

$f(\boldsymbol{X})$ **n元标量函数**

$$\nabla_X f(\boldsymbol{X})=\begin{pmatrix}\dfrac{\partial f(\boldsymbol{X})}{\partial x_1}\\ \dfrac{\partial f(\boldsymbol{X})}{\partial x_2}\\ \mathrm{M}\\ \dfrac{\partial f(\boldsymbol{X})}{\partial x_n}\end{pmatrix}$$

$f(\boldsymbol{X})$对$\boldsymbol{X}$的梯度

$$= \ gradf(\boldsymbol{X})$$

梯度下降法

若干数学预备知识

$A=[a_{ij}]_{n\times m}$　　$n\times m$阶矩阵

$Y=(y_1,y_2,\dots y_n)^T$　　与X无关的n维列向量

$$\nabla_X(Y^TX)=\nabla_X(X^TY)=Y$$

$$\nabla_X(X^TAY)=AY$$

$$\nabla_X(Y^TAX)=A^TY$$

$$\nabla_X(X^TAX)=(A+A^T)X$$

当A为对称阵时，有：　$\nabla_X(X^TAX)=2AX$

§ 2.3 线性可分情况下的几何分类法

梯度下降法

若干数学预备知识

$$\boldsymbol{A}=[a_{ij}]_{n\times m} \qquad \boldsymbol{n\times m}\text{阶矩阵}$$

$$f(\boldsymbol{A}) \qquad \boldsymbol{n\times m}\text{元标量函数}$$

$$\frac{df(\boldsymbol{A})}{d\boldsymbol{A}}=\begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{12}} & \cdots & \frac{\partial f}{\partial a_{1m}} \\ \frac{\partial f}{\partial a_{21}} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \frac{\partial f}{\partial a_{n1}} & \cdots & \cdots & \frac{\partial f}{\partial a_{nm}} \end{bmatrix}$$

$$\frac{d\det\boldsymbol{A}}{d\boldsymbol{A}}=\det\boldsymbol{A}(\boldsymbol{A}^T)^{-1}$$

梯度下降法

若干数学预备知识

$$\frac{d\boldsymbol{A}(x)}{dx}=\begin{bmatrix}\frac{da_{11}}{dx} & \frac{da_{12}}{dx} & \cdots & \frac{da_{1m}}{dx}\\ \frac{da_{21}}{dx} & \cdots & \cdots & \vdots\\ \vdots & \cdots & \cdots & \vdots\\ \frac{da_{n1}}{dx} & \cdots & \cdots & \frac{da_{nm}}{dx}\end{bmatrix}$$

$$\frac{d}{dx}[\boldsymbol{A}(x)\boldsymbol{B}(x)]=\frac{d\boldsymbol{A}(x)}{dx}\boldsymbol{B}(x)+\boldsymbol{A}(x)\frac{d\boldsymbol{B}(x)}{dx}$$

§ 2.3 线性可分情况下的几何分类法

梯度下降法

若干数学预备知识

$$\left(\frac{d\boldsymbol{A}(x)}{dx}\right)^T = \frac{d\boldsymbol{A}^T(x)}{dx}$$

$$\frac{d}{dx}[\alpha\boldsymbol{A}(x)+\beta\boldsymbol{B}(x)] = \alpha\frac{d\boldsymbol{A}(x)}{dx}+\beta\frac{d\boldsymbol{B}(x)}{dx}$$

$$\frac{d\boldsymbol{A}^{-1}(x)}{dx} = -\boldsymbol{A}^{-1}(x)\frac{d\boldsymbol{A}(x)}{dx}\boldsymbol{A}^{-1}(x)$$

$$\frac{d}{dx}[\boldsymbol{A}(x)\boldsymbol{B}(x)] = \frac{d\boldsymbol{A}(x)}{dx}\boldsymbol{B}(x)+\boldsymbol{A}(x)\frac{d\boldsymbol{B}(x)}{dx}$$

§ 2.3 线性可分情况下的几何分类法

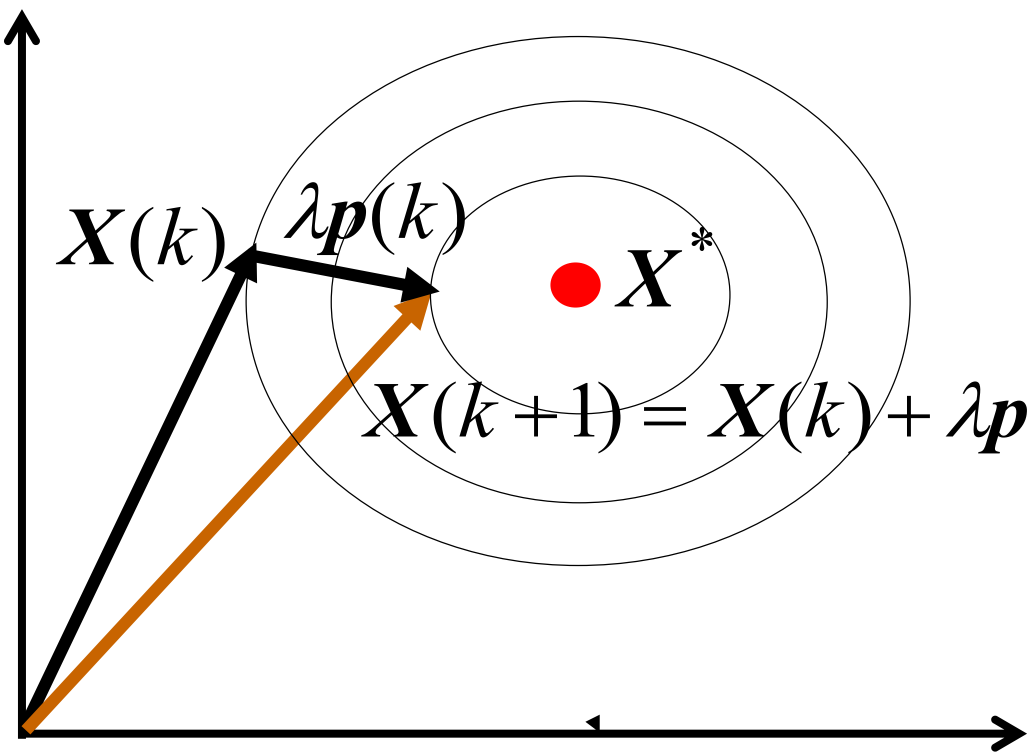
梯度下降法

无约束极值问题

目标函数

$$Minimise\ f(\boldsymbol{X}), \quad \boldsymbol{X} \in \mathscr{E}^n$$

➡ 确定使$f(X)$取得极值的X^*　　解析解 ➡ 迭代求解



$$\boldsymbol{X}(k+1) = \boldsymbol{X}(k) + \lambda\boldsymbol{p}(k), \qquad \lambda \geq 0$$

§ 2.3 线性可分情况下的几何分类法

梯度下降法

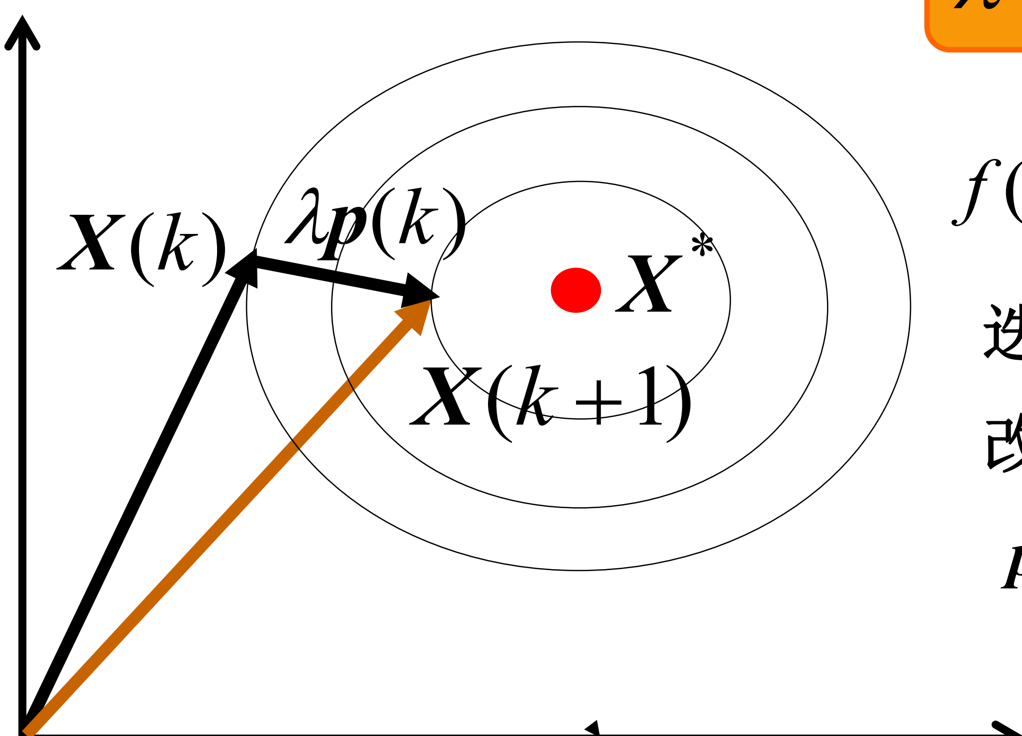
$$\lim_{\lambda \to 0} \frac{o(\lambda)}{\lambda} = 0$$

$$\boldsymbol{f}(\boldsymbol{X}(k+1)) = \boldsymbol{f}(\boldsymbol{X}(k) + \lambda \boldsymbol{p}(k))$$

$$= \boldsymbol{f}(\boldsymbol{X}(k)) + \lambda \nabla \boldsymbol{f}(\boldsymbol{X}(k))^T \boldsymbol{p}(k) + \boldsymbol{o}(\lambda)$$

$$\boxed{\lambda \nabla f(\boldsymbol{X}(k))^T \boldsymbol{p}(k) < 0}$$

$$f(\boldsymbol{X}(k) + \lambda \boldsymbol{p}(k)) < f(\boldsymbol{X}(k))$$



迭代后$f(X)$得到改善

改善程度随 $\boldsymbol{p}(k)$变化

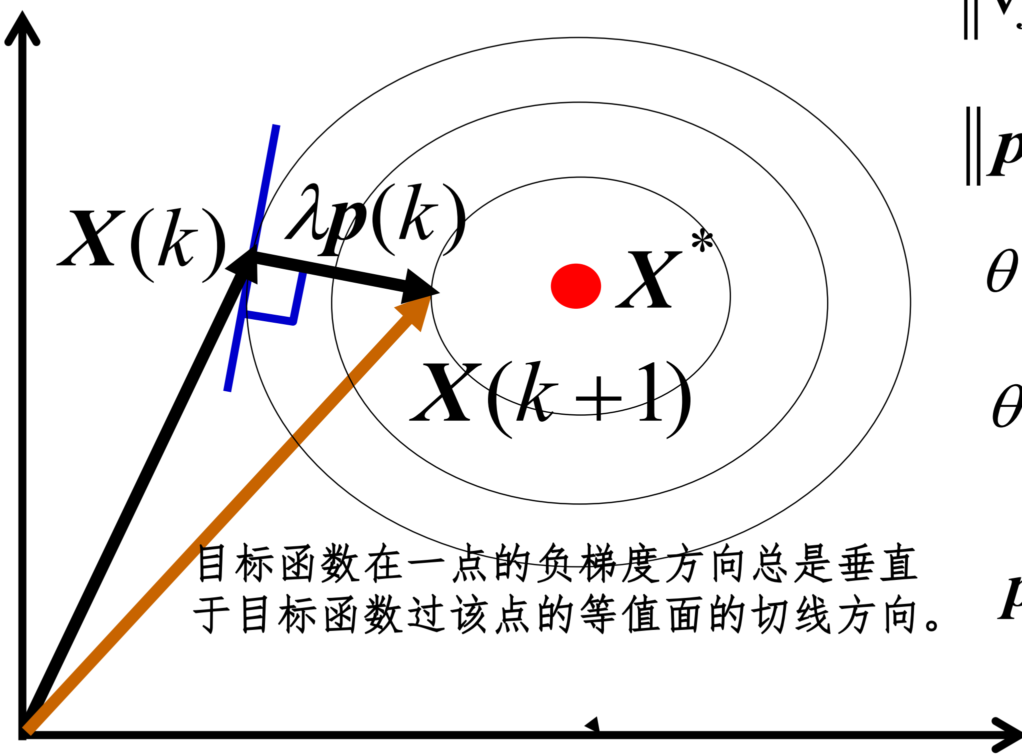
$\boldsymbol{p}(k)$存在最大改善程度方向？

§ 2.3 线性可分情况下的几何分类法

梯度下降法

目标函数值改善程度的计算

$$\nabla f(\boldsymbol{X}(k))^T \boldsymbol{p}(k) = \|\nabla f(\boldsymbol{X}(k))\| \|\boldsymbol{p}(k)\| \cos\theta$$



$\|\nabla f(\boldsymbol{X}(k))\|$ 梯度幅度，定值。

$\|\boldsymbol{p}(k)\|$ 向量幅度，定值。

θ $\nabla f(\boldsymbol{X}(k))$和$\boldsymbol{p}(k)$之间的夹角。

$\theta = 180°$ 时，有最小值。

$\boldsymbol{p}(k) = -\nabla f(\boldsymbol{X}(k))$ 负梯度方向

§ 2.3 线性可分情况下的几何分类法

梯度下降法

步长λ的选择问题

- 试探法　　步长值由大到小试探

$$f(\boldsymbol{X}(k)-\lambda\nabla f(\boldsymbol{X}(k)))<f(\boldsymbol{X}(k))$$

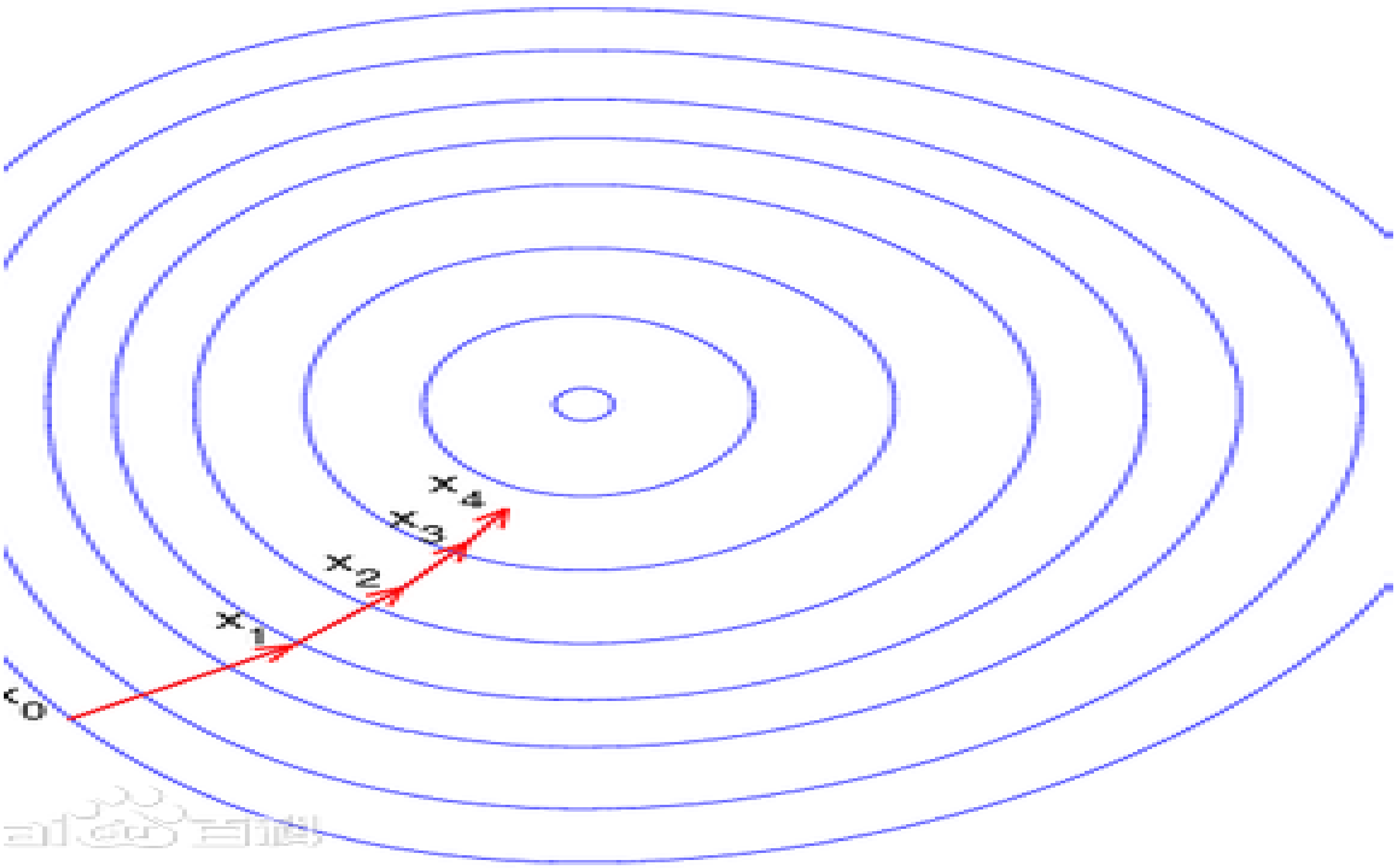
- 最佳步长法　沿负梯度方向作一维搜索并选择使目标函数值取最小值的步长

负梯度方向是理想的搜索方向?

