

模式识别

总复习

2018~2019学年



内容安排

一、绪论、数学基础（第1讲）

二、聚类分析（第2讲）

三、判别函数分类法（几何分类法）（第3、4讲）

四、统计决策分类法（概率分类法）（第5、6讲）

五、特征提取与选择（第7讲）

六、模糊模式识别（第8讲）

七、神经网络模式识别（第9讲）

期末考试（平时作业：30%，期末考试：70%）

一、绪论

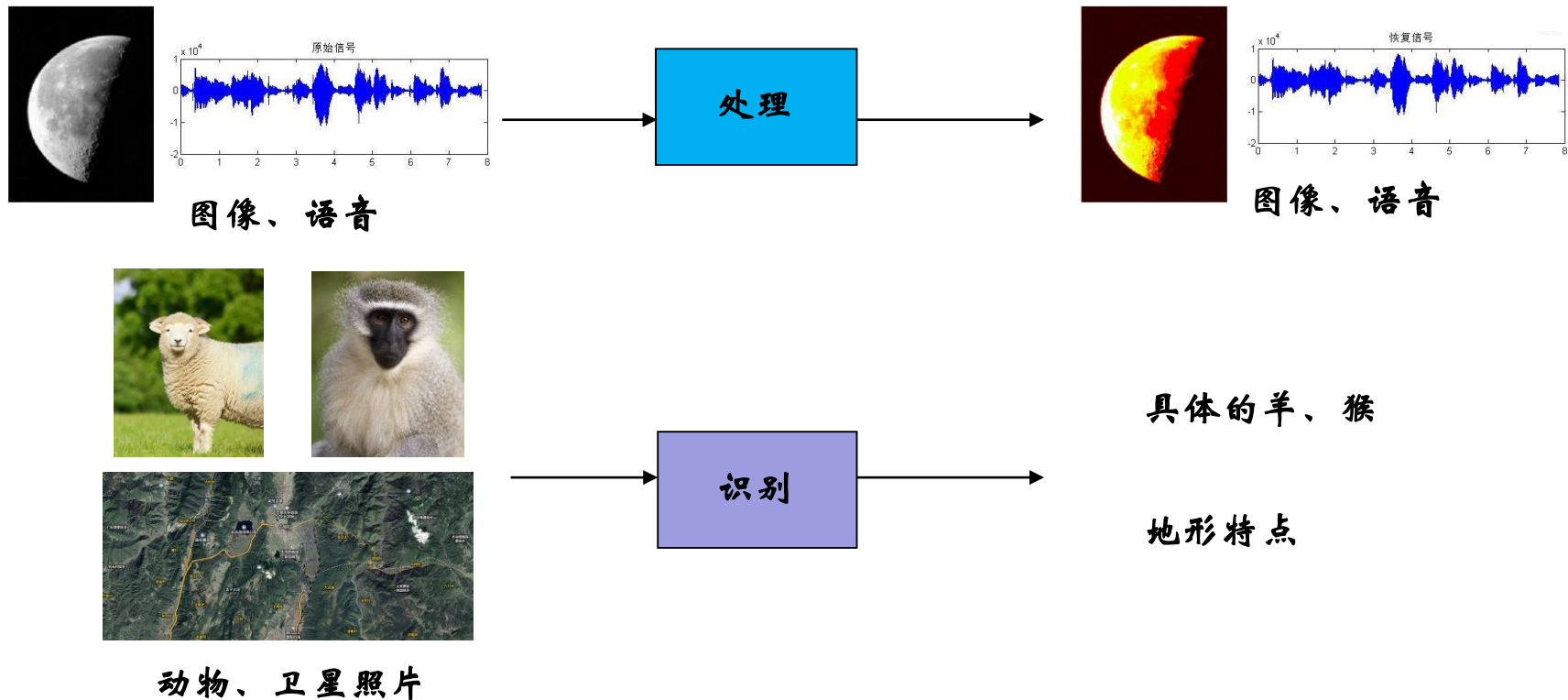
- **样本** (sample)：一个具体的研究对象，具有一个或多个可观测量。
- **特征** (features)：能从某个方面对样本进行描述、刻画或表达的可观测量（数值型、结构型）。多个特征通常用矢量表示——特征矢量。
- **模式** (pattern)：样本特征矢量的观测值，是抽象样本的数值代表。在这个意义上，“样本”与“模式”说的是同一件事情。
- **类别、类属、模式类** (class)：指在一定合理颗粒度下、有实际区分意义的基础上，主观或客观地被归属于“同一类”的客观对象（样本、模式）的**类别代号**。数学上一般处理为整数。
- **样本集** (sample set)：多个样本的集合。训练集、验证集、测试集。
- **已知样本** (known samples, example)：事先知道类别标号的样本。
- **未知样本** (unknown samples)：指特征已知、但类别标号未知的样本，也称：待分样本、待识样本。
- ——机器学习：称已知样本为**样例**，未知样本为**实例**。
- **模式识别** (pattern recognition)：基于样本和应用目的设计分类器；并对性能进行验证，进行必要优化；最后用分类器对未知类属样本进行类属判定。**识别**，不是宽泛的认识或理解，而是对类别判定。

模式识别系统的组成

- **训练集**：样本的集合，用它来设计开发模式分类器。在有监督分类中，已知类别样本的集合，在无监督分类中，是未知类别样本的集合。但都被用来提取“关于各类之间的分类知识”，即使这些样本本身没有类别信息，仍然隐含着某种分类信息！
- **验证集**：既可以认为是宽泛的训练集中的一部分，也可以理解为独立于训练集的有监督样本集。在一次特定的模型训练过程中，两者不重合；但是在多次模型训练（伴随不同模型超参数设置），两者之间是可以交换部分数据的。其作用在于：通过对特定超参数或特定子模型的设置后，对训练后的模型进行“交付前测试”，以便从多个交付模型中选择泛化性能最优者。
- **测试集**：在设计与调优分类系统时没有用过的独立样本集。
- **系统评价原则**：为了更好地对模式识别系统性能进行评价，必须使用一组独立于训练集的测试集对系统进行测试与评价。

模式识别系统的组成

“处理”与“识别”两个概念的区别

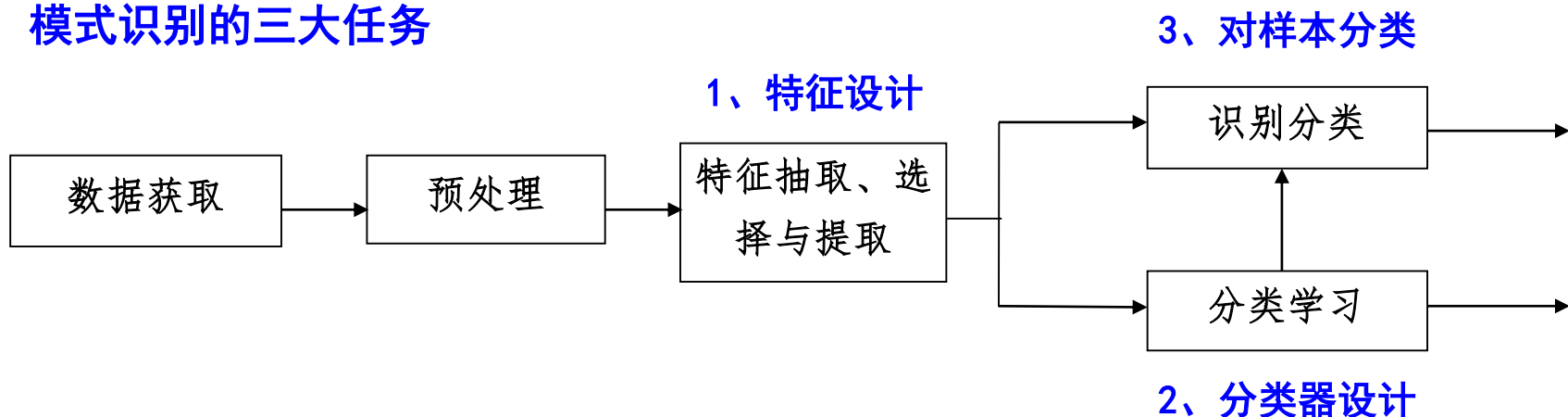


处理：输入与输出是同样性质的数值或特征。

识别：输入的是对象特征，输出的是它的分类和描述。

模式识别系统的组成

模式识别的三大任务



模式识别系统概念框图

模式识别系统的主要环节：

- **特征抽取**：特征数据的采集。如长度测量、波形测量、...
- **特征选择**：选择尽可能少、且更有利于分类的特征。含特征变换（提取）
- **分类学习**：利用样本集建立分类和识别规则，即分类器设计
- **识别分类**：对所获得样本按建立的分类规则进行分类识别

2.1 相似性聚类的概念

- **无监督学习**：使用**不知类别**的样本集进行分类器设计
 - 基于概率密度函数估计的方法（不讲）
 - 基于样本间相似性度量的方法（**聚类分析**）
- **聚类分析**：是指在没有太多先验知识的情况下，按“**物以类聚**”思想，根据模式间的**某种相似性**，对样本进行分类。因此也称为**相似性聚类**
 - 训练前，甚至没有确切的类别数目和类别定义，需要根据待分类样本集的实际特征分布情况与分类活动的应用目的，通过训练样本来**学习出类别数目和“类别的操作定义”**，同时为训练样本分配类别
 - 一般需要迭代多次才会得出有意义的结果
 - “物以类聚”原则展开说就是：
“**同类样本间的相似性 大于 不同类样本间的相似性**”。
 - 聚类方法的有效性：取决于**分类算法**与样本**特征分布**的**匹配**
 - 从方法丰富性与成熟度看，与判别分析（分类）方法相比，聚类方法更多利用直观思想和启发式方法，方法较多，但是缺乏完整稳定的聚类理论基础

1. 欧氏距离（Euclidean, 欧几里德）

设 \mathbf{X}_1 、 \mathbf{X}_2 为两个 n 维模式样本

$$\mathbf{X}_1 = [x_{11}, x_{12}, \dots, x_{1n}]^T \quad \mathbf{X}_2 = [x_{21}, x_{22}, \dots, x_{2n}]^T$$

欧氏距离定义为：

$$\begin{aligned} D(\mathbf{X}_1, \mathbf{X}_2) &= \|\mathbf{X}_1 - \mathbf{X}_2\| = \sqrt{(\mathbf{X}_1 - \mathbf{X}_2)^T (\mathbf{X}_1 - \mathbf{X}_2)} \\ &= \sqrt{(x_{11} - x_{21})^2 + \dots + (x_{1n} - x_{2n})^2} \end{aligned}$$

距离越小，越相似。

注意：

- 各特征维上应当是相同的物理量；
- 注意同类物理量的量纲应该一样。

2. 马氏距离 (Mahalanobis, 马哈拉诺比斯)

平方表达式: $D^2 = (\mathbf{X} - \mathbf{M})^T \mathbf{C}^{-1} (\mathbf{X} - \mathbf{M})$

式中, \mathbf{X} : 模式向量; \mathbf{M} : 均值向量;

\mathbf{C} : 该类模式总体的协方差矩阵。

对 n 维向量: $\mathbf{X} = \begin{bmatrix} x_1 \\ \mathbf{M} \\ x_n \end{bmatrix} \quad \mathbf{M} = \begin{bmatrix} m_1 \\ \mathbf{M} \\ m_n \end{bmatrix}$

$$\begin{aligned} \mathbf{C} &= E\{(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T\} \\ &= E\left\{ \begin{bmatrix} (x_1 - m_1) \\ (x_2 - m_2) \\ \mathbf{M} \\ (x_n - m_n) \end{bmatrix} \begin{bmatrix} (x_1 - m_1) & (x_2 - m_2) & \Lambda & (x_n - m_n) \end{bmatrix} \right\} \end{aligned}$$