NO! 只有当目标函数的等值面为超球面时，负梯度方向才指向极小点。当等值面不是超球面时，沿负梯度方向搜索一般并不能保证一步到位找到极小点。

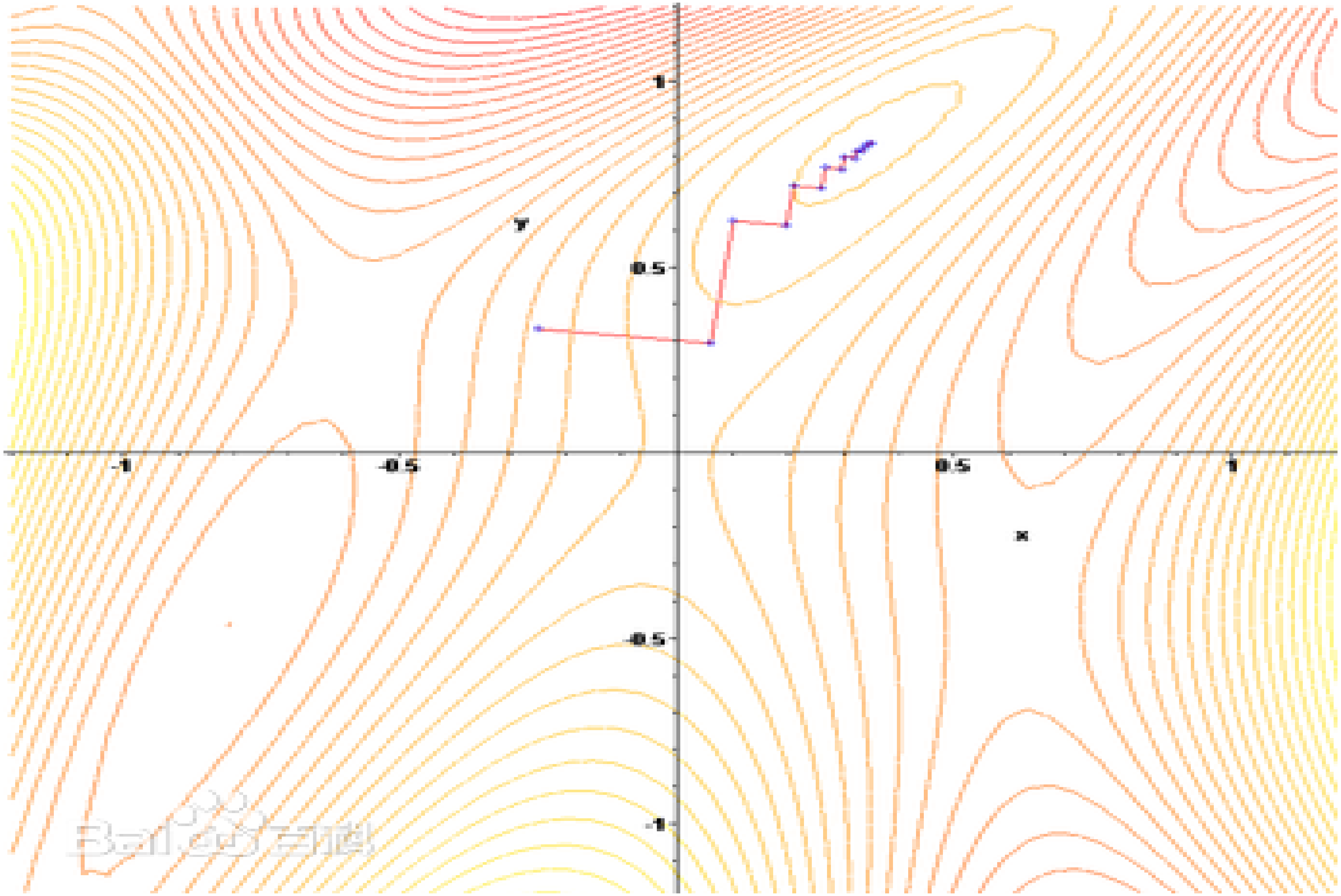
§ 2.3 线性可分情况下的几何分类法

梯度下降法

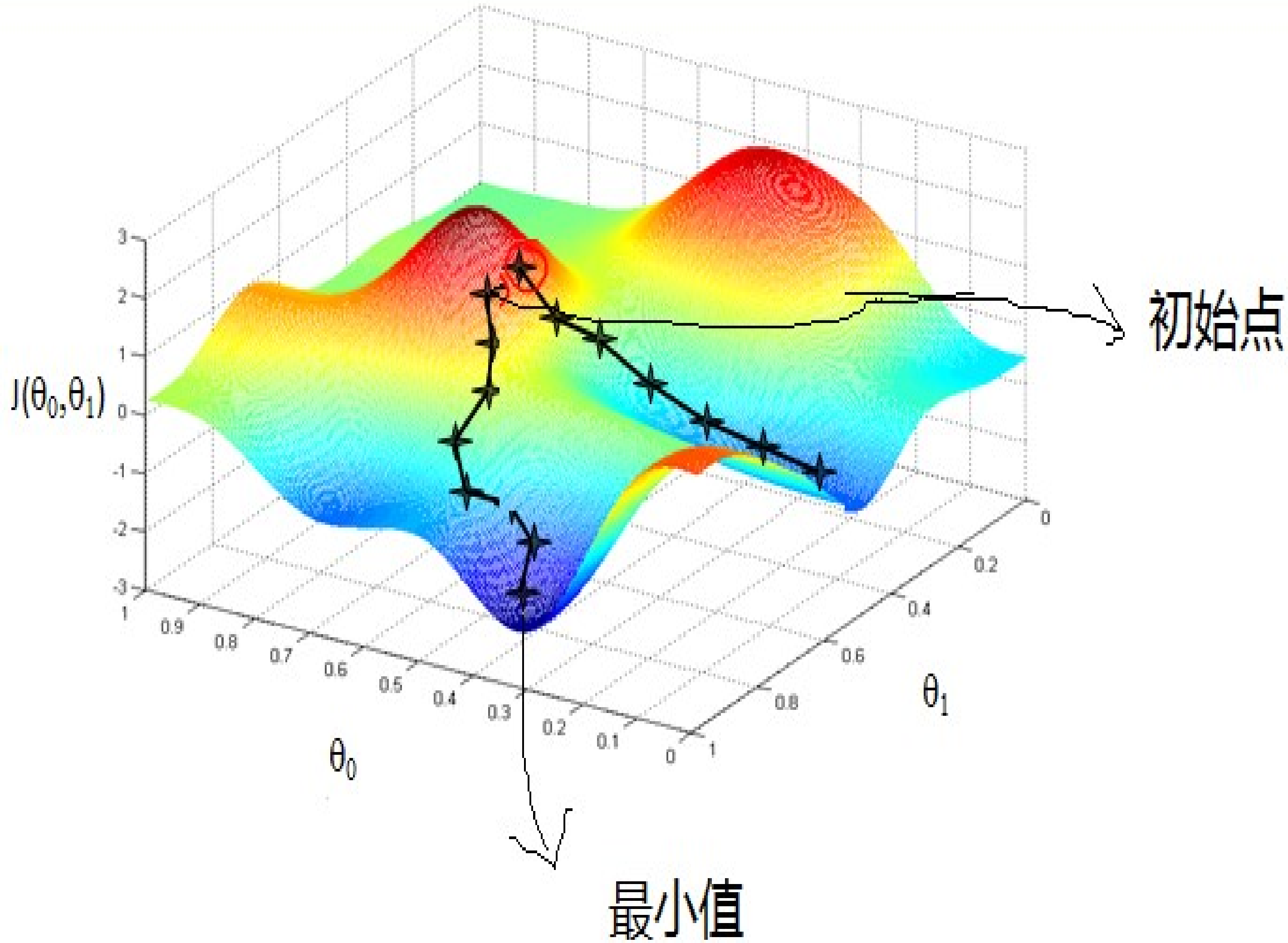


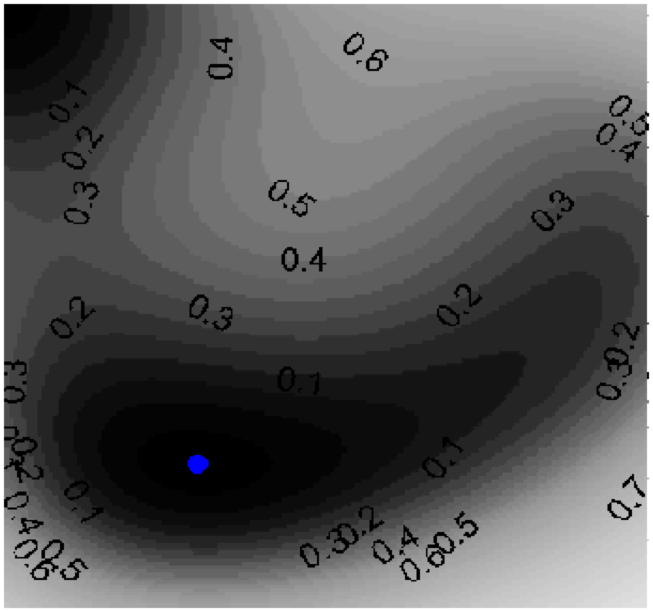
§ 2.3 线性可分情况下的几何分类法

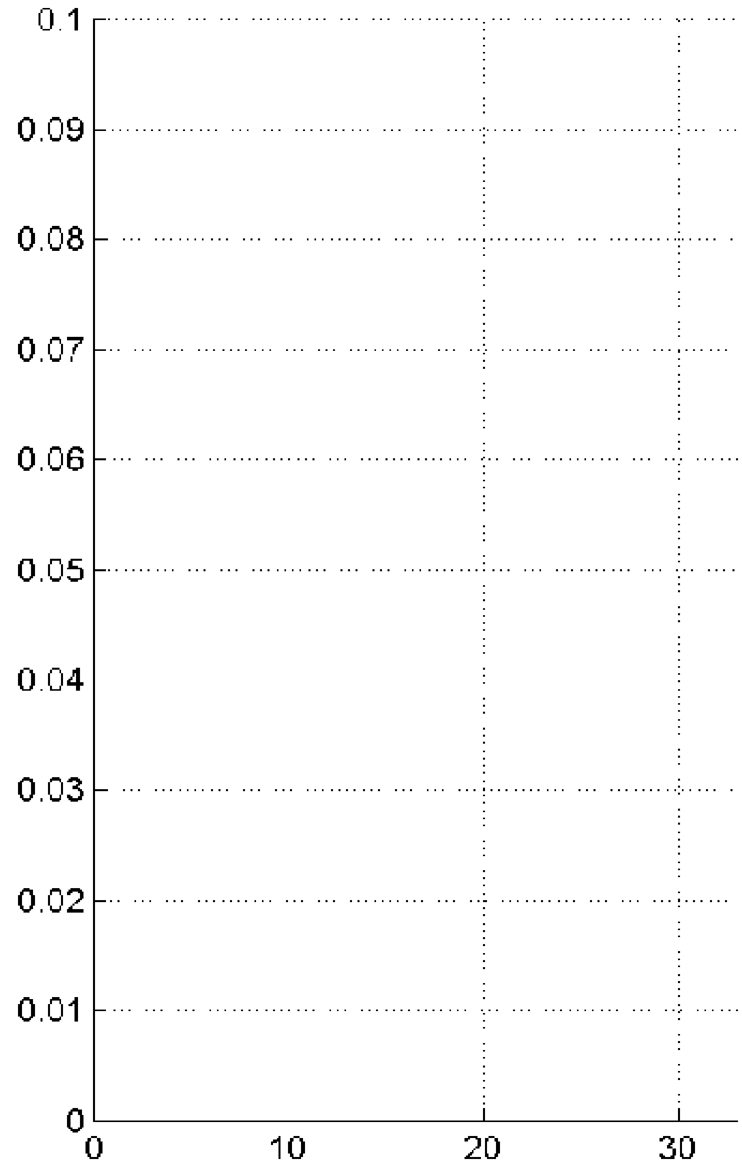
梯度下降法



梯度下降法






0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30

用梯度下降法求解解权向量

$$J(\boldsymbol{W},\boldsymbol{X}) = c(|\boldsymbol{W}^T\boldsymbol{X}| - \boldsymbol{W}^T\boldsymbol{X})$$

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(k) - \rho\nabla J(\boldsymbol{W}(k),\boldsymbol{X}), \qquad \rho > 0$$

感知器算法和梯度下降法之间的关系

$$J(\boldsymbol{W},\boldsymbol{X}) = \frac{1}{2}(|\boldsymbol{W}^T\boldsymbol{X}| - \boldsymbol{W}^T\boldsymbol{X})$$

$$= \frac{1}{2}(\mathrm{sgn}(\boldsymbol{W}^T\boldsymbol{X})\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{W}^T\boldsymbol{X})$$

$$\mathrm{sgn}(\boldsymbol{W}^T\boldsymbol{X}) = \begin{cases} 1 & \text{若 } \boldsymbol{W}^T\boldsymbol{X} > 0 \\ -1 & \text{否则} \end{cases}$$

用梯度下降法求解解权向量

$$\nabla J(\boldsymbol{W},\boldsymbol{X})=\frac{\partial J}{\partial \boldsymbol{W}}=\frac{1}{2}(\mathrm{sgn}(\boldsymbol{W}^T\boldsymbol{X})\boldsymbol{X}-\boldsymbol{X})$$

$$\boldsymbol{W}(k+1)=\boldsymbol{W}(k)-\frac{\rho}{2}(\mathrm{sgn}(\boldsymbol{W}(k)^T\boldsymbol{X})\boldsymbol{X}-\boldsymbol{X}),\qquad \lambda\geq 0$$

$$=\begin{cases}\boldsymbol{W}(k) & \text{若}\,\boldsymbol{W}(k)^T\boldsymbol{X}>0\\ \boldsymbol{W}(k)+\rho\boldsymbol{X} & \text{否则}\end{cases}$$

当ρ>0为常数时，更新公式与基于固定增量的感知器算法完全一致。

当ρ>0的取值随迭代步数而变时，称为可变增量法。

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

感知器算法和梯度下降法的缺点：

不能对训练样本集是否线性可分作出判断。

希望

- 能对输入训练样本集是否线性可分作出判断；
- 训练样本集线性不可分时，能够出某种准则下的最优解。例如，被误分样本个数最少。

➡ 最小二乘法求W的伪逆解

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

最小二乘法求$\boldsymbol{W}$的伪逆解

$$\mathscr{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_N\}\begin{cases}\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_{n_i}\in\omega_i\\ \boldsymbol{X}_{n_i+1},\boldsymbol{X}_{n_i+2},\ldots,\boldsymbol{X}_N\in\omega_j\end{cases}$$

根据预分类结果，$\boldsymbol{W}^*$满足

$$\boldsymbol{W}^T\boldsymbol{X}_k>0,\quad k=1,2,3,\ldots,n_i\qquad \boldsymbol{W}^T(-\boldsymbol{X}_k)>0,\quad k=n_i+1,\ldots,N$$

$$[\boldsymbol{X}]=\begin{bmatrix}\boldsymbol{X}_1^{\mathrm{T}}\\ \vdots\\ \boldsymbol{X}_{n_i}^{\mathrm{T}}\\ \boldsymbol{X}_{n_i+1}^{\mathrm{T}}\\ \vdots\\ \boldsymbol{X}_N^{\mathrm{T}}\end{bmatrix}=\begin{bmatrix}x_{1,1} & x_{1,2} & \cdots & x_{1,d} & 1\\ \vdots & & & & \\ x_{n_i,1} & x_{n_i,2} & \cdots & x_{n_i,d} & 1\\ -x_{n_i+1,1} & -x_{n_i+1,2} & \cdots & -x_{n_i+1,d} & -1\\ \vdots & & & & \\ -x_{N,1} & -x_{N,2} & \cdots & -x_{N,d} & -1\end{bmatrix}$$

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

最小二乘法求*W*的伪逆解

所有的分量值均大于0

$$[X]W > 0 \Rightarrow [X]W = \boldsymbol{b}$$

平方误差准则函数

定义 $\boldsymbol{e} = [X]W - \boldsymbol{b}$

$$J_s(W,[X],\boldsymbol{b}) = \|\boldsymbol{e}\|^2 = \|[X]W - \boldsymbol{b}\|^2 = \sum_{i=1}^{N}(W^T X_i - b_i)^2$$

极小化平方误差准则函数

$$\nabla_W J_s(W,[X],\boldsymbol{b}) = \frac{\partial J_s(W,[X],\boldsymbol{b})}{\partial W} = \frac{\partial \|[X]W - \boldsymbol{b}\|^2}{\partial W}$$

$$= \frac{\partial}{\partial W}\left(([X]W - \boldsymbol{b})^{\mathrm{T}}([X]W - \boldsymbol{b})\right) = \frac{\partial}{\partial W}\left(W^{\mathrm{T}}[X]^{\mathrm{T}}[X]W - 2W^{\mathrm{T}}[X]^{\mathrm{T}}\boldsymbol{b}\right)$$

$$= 2[X]^{\mathrm{T}}([X]W - \boldsymbol{b}) = 0$$

$$\Rightarrow [X]^{\mathrm{T}}[X]W = [X]^{\mathrm{T}}\boldsymbol{b}$$

$$\nabla_X(X^T AX) = (A + A^T)X$$

$$\nabla_X(X^T AY) = AY$$

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

最小二乘法求W的伪逆解

$$[X]^{\mathrm{T}}[X]W=[X]^{\mathrm{T}}b$$

$[X]^T[X]$何物？

$$[X]=\begin{bmatrix} X_1^T \\ \vdots \\ X_{n_i}^T \\ X_{n_i+1}^T \\ \vdots \\ X_N^T \end{bmatrix}=\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} & 1 \\ \vdots & & & & \\ x_{n_i,1} & x_{n_i,2} & \cdots & x_{n_i,d} & 1 \\ -x_{n_i+1,1} & -x_{n_i+1,2} & \cdots & -x_{n_i+1,d} & -1 \\ \vdots & & & & \\ -x_{N,1} & -x_{N,2} & \cdots & -x_{N,d} & -1 \end{bmatrix}$$

答案：*(d+1)* ×*(d+1)*的方阵。

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

最小二乘法求*W*的伪逆解

$$[X]^{\mathrm{T}}[X]\boldsymbol{W}=[X]^{\mathrm{T}}\boldsymbol{b}$$

如果 $[X]^T[X]$ 非奇异，则

$$\boldsymbol{W}=[[X]^T[X]]^{-1}[X]^T\boldsymbol{b}=[X]^{\#}\boldsymbol{b}$$

$$[X]^{\#}=[[X]^T[X]]^{-1}[X]^T \quad \text{伪逆}$$

结论：
- *W*的伪逆解依赖于*b*
- 误差向量*e*也依赖于*b*。

$$\boldsymbol{e}=[X]\boldsymbol{W}-\boldsymbol{b}=[X][X]^{\#}\boldsymbol{b}-\boldsymbol{b}=([X][X]^{\#}-\boldsymbol{I})\boldsymbol{b}$$