2. 马氏距离

$$= \begin{bmatrix} E(x_1 - m_1)(x_1 - m_1) & E(x_1 - m_1)(x_2 - m_2) & \Lambda & E(x_1 - m_1)(x_n - m_n) \\ E(x_2 - m_2)(x_1 - m_1) & E(x_2 - m_2)(x_2 - m_2) & \Lambda & \Lambda \\ \text{M} & \text{M} & \text{M} & \text{M} \\ E(x_n - m_n)(x_1 - m_1) & \Lambda & \Lambda & E(x_n - m_n)(x_n - m_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \Lambda & \sigma_{1n}^2 \\ \sigma_{21}^2 & \text{O} & \sigma_{jk}^2 & \text{M} \\ \text{M} & \text{M} & \sigma_{kk}^2 & \text{M} \\ \sigma_{n1}^2 & \Lambda & \Lambda & \sigma_{nn}^2 \end{bmatrix}$$

马氏距离：在各分量特征维度计算样本模式与均值模式的距离上，剔除了该维度上模式类的方差影响——方差大，说明模式取值变化大，因此计算距离时要用方差归一化，这样得出的分量模式与均值模式的距离才具有比较意义。

优点：排除了模式样本之间的相关性影响。

特例：当 $\mathbf{C} = \mathbf{I}$ 时，马氏距离退化为欧氏距离。

3. 明氏距离(Minkowski , 闵可夫斯基)

n 维模式向量 X_i 、 X_j 间的明氏距离表示为：

$$D_m(X_i, X_j) = \left[\sum_{k=1}^n |x_{ik} - x_{jk}|^m \right]^{1/m}$$

式中， x_{ik} 、 x_{jk} 分别表示 X_i 和 X_j 的第 k 个分量。也称为 **m-范数**。

当 $m=2$ 时，明氏距离为欧氏距离。

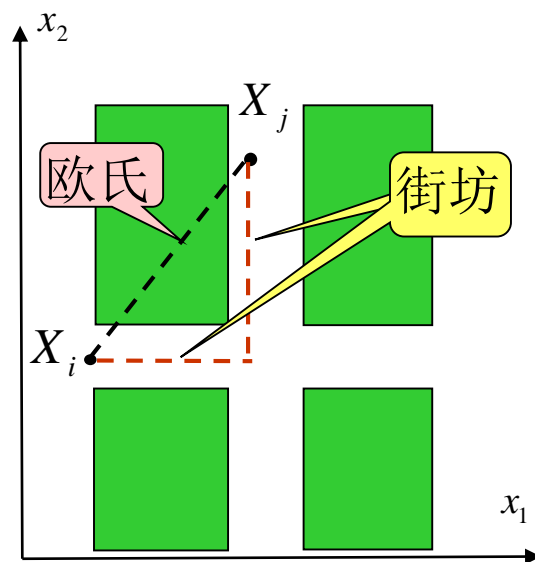
当 $m=1$ 时：

$$D_1(X_i, X_j) = \sum_{k=1}^n |x_{ik} - x_{jk}|$$

称为“**街坊**”距离 (“City block” distance)。

示例：当 $k=2$ 时：图中

$$D_1(X_i, X_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|$$



4. 汉明距离 (Hamming)

设 \mathbf{X}_i 、 \mathbf{X}_j 为 n 维二值模式（分量取值1或-1）样本向量，则

汉明距离：
$$D_h(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{2} \left(n - \sum_{k=1}^n x_{ik} \cdot x_{jk} \right)$$

式中， x_{ik} 、 x_{jk} 分别表示 \mathbf{X}_i 和 \mathbf{X}_j 的第 k 个分量。

说明：汉明距离中的**求和式**，表达的是两个二值向量之间，**同值分量数与不同值分量数之差**；显然，这个求和式表示的**差值**越大，表示两个向量越相似，于是**求和式的负值也就越小**，因此**求和式的负值可表示距离**，取值于 **$(-n, n)$** 。为更直观表达，用最大值 **n** 加上这个差值并除以**2**，取值于 **$(0, n)$** 。

——0：表示两模式向量的各个分量都相同；

—— n ：表达两模式向量的各个分量都不同；

—— $n/4$ ：表示两模式向量中，取值相同的分量数比取值不同的分量数多 **$n/2$** 。因为此时：取值相同的分量数为 **$(3/4)n$** ，取值相同的分量数为 **$(1/4)n$** 。

5. 角度相似性函数

$$S(\mathbf{X}_i, \mathbf{X}_j) = \frac{\mathbf{X}_i^T \mathbf{X}_j}{\|\mathbf{X}_i\| \cdot \|\mathbf{X}_j\|}$$

定义为：模式向量 \mathbf{X}_i ， \mathbf{X}_j 之间夹角的余弦。

测度基础：以两矢量的方向是否相近作为考虑的基础,矢量长度并不重要。设

$$\bar{x} = (x_1, x_2, \dots, x_n)', \bar{y} = (y_1, y_2, \dots, y_n)'$$

注意：该测度对坐标系的旋转和尺度的缩放是不变的，但对一般的线形变换和坐标系的平移不具有不变性。

聚类准则

聚类准则：根据相似性测度确定的、衡量模式聚类结果中得到的聚类是否满足某种优化目标的一个判断标准或方法。

确定聚类准则的两种方式：

1. 阈值准则：根据规定的距离阈值进行判断。
2. 函数准则：利用聚类准则函数进行判断。

聚类准则函数：在聚类分析中，表示聚类过程中，所产生的中间分类结果质量的一种度量函数。

聚类准则函数应是模式样本集 $\{X\}$ 和模式类别 $\{S_j, j=1, 2, \dots, c\}$ 的函数。可使聚类分析转化为寻找准则函数极值的优化问题。一种常用的指标是误差（距离）平方和。

聚类准则

聚类准则函数：

$$J = \sum_{j=1}^c \sum_{X \in S_j} \|X - M_j\|^2$$

式中： c 为聚类类别的数目，

$$M_j = \frac{1}{N_j} \sum_{X \in S_j} X \text{ 为属于 } S_j \text{ 集的样本的均值向量，}$$

N_j 为模式类 S_j 中样本数目。

J 代表了分属于 c 个聚类类别的全部模式样本与其相应类别模式均值样本之间的误差（距离）平方和。

显然，这里的 J 是类别数 c 的单调减函数，因此这样的准则函数，如果不加控制，容易把任何模式集分为更多的类。

适用范围：

适用于各类样本密集且数目相差不多，而不同类间的样本又明显分开的情况。

2.3.2 最大最小距离算法（小中取大距离算法）

1. 问题：已知 N 个待分类的模式 $\{X_1, X_2, \dots, X_N\}$
分类到聚类中心 Z_1, Z_2, \dots, Z_k 对应的类别中。

2. 算法描述

- ① 选任意一模式样本做为第一聚类中心 Z_1 。
- ② 选择离 Z_1 距离最远的样本作为第二聚类中心 Z_2 。
- ③ 逐个计算各模式样本 X_i 与已确定的所有聚类中心 Z_j 之间的距离，并选出其中的最小距离。例如：当目前聚类中心数 $k=2$ 时，计算

$$D_{i1} = \|X_i - Z_1\| \qquad D_{i2} = \|X_i - Z_2\|$$

$$\min(D_{i1}, D_{i2}), \quad i=1, \dots, N \quad (N \text{ 个最小距离})$$

2.3.2 最大最小距离算法（小中取大距离算法）

④ 在所有最小距离中选出最大距离，如该最大值达到 $\|\mathbf{Z}_1 - \mathbf{Z}_2\|$ 的一定分数比值(阈值 T) 以上，则相应的样本点取为新的聚类中心，返回③；否则，寻找聚类中心的工作结束。

例 $k=2$ 时

若 $\max\{\min(D_{i1}, D_{i2}), i = 1, 2, \dots, N\} > \theta \|\mathbf{Z}_1 - \mathbf{Z}_2\|, 0 < \theta < 1$

则 \mathbf{Z}_3 存在。（ θ ：用试探法取为一固定分数，如 $1/2$ 。）

⑤ 重复步骤③④，直到没有新的聚类中心出现为止。

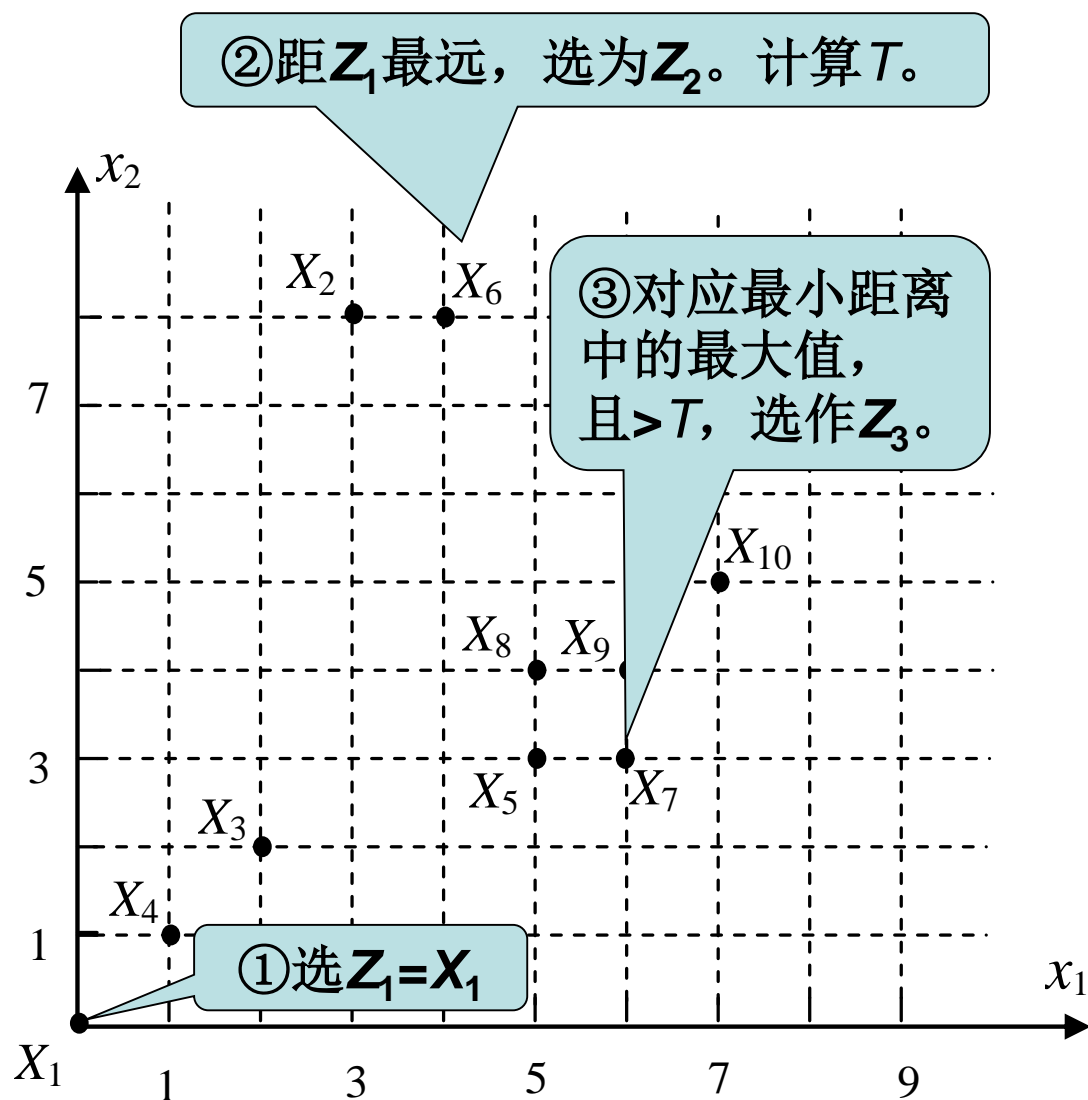
⑥ 将样本 $\{\mathbf{X}_i, i = 1, 2, \dots, N\}$ 按最近距离划分到相应聚类中心对应的类别中。