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

关于伪逆解W的几个结论

- 即使在训练样本集线性可分的情况下，对于某些b的选择，W的伪逆解也不一定构成正确的解权向量。

- 由于W的伪逆解依赖于b，不同b所导致的W对已知样本的误分情况是不一样的。

特别是某个b导致的伪逆解对已知训练样本的 分类结果不理想,并不意味着不存在其它b导致的伪逆解可对已知训练样本做到正确分类。

- 伪逆解是在b一定情况下极小化平方误差准则函数的结果。不管是在什么情况下，都会提供一个对分类有益的、最小二乘意义上的最优解。

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

针对伪逆解W的缺点，对其进行改造：

增加约束条件

$$W^{T}X_{i}-b_{i}\geq 0,\qquad i=1,2,\ldots,N \Rightarrow [X]W-b\geq 0$$

$\Downarrow$

LMSE（Least Mean Square Error）算法

Ho-Kashyap算法

H-K算法

最小平方误差法

LMSE（Least Mean Square Error）算法

- 求$\boldsymbol{W}$的伪逆解。
- 对$\boldsymbol{W}$和$\boldsymbol{b}$同时进行迭代优化，得到真正可用的$\boldsymbol{W}$。

用梯度下降法求$\boldsymbol{b}$的迭代解

增量向量

$$\boldsymbol{b}(k+1)=\boldsymbol{b}(k)+\delta\boldsymbol{b}(k)$$

$$=\boldsymbol{b}(k)-\rho\nabla_{\boldsymbol{b}}\boldsymbol{J}_{\mathrm{S}}(\boldsymbol{W},[\boldsymbol{X}],\boldsymbol{b})\Big|_{\boldsymbol{b}=\boldsymbol{b}(k)}$$

$$=\boldsymbol{b}(k)-\rho\frac{\partial \boldsymbol{J}_{\mathrm{S}}(\boldsymbol{W},[\boldsymbol{X}],\boldsymbol{b})}{\partial \boldsymbol{b}}\Big|_{\boldsymbol{b}=\boldsymbol{b}(k)}$$

最小平方误差法

LMSE（Least Mean Square Error）算法

$$b(k+1) = b(k) - \rho \frac{\partial J_s(W,[X],b)}{\partial b}\Big|_{b=b(k)}$$

$$\frac{\partial J_s(W,[X],b)}{\partial b}\Big|_{b=b(k)}$$

$$= \frac{\partial}{\partial b}\left(([X]W - b)^{\mathrm{T}}([X]W - b)\right)\Big|_{b=b(k)}$$

$$= \frac{\partial}{\partial b}\left(([X]W)^{\mathrm{T}}[X]W - ([X]W)^{\mathrm{T}} b - b^{\mathrm{T}}[X]W + b^{\mathrm{T}} b\right)\Big|_{b=b(k)}$$

$$= \frac{\partial}{\partial b}\left(([X]W)^{\mathrm{T}}[X]W - 2([X]W)^{\mathrm{T}} b + b^{\mathrm{T}} I b\right)\Big|_{b=b(k)}$$

$$= -2([X]W(k) - b(k))$$

$$\nabla_X(X^T A X) = (A + A^T)X$$

$$\nabla_X(Y^T X) = \nabla_X(X^T Y) = Y$$

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

LMSE（Least Mean Square Error）算法

$$\boldsymbol{b}(k+1)=\boldsymbol{b}(k)+\delta\boldsymbol{b}(k)$$

$$=\boldsymbol{b}(k)-\rho\frac{\partial \boldsymbol{J}_s(\boldsymbol{W},[\boldsymbol{X}],\boldsymbol{b})}{\partial \boldsymbol{b}}\bigg|_{\boldsymbol{b}=\boldsymbol{b}(\boldsymbol{k})}$$

$$=\boldsymbol{b}(k)+2\rho\left([\boldsymbol{X}]\boldsymbol{W}(k)-\boldsymbol{b}(k)\right)$$

$$\boldsymbol{b}(k+1)\geq\boldsymbol{b}(k)>\boldsymbol{0}$$

$$[\boldsymbol{X}]\boldsymbol{W}(k)-\boldsymbol{b}(k)\geq 0$$

$$\delta\boldsymbol{b}(\boldsymbol{k})=\begin{cases}0 & \text{若 } [\boldsymbol{X}]\boldsymbol{W}(\mathrm{k})-\boldsymbol{b}(\mathrm{k})\leq\boldsymbol{0}\\ 2\rho\left([\boldsymbol{X}]\boldsymbol{W}(\mathrm{k})-\boldsymbol{b}(\mathrm{k})\right) & \text{若 } [\boldsymbol{X}]\boldsymbol{W}(\mathrm{k})-\boldsymbol{b}(\mathrm{k})>\boldsymbol{0}\end{cases}$$

由定义 $\boldsymbol{e}(k)=[\boldsymbol{X}]\boldsymbol{W}(k)-\boldsymbol{b}(k)$

有表达 $\delta\boldsymbol{b}(k)=\rho\left(\boldsymbol{e}(k)+\left|\boldsymbol{e}(k)\right|\right)$

所有分量取绝对值后形成的新向量

§ 2.3 线性可分情况下的几何分类法

最小平方误差法

LMSE（Least Mean Square Error）算法

迭代公式

$$\delta \boldsymbol{b}(k)=\rho(\boldsymbol{e}(k)+|\boldsymbol{e}(k)|)$$

$$\Rightarrow \quad \boldsymbol{b}(k+1)=\boldsymbol{b}(k)+\delta \boldsymbol{b}(k)$$

$$\begin{aligned}\boldsymbol{W}(k+1)&=[\boldsymbol{X}]^{\#}\boldsymbol{b}(k+1)\\&=[\boldsymbol{X}]^{\#}(\boldsymbol{b}(k)+\delta \boldsymbol{b}(k))\\&=\boldsymbol{W}(k)+[\boldsymbol{X}]^{\#}\delta \boldsymbol{b}(k)\end{aligned}$$

$$\Rightarrow \quad \begin{cases}\boldsymbol{e}(k)=[\boldsymbol{X}]\boldsymbol{W}(k)-\boldsymbol{b}(k)\\ \delta \boldsymbol{b}(k)=\rho(\boldsymbol{e}(k)+|\boldsymbol{e}(k)|)\\ \boldsymbol{b}(k+1)=\boldsymbol{b}(k)+\delta \boldsymbol{b}(k)\\ \boldsymbol{W}(k+1)=\boldsymbol{W}(k)+[\boldsymbol{X}]^{\#}\delta \boldsymbol{b}(k)\end{cases}$$

§ 2.3 线性可分情况下的几何分类法

最小平方误差法：LMSE（Least Mean Square Error）算法

Step1. 由训练样本集构建 $[X]$；并计算 $[X]^{\#} = [[X]^T[X]]^{-1}[X]^T$

Step2. 赋初值： $k = 0,\ 0 \le \rho \le 1$

$\boldsymbol{b}(0) =$ 任一正向量, $\boldsymbol{W}(0) = [\boldsymbol{X}]^{\#}\boldsymbol{b}(0)$ 。

Step3. 计算 $\boldsymbol{e}(k) = [\boldsymbol{X}]\boldsymbol{W}(k) - \boldsymbol{b}(k)$ 。

Step4. 根据$\boldsymbol{e}(k)$ 取值情况作进一步处理：

4.1 若各分量取值全部大于或等于0，则判定算法完成；

4.2 若各分量取值有正有负，则继续下一步骤；

4.3 若各分量取值小于或等于0，但不是全部为0，则中止迭代，判定训练样本集不是线性可分的。

Step5. 完成更新计算

$$
\begin{cases}
\delta\boldsymbol{b}(k) = \rho(\boldsymbol{e}(k) + |\boldsymbol{e}(k)|) \\
\boldsymbol{b}(k+1) = \boldsymbol{b}(k) + \delta\boldsymbol{b}(k) \\
\boldsymbol{W}(k+1) = \boldsymbol{W}(k) + [\boldsymbol{X}]^{\#}\delta\boldsymbol{b}(k)
\end{cases}
$$

Step6. 置$k=k+1$，返回**step3**。

§ 2.3 线性可分情况下的几何分类法

最小平方误差法：LMSE（Least Mean Square Error）算法

Step4的进一步讨论

$e(k)$：

- 各分量取值≥0 ⇒ $e(k)=[X]W(k)-b(k)\geq 0$
- 各分量取值有正有负　情况不定
- 各分量取值≤ 0，但$e(k)\neq 0$ ⇒ 训练样本集线性不可分

$e(k)+|e(k)|=0$ ⇒ W和b在其后将保持不变！

训练样本集线性可分吗？

$[X]W(k)-b(k)\leq 0$

从逻辑层面看，没有答案！

$$\delta b(k)=\rho(e(k)+|e(k)|)$$
$$b(k+1)=b(k)+\delta b(k)$$
$$W(k+1)=W(k)+[X]^{\#}\delta b(k)$$

最小平方误差法：LMSE（Least Mean Square Error）算法

结论：当$e(k)$各分量取值≤ 0，但$e(k)\ne 0$时，训练样本集线性不可分 。

[证明] 用反证法。

假定训练样本集线性可分，则根据定义，可知：

存在 $\hat{W}(k)$ 和 $\hat{b}(k)>0$，使得

$$[X]\hat{W}(k)=\hat{b}(k)$$

成立。

$$[X]^T[X]\hat{W}(k)=[X]^T\hat{b}(k)$$

$$[X]^T\left([X]\hat{W}(k)-\hat{b}(k)\right)=0$$

$$[X]^T\hat{e}(k)=0$$

$$\hat{e}^T(k)[X]=0^T$$

$$\hat{e}^T(k)[X]\hat{W}(k)=\hat{e}^T(k)\hat{b}(k)=0$$

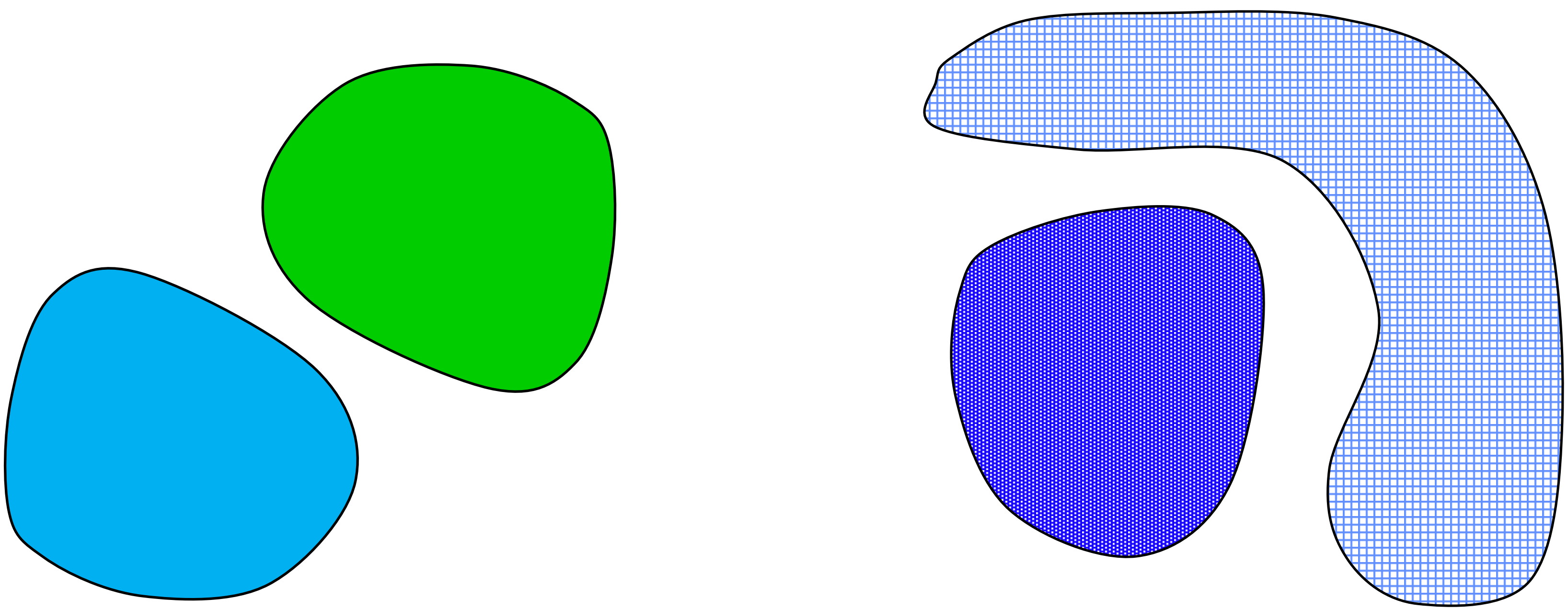
易证 $\hat{e}^T(k)\hat{b}(k)<0$，与上述假设推出的结论矛盾。证毕。

最小平方误差法

- **LMSE算法小结**
  - ✓ 能对输入训练样本集的线性可分性作出正确判断。
  - ✓ 当训练样本集线性可分时，能给出正确的解权向量。
  - ✓ 当训练样本集线性不可分时，能给出**LMSE**准则下的一个最优解。

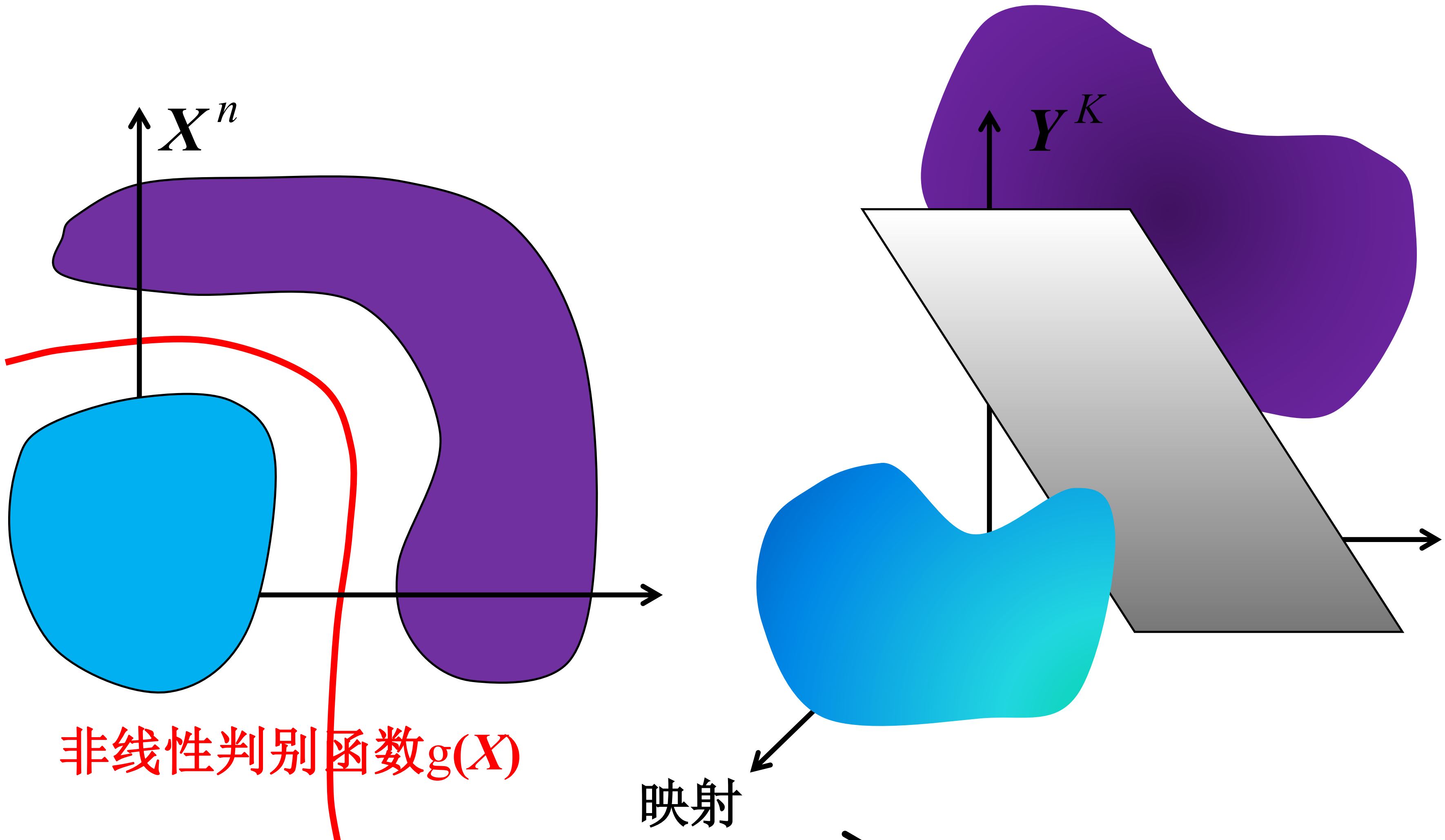
§ 2.4 非线性可分情况下的几何分类法

- 线性可分
- 非线性可分
  - 广义线性判别函数法
  - 分段线性判别函数近似法
  - 非线性判别函数法



§ 2.4 非线性可分情况下的几何分类法

广义线性判别函数法



§ 2.4 非线性可分情况下的几何分类法

广义线性判别函数法

$$\boldsymbol{X}^n \xrightarrow{\text{映射}} \boldsymbol{Y}^K$$

$$\boldsymbol{X} = (x_1, x_2, \ldots, x_n)^T \qquad \boldsymbol{Y} = (y_1, y_2, \ldots, y_K)^T$$

若$g(\boldsymbol{X})$由形式已知的有限个非线性函数所构成，则

$$g(\boldsymbol{X}) = w_1 f_1(\boldsymbol{X}) + w_2 f_2(\boldsymbol{X}) + \Lambda + w_K f_K(\boldsymbol{X}) + w_{K+1}$$

$f_i(\boldsymbol{X}), i = 1,2,\ldots K$　　特征空间中的模式向量X的单值实函数

$w_i, i = 1,2,\ldots, K+1$　　加权系数

引入变换　　$y_i = f_i(\boldsymbol{X}),\qquad i = 1,2,\ldots K$

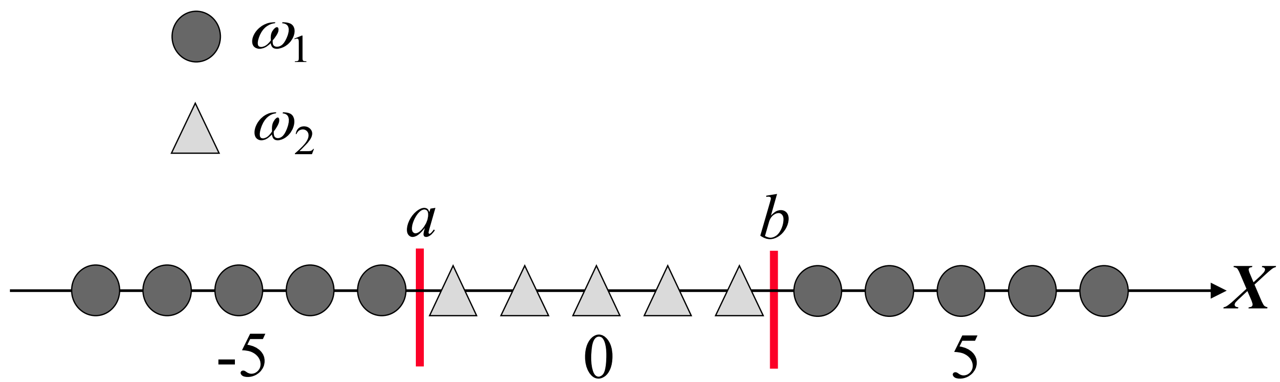
$$\boldsymbol{Y} = (y_1, y_2, \ldots, y_K)^T \Rightarrow \boldsymbol{Y}^K$$

$$g(\boldsymbol{Y}) = w_1 y_1 + w_2 y_2 + \Lambda + w_K y_K + w_{K+1}$$　　线性判别函数

施行非线性变换的过程往往伴随着特征空间维数的增加！

§ 2.4 非线性可分情况下的几何分类法

广义线性判别函数法



决策面函数

$$g(x)=(x-a)(x-b)=\boxed{x^2}-(a+b)\boxed{x}+ab$$

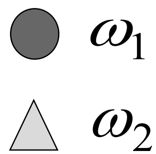
判决规则

$$\begin{cases}\text{若}x<a\text{或}x>b, & \text{则}x\in\omega_1\\ \text{若}a<x<b, & \text{则}x\in\omega_2\end{cases}$$

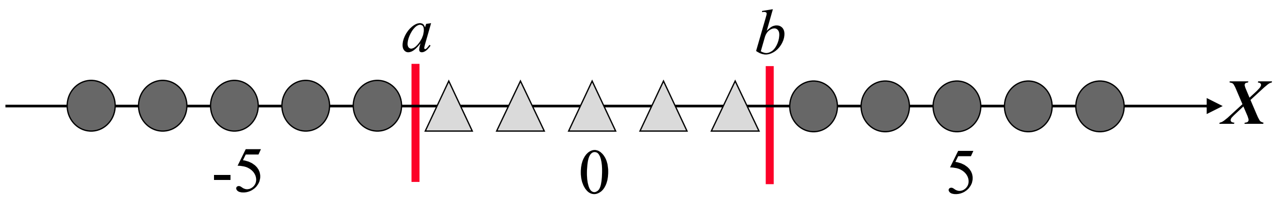
$$\begin{cases}g(x)>0\\ g(x)<0\end{cases}\Rightarrow\begin{cases}x\in\omega_1\\ x\in\omega_2\end{cases}$$

§ 2.4 非线性可分情况下的几何分类法

广义线性判别函数法



$$g(x)=(x-a)(x-b)=\boxed{x^2}-(a+b)\boxed{x}+ab$$

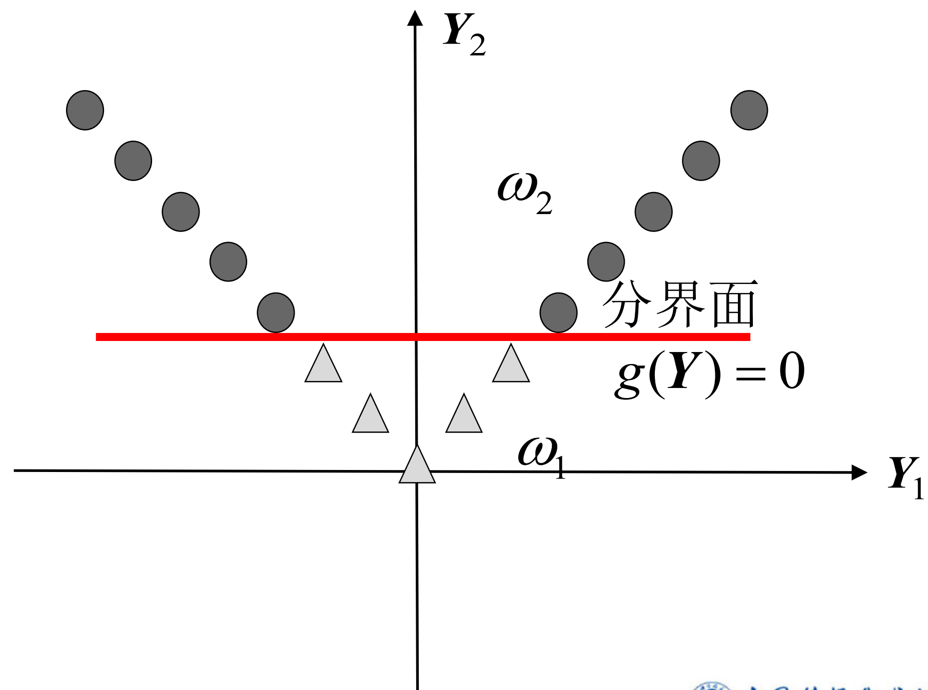


定义变换

$$\begin{cases} y_1 = x^2 \\ y_2 = x \end{cases}$$

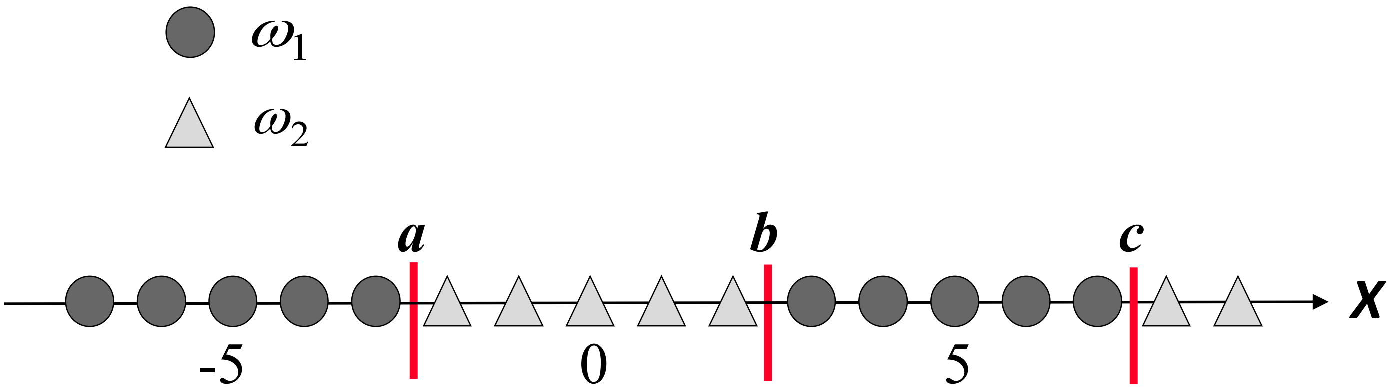
决策面函数

$$g(Y)=w_1y_1+w_2y_2+w_3$$



§ 2.4 非线性可分情况下的几何分类法

广义线性判别函数法



$$g(x)=(x-a)(x-b)(x-c)=x^3-(a+b+c)x^2+(ab+ac+bc)x-abc$$

$$y_1=x^3,\quad y_2=x^2,\quad y_3=x$$

$$g(\boldsymbol{Y})=w_1y_1+w_2y_2+w_3y_3+w_4$$

维数灾

如何确定非线性变换？

- 先验知识
- 函数逼近

如果待求非线性决策面函数可用一个连续函数表示，则根据函数逼近理论，只要多项式次数取得足够高，总可以在预定容限范围内任意地逼近该函数。

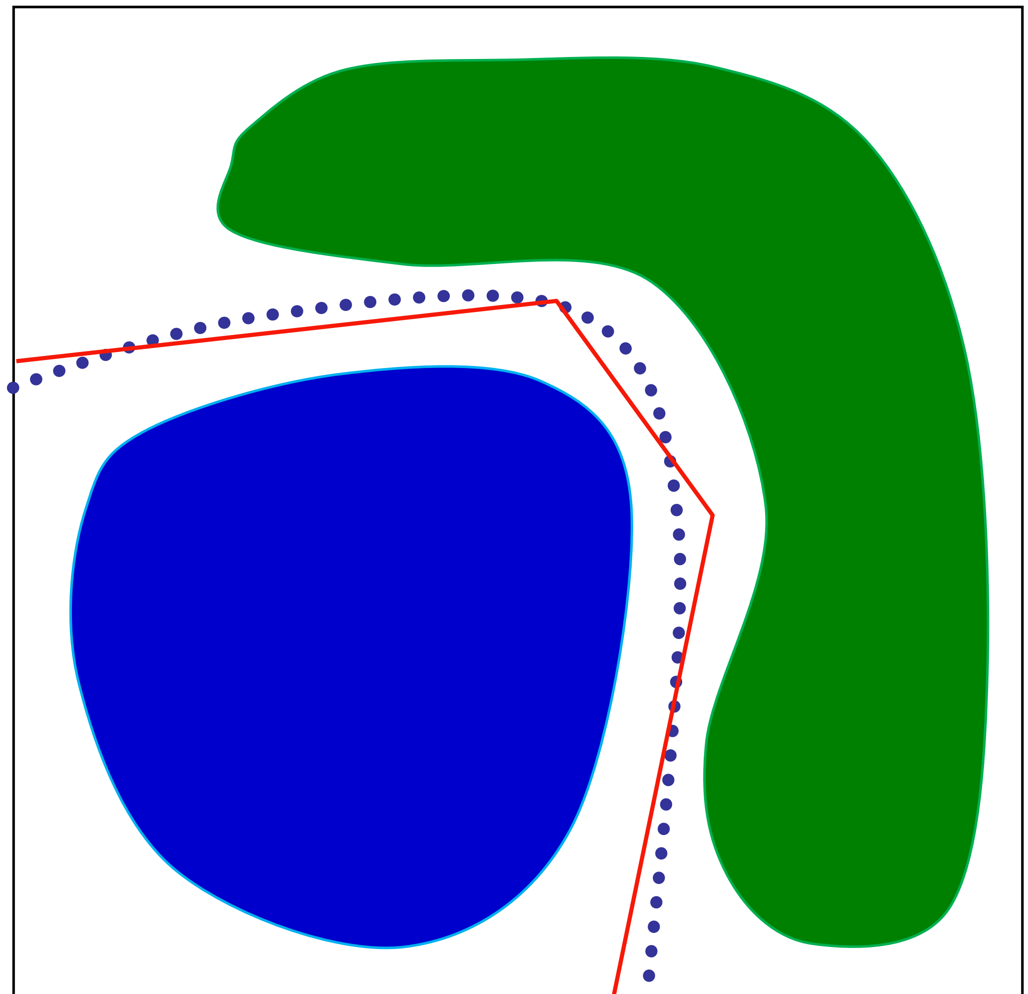
§ 2.4 非线性可分情况下的几何分类法

分段线性判别函数法

避免“维数灾”问题的方策1：分段线性逼近

问题:
在二维情况下，如何用折线近似曲线?

推而广之:
在多维情况下，如何用超平面近似超曲面?