思路总结：

二步法。先找全部中心，然后再对剩余模式归类。关键：怎样开新类，聚类中心如何定。

为使聚类中心更有代表性，可取各类的样本均值作为聚类中心。

例2.1 对图示模式样本用最大最小距离算法进行聚类分析



$$③ T = \frac{1}{2} \|Z_1 - Z_2\| = \frac{1}{2} \sqrt{80}$$

10个最小距离中, X_7 对应的距离 $> T$,

$$\therefore Z_3 = X_7$$

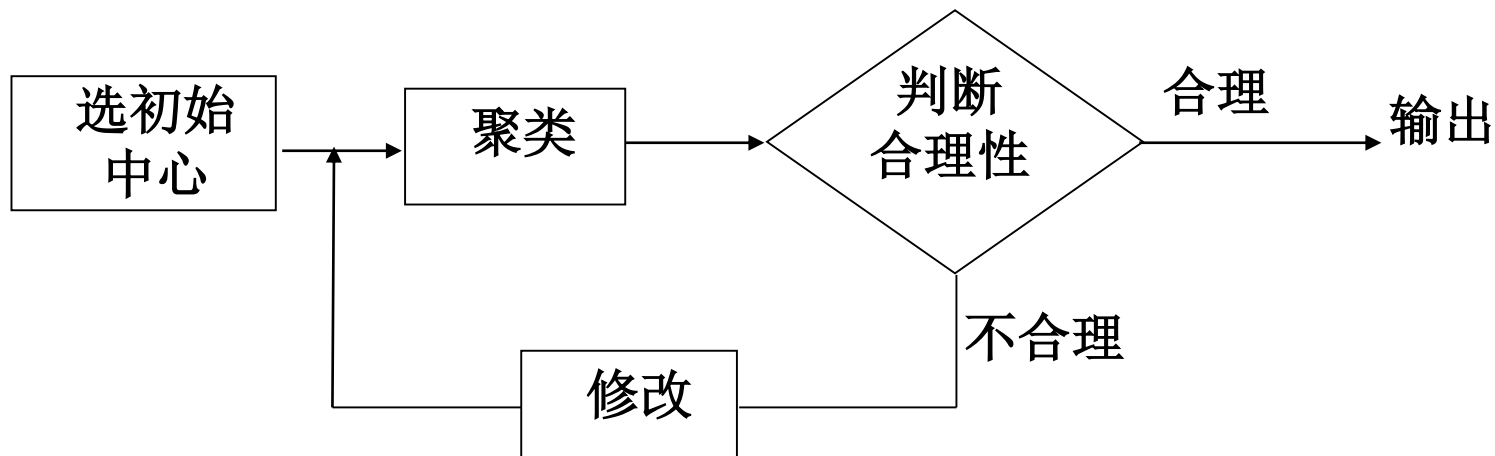
④ 用全体模式对三个聚类中心计算最小距离中的最大值, 无 $> T$ 情况, 停止寻找中心。

结果: $Z_1 = X_1$; $Z_2 = X_6$;

$$Z_3 = X_7。$$

⑤ 对剩余模式归类 (聚类)

2.5 动态聚类法



两种常用算法：

- * **K-均值算法**（或C-均值算法）
- * **ISODATA算法**（迭代自组织数据分析算法）
（Iterative Self-Organizing Data Analysis Techniques Algorithm）

2.5.1 K-均值算法

基于使聚类准则函数最小化，其准则函数：

聚类集中各点到该聚类集中心的距离平方和。

对于第 j 个聚类集，准则函数定义为

$$J_j = \sum_{i=1}^{N_j} \| \mathbf{X}_i - \mathbf{Z}_j \|^2, \quad \mathbf{X}_i \in S_j$$

S_j : 第 j 个聚类集（域），聚类中心为 \mathbf{Z}_j ；

N_j : 第 j 个聚类集 S_j 中所包含的样本个数。

对所有 K 个模式类有

$$J = \sum_{j=1}^K \sum_{i=1}^{N_j} \| \mathbf{X}_i - \mathbf{Z}_j \|^2, \quad \mathbf{X}_i \in S_j$$

K-均值算法的聚类准则：聚类中心的选择，应使准则函数 J 极小，
也即：使 J_j 的值极小。

2.5.1 K-均值算法

应有 $\frac{\partial J_j}{\partial \mathbf{Z}_j} = 0$

即
$$\frac{\partial}{\partial \mathbf{Z}_j} \sum_{i=1}^{N_j} \|\mathbf{X}_i - \mathbf{Z}_j\|^2 = \frac{\partial}{\partial \mathbf{Z}_j} \sum_{i=1}^{N_j} (\mathbf{X}_i - \mathbf{Z}_j)^T (\mathbf{X}_i - \mathbf{Z}_j) = 0$$

可解得
$$\mathbf{Z}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{X}_i, \quad \mathbf{X}_i \in S_j$$

上式表明， S_j 类的聚类中心应选为该类样本的均值向量。

1. 算法描述

(1) 任选 K 个初始聚类中心： $\mathbf{Z}_1(1)$ ， $\mathbf{Z}_2(1)$ ， \dots ， $\mathbf{Z}_K(1)$

括号内序号：迭代运算的次序号。

2.5.1 K-均值算法

(2) 按最小距离原则将其余样本**分配**到**K**个聚类中心中的某一个，即：

若 $\min \{ \|X - Z_i(k)\|, i=1,2,\dots,K \} = \|X - Z_j(k)\| = D_j(k)$ ，则 $X \in S_j(k)$

注意： k ——迭代运算次序号； K ——聚类中心的个数。

(3) 再次**计算**各个聚类中心的新向量值： $Z_j(k+1) \quad j=1,2,\dots,K$

$$Z_j(k+1) = \frac{1}{N_j} \sum_{X \in S_j(k)} X \quad j=1,2,\dots,K$$

N_j ：第 j 类的样本数。

这里：分别计算**K**个聚类中的样本均值向量，故称**K-均值算法**。

(4) 如果 $Z_j(k+1) \neq Z_j(k) \quad j=1,2,\dots,K$ ，则回到 (2)，将模式样本逐个重新分类，重复迭代计算。

如果 $Z_j(k+1) = Z_j(k) \quad j=1,2,\dots,K$ ，算法收敛，计算完毕。

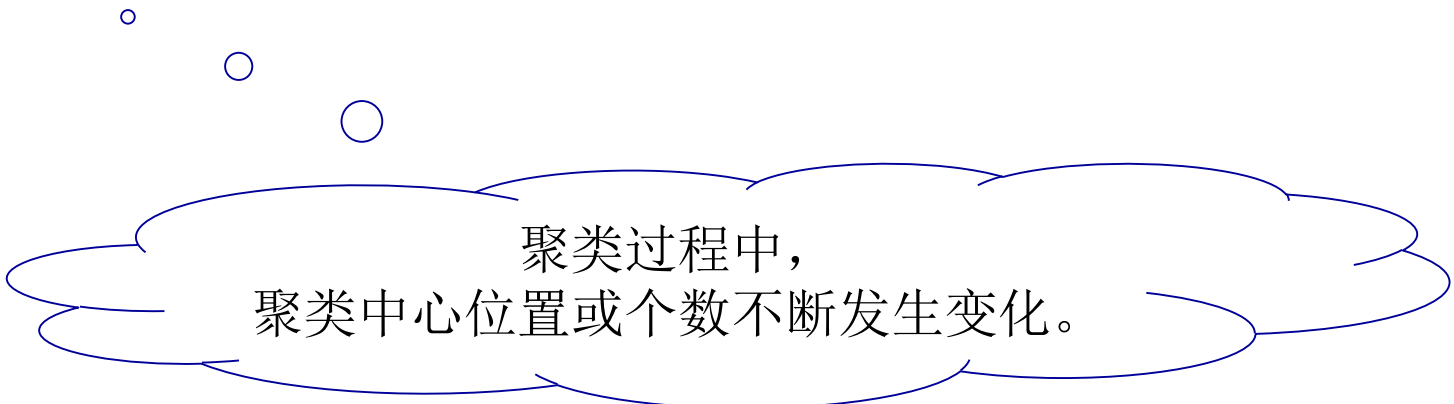
2.5.1 K-均值算法

“动态”聚类法

。

○

○



聚类过程中，
聚类中心位置或个数不断发生变化。

2. 算法讨论

结果很容易受到所选**聚类中心个数**和其**初始位置**，以及**模式样本的几何性质**及**读入次序**等的**影响**。

实际应用中，需要试探不同的K值和选择不同的聚类中心起始值。

Algorithm 8.2: $KMedoids(D, K, Dis)$ – K -medoids clustering using arbitrary distance metric Dis .

Input : data $D \subseteq \mathcal{X}$; number of clusters $K \in \mathbb{N}$;
distance metric $Dis: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

Output : K medoids $\mu_1, \dots, \mu_K \in D$, representing a predictive clustering of \mathcal{X} .

```
1 randomly pick  $K$  data points  $\mu_1, \dots, \mu_K \in D$ ;  
2 repeat  
3   assign each  $x \in D$  to  $\arg\min_j Dis(x, \mu_j)$ ;  
4   for  $j = 1$  to  $k$  do  
5      $D_j \leftarrow \{x \in D \mid x \text{ assigned to cluster } j\}$ ;  
6      $\mu_j = \arg\min_{x \in D_j} \sum_{x' \in D_j} Dis(x, x')$ ;  
7   end  
8 until no change in  $\mu_1, \dots, \mu_K$ ;  
9 return  $\mu_1, \dots, \mu_K$ ;
```

3.4 线性判别函数的几何性质

3.4.1 模式空间与超平面

1. 概念

模式空间：以 n 维模式向量 \mathbf{X} 的 n 个分量为坐标变量的欧氏空间。

模式向量：点、有向线段。

线性分类：用 $d(\mathbf{X})$ 进行分类，相当于用超平面 $d(\mathbf{X})=0$ 把模式空间分成不同的决策区域。

2. 讨论

设判别函数：
$$d(\mathbf{X}) = \mathbf{W}_0^T \mathbf{X} + w_{n+1}$$

式中， $\mathbf{W}_0 = [w_1, w_2, \dots, w_n]^T$ ， $\mathbf{X} = [x_1, x_2, \dots, x_n]^T$ 。

超平面：
$$d(\mathbf{X}) = \mathbf{W}_0^T \mathbf{X} + w_{n+1} = 0$$

(1) 模式向量 X_1 和 X_2 在超平面上

$$\mathbf{W}_0^T \mathbf{X}_1 + w_{n+1} = \mathbf{W}_0^T \mathbf{X}_2 + w_{n+1}$$