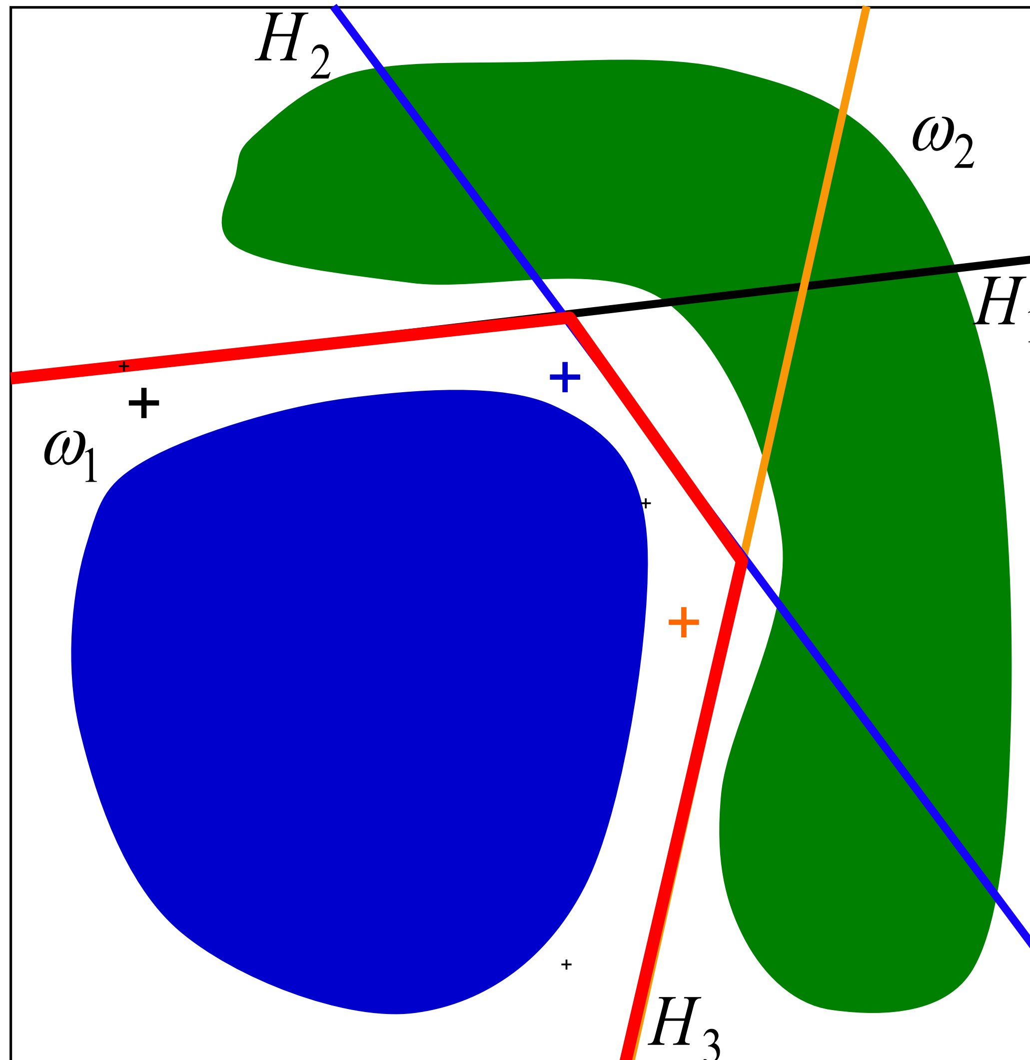
§ 2.4 非线性可分情况下的几何分类法

分段线性判别函数法:单超平面法

起始时整个特征空间为一个区域。

对每一个包含两个不同类别样本的区域递归地使用超平面进行再分类，直到每一个区域只包含同一类别的样本为止。

对得到的所有分类区域进行综合形成最终的分段超平面。



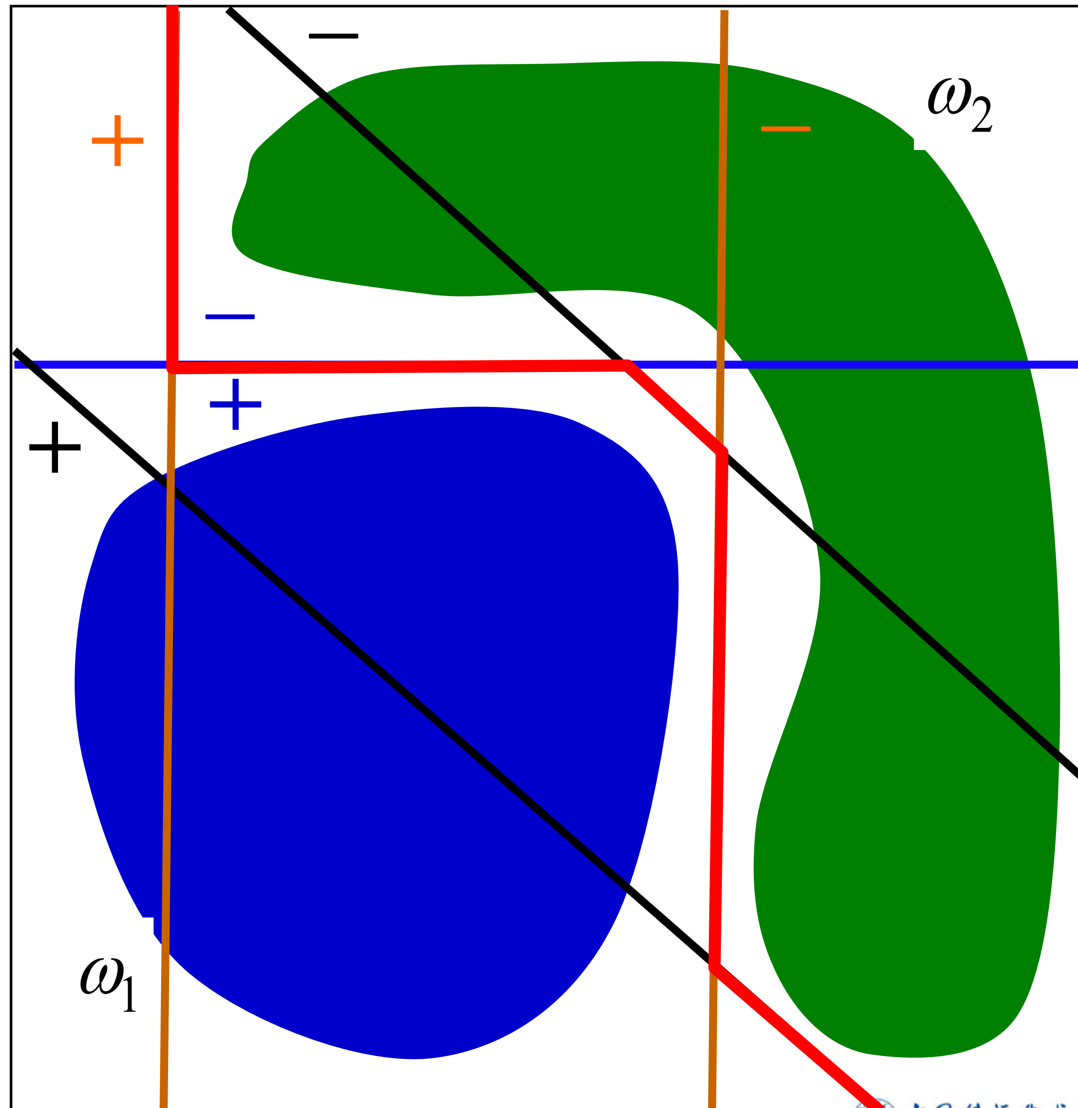
§ 2.4 非线性可分情况下的几何分类法

分段线性判别函数法:双超平面法

起始时整个特征空间为一个区域。

对每一个包含两个不同类别样本的区域递归地使用相互平行的两个超平面进行再分类，直到两个超平面所夹内侧区域区域只包含同一类别的样本为止。

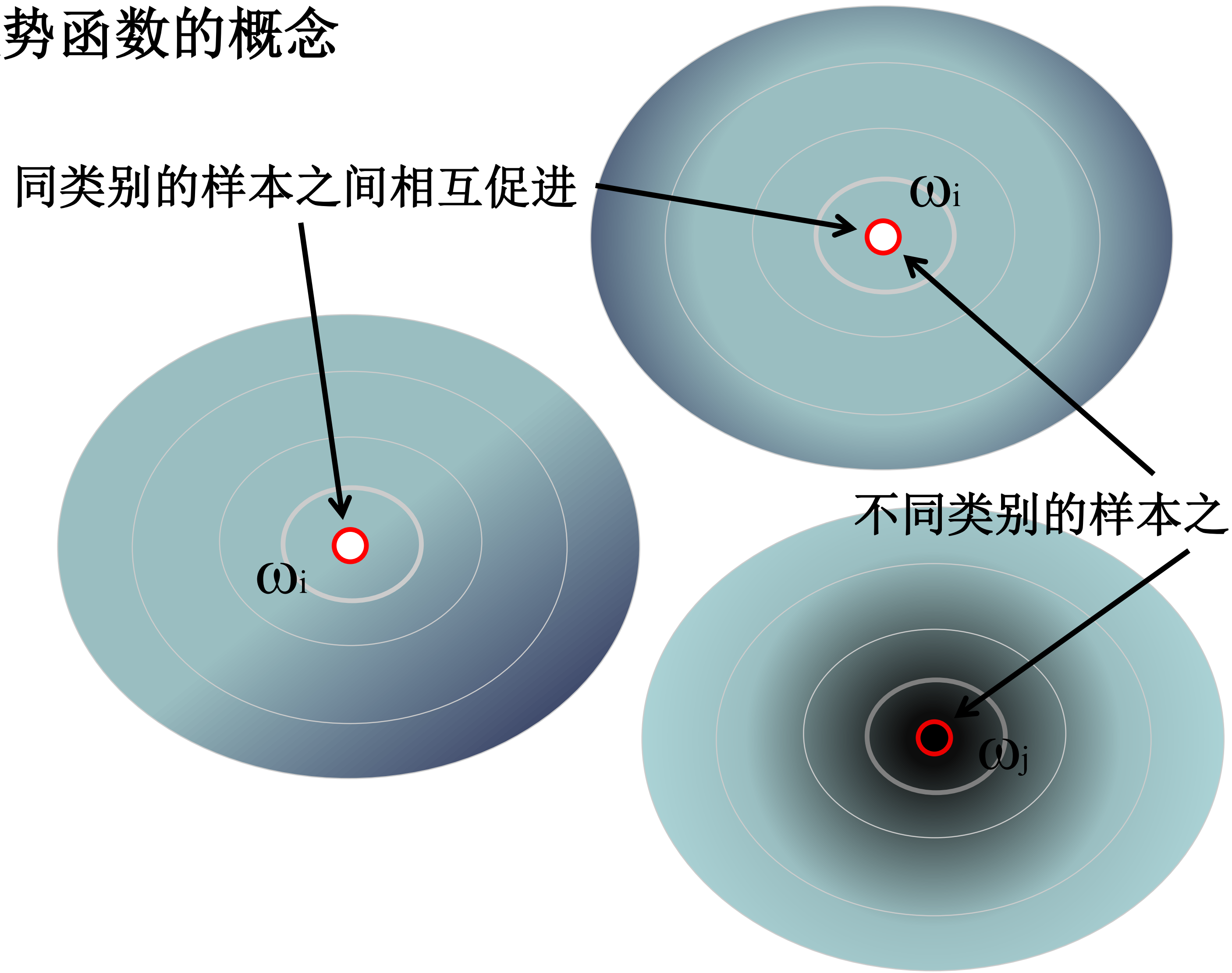
对得到的所有分类区域进行综合形成最终的分段超平面。



§ 2.4 非线性可分情况下的几何分类法

非线性判别函数法：位势函数法

位势函数的概念



§ 2.4 非线性可分情况下的几何分类法

非线性判别函数法：位势函数法

位势函数的概念

$K(X, X_n)$ → 样本点

样本点X_n在X处引起的电位值

位势函数的条件

1. $K(X, X_n) = K(X_n, X)$
2. 当且仅当$X = X_n$时，$K(X,X_n)$达到极值。
3. 当X和X_n之间的距离$\to \infty$时，$K(X,X_n) \to 0$。
4. $K(X,X_n)$是X和X_n之间距离的单调和光滑函数。

§ 2.4 非线性可分情况下的几何分类法

非线性判别函数法：位势函数法

位势函数举例

1. $K(\boldsymbol{X},\boldsymbol{X}_n)=\sum_{i=1}^{m}\phi_i(\boldsymbol{X})\phi_i(\boldsymbol{X}_n)$ 完备正交多项式

2. $K(\boldsymbol{X},\boldsymbol{X}_n)=\exp\{-\boxed{\alpha}\|\boldsymbol{X}-\boldsymbol{X}_n\|^2\}$

3. $K(\boldsymbol{X},\boldsymbol{X}_n)=\dfrac{1}{1+\boxed{\alpha}\|\boldsymbol{X}-\boldsymbol{X}_n\|^2}$

4. $K(\boldsymbol{X},\boldsymbol{X}_n)=\left|\dfrac{\sin\boxed{\alpha}\|\boldsymbol{X}-\boldsymbol{X}_n\|^2}{\alpha\|\boldsymbol{X}-\boldsymbol{X}_n\|^2}\right|$

α 用于控制位势函数衰减速度的常数

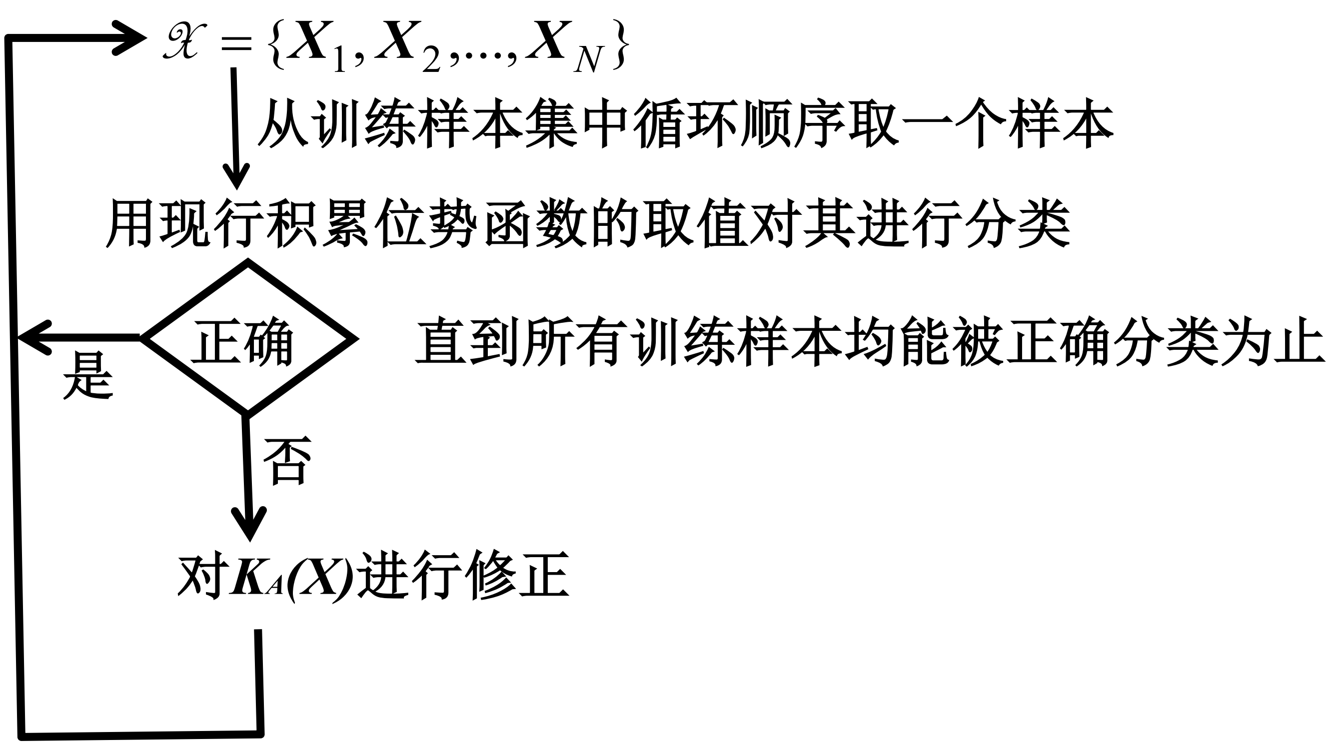
§ 2.4 非线性可分情况下的几何分类法

非线性判别函数法：位势函数法

积累位势函数$K_A(X)$ $\longrightarrow$ 判别函数

由相关样本点在任一点X处产生的位势总和。

注意：不是各样本点位势的简单叠加，而是一种加权和。



§ 2.4 非线性可分情况下的几何分类法

非线性判别函数法：位势函数法

积累位势函数$K_A(X)$的修正规则

$$\mathscr{X} = \{X_1, X_2, \ldots, X_N\} \begin{cases} \omega_i \\ \omega_j \end{cases}$$

$K_{A,k}(X)$，使对所有训练样本正确分类

第 k步积累位势函数

$$K_{A,k}(X) \begin{cases} > 0 \\ = 0 \\ < 0 \end{cases}$$

正确分类

$$\begin{cases} X \in \omega_i 时, K_{A,k}(X) > 0 \\ X \in \omega_j 时, K_{A,k}(X) < 0 \end{cases}$$

错误分类

$$\begin{cases} X \in \omega_i 时, K_{A,k}(X) \le 0 \\ X \in \omega_j 时, K_{A,k}(X) \ge 0 \end{cases}$$

§ 2.4 非线性可分情况下的几何分类法

非线性判别函数法：位势函数法

积累位势函数$K_A(X)$的修正规则

$$K_{A,0}(\boldsymbol{X}) = 0$$

$$K_{A,k}(\boldsymbol{X}) = K_{A,k-1}(\boldsymbol{X}) + \gamma_k K(\boldsymbol{X}, \boldsymbol{X}_k)$$

$$\gamma_k = \begin{cases} 0 & \text{若 } \boldsymbol{X}_k \in \omega_i, \text{ 且} K_{A,k-1}(\boldsymbol{X}_k) > 0 \\ 0 & \text{若 } \boldsymbol{X}_k \in \omega_j, \text{ 且} K_{A,k-1}(\boldsymbol{X}_k) < 0 \\ 1 & \text{若 } \boldsymbol{X}_k \in \omega_i, \text{ 且} K_{A,k-1}(\boldsymbol{X}_k) \le 0 \\ -1 & \text{若 } \boldsymbol{X}_k \in \omega_j, \text{ 且} K_{A,k-1}(\boldsymbol{X}_k) \ge 0 \end{cases}$$

§ 2.4 非线性可分情况下的几何分类法

非线性判别函数法：位势函数法

Step1. 赋初值：k=0，$K_{A,0}(\boldsymbol{X})=0$，$N_c=0$；

Step2. 读入训练样本集合 $\mathscr{X}=\{\boldsymbol{X}_0,\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{N-1}\}$；

Step3. 取样本$\boldsymbol{X}'=\boldsymbol{X}_{[k]_N}$ $[k]_N=k \mod(N)$；

Step4. 根据修正规则，更新积累位势函数：

若 $\boldsymbol{X}'\in\omega_i$，且$K_{A,k}(\boldsymbol{X}')>0$，则$K_{A,k+1}(\boldsymbol{X})=K_{A,k}(\boldsymbol{X})$，$\boldsymbol{N}_c+=1$;

若 $\boldsymbol{X}'\in\omega_i$，且$K_{A,k}(\boldsymbol{X}')\le 0$，则$K_{A,k+1}(\boldsymbol{X})=K_{A,k}(\boldsymbol{X})+\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}')$，$\boldsymbol{N}_c=0$;

若 $\boldsymbol{X}'\in\omega_j$，且$K_{A,k}(\boldsymbol{X}')<0$，则$K_{A,k+1}(\boldsymbol{X})=K_{A,k}(\boldsymbol{X})$，$\boldsymbol{N}_c+=1$;

若 $\boldsymbol{X}'\in\omega_j$，且$K_{A,k}(\boldsymbol{X}')\ge 0$，则$K_{A,k+1}(\boldsymbol{X})=K_{A,k}(\boldsymbol{X})-\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}')$，$\boldsymbol{N}_c=0$;

Step5. 若 $N_c\ge N$，输出 $K_{A,k+1}(\boldsymbol{X})$，算法结束；

否则 $k=k+1$，返回**step3**。

§ 2.4 非线性可分情况下的几何分类法

非线性判别函数法：位势函数法

小结

- 分类能力很强
- 存储量和计算量大
- 分界面可能相当复杂

§ 2.5 线性可分问题的非迭代解法

已介绍算法的特点？

迭代
- 冗长
- 不方便分析和评估

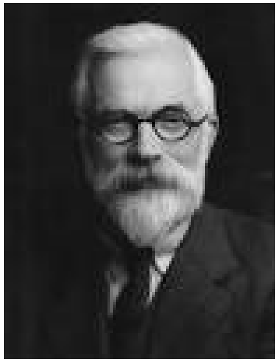
VS

非迭代
- 直接
- 方便分析和评估

Fisher线性判别法

先驱者，R. A. Fisher (1890～1962)





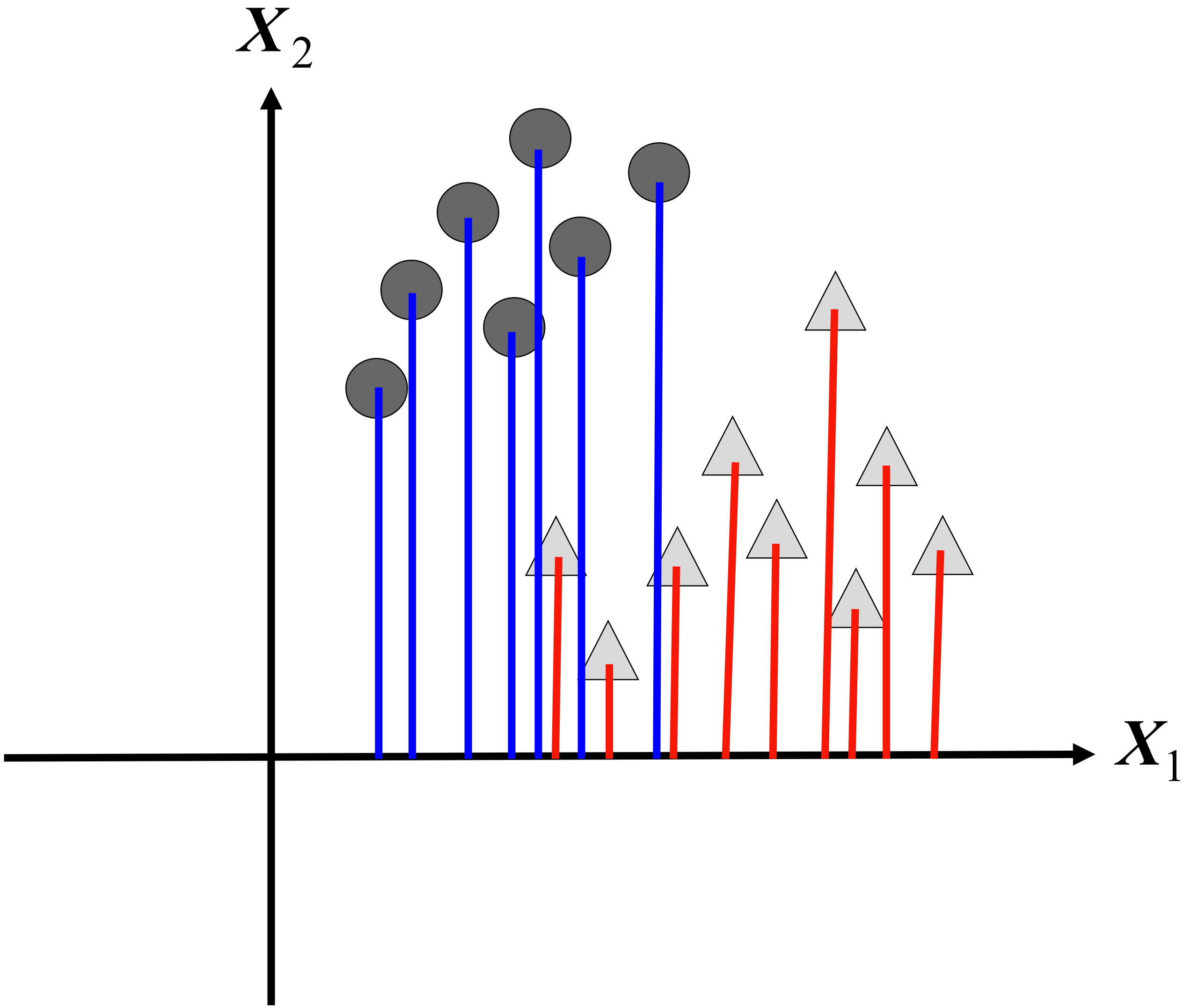
§ 2.5 线性可分问题的非迭代解法

Fisher线性判别法

- 把高维特征空间中的样本投影到一条直线上，实现从高维到一维的数据压缩。
- 改变直线的空间取向，得到不同的投影结果。
- 如果在投影后的直线上训练样本具有很好的分布，则可以通过简单操作实现对输入样本的分类。