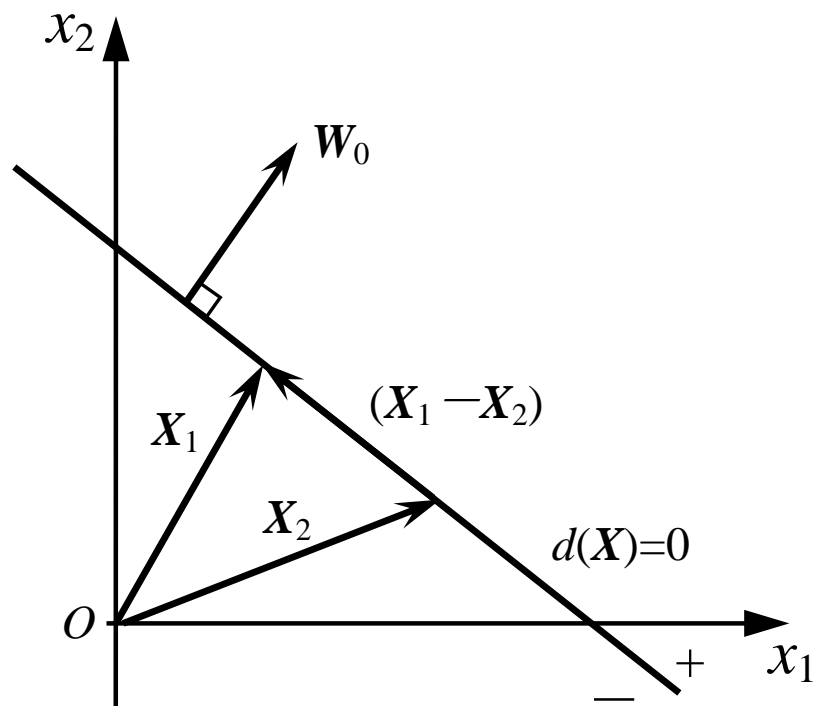
$$\mathbf{W}_0^T (\mathbf{X}_1 - \mathbf{X}_2) = 0$$

可见： \mathbf{W}_0 是超平面的法向量，
所指方向即为超平面的正侧。

记超平面的单位法线向量为 \mathbf{U} ：

$$\mathbf{U} = \frac{\mathbf{W}_0}{\|\mathbf{W}_0\|}$$

$$\|\mathbf{W}_0\| = \sqrt{w_1^2 + w_2^2 + \dots + w_n^2}$$



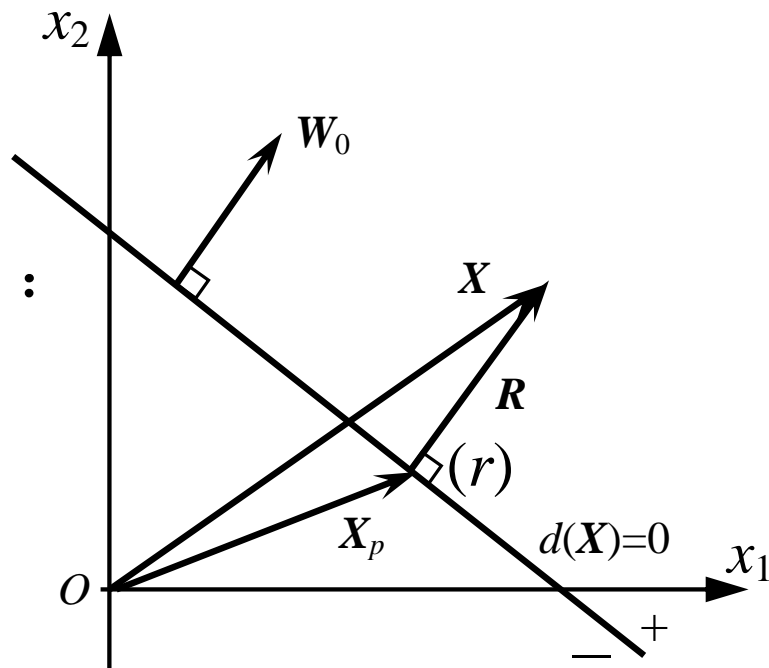
(2) X 不在超平面上

将 X 向超平面投影得向量 X_p ,
构造向量 \mathbf{R} (投影点 X_p 到 X 的向量) :

$$\mathbf{R} = r \cdot \mathbf{U} = r \frac{\mathbf{W}_0}{\|\mathbf{W}_0\|}$$

r : X 到超平面的代数距离。有

$$\mathbf{X} = \mathbf{X}_p + \mathbf{R} = \mathbf{X}_p + r \frac{\mathbf{W}_0}{\|\mathbf{W}_0\|}$$



$$d(\mathbf{X}) = \mathbf{W}_0^T \mathbf{X} + w_{n+1}$$

$$\begin{aligned} d(\mathbf{X}) &= \mathbf{W}_0^T \left(\mathbf{X}_p + r \frac{\mathbf{W}_0}{\|\mathbf{W}_0\|} \right) + w_{n+1} = (\mathbf{W}_0^T \mathbf{X}_p + w_{n+1}) + \mathbf{W}_0^T \cdot r \frac{\mathbf{W}_0}{\|\mathbf{W}_0\|} \\ &= r \|\mathbf{W}_0\| \end{aligned}$$

可见：判别函数 $d(X)$ 正比于点 X 到超平面的代数距离。

X 到超平面的代数距离: $r \triangleq \frac{d(X)}{\|\mathbf{W}_0\|}$

换言之: 点 X 到超平面的代数距离正比于 $d(X)$ 函数值。

(3) X 在原点

$$d(\mathbf{X}) = \mathbf{W}_0^T \mathbf{X} + w_{n+1} = w_{n+1}$$

得: 原点到超平面的代数距离

$$r_0 = \frac{w_{n+1}}{\|\mathbf{W}_0\|}$$

可见: 原点在超平面的正负侧位置由阈值权 w_{n+1} 决定:

$w_{n+1} > 0$ 时, 原点在超平面的正侧;

$w_{n+1} < 0$ 时, 原点在超平面负侧;

$w_{n+1} = 0$ 时, 超平面通过原点。

3.6 感知器算法

对线性判别函数，当模式维数已知时，判别函数的形式实际上已经确定，如：三维时

$$d(\mathbf{X}) = w_1x_1 + w_2x_2 + w_3x_3 + w_4 = \mathbf{W}^T \mathbf{X}$$

$$\mathbf{X} = [x_1, x_2, x_3, 1]^T \quad \mathbf{W} = [w_1, w_2, w_3, w_4]^T$$

只要求出权向量，分类器的设计即告完成。本节开始介绍如何通过各种算法，利用已知类别的模式样本训练权向量 \mathbf{W} 。

1. 概念理解

1) 训练与学习

训练：用已知类别的模式样本指导机器对分类规则进行反复修改，最终使分类结果与已知类别信息完全相同的过程。

学习：从分类器的角度讲 $\left\{ \begin{array}{l} \text{非监督学习} \\ \text{有监督学习} \end{array} \right. \longleftrightarrow \text{训练}$

2) 确定性分类器

处理确定可分情况的分类器。通过几何方法将特征空间分解为对应不同类的子空间，又称为**几何分类器**。

3) **感知器 (Perceptron)**

一种早期神经网络分类学习模型，属于有关机器学习的仿生学领域中的问题，由于**无法实现非线性分类**而下马（Minsky and Papert）。但“赏罚概念（reward-punishment）”得到广泛应用。

2. 感知器算法

两类线性可分的模式类： ω_1, ω_2 ，设 $d(\mathbf{X}) = \mathbf{W}^T \mathbf{X}$

其中， $\mathbf{W} = [w_1, w_2, w_3, w_4]^T$ ， $\mathbf{X} = [x_1, x_2, \dots, x_n, 1]^T$

应具有性质 $d(\mathbf{X}) = \mathbf{W}^T \mathbf{X} \begin{cases} > 0, & \text{若 } \mathbf{X} \in \omega_1 \\ < 0, & \text{若 } \mathbf{X} \in \omega_2 \end{cases}$

对样本进行规范化处理，即 ω_2 类样本全部乘以 (-1) ，则有：

$$d(\mathbf{X}) = \mathbf{W}^T \mathbf{X} > 0$$

感知器算法的基本思想：用训练模式验证当前权向量的合理性，如果不合理，就根据误差进行反向纠正，直到全部训练样本都被合理分类。本质上是梯度下降方法类。

感知器算法步骤：

(1) 选择 N 个分属于 ω_1 和 ω_2 类的模式样本构成训练样本集

$$\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$$

构成增广向量形式，并进行规范化处理。任取权向量初始值 $\mathbf{W}(1)$ ，开始迭代。初始迭代次数 $k=1$ 。

(2) 用全部训练样本进行一轮迭代，计算 $\mathbf{W}^T(k)\mathbf{X}_i$ 的值，并修正权向量。

分两种情况，更新权向量的值：

① 若 $\mathbf{W}^T(k)\mathbf{X}_i \leq 0$ ，说明分类器对第 i 个模式做了错误分类，

权向量校正为： $\mathbf{W}(k+1) = \mathbf{W}(k) + c\mathbf{X}_i$

c ：正的校正增量（步长）。

② 若 $\mathbf{W}^T(k)\mathbf{X}_i > 0$ ，分类正确，权向量不变：

$$\mathbf{W}(k+1) = \mathbf{W}(k)$$

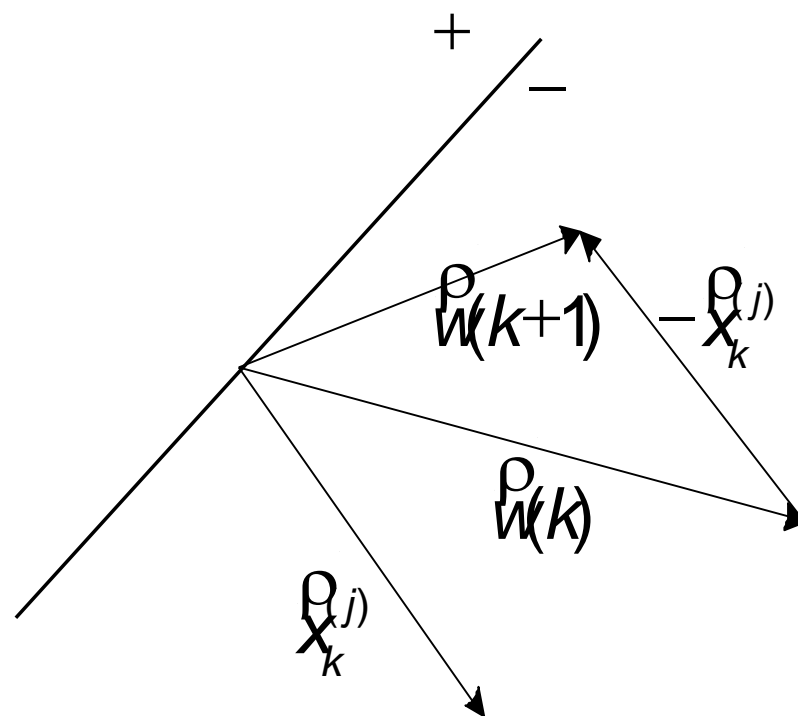
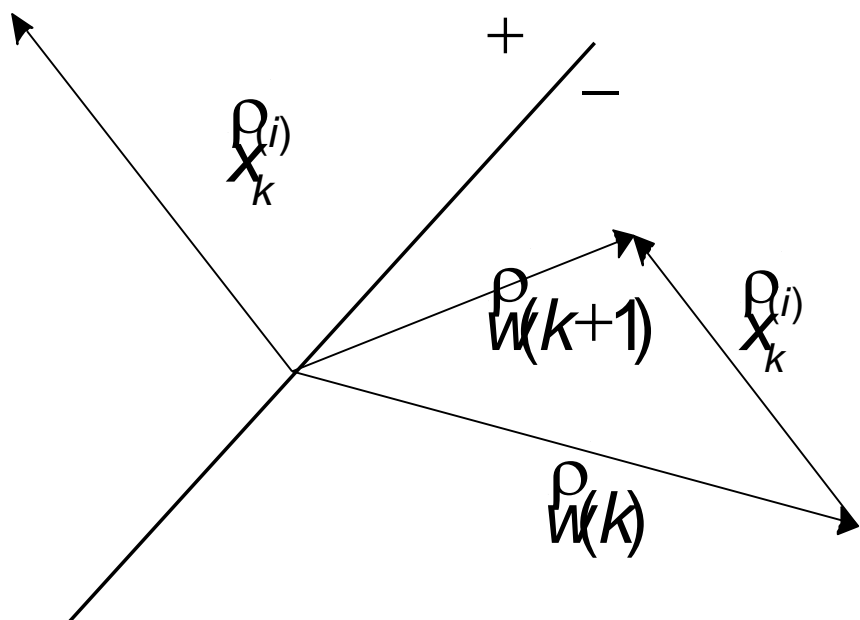
统一写为：

$$\mathbf{W}(k+1) = \begin{cases} \mathbf{W}(k) & \text{若 } \mathbf{W}^T(k)\mathbf{X}_i > 0 \\ \mathbf{W}(k) + c\mathbf{X}_i & \text{若 } \mathbf{W}^T(k)\mathbf{X}_i \leq 0 \end{cases}$$

(3) 分析分类结果：只要有一个错误分类，回到（2），直至对所有样本正确分类。

感知器算法是一种赏罚过程：

- 分类**正确时**对权向量“赏”——**不罚**，即权向量不变；
- 分类**错误时**对权向量“罚”——**修改**，向正确的方向转换。



权空间中感知器算法权矢量校正过程示意图

注意：右边的 $x_k^{(j)}$ 是未进行符号规范化的属于 j -类矢量，符号规范化后，与权矢量的夹角大于 90° ，内积为负，也需要调整权向量！左边的 $x_k^{(i)}$ 是 i -类矢量，也需要校正权矢量。

例3.8 已知两类训练样本

$$\omega_1 : \mathbf{X}_1 = [0, 0]^T \quad \mathbf{X}_2 = [0, 1]^T$$

$$\omega_2 : \mathbf{X}_3 = [1, 0]^T \quad \mathbf{X}_4 = [1, 1]^T$$

用感知器算法求出将模式分为两类的权向量解和判别函数。

解：所有样本写成增广向量形式；

进行规范化处理，属于 ω_2 的样本乘以 (-1) 。

$$\mathbf{X}_1 = [0, 0, 1]^T \quad \mathbf{X}_2 = [0, 1, 1]^T \quad \mathbf{X}_3 = [-1, 0, -1]^T \quad \mathbf{X}_4 = [-1, -1, -1]^T$$

任取 $\mathbf{W}(1)=\mathbf{0}$ ，取 $c=1$ ，迭代过程为：

第一轮：

$$\mathbf{W}^T(1)\mathbf{X}_1 = [0,0,0] \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 0, \leq 0, \quad \text{故 } \mathbf{W}(2) = \mathbf{W}(1) + \mathbf{X}_1 = [0,0,1]^T$$

$$\mathbf{W}^T(2)\mathbf{X}_2 = [0,0,1] \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = 1, > 0, \quad \text{故 } \mathbf{W}(3) = \mathbf{W}(2) = [0,0,1]^T$$

$$\mathbf{W}^T(3)\mathbf{X}_3 = [0,0,1] \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix} = -1, \leq 0, \quad \text{故 } \mathbf{W}(4) = \mathbf{W}(3) + \mathbf{X}_3 = [-1,0,0]^T$$

$$\mathbf{W}^T(4)\mathbf{X}_4 = [-1,0,0] \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} = 1, > 0, \quad \text{故 } \mathbf{W}(5) = \mathbf{W}(4) = [-1,0,0]^T$$

有两个 $\mathbf{W}^T(k)\mathbf{X}_i \leq 0$ 的情况（错判），进行第二轮迭代。

第二轮: $W^T(5)X_1 = 0 \leq 0$, 故 $W(6) = W(5) + X_1 = [-1, 0, 1]^T$
 $W^T(6)X_2 = 1 > 0$, 故 $W(7) = W(6) = [-1, 0, 1]^T$
 $W^T(7)X_3 = 0 \leq 0$, 故 $W(8) = W(7) + X_3 = [-2, 0, 0]^T$
 $W^T(8)X_4 = 2 > 0$, 故 $W(9) = W(8) = [-2, 0, 0]^T$

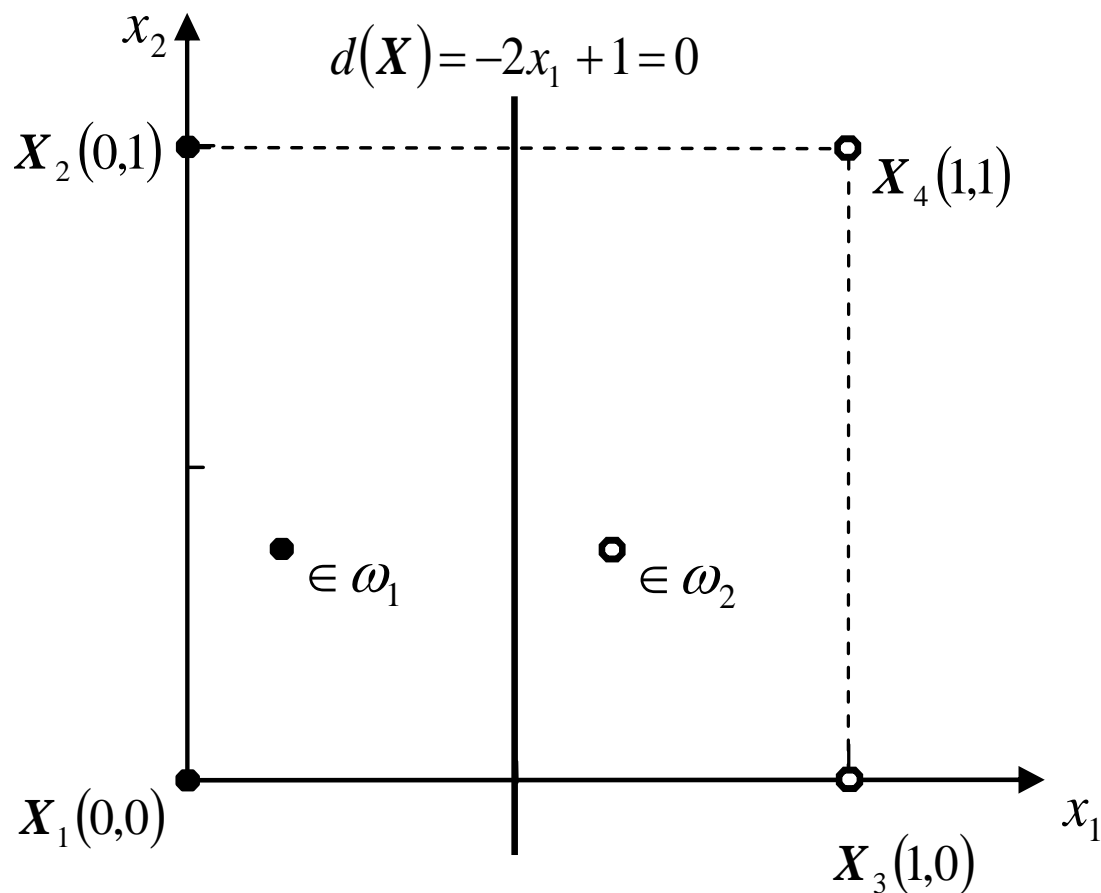
第三轮: $W^T(9)X_1 = 0 \leq 0$, 故 $W(10) = W(9) + X_1 = [-2, 0, 1]^T$
 $W^T(10)X_2 = 1 > 0$, 故 $W(11) = W(10)$
 $W^T(11)X_3 = 1 > 0$, 故 $W(12) = W(11)$
 $W^T(12)X_4 = 1 > 0$, 故 $W(13) = W(12)$

第四轮: $W^T(13)X_1 = 1 > 0$, 故 $W(14) = W(13)$
 $W^T(14)X_2 = 1 > 0$, 故 $W(15) = W(14)$
 $W^T(15)X_3 = 1 > 0$, 故 $W(16) = W(15)$
 $W^T(16)X_4 = 1 > 0$, 故 $W(17) = W(16)$

该轮迭代的分类结果全部正确，故解向量 $\mathbf{W} = [-2, 0, 1]^T$

相应的判别函数为： $d(\mathbf{X}) = -2x_1 + 1$

判别界面 $d(\mathbf{X})=0$ 如图示。法向量为 $(-2, 0)$ ，指向 $-x_1$ 方向。



当 c 、 $\mathbf{W}(1)$ 取其他值时，
结果可能不一样，所以
感知器算法的解不唯一。

4.2 贝叶斯决策

4.2.1 最小错误率贝叶斯决策

1. 问题分析

讨论模式集的分类，目的是确定 \mathbf{X} 属于那一类，所以要看 \mathbf{X} 来自哪类的概率大。在下列三种概率中：

- ① 先验概率 $P(\omega_i)$
- ② 类(条件)概率密度 $p(\mathbf{X} | \omega_i)$
- ③ 后验概率 $P(\omega_i | \mathbf{X})$

采用哪种概率进行分类最合理？

后验概率 $P(\omega_i | \mathbf{X})$

2. 决策规则

设有 M 类模式，

若 $P(\omega_i | \mathbf{X}) = \max \{P(\omega_j | \mathbf{X})\}, j = 1, 2, \dots, M$ 则 $\mathbf{X} \in \omega_i$ 类 (4-6)

—— 最小错误率贝叶斯决策规则

若 $P(\omega_i | \mathbf{X}) = \max \{P(\omega_j | \mathbf{X})\}, j = 1, 2, \dots, M$

则 $\mathbf{X} \in \omega_i$ 类

几种等价形式:

虽然后验概率 $P(\omega_i | \mathbf{X})$ 可以提供有效的分类信息，但直接获得后验概率是困难的。根据Bayes定理，后验概率可以由类概率密度函数和先验概率计算得出。实际上，先验概率 $P(\omega_i)$ 和类概率密度函数 $p(\mathbf{X} | \omega_i)$ 更容易从统计资料中容易获得。

$$P(\omega_i | \mathbf{X}) = \frac{p(\mathbf{X} | \omega_i) P(\omega_i)}{\sum_{i=1}^M p(\mathbf{X} | \omega_i) P(\omega_i)}$$

可知，分母与 i 无关，即与分类无关，故分类规则又可表示为：

若 $p(\mathbf{X} | \omega_i) P(\omega_i) = \max \{p(\mathbf{X} | \omega_j) P(\omega_j)\} \quad j = 1, 2, \dots, M$ ， 则 $\mathbf{X} \in \omega_i$ 类

对两类问题，(4-7)式相当于

若 $p(\mathbf{X} | \omega_1)P(\omega_1) > p(\mathbf{X} | \omega_2)P(\omega_2)$, 则 $\mathbf{X} \in \omega_1$

若 $p(\mathbf{X} | \omega_1)P(\omega_1) < p(\mathbf{X} | \omega_2)P(\omega_2)$, 则 $\mathbf{X} \in \omega_2$