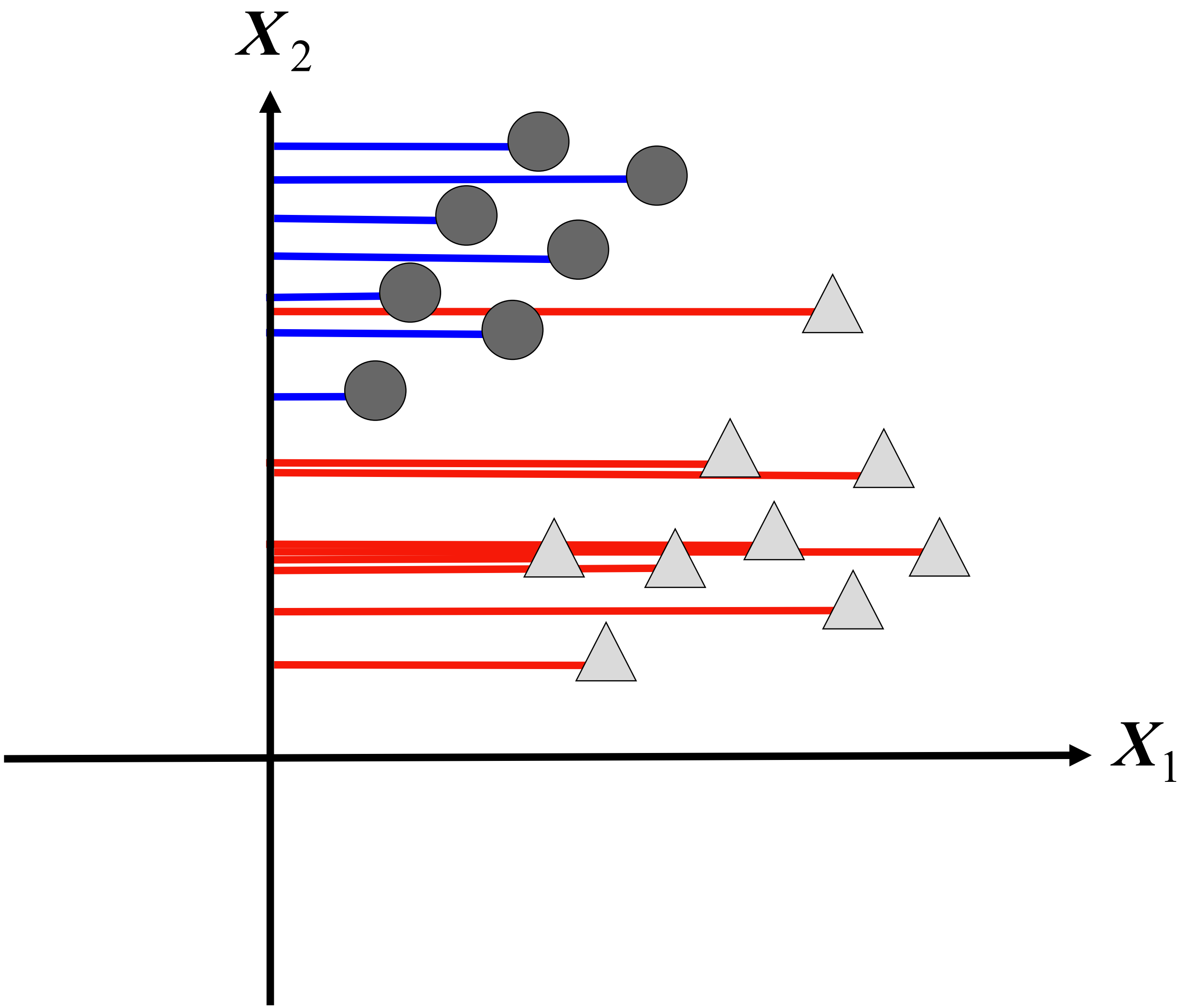
§ 2.5 线性可分问题的非迭代解法

Fisher线性判别法



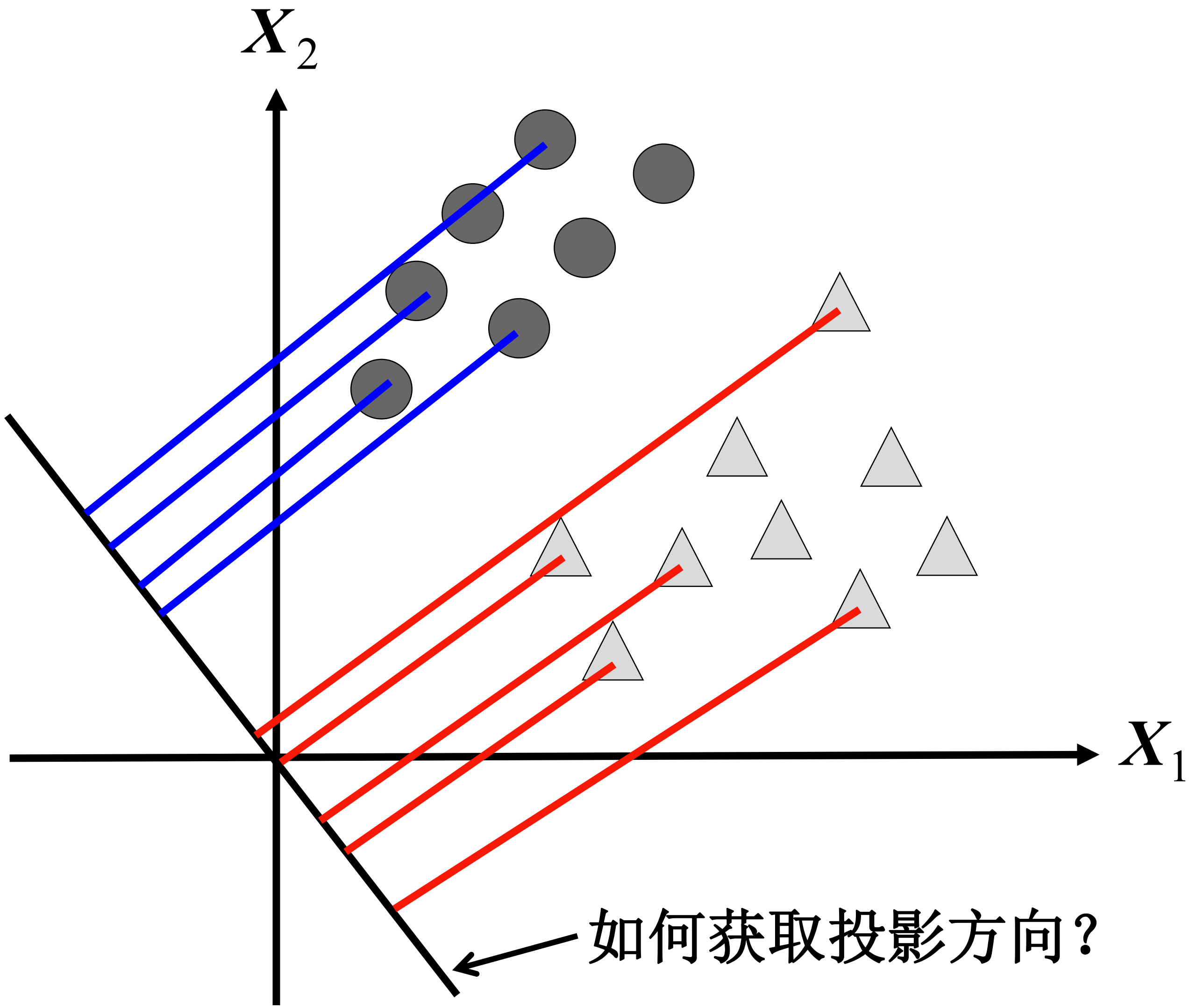
§ 2.5 线性可分问题的非迭代解法

Fisher线性判别法



§ 2.5 线性可分问题的非迭代解法

Fisher线性判别法



§ 2.5 线性可分问题的非迭代解法

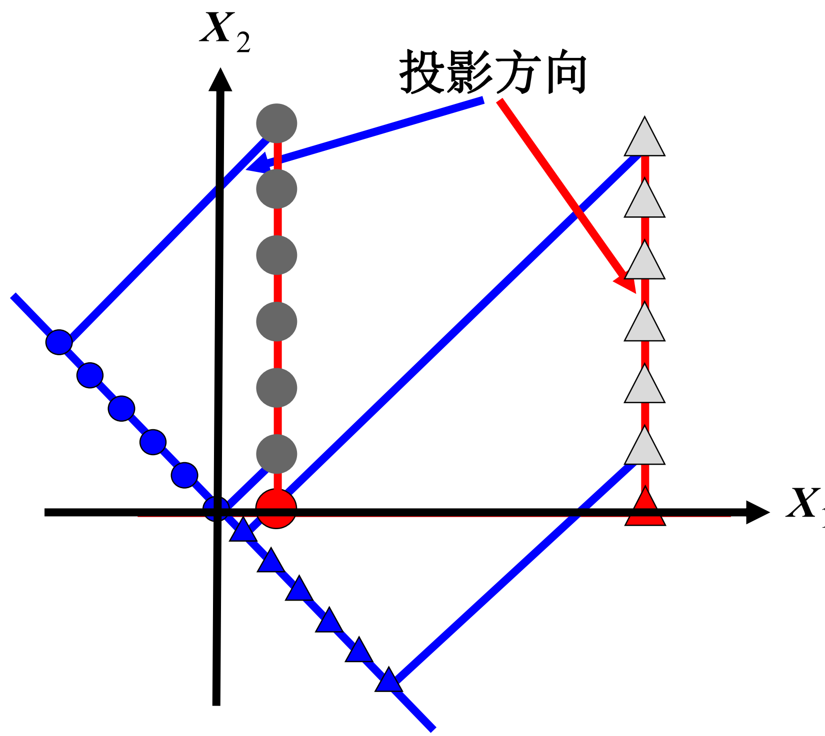
Fisher线性判别法

$$\mathscr{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_N\} \begin{cases} \boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_{n_i} \in \omega_i \\ \boldsymbol{X}_{n_i+1}, \boldsymbol{X}_{n_i+2}, \dots, \boldsymbol{X}_N \in \omega_j \end{cases}$$

$\boldsymbol{X}_k, k = 1,2,\dots N$ 在$\boldsymbol{W}$方向上的投影

$$y_k = \boldsymbol{W}^T \boldsymbol{X}_k, \qquad k = 1,2,\dots,N$$

区分超平面的法线方向



如何获取最佳的投影方向？

引入准则函数

§ 2.5 线性可分问题的非迭代解法

Fisher线性判别法

准则

- 投影后不同类别的样本均值的差别应尽可能大；

  ➡ 类间距离应尽可能大

- 投影后相同类别的样本方差应尽可能小。

  ➡ 类内距离应尽可能小。

投影前后两个类别的样本均值

$$\boldsymbol{m}_x^l = \frac{1}{\boldsymbol{n}_l} \sum_{X_k \in \omega_l} \boldsymbol{X}_k, \qquad \boldsymbol{l} = \boldsymbol{i}, \boldsymbol{j}$$

$$m_y^l = \frac{1}{\boldsymbol{n}_l} \sum_{y_k \in \omega_l} \boldsymbol{y}_k, \qquad \boldsymbol{l} = \boldsymbol{i}, \boldsymbol{j}$$

投影前后类别样本均值之间的关系

$$m_y^l = \frac{1}{n_l} \sum_{y_k \in \omega_l} y_k = \frac{1}{n_l} \sum_{X_k \in \omega_l} \boldsymbol{W}^T \boldsymbol{X}_k = \boldsymbol{W}^T \left(\frac{1}{n_l} \sum_{X_k \in \omega_l} \boldsymbol{X}_k\right) = \boldsymbol{W}^T \boldsymbol{m}_x^l, \qquad l = i, j$$

Fisher线性判别法

投影后两个类别的类间离散度

$$\left|m_y^i - m_y^j\right|^2 = \left\|W^T m_x^i - W^T m_x^j\right\|^2 = \left\|W^T (m_x^i - m_x^j)\right\|^2$$

$$= W^T (m_x^i - m_x^j)[W^T (m_x^i - m_x^j)]^T = W^T \boxed{(m_x^i - m_x^j)(m_x^i - m_x^j)^T} W$$

$$= W^T \boxed{S_b} W$$

大

S_b ⇒ 投影前两个类别的类间离散度矩阵

投影后每个类别的类内离散度

$$\bar{S}_l^2 = \sum_{y_k \in \omega_l} (y_k - m_y^l)^2 = \sum_{X_k \in \omega_l} (W^T X_k - W^T m_x^l)^2, \qquad l = i, j$$

$$\bar{S}_i^2 + \bar{S}_j^2 = \sum_{X_k \in \omega_i} (W^T X_k - W^T m_x^i)^2 + \sum_{X_k \in \omega_j} (W^T X_k - W^T m_x^j)^2$$

$$= W^T \boxed{[\sum_{X_k \in \omega_i} (X_k - m_x^i)(X_k - m_x^i)^T + \sum_{X_k \in \omega_j} (X_k - m_x^j)(X_k - m_x^j)^T]} W$$

$$= W^T \boxed{S_w} W$$

小

S_w ⇒ 样本的类内总离散度矩阵

§ 2.5 线性可分问题的非迭代解法

Fisher线性判别法

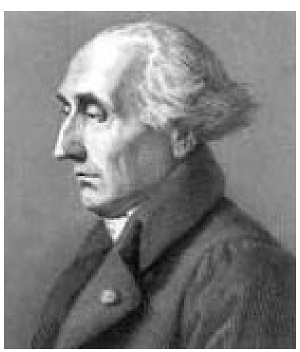
Fisher准则函数

$$J_F(\boldsymbol{W}) = \frac{\left|m_y^i - m_y^j\right|^2}{\bar{S}_i^2 + \bar{S}_j^2} = \frac{\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}}$$

最大化

求解多元函数极值

Lagrange乘子法



Lagrange（拉格朗日，**1736～1813**），**18**世纪最伟大的数学家之一。

拉格朗日一生才华横溢，在数学、物理学和天文学等领域做出了很多重大的贡献，包括著名的拉格朗日中值定理和求解多元函数极值问题的拉格朗日乘子方法（ **Lagrange multiplier method** ）等。

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 单变量（一元）函数极值的定义

设函数$y=f(x)$在点x_0的邻域$U(\delta)$有定义，$\forall x\in U(\delta)$,

$$\left.\begin{array}{l}\text{若}f(x)<f(x_0),\text{则称函数在}x_0\text{有极大值；}\\ \text{若}f(x)>f(x_0),\text{则称函数在}x_0\text{有极小值；}\end{array}\right\}\text{统称为极值}$$

使函数取得极值的点称为极值点。

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 多元函数极值的定义

✓二元情形

设函数$z=f(x, y)$在点(x_0,y_0)的邻域$U(\delta)$有定义，$\forall(x,y)\in U(\delta)$，

若$f(x,y)<f(x_0,y_0)$,则称$f(x, y)$在(x_0,y_0)有极大值；

若$f(x,y)>f(x_0,y_0)$,则称$f(x, y)$在(x_0,y_0)有极小值。

✓一般n元情形　$X=(x_1,x_2,x_3, \ldots,x_n)$

设函数$z=f(X)$在点X_0的邻域$U(\delta)$有定义，$\forall X\in U(\delta)$，

若$f(X) < f(X_0)$,则称$f(X)$在X_0有极大值；

若$f(X) > f(X_0)$,则称$f(X)$在X_0有极小值。

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 多元函数取得极值的必要条件

✓二元情形

设函数$z=f(x, y)$在点(x_0,y_0)处具有连续的一阶偏导数，则它在该点取得极值的必要条件是: $f_x(x_0,y_0)=0$和$f_y(x_0,y_0)=0$ 。

证明：不妨设$z=f(x, y)$在点(x_0,y_0)处有极大值，则对于包含(x_0,y_0)的小邻域内的任意点$(x, y)\neq(x_0,y_0)$，均有$f(x,y)<f(x_0,y_0)$。

故当$y=y_0$， $x\neq x_0$时处有$f(x, y_0)<f(x_0,y_0)$。

说明$f(x, y_0)$在 $x = x_0$处有极大值，必有$f_x(x_0,y_0)=0$ 。

同理可证：$f_x(x_0,y_0)=0$ 。

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 多元函数取得极值的必要条件

✓一般n元情形

设函数$z=f(x_1, x_2, \ldots, x_n)$在点$(x_{10}, x_{20}, \ldots, x_{n0})$处具有连续的一阶偏导数，则它在该点取得极值的必要条件是:

$$f_{x_1}(x_{10}, x_{20}, \ldots, x_{n0}) = \frac{\partial}{\partial x_1}\left(f(x_{10}, x_{20}, \ldots, x_{n0})\right) = 0$$

$$f_{x_2}(x_{10}, x_{20}, \ldots, x_{n0}) = \frac{\partial}{\partial x_2}\left(f(x_{10}, x_{20}, \ldots, x_{n0})\right) = 0$$

$$\cdots\cdots\cdots\cdots$$

$$f_{x_n}(x_{10}, x_{20}, \ldots, x_{n0}) = \frac{\partial}{\partial x_n}\left(f(x_{10}, x_{20}, \ldots, x_{n0})\right) = 0$$

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 多元函数取得极值的必要条件

  ✓一般n元情形(向量形式)

设函数$z=f(X)$在点X_0处具有连续的一阶偏导数，则它在该点取得极值的必要条件是:

$$\frac{\partial}{\partial x_1}\left(f(X_0)\right)=0$$

$$\frac{\partial}{\partial x_2}\left(f(X_0)\right)=0$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$\frac{\partial}{\partial x_n}\left(f(X_0)\right)=0$$

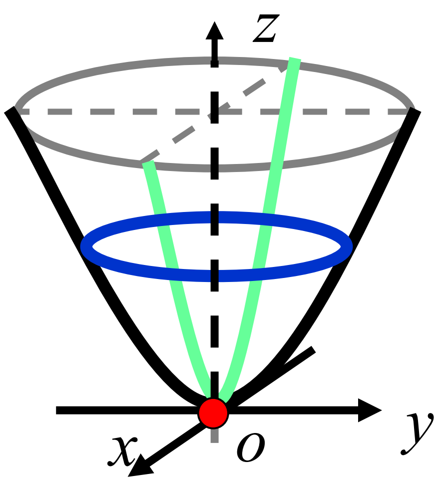
或等价地写成 $$\nabla_X\left(f(X_0)\right)=0$$

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

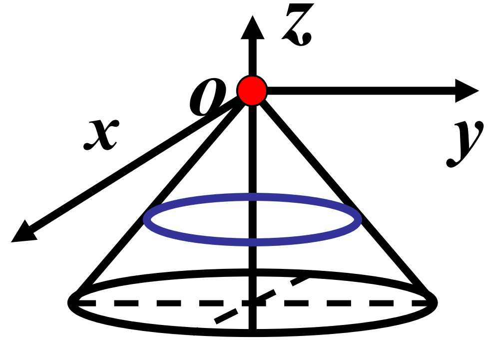
- 多元函数驻点和极值点
  - ✓ 使所有一阶偏导数同时为零的点，称为多元函数的驻点。

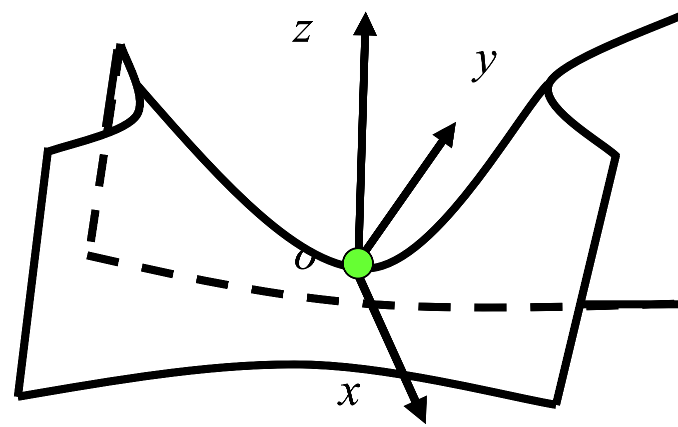
驻点 $\overset{?}{\Longleftarrow}$ $\not\Longrightarrow$ 极值点



$$z=3x^2+4y^2$$

$$z=-\sqrt{x^2+y^2}$$





$$z=xy$$

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 多元函数极值点的判定条件

✓二元情形

设函数z=$f(x, y)$在点(x_0,y_0)的邻域内连续，且有一阶和二阶连续偏导数，又(x_0,y_0)是z=$f(x, y)$的驻点，满足

$$f_x(x_0,y_0)=0 \text{和} f_y(x_0,y_0)=0$$

记$f_{xx}(x_0,y_0)=\boldsymbol{A}$ 、 $f_{xy}(x_0,y_0)=\boldsymbol{B}$ 和 $f_{yy}(x_0,y_0)=\boldsymbol{C}$ ，则点(x_0,y_0)是否极值点的判定条件如下：

1. 若$AC-B^2>0$，则(x_0,y_0)是$f(x, y)$的极值点，且当$A<0$时为极大值，$A>0$时为极小值。
2. 若$AC-B^2<0$，则(x_0,y_0)不是$f(x, y)$的极值点 。
3. 若$AC-B^2=0$，则(x_0,y_0)不确知是否为 $f(x, y)$的极值点 。

回顾：多元函数的极值及其求法

- 多元函数极值点的判定条件

✓ n元情形

$f(X)$的海森（**Hessian**）矩阵：

$$
H\left(f(X)\right)=\begin{bmatrix}
\dfrac{\partial^2 f(X)}{\partial x_1^2} & \dfrac{\partial^2 f(X)}{\partial x_1 x_2} & \cdots & \dfrac{\partial^2 f(X)}{\partial x_1 x_n} \\
\dfrac{\partial^2 f(X)}{\partial x_2 x_1} & \dfrac{\partial^2 f(X)}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(X)}{\partial x_2 x_n} \\
\vdots & \cdots & \cdots & \vdots \\
\dfrac{\partial^2 f(X)}{\partial x_n x_1} & \cdots & \cdots & \dfrac{\partial^2 f(X)}{\partial x_n^2}
\end{bmatrix}
$$

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 多元函数极值点的判定条件

✓ n元情形

设函数$z=f(X)$在点X_0的邻域内连续，且有一阶和二阶连续偏导数，又X_0是$z=f(X)$的驻点，满足

$$\nabla_X\big(f(X_0)\big)=0$$

若$H(f(X))$为$f(X)$的海森矩阵，则X_0是否为极值点的判定条件

1. 若$H(f(X_0))$是正定矩阵，则X_0是$f(X)$的极小值点。
2. 若$H(f(X_0))$是负定矩阵，则X_0是$f(X)$的极大值点。
3. 若$H(f(X_0))$是不定矩阵，则X_0不是$f(X)$的极值点。

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

海森矩阵 $\boldsymbol{H}\left(\boldsymbol{f}(\boldsymbol{X})\right)$ ⇨ 对称矩阵

赫尔维茨定理

对称矩阵 A 为正定的充分必要条件是:

A 的各阶（顺序）主子式都为正;

对称矩阵 A为负定的充分必要条件是:

奇数阶（顺序）主子式为负,而偶数阶（顺序）主子式为正。

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

设A 表为

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & \vdots & & \\ a_{n1} & \cdots & & a_{nn} \end{bmatrix}$$

则A 的r阶（顺序）主子式定义为

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ & \vdots & & \\ a_{r1} & \cdots & & a_{rr} \end{vmatrix} \qquad r=1,2,\ldots,n$$

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 条件极值及拉格朗日乘数法

■ 无约束极值问题

$$Min\ f(X), \quad X \in \mathscr{E}^n$$

■ 等式约束条件极值问题

$$Min\ f(X), \quad X \in \mathscr{E}^n$$

$$s.t. \quad g_i(X) = 0, \quad X \in \mathscr{E}^n, i = 1,2,\ldots,m$$

■ 有不等式约束的条件极值问题

$$Min\ f(X), \quad X \in \mathscr{E}^n$$

$$\begin{aligned} s.t. \quad & g_i(X) \le 0, \quad X \in \mathscr{E}^n, i = 1,2,\ldots,l \\ & h_j(X) = 0, \quad X \in \mathscr{E}^n, j = 1,2,\ldots,m \end{aligned}$$

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

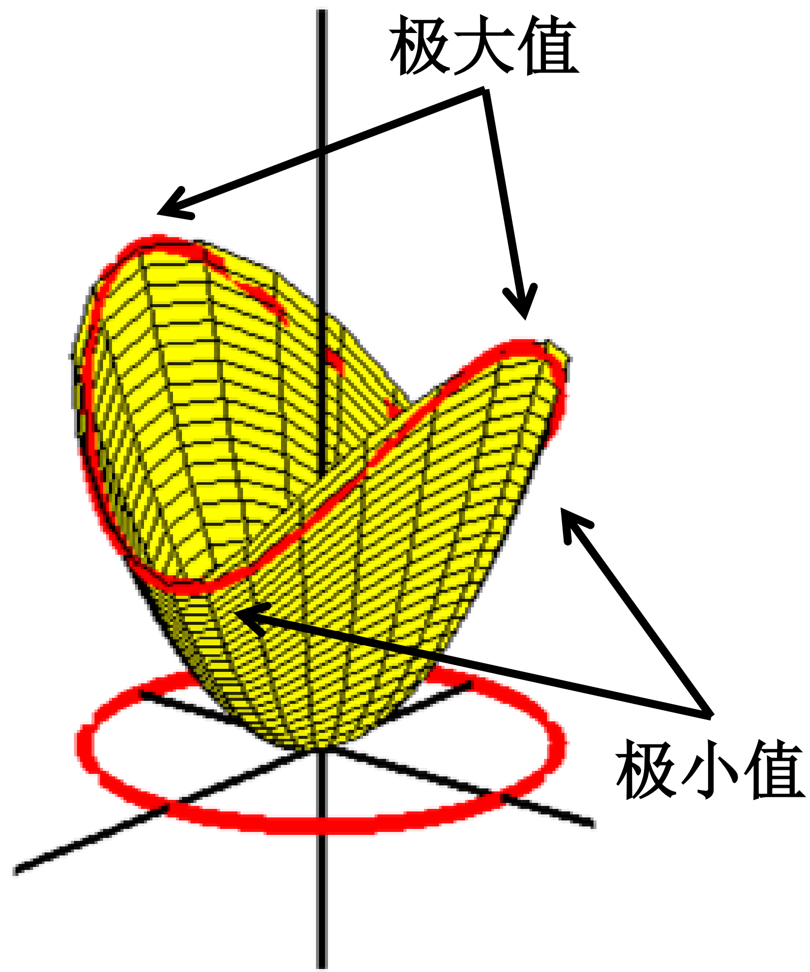
- 有约束条件极值示例

目标函数

$$f(x,y)=3x^2+4y^2$$

约束条件

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



§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 条件极值及拉格朗日乘数法

目标函数 $z = f(x, y)$

约束条件 $g(x, y) = 0$

考虑$f(x, y)$在点(x_0, y_0)处取得条件极值的必要条件

分析：因为(x_0, y_0)是条件极值点，故应有 $g(x_0, y_0) = 0$

又由隐函数存在定理，可确定一个有连续导数的函数$\boldsymbol{y=y(x)}$，使得 $z = f(x, y(x))$

故$f(x, y)$在(x_0, y_0)处的条件极值，可由一元可导函数 $z = f(x, y(x))$ 在x_0处的极值给出：

$$\left.\frac{dz}{dx}\right|_{x=x_0} = f_x(x_0, y_0) + f_y(x_0, y_0)\left.\frac{dy}{dx}\right|_{x=x_0} = 0$$

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 条件极值及拉格朗日乘数法

$$\left.\frac{dz}{dx}\right|_{x=x_0} = f_x(x_0, y_0) + f_y(x_0, y_0)\left.\frac{dy}{dx}\right|_{x=x_0} = 0$$

由隐函数求导公式，有 $\left.\dfrac{dy}{dx}\right|_{x=x_0} = -\dfrac{g_x(x_0, y_0)}{g_y(x_0, y_0)}$

故有 $f_x(x_0, y_0) - \dfrac{f_y(x_0, y_0) g_x(x_0, y_0)}{g_y(x_0, y_0)} = 0$ 记 $\lambda = -\dfrac{f_y(x_0, y_0)}{g_y(x_0, y_0)}$

则取得条件极值的必要条件

$$\begin{cases} f_x(x_0, y_0) + \lambda g_x(x_0, y_0) = 0 \\ f_y(x_0, y_0) + \lambda g_y(x_0, y_0) = 0 \\ g(x_0, y_0) = 0 \end{cases} \Rightarrow 驻点$$

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 拉格朗日乘数法小结（二元情况）

问题：求函数 $z=f(x,y)$ 在条件约束 $g(x,y)=0$ 下的极值点。

解法：构造拉格朗日函数

$$L(x,y,\lambda)=f(x,y)+\lambda g(x,y)$$

其中，λ称为拉格朗日乘数。

求解

$$\begin{cases} L_x(x,y,\lambda)=f_x(x,y)+\lambda g_x(x,y)=0 \\ L_y(x,y,\lambda)=f_y(x,y)+\lambda g_y(x,y)=0 \\ L_\lambda(x,y,\lambda)=g(x,y)=0 \end{cases}$$

所得（x,y）即为候选极值点。

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- **拉格朗日乘数法小结（n元m约束情况）**

问题：求函数 $z = f\left(x_1, x_2, \ldots, x_n\right)$ 在 $g_i\left(x_1, x_2, \ldots, x_n\right) = 0, i = 1, 2, \ldots, m$ 下的极值点。

解法：构造拉格朗日函数

$$L(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2, \ldots, \lambda_m) = f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{m} \lambda_i g_i\left(x_1, x_2, \ldots, x_n\right)$$

求解

$$\begin{cases}
\dfrac{\partial}{\partial x_1} L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = \dfrac{\partial}{\partial x_1} f\left(x_1, \ldots, x_n\right) + \sum\limits_{i=1}^{m} \lambda_i \dfrac{\partial}{\partial x_1} g_i\left(x_1, \ldots, x_n\right) = 0 \\
\vdots \\
\dfrac{\partial}{\partial x_n} L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = \dfrac{\partial}{\partial x_n} f\left(x_1, \ldots, x_n\right) + \sum\limits_{i=1}^{m} \lambda_i \dfrac{\partial}{\partial x_n} g_i\left(x_1, \ldots, x_n\right) = 0 \\
\dfrac{\partial}{\partial \lambda_1} L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = g_1\left(x, y\right) = 0 \\
\vdots \\
\dfrac{\partial}{\partial \lambda_m} L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = g_1\left(x, y\right) = 0
\end{cases}$$

所得 $(x_1, x_2, \ldots, x_n)$ 即为候选极值点。

§ 2.5 线性可分问题的非迭代解法

回顾：多元函数的极值及其求法

- 拉格朗日乘数法小结（向量表示形式）

问题：求函数 $z = f(X)$ 在约束条件

$$g_i(\boldsymbol{X}) = 0, i = 1,2,\ldots,m$$

下的极值点。

解法：构造拉格朗日函数

$$L(\boldsymbol{X},\boldsymbol{\lambda}) = f(\boldsymbol{X}) + \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{X})$$

$$\boldsymbol{g}(\boldsymbol{X}) = (g_1(\boldsymbol{X}), g_2(\boldsymbol{X}), \ldots, g_m(\boldsymbol{X}))^T$$

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^T$$

求解

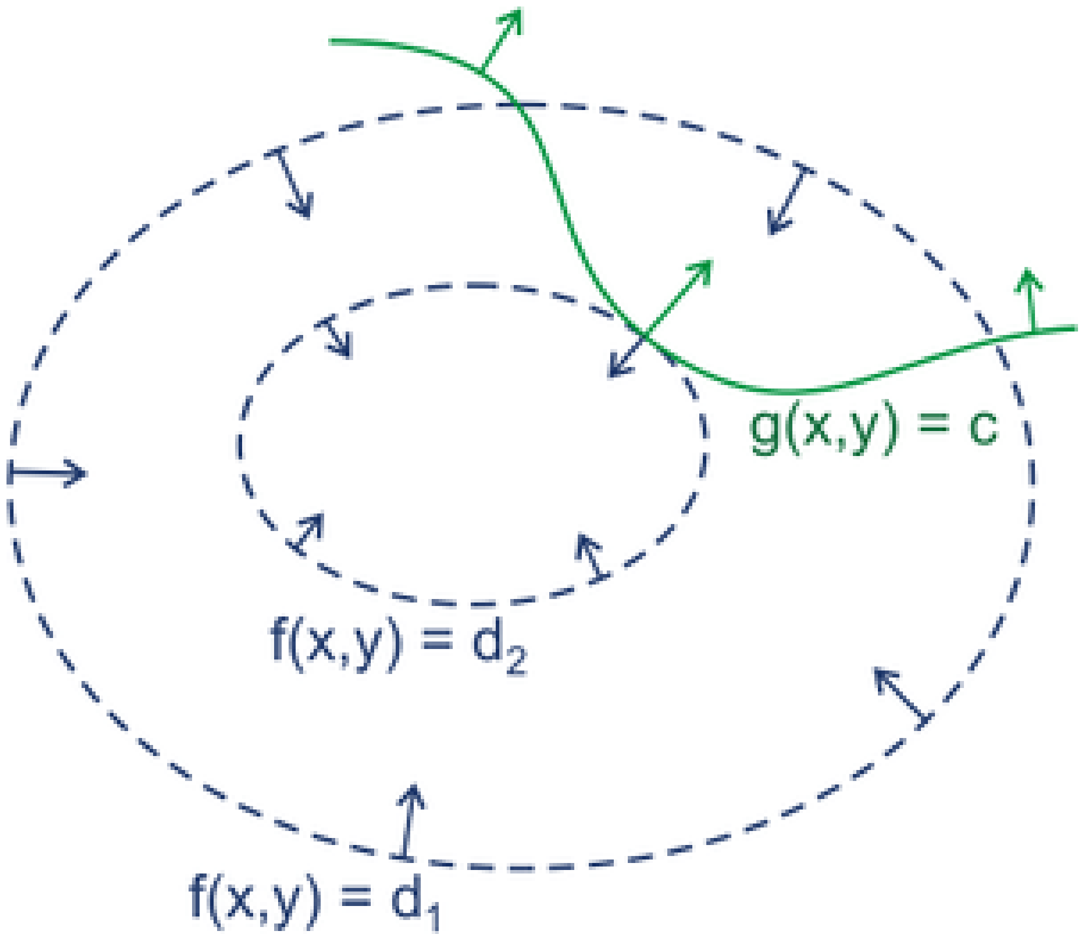
$$\nabla_X L(\boldsymbol{X},\boldsymbol{\lambda}) = \nabla_X f(\boldsymbol{X}) + \nabla_X \left(\boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{X})\right) = 0$$

$$\nabla_\lambda L(\boldsymbol{X},\boldsymbol{\lambda}) = \nabla_\lambda \left(\boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{X})\right) = \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0}$$

所得 $\boldsymbol{X}$ 即为候选极值点。

§ 2.5 线性可分问题的非迭代解法

Fisher线性判别法



§ 2.5 线性可分问题的非迭代解法

用拉格朗日乘数法求解最佳投影方向问题

Fisher准则函数

$$J_F(W)=\frac{\left|m_y^i-m_y^j\right|^2}{\bar{S}_i^2+\bar{S}_j^2}=\frac{\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}}{\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}}$$