可改写为:

$$\text{若 } l_{12}(\mathbf{X}) = \frac{p(\mathbf{X} | \omega_1)}{p(\mathbf{X} | \omega_2)} \gtrless \frac{P(\omega_2)}{P(\omega_1)}, \text{ 则 } \mathbf{X} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \quad (4-8)$$

统计学中称 $l_{12}(\mathbf{X})$ 为似然比, $P(\omega_2)/P(\omega_1)$ 为似然比阈值。

对(4-9)式取自然对数, 有:

若 $h(\mathbf{X}) = \ln l_{12}(\mathbf{X})$

$$= \ln p(\mathbf{X} | \omega_1) - \ln p(\mathbf{X} | \omega_2) \gtrless \ln \frac{P(\omega_2)}{P(\omega_1)}, \text{ 则 } \mathbf{X} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \quad (4-9)$$

(4-7), (4-8), (4-9)都是最小错误率贝叶斯决策规则的等价形式。

例4.1 假定在细胞识别中，病变细胞的先验概率和正常细胞的先验概率分别为 $P(\omega_1)=0.05$ ， $P(\omega_2)=0.95$ 。现有一待识别细胞，其观察值为 \mathbf{X} ，从类条件概率密度发布曲线上查得：

$$p(\mathbf{X} | \omega_1) = 0.5 \quad p(\mathbf{X} | \omega_2) = 0.2$$

试对细胞 \mathbf{X} 进行分类。

解：[方法1] 通过后验概率计算。

$$\begin{aligned} P(\omega_1 | \mathbf{X}) &= \frac{p(\mathbf{X} | \omega_1)P(\omega_1)}{\sum_{i=1}^2 p(\mathbf{X} | \omega_i)P(\omega_i)} \\ &= \frac{0.5 \times 0.05}{0.05 \times 0.5 + 0.95 \times 0.2} \approx 0.16 \\ P(\omega_2 | \mathbf{X}) &= \frac{0.2 \times 0.95}{0.05 \times 0.5 + 0.95 \times 0.2} \approx 0.884 \end{aligned}$$

$$\ominus P(\omega_2 | \mathbf{X}) > P(\omega_1 | \mathbf{X}) \quad \therefore \mathbf{X} \in \omega_2$$

[方法2]: 利用先验概率和类概率密度计算。

$$p(\mathbf{X} | \omega_1)P(\omega_1) = 0.5 \times 0.05 = 0.025$$

$$p(\mathbf{X} | \omega_2)P(\omega_2) = 0.2 \times 0.95 = 0.19$$

$$\ominus p(\mathbf{X} | \omega_2)P(\omega_2) > p(\mathbf{X} | \omega_1)P(\omega_1)$$

$\therefore \mathbf{X} \in \omega_2$, 是正常细胞。

4.2.2 最小风险贝叶斯决策

1. 风险的概念

- * 自动灭火系统:

- * 疾病诊断:

不同的错判造成的损失不同，因此风险不同，两者紧密相连。

考虑到对某一类的错判要比对另一类的错判更为关键，把最小错误率的贝叶斯判决做一些修改，提出了“**条件风险**”的概念。

最小风险贝叶斯决策基本思想：

以各种错误分类所造成的**条件风险**（后验风险）最小为规则，进行分类决策。

2. 决策规则

对 M 类问题，如果把观察样本 \mathbf{X} 判定为 ω_i 类记为决策 a_i ，则**决策 a_i 的条件(平均)风险 $r_i(\mathbf{X})$** 是指此时决策 a_i 造成的**平均损失**（**条件**：指观察到 \mathbf{X} 。**风险**：指**平均损失**。这里指，把 \mathbf{X} 实属不同类别却被硬判为 ω_i 类时引起的**损失**，通过后验概率进行加权**平均**）

$$r_i(\mathbf{X}) = \sum_{j=1}^M L_{ij}(\mathbf{X}) P(\omega_j | \mathbf{X})$$

L_{ij} 对 P 作加权平均

式中，

i ——分类判决后指定的判决号；

j ——样本实际属于的类别号；

L_{ij} ——将自然属性是 ω_j 类的样本决策为 ω_i 类时的是非代价，即损失函数。注意下标顺序及其含义： **$L_{i \leftarrow j}$**

$$L_{ij}(\mathbf{X}) = \begin{cases} 0 \text{ 或 负值} & i = j \text{ 时} \\ \text{正值} & i \neq j \text{ 时} \end{cases}$$

自然属性为 j 类的样本，被划分到 i 类中，在 i 类中产生一错误分类，风险增加。

注：这里的 L_{ij} 与孙即祥书中的 λ_{ij} 的含义相反！

对每个 \mathbf{X} 有 M 种可能的类别划分， \mathbf{X} 被判决为每一类的条件均风险分别为 $r_1(\mathbf{X})$, $r_2(\mathbf{X})$, ..., $r_M(\mathbf{X})$ 。决策规则：

$$\text{若 } r_k(\mathbf{X}) = \min \{ r_i(\mathbf{X}), \quad i = 1, \dots, M \} \quad \text{则 } \mathbf{X} \in \omega_k$$

每个 \mathbf{X} 都按条件风险最小决策，则总的条件风险也最小。
总的条件风险称为**总的风险**。

条件风险 与 总的风险 的区别	{	条件风险：对某个样本而言。 总的风险：对模式总体而言。
--------------------	---	--------------------------------

1) 多类情况

设有 M 类，对于任一 \mathbf{X} 对应 M 个条件风险（后验风险）：

$$r_i(\mathbf{X}) = \sum_{j=1}^M L_{ij}(\mathbf{X}) P(\omega_j | \mathbf{X}) \quad , \quad i=1, 2, \dots, M$$

用先验概率和条件概率的形式：

$$\begin{aligned} r_i(\mathbf{X}) &= \sum_{j=1}^M L_{ij} P(\omega_j | \mathbf{X}) = \sum_{j=1}^M L_{ij} \frac{p(\mathbf{X} | \omega_j) P(\omega_j)}{p(\mathbf{X})} \\ &= \frac{1}{p(\mathbf{X})} \sum_{j=1}^M L_{ij} p(\mathbf{X} | \omega_j) P(\omega_j) \end{aligned}$$

\because $p(\mathbf{X})$ 对所有类别一样，不提供分类信息。

$$\therefore r_i(\mathbf{X}) = \sum_{j=1}^M L_{ij} p(\mathbf{X} | \omega_j) P(\omega_j) \quad , \quad i=1,2,\dots,M$$

决策规则为：

若 $r_k(\mathbf{X}) < r_i(\mathbf{X})$, $i = 1, 2, \dots, M$; $i \neq k$, 则 $\mathbf{X} \in \omega_k$

2) 两类情况：对样本 \mathbf{X}

当 \mathbf{X} 被判为 ω_1 类时：

$$r_1(\mathbf{X}) = L_{11}p(\mathbf{X} | \omega_1)P(\omega_1) + L_{12}p(\mathbf{X} | \omega_2)P(\omega_2)$$

当 \mathbf{X} 被判为 ω_2 类时：

$$r_2(\mathbf{X}) = L_{21}p(\mathbf{X} | \omega_1)P(\omega_1) + L_{22}p(\mathbf{X} | \omega_2)P(\omega_2)$$

决策规则：

$$\text{若 } r_1(\mathbf{X}) < r_2(\mathbf{X}) \quad \text{则 } \mathbf{X} \in \omega_1 \quad (4-15)$$

$$\text{若 } r_1(\mathbf{X}) > r_2(\mathbf{X}) \quad \text{则 } \mathbf{X} \in \omega_2 \quad (4-16)$$

由 (4-15) 式：

$$\underline{L_{11}p(\mathbf{X} | \omega_1)P(\omega_1) + L_{12}p(\mathbf{X} | \omega_2)P(\omega_2)} < \underline{L_{21}p(\mathbf{X} | \omega_1)P(\omega_1) + L_{22}p(\mathbf{X} | \omega_2)P(\omega_2)}$$

$$(L_{12} - L_{22})p(\mathbf{X} | \omega_2)P(\omega_2) < (L_{21} - L_{11})p(\mathbf{X} | \omega_1)P(\omega_1)$$

$$\frac{p(\mathbf{X} | \omega_1)}{p(\mathbf{X} | \omega_2)} > \frac{(L_{12} - L_{22})P(\omega_2)}{(L_{21} - L_{11})P(\omega_1)}$$

$$\therefore r_i(\mathbf{X}) = \sum_{j=1}^M L_{ij}p(\mathbf{X} | \omega_j)P(\omega_j)$$

$$\frac{p(\mathbf{X} | \omega_1)}{p(\mathbf{X} | \omega_2)} > \frac{(L_{12} - L_{22})P(\omega_2)}{(L_{21} - L_{11})P(\omega_1)}$$

令： $l_{12}(\mathbf{X}) = \frac{p(\mathbf{X} | \omega_1)}{p(\mathbf{X} | \omega_2)}$ ，称似然比；

$\theta_{12} = \frac{(L_{12} - L_{22})P(\omega_2)}{(L_{21} - L_{11})P(\omega_1)}$ ，为阈值。

判别步骤：

- ① 定义损失函数 L_{ij} 。
- ② 计算似然比阈值 θ_{12} 。
- ③ 计算似然比 $l_{12}(\mathbf{X})$ 。
- ④ 若 $l_{12}(\mathbf{X}) > \theta_{12}$ ， 则 $\mathbf{X} \in \omega_1$
 若 $l_{12}(\mathbf{X}) < \theta_{12}$ ， 则 $\mathbf{X} \in \omega_2$
 若 $l_{12}(\mathbf{X}) = \theta_{12}$ ， 任意判决

类概率密度函数
 $p(\mathbf{X} | \omega_i)$ 也称 ω_i
 的似然函数

例4.2 在细胞识别中，病变细胞和正常细胞的先验概率 分别为

$$P(\omega_1) = 0.05, \quad P(\omega_2) = 0.95$$

现有一待识别细胞，观察值为 \mathbf{X} ，从类概率密度分布曲线上查得

$$p(\mathbf{X} | \omega_1) = 0.5, \quad p(\mathbf{X} | \omega_2) = 0.2$$

损失函数分别为 $L_{11}=0$ ， $L_{21}=10$ ， $L_{22}=0$ ， $L_{12}=1$ 。按最小风险贝叶斯决策分类。

解：计算 θ_{12} 和 $l_{12}(\mathbf{X})$ 得：

$$\theta_{12} = \frac{(L_{12} - L_{22})P(\omega_2)}{(L_{21} - L_{11})P(\omega_1)} = \frac{(1-0) \times 0.95}{(10-0) \times 0.05} = 1.9$$

$$l_{12}(\mathbf{X}) = \frac{p(\mathbf{X} | \omega_1)}{p(\mathbf{X} | \omega_2)} = \frac{0.5}{0.2} = 2.5$$

$$\because l_{12}(\mathbf{X}) > \theta_{12} \quad \therefore \mathbf{X} \in \omega_1$$

为病变细胞。

思考：为什么同样检查结果，这里被判为病变细胞？

5.4 基于K-L变换的多类模式特征提取

特征提取的目的：

对一类模式：维数压缩。

对多类模式：维数压缩，突出类别的可分性。

卡洛南-洛伊（Karhunen-Loeve）变换（K-L变换）：

- * 一种常用的特征提取方法；
- * 最小均方误差逼近意义下的最优正交变换；
- * 适用于任意的概率密度函数；
- * 在消除模式特征之间的相关性、突出差异性方面有最优的效果。

分为：连续K-L变换 离散K-L变换

1. K-L展开式

设 $\{\mathbf{X}\}$ 是 n 维随机模式向量 \mathbf{X} 的集合, 对每一个 \mathbf{X} 可以用确定的完备归一化正交向量系 $\{\mathbf{u}_j\}$ 中的正交向量**展开式**:

$$\mathbf{X} = \sum_{j=1}^n a_j \mathbf{u}_j \quad a_j: \text{随机系数};$$

用有限项 ($d < n$) 估计 \mathbf{X} 时: $\hat{\mathbf{X}} = \sum_{j=1}^d a_j \mathbf{u}_j$

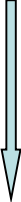
引起的均方误差: $\xi = E[(\mathbf{X} - \hat{\mathbf{X}})^T (\mathbf{X} - \hat{\mathbf{X}})]$

代入 \mathbf{X} 、 $\hat{\mathbf{X}}$, 利用 $\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$



$$\xi = E\left[\sum_{j=d+1}^n a_j^2\right]$$

$$\xi = E\left[\sum_{j=d+1}^n a_j^2\right]$$

由 $\mathbf{X} = \sum_{j=1}^n a_j \mathbf{u}_j$ 两边  左乘 \mathbf{u}_j^T 得 $a_j = \mathbf{u}_j^T \mathbf{X}$ 。

$$\xi = E\left[\sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{X} \mathbf{X}^T \mathbf{u}_j\right]$$

$$= \sum_{j=d+1}^n \mathbf{u}_j^T E[\mathbf{X} \mathbf{X}^T] \mathbf{u}_j$$

\mathbf{u}_j 为确定性向量

$$= \sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j \quad \mathbf{R}: \text{自相关矩阵。}$$

不同的 $\{\mathbf{u}_j\}$ 对应不同的均方误差， \mathbf{u}_j 的选择应使 ξ 最小。

利用拉格朗日乘数法求使 ξ 最小的正交系 $\{\mathbf{u}_j\}$ ，令

$$g(\mathbf{u}_j) = \sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j - \sum_{j=d+1}^n \lambda_j (\mathbf{u}_j^T \mathbf{u}_j - 1) \quad \lambda_j: \text{拉格朗日乘数}$$

$$g(\mathbf{u}_j) = \sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j - \sum_{j=d+1}^n \lambda_j (\mathbf{u}_j^T \mathbf{u}_j - 1)$$

用函数 $g(\mathbf{u}_j)$ 对 \mathbf{u}_j 求导，并令导数为零，得

$$(\mathbf{R} - \lambda_j \mathbf{I}) \mathbf{u}_j = 0 \quad j = d + 1, \dots, n$$

——正是矩阵 \mathbf{R} 与其特征值和对应特征向量的关系式。

说明：当用 \mathbf{X} 的自相关矩阵 \mathbf{R} 的特征值对应的特征向量展开 \mathbf{X} 时，截断误差最小。

选前 d 项估计 \mathbf{X} 时引起的均方误差为

$$\xi = \sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j = \sum_{j=d+1}^n \text{Tr}[\mathbf{u}_j \mathbf{R} \mathbf{u}_j^T] = \sum_{j=d+1}^n \lambda_j$$

λ_j 决定截断的均方误差， λ_j 的值小，那么 ξ 也小。

因此，当用 \mathbf{X} 的正交展开式中的前 d 项估计 \mathbf{X} 时，展开式中的 \mathbf{u}_j 应当是前 d 个较大的特征值对应的特征向量。

K-L变换具体方法:

对 R 的特征值由大到小进行排队: $\lambda_1 \geq \lambda_2 \geq \Lambda \geq \lambda_d \geq \lambda_{d+1} \geq \Lambda$

均方误差最小的 \mathbf{X} 的近似式: $\mathbf{X} = \sum_{j=1}^d a_j \mathbf{u}_j$ —— K-L展开

矩阵形式: $\mathbf{X} = \mathbf{U}\mathbf{a}$ (5-49)

式中, $\mathbf{a} = [a_1, a_2, \dots, a_d]^T$, $\mathbf{U}_{n \times d} = [\mathbf{u}_1, \dots, \mathbf{u}_j, \dots, \mathbf{u}_d]$ 。

其中: $\mathbf{u}_j = [u_{j1}, u_{j2}, \dots, u_{jn}]^T$

$$\mathbf{U}^T \mathbf{U} = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \dots \\ \mathbf{u}_d^T \end{bmatrix} [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_d] = \mathbf{I}$$

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$$

对式(5-49)两边左乘 \mathbf{U}^T : $\mathbf{a} = \mathbf{U}^T \mathbf{X}$ —— K-L变换

系数向量 \mathbf{a} 就是变换后的模式向量。

2. 利用自相关矩阵的K-L变换进行特征提取

设 \mathbf{X} 是 n 维模式向量， $\{\mathbf{X}\}$ 是来自 M 个模式类的样本集，总样本数目为 N 。将 \mathbf{X} 变换为 d 维 ($d < n$) 向量的方法：

第一步：求样本集 $\{\mathbf{X}\}$ 的总体自相关矩阵 \mathbf{R} 。

$$\mathbf{R} = E[\mathbf{X}\mathbf{X}^T] \approx \frac{1}{N} \sum_{j=1}^N \mathbf{X}_j \mathbf{X}_j^T$$

决定压缩
后的维数 d

第二步：求 \mathbf{R} 的特征值 λ_j ， $j = 1, 2, \dots, n$ 。对特征值由大到小进行排队，选择前 d 个较大的特征值。

第三步：计算 d 个特征值对应的特征向量 \mathbf{u}_j ， $j = 1, 2, \dots, d$ ，归一化后构成变换矩阵 \mathbf{U} 。

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d]$$

第四步：对 $\{\mathbf{X}\}$ 中的每个 \mathbf{X} 进行 K-L 变换，得变换后向量 \mathbf{X}^* ：

$$\mathbf{X}^* = \mathbf{U}^T \mathbf{X}$$

d 维向量 \mathbf{X}^* 就是代替 n 维向量 \mathbf{X} 进行分类的模式向量。

3. 使用不同散布矩阵进行K-L变换

根据不同的散布矩阵进行K-L变换，对保留分类鉴别信息的效果不同。

1) 采用多类类内散布矩阵 S_w 作 K-L 变换

多类类内散布矩阵：

$$S_w = \sum_{i=1}^c P(\omega_i) E[(X - M_i)(X - M_i)^T | X \in \omega_i]$$

若要突出各类模式的主要特征分量的分类作用：

选用对应于大特征值的特征向量组成变换矩阵；

若要使同一类模式聚集于最小的特征空间范围：

选用对应于小特征值的特征向量组成变换矩阵。

2) 采用类间散布矩阵 S_b 作 K-L 变换

类间散布矩阵：

$$S_b = \sum_{i=1}^c P(\omega_i) (M_i - M_0)(M_i - M_0)^T$$

适用于类间距离比类内距离大得多的多类问题，选择与大特征值对应的特征向量组成变换矩阵。

3) 采用总体散布矩阵 S_t 作 K-L 变换

把多类模式合并起来看成一个总体分布。

总体散布矩阵: $S_t = E[(X - M_0)(X - M_0)^T] = S_b + S_w$

适合于多类模式在总体分布上具有良好的可分性的情况。

采用大特征值对应的特征向量组成变换矩阵，能够保留模式原有分布的主要结构。

利用K-L变换进行特征提取的优点:

- 1) 在均方逼近误差最小的意义下使新样本集 $\{X^*\}$ 逼近原样本集 $\{X\}$ 的分布，既压缩了维数、又保留了数据集的分布信息和类别鉴别信息。

2) 变换后的新模式向量各分量相对总体均值的方差等于原样本集总体自相关矩阵的大特征值，表明变换加强了模式类之间的差异性。

$$C^* = E\{(X^* - M^*)(X^* - M^*)^T\} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{bmatrix}$$

3) C^* 为对角矩阵说明了变换后样本各分量特征互不相关，即消除了原分量特征间的相关性，便于进一步进行特征的选择。

K-L变换的不足之处:

- 1) 对两类问题容易得到较满意的结果。类别愈多，效果愈差。
- 2) 需要通过足够多的样本估计样本集的协方差矩阵或其它类型的散布矩阵。当样本数不足时，矩阵的估计会变得十分粗略，变换的优越性也就不能充分地显示出来。
- 3) 矩阵的本征值和本征向量缺乏统一的快速算法，计算较困难。

例5.3 两个二维模式类的样本分别为【注意：第三步的计算!】

$$\omega_1: \mathbf{X}_1=[2, 2]^T, \mathbf{X}_2=[2, 3]^T, \mathbf{X}_3=[3, 3]^T$$

$$\omega_2: \mathbf{X}_4=[-2, -2]^T, \mathbf{X}_5=[-2, -3]^T, \mathbf{X}_6=[-3, -3]^T$$

利用总体自相关矩阵 \mathbf{R} 作K-L变换，把原样本压缩成一维样本。

解：第一步：计算总体自相关矩阵 \mathbf{R} 。

$$\mathbf{R} = E\{\mathbf{X}\mathbf{X}^T\} = \frac{1}{6} \sum_{j=1}^6 \mathbf{X}_j \mathbf{X}_j^T = \begin{bmatrix} 5.7 & 6.3 \\ 6.3 & 7.3 \end{bmatrix}$$

第二步：计算 \mathbf{R} 的本征值，并选择较大者。由 $|\mathbf{R} - \lambda \mathbf{I}| = 0$ 得

$$\lambda_1 = 12.85, \lambda_2 = 0.15, \text{ 选择 } \lambda_1。$$

第三步：根据 $\mathbf{R}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$ 计算 λ_1 对应的特征向量 \mathbf{u}_1 ，令第1分量为1，归一化后为：