构造拉格朗日函数

$$\boldsymbol{L}(\boldsymbol{W},\lambda)=\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}-\lambda(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}-\boldsymbol{c})$$

求解

$$\frac{\partial}{\partial \boldsymbol{W}}L(\boldsymbol{W},\lambda)=2(\boldsymbol{S}_b\boldsymbol{W}-\lambda\boldsymbol{S}_w\boldsymbol{W})=0 \qquad \frac{\partial}{\partial\lambda}L(\boldsymbol{W},\lambda)=\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}-\boldsymbol{c}=0$$

$$\boxed{\boldsymbol{S}_b}\boldsymbol{W}^*=\lambda\boldsymbol{S}_w\boldsymbol{W}^*$$

$$\boldsymbol{S}_b\boldsymbol{W}^*=[(\boldsymbol{m}_x^i-\boldsymbol{m}_x^j)(\boldsymbol{m}_x^i-\boldsymbol{m}_x^j)^T]\boldsymbol{W}^*=(\boldsymbol{m}_x^i-\boldsymbol{m}_x^j)\boxed{[(\boldsymbol{m}_x^i-\boldsymbol{m}_x^j)^T\boldsymbol{W}^*]}$$

$$\left.\begin{matrix}\boldsymbol{S}_b\boldsymbol{W}^*\\(\boldsymbol{m}_x^i-\boldsymbol{m}_x^j)\end{matrix}\right\}\text{方向一致或相反}$$

标量,记为r

§ 2.5 线性可分问题的非迭代解法

用拉格朗日乘数法求解最佳投影方向问题

$$S_b W^* = \lambda S_w W^*$$

$$S_b W^* = r(m_x^i - m_x^j)$$

当样本数足够多时，S_w一般是非奇异的。

$$W^* = \frac{r}{\lambda} S_w^{-1}(m_x^i - m_x^j)$$ Fisher线性判别式给出最佳投影方向

$$\mathscr{X} = \{X_1, X_2, \ldots, X_N\} \Rightarrow m_x^l = \frac{1}{n_l} \sum_{X_k \in \omega_l} X_k, \qquad l = i, j$$

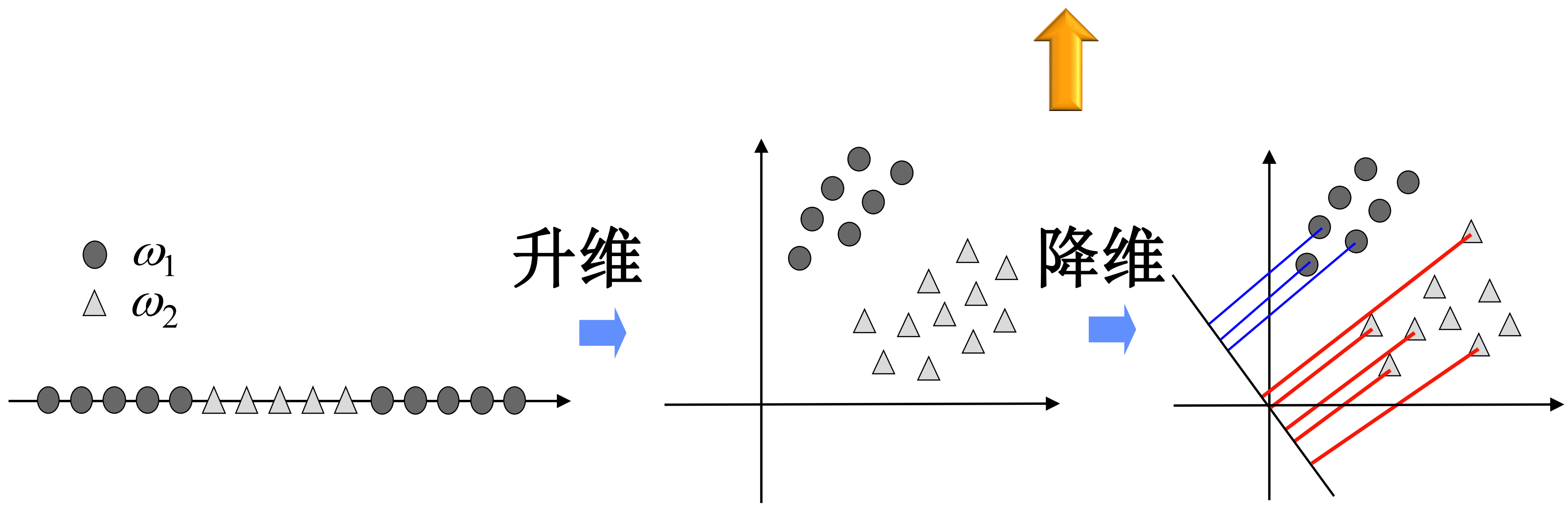
$$S_w = \sum_{X_k \in \omega_i} (X_k - m_x^i)(X_k - m_x^i)^T + \sum_{X_k \in \omega_j} (X_k - m_x^j)(X_k - m_x^j)^T$$

$$W^* = S_w^{-1}(m_x^i - m_x^j)$$

§ 2.5 线性可分问题的非迭代解法

讨论

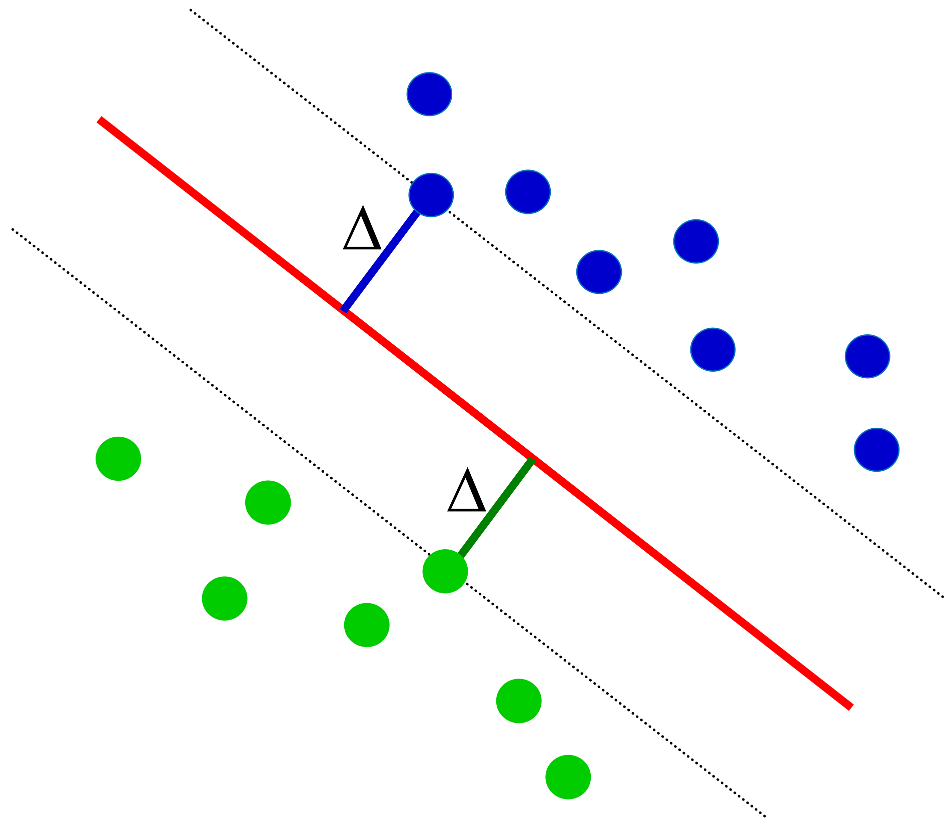
避免了维数灾问题



非线性可分 → 线性可分 → 一维线性可分

§ 2.6 最优分类超平面

优化准则：
寻找与两类中最靠近的样本点的距离同时达到最大的超平面。



§ 2.6 最优分类超平面

如何得到最优分类超平面？

$$\mathscr{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$$ 样本集合

$$\downarrow \quad \downarrow \quad \quad \downarrow$$

$$(y_1, y_2, \ldots, y_N)$$ 标签集合

$$\begin{cases} \boldsymbol{X}_{\mathrm{k}} \in \omega_{\mathrm{i}} & y_k = 1 \\ \boldsymbol{X}_{\mathrm{k}} \in \omega_{\mathrm{j}} & y_k = -1 \end{cases}$$

$$\boldsymbol{W}^T \boldsymbol{X} + w_{d+1} = 0$$ 最优区分超平面

最靠近样本点的距离 Δ

§ 2.6 最优分类超平面

如何得到最优分类超平面？

最优区分超平面 $\boldsymbol{W}^T\boldsymbol{X}+w_{d+1}=0$

判决规则

$$\begin{cases}\boldsymbol{W}^T\boldsymbol{X}+w_{d+1}\geq\Delta\\ \boldsymbol{W}^T\boldsymbol{X}+w_{d+1}\leq-\Delta\end{cases}\Rightarrow\begin{cases}y=1\\ y=-1\end{cases}$$

归一化处理后

$$\begin{cases}\boldsymbol{W}^T\boldsymbol{X}+w_{d+1}\geq 1\\ \boldsymbol{W}^T\boldsymbol{X}+w_{d+1}\leq-1\end{cases}\Rightarrow\begin{cases}y=1\\ y=-1\end{cases}$$

统一表为

$$y[\boldsymbol{W}^T\boldsymbol{X}+w_{d+1}]-1\geq 0$$

$$y_k[\boldsymbol{W}^T\boldsymbol{X}_k+w_{d+1}]-1\geq 0,\quad k=1,2,\ldots N$$

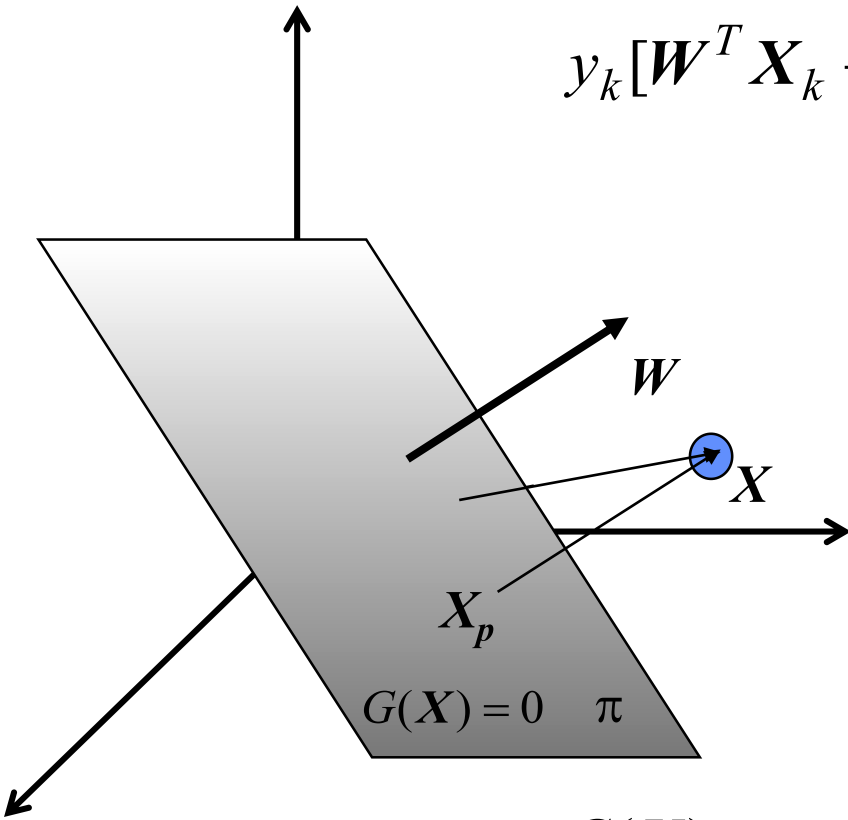
任务：寻找最优 $\boldsymbol{W}^*$

§ 2.6 最优分类超平面

如何得到最优分类超平面？

最优区分超平面　$\boldsymbol{W}^T\boldsymbol{X}+w_{d+1}=0$

$$y_k[\boldsymbol{W}^T\boldsymbol{X}_k+w_{d+1}]-1\geq 0,\quad k=1,2,\ldots N$$



$$\boldsymbol{X}-\boldsymbol{X}_p=r\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|}\qquad r=\frac{G(\boldsymbol{X})}{\|\boldsymbol{W}\|}$$

最靠近样本间的间隔

$$\frac{1}{\|\boldsymbol{W}\|}-\frac{-1}{\|\boldsymbol{W}\|}=\frac{2}{\|\boldsymbol{W}\|}$$

使 $\|\boldsymbol{W}\|$ 最小的超平面

有不等式约束的极值问题

§ 2.6 最优分类超平面

如何得到最优分类超平面？

构造拉格朗日函数 $L(\boldsymbol{W}, w_{n+1}, \lambda) = \frac{1}{2}\boldsymbol{W}^T\boldsymbol{W} - \sum_{k=1}^{N}\lambda_k[y_k[\boldsymbol{W}^T\boldsymbol{X}_k + w_{d+1}] - 1]$

求对偶形式

$$\frac{\partial}{\partial \boldsymbol{W}} L(\boldsymbol{W}, w_{d+1}, \lambda)\Big|_{\boldsymbol{W}=\boldsymbol{W}^*} = \boldsymbol{W}^* - \sum_{k=1}^{N}\lambda_k y_k \boldsymbol{X}_k = 0$$

$$\frac{\partial}{\partial w_{d+1}} L(\boldsymbol{W}, w_{d+1}, \lambda)\Big|_{w_{d+1}=w_{d+1}^*} = -\sum_{k=1}^{N}\lambda_k y_k = 0$$

$$\Rightarrow \quad \boldsymbol{W}^* = \sum_{k=1}^{N}\lambda_k y_k \boldsymbol{X}_k \qquad \sum_{k=1}^{N}\lambda_k y_k = 0$$

$$L_D(\lambda) = \frac{1}{2}\boldsymbol{W}^T\boldsymbol{W} - \sum_{k=1}^{N}\lambda_k[y_k[\boldsymbol{W}^T\boldsymbol{X}_k + w_{d+1}] - 1]$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N}\lambda_i y_i \boldsymbol{X}_i^T\right)\left(\sum_{j=1}^{N}\lambda_j y_j \boldsymbol{X}_j\right) - \sum_{k=1}^{N}\lambda_k y_k(\boldsymbol{W}^T\boldsymbol{X}_k + w_{d+1}) + \sum_{k=1}^{N}\lambda_k$$

§ 2.6 最优分类超平面

如何得到最优分类超平面？

$$L_D(\lambda)=\frac{1}{2}\left(\sum_{i=1}^{N}\lambda_i y_i X_i^T\right)\left(\sum_{j=1}^{N}\lambda_j y_j X_j\right)-\boxed{\sum_{k=1}^{N}\lambda_k y_k\left(W^T X_k+w_{d+1}\right)}+\sum_{k=1}^{N}\lambda_k$$

$$\sum_{k=1}^{N}\lambda_k y_k\left(W^T X_k+w_{d+1}\right)=\sum_{i=1}^{N}\lambda_i y_i\left(\sum_{j=1}^{N}\lambda_j y_j X_j^T X_i+w_{d+1}\right)$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j X_i^T X_j+w_{d+1}\sum_{i=1}^{N}\lambda_i y_i$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j X_i^T X_j$$

$$L_D(\lambda)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j X_i^T X_j+\sum_{k=1}^{N}\lambda_k$$

§ 2.6 最优分类超平面

如何得到最优分类超平面？

$$L_D(\lambda)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \boldsymbol{X}_i^T\boldsymbol{X}_j+\sum_{k=1}^{N}\lambda_k$$

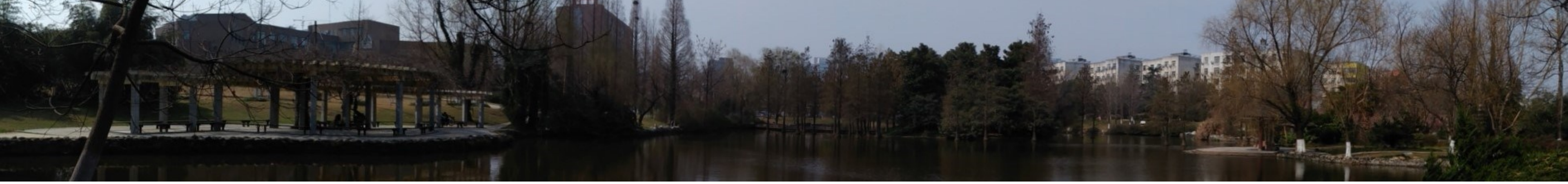
$$Maximise\ \ L_D(\lambda)=Maximise\ \left\{\sum_{k=1}^{N}\lambda_k-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \boldsymbol{X}_i^T\boldsymbol{X}_j\right\}$$

Subject to $\lambda_k\geq 0,\ k=1,2,\ldots N$ $\quad \sum_{k=1}^{N}\lambda_k y_k=0$

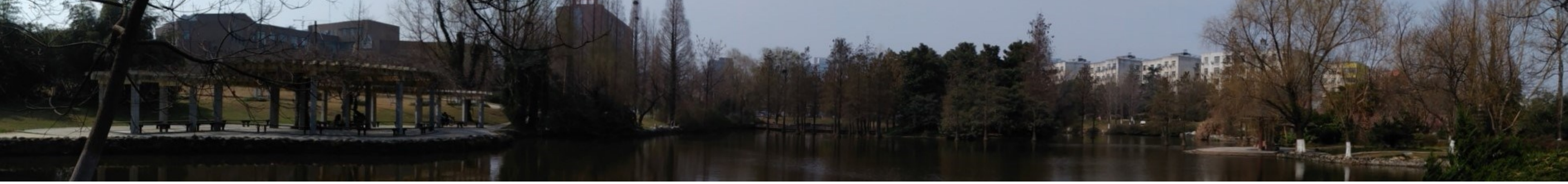
$\lambda_k\geq 0,\ k=1,2,\ldots N$ $\quad \lambda_k>0, k=1,2,\ldots N$ $\boldsymbol{X}_k$,支持向量

$$\boldsymbol{W}^*=\sum_{k=1}^{N}\lambda_k y_k \boldsymbol{X}_k$$

定出 w_{d+1}



若干图片材料取自网络，特此致谢。



谢谢聆听！

中国科学技术大学