$$\mathbf{u}_1 = \frac{1}{\sqrt{2.3}} [1, 1.14]^T = [0.66, 0.75]^T$$

$$\mathbf{u}_1 = [0.66, 0.75]^T$$

变换矩阵为 $\mathbf{U} = [\mathbf{u}_1] = \begin{bmatrix} 0.66 \\ 0.75 \end{bmatrix}$

第四步：利用 \mathbf{U} 对样本集中每个样本进行 K-L 变换。

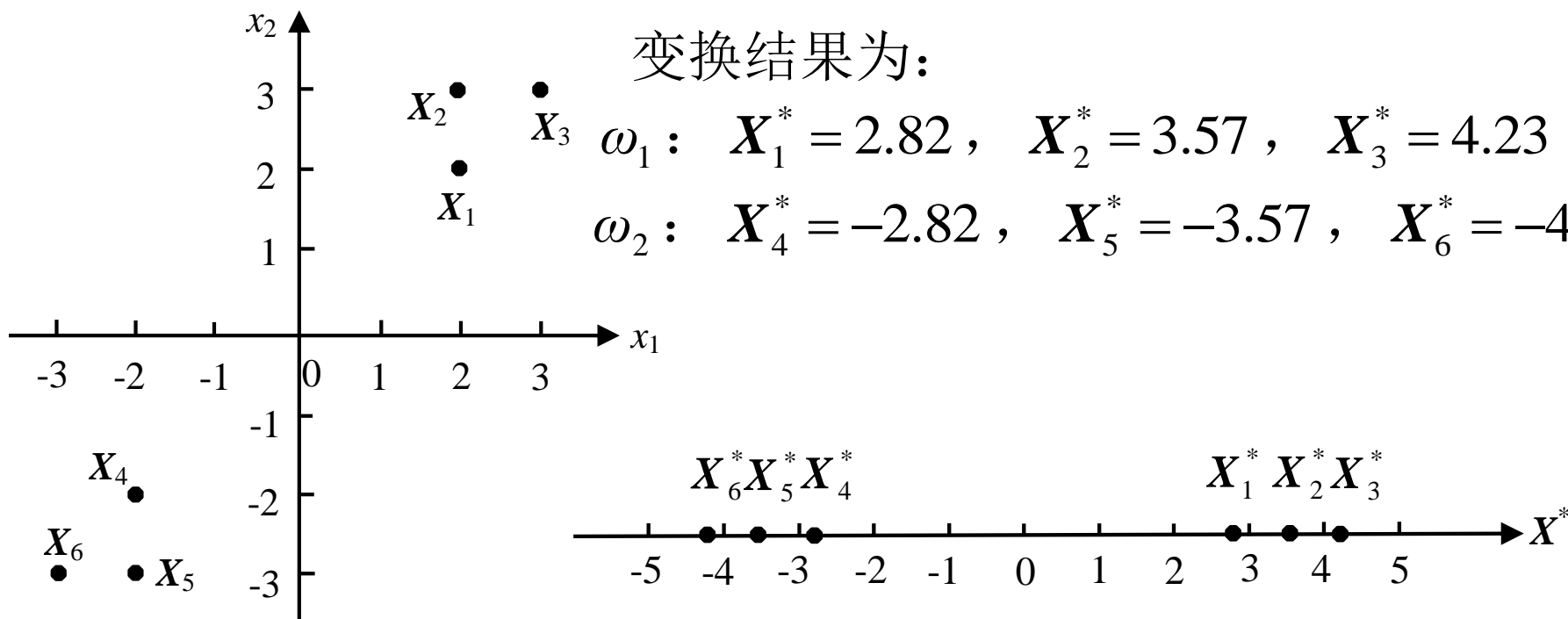
$$\mathbf{X}_1^* = \mathbf{U}^T \mathbf{X}_1 = [0.66 \ 0.75] \begin{bmatrix} 2 \\ 2 \end{bmatrix} = 2.82$$

.....

变换结果为：

$$\omega_1 : \mathbf{X}_1^* = 2.82, \mathbf{X}_2^* = 3.57, \mathbf{X}_3^* = 4.23$$

$$\omega_2 : \mathbf{X}_4^* = -2.82, \mathbf{X}_5^* = -3.57, \mathbf{X}_6^* = -4.23$$



六、模糊模式识别

6.1 模糊数学概述

6.2 模糊集合

6.3 模糊关系与模糊矩阵

6.4 模糊模式分类的直接和间接方法

6.2.3 模糊集合的运算

1. 基本运算

两个模糊子集间的运算：

逐点对隶属函数作相应的运算，得到新的隶属函数。
在此过程中，论域保持不变。

设 $\underset{\sim}{A}$ 、 $\underset{\sim}{B}$ 、 $\underset{\sim}{C}$ 、 $\underset{\sim}{\bar{A}}$ 为论域中的模糊集合，定义如下几个集合关系

(1)两个模糊子集相等：

若 $\forall x \in X$ ，均有 $\mu_{\underset{\sim}{A}}(x) = \mu_{\underset{\sim}{B}}(x)$ ，则称 $\underset{\sim}{A}$ 和 $\underset{\sim}{B}$ 相等。

即：
$$\underset{\sim}{A} = \underset{\sim}{B} \Leftrightarrow \mu_{\underset{\sim}{A}}(x) = \mu_{\underset{\sim}{B}}(x)$$

(2)包含：若 $\forall x \in X$ ，均有 $\mu_{\tilde{A}}(x) \leq \mu_{\tilde{B}}(x)$ ，称 \tilde{B} 包含 \tilde{A} 。

$$\text{即： } \tilde{A} \subseteq \tilde{B} \Leftrightarrow \mu_{\tilde{A}}(x) \leq \mu_{\tilde{B}}(x)$$

(3)补集：

若 $\forall x \in X$ ，均有 $\mu_{\tilde{\bar{A}}}(x) = 1 - \mu_{\tilde{A}}(x)$ ，称 $\tilde{\bar{A}}$ 为 \tilde{A} 的补集。

$$\text{即： } \tilde{\bar{A}} \Leftrightarrow \mu_{\tilde{\bar{A}}}(x) = 1 - \mu_{\tilde{A}}(x)$$

(4)空集：若 $\forall x \in X$ ，均有 $\mu_{\tilde{A}}(x) = 0$ ，称 \tilde{A} 为空集。

$$\text{即： } \tilde{A} = \phi \Leftrightarrow \mu_{\tilde{A}}(x) = 0$$

(5)全集：若 $\forall x \in X$ ，均有 $\mu_{\tilde{A}}(x) = 1$ ，称 \tilde{A} 为全集，记作 Ω 。

$$\text{即： } \tilde{A} = \Omega \Leftrightarrow \mu_{\tilde{A}}(x) = 1$$

(6) 并集：若 $\forall x \in X$ ，均有 $\mu_{\underset{\sim}{C}}(x) = \max\left(\mu_{\underset{\sim}{A}}(x), \mu_{\underset{\sim}{B}}(x)\right)$ ，
 则称 $\underset{\sim}{C}$ 为 $\underset{\sim}{A}$ 与 $\underset{\sim}{B}$ 的并集。

$$\text{即： } \underset{\sim}{C} = \underset{\sim}{A} \cup \underset{\sim}{B} \Leftrightarrow \mu_{\underset{\sim}{C}}(x) = \max\left(\mu_{\underset{\sim}{A}}(x), \mu_{\underset{\sim}{B}}(x)\right)$$

$$\text{或 } \underset{\sim}{C} = \underset{\sim}{A} \cup \underset{\sim}{B} \Leftrightarrow \mu_{\underset{\sim}{C}}(x) = \vee\left(\mu_{\underset{\sim}{A}}(x), \mu_{\underset{\sim}{B}}(x)\right) = \mu_{\underset{\sim}{A}}(x) \vee \mu_{\underset{\sim}{B}}(x)$$

(7) 交集：若 $\forall x \in X$ ，均有 $\mu_{\underset{\sim}{C}}(x) = \min\left(\mu_{\underset{\sim}{A}}(x), \mu_{\underset{\sim}{B}}(x)\right)$ ，
 则称 $\underset{\sim}{C}$ 为 $\underset{\sim}{A}$ 与 $\underset{\sim}{B}$ 的交集。

$$\text{即： } \underset{\sim}{C} = \underset{\sim}{A} \cap \underset{\sim}{B} \Leftrightarrow \mu_{\underset{\sim}{C}}(x) = \min\left(\mu_{\underset{\sim}{A}}(x), \mu_{\underset{\sim}{B}}(x)\right)$$

$$\text{或 } \underset{\sim}{C} = \underset{\sim}{A} \cap \underset{\sim}{B} \Leftrightarrow \mu_{\underset{\sim}{C}}(x) = \wedge\left(\mu_{\underset{\sim}{A}}(x), \mu_{\underset{\sim}{B}}(x)\right) = \mu_{\underset{\sim}{A}}(x) \wedge \mu_{\underset{\sim}{B}}(x)$$

例 ① $\underset{\sim}{A} = \{ (0.2, 1) \}, \quad \underset{\sim}{B} = \{ (0.2, 1) \}: \quad \underset{\sim}{A} = \underset{\sim}{B}$

② $\underset{\sim}{C} = \{ (0.7, 1) \}, \quad \underset{\sim}{D} = \{ (0.0, 1) \}: \quad$

$$\underset{\sim}{C} \supseteq \underset{\sim}{D}, \quad \underset{\sim}{D} = \phi, \quad \underset{\sim}{\overline{C}} = \{ (0.3, 1) \}$$

③ $\underset{\sim}{E} = \{ (1.0, 1) \}: \quad \underset{\sim}{E} = \Omega$

例 $\underset{\sim}{A} = \{ (0.1, 1) \}, \quad \underset{\sim}{B} = \{ (0.5, 1) \},$

则: $\underset{\sim}{A} \cup \underset{\sim}{B} = \{ (0.1 \vee 0.5, 1) \} = \{ (0.5, 1) \}$

$$\underset{\sim}{A} \cap \underset{\sim}{B} = \{ (0.1 \wedge 0.5, 1) \} = \{ (0.1, 1) \}$$

2. 运算的基本性质

(1) 自反律: $\underline{\underline{A}} \subseteq \underline{\underline{A}}$

(2) 反对称律: 若 $\underline{\underline{A}} \subseteq \underline{\underline{B}}$, $\underline{\underline{B}} \subseteq \underline{\underline{A}}$, 则 $\underline{\underline{A}} = \underline{\underline{B}}$

(3) 交换律: $\underline{\underline{A}} \cup \underline{\underline{B}} = \underline{\underline{B}} \cup \underline{\underline{A}}$, $\underline{\underline{A}} \cap \underline{\underline{B}} = \underline{\underline{B}} \cap \underline{\underline{A}}$

(4) 结合律: $(\underline{\underline{A}} \cup \underline{\underline{B}}) \cup \underline{\underline{C}} = \underline{\underline{A}} \cup (\underline{\underline{B}} \cup \underline{\underline{C}})$

$$(\underline{\underline{A}} \cap \underline{\underline{B}}) \cap \underline{\underline{C}} = \underline{\underline{A}} \cap (\underline{\underline{B}} \cap \underline{\underline{C}})$$

(5) 分配律: $\underline{\underline{A}} \cup (\underline{\underline{B}} \cap \underline{\underline{C}}) = (\underline{\underline{A}} \cup \underline{\underline{B}}) \cap (\underline{\underline{A}} \cup \underline{\underline{C}})$

$$\underline{\underline{A}} \cap (\underline{\underline{B}} \cup \underline{\underline{C}}) = (\underline{\underline{A}} \cap \underline{\underline{B}}) \cup (\underline{\underline{A}} \cap \underline{\underline{C}})$$

(6) 传递律: 若 $\underline{\underline{A}} \subseteq \underline{\underline{B}}$, $\underline{\underline{B}} \subseteq \underline{\underline{C}}$, 则 $\underline{\underline{A}} \subseteq \underline{\underline{C}}$

(7) 幂等律: $\underline{\underline{A}} \cup \underline{\underline{A}} = \underline{\underline{A}}, \quad \underline{\underline{A}} \cap \underline{\underline{A}} = \underline{\underline{A}}$

(8) 吸收律: $(\underline{\underline{A}} \cap \underline{\underline{B}}) \cup \underline{\underline{A}} = \underline{\underline{A}}, \quad (\underline{\underline{A}} \cup \underline{\underline{B}}) \cap \underline{\underline{A}} = \underline{\underline{A}}$

(9) 对偶律: $\overline{\underline{\underline{A}} \cup \underline{\underline{B}}} = \overline{\underline{\underline{A}}} \cap \overline{\underline{\underline{B}}}, \quad \overline{\underline{\underline{A}} \cap \underline{\underline{B}}} = \overline{\underline{\underline{A}}} \cup \overline{\underline{\underline{B}}},$

也称德·摩根定律。

(10) 对合律: $\overline{\underline{\underline{A}}} = \underline{\underline{A}},$ 即双重否定定律。

(11) 定常律: $\underline{\underline{A}} \cup \Omega = \Omega, \quad \underline{\underline{A}} \cap \Omega = \underline{\underline{A}}$

$$\underline{\underline{A}} \cup \varphi = \underline{\underline{A}}, \quad \underline{\underline{A}} \cap \varphi = \varphi$$

(12) 一般地互补律不成立: $\underline{\underline{A}} \cup \overline{\underline{\underline{A}}} \neq \Omega, \quad \underline{\underline{A}} \cap \overline{\underline{\underline{A}}} \neq \varphi$

例 $\underset{\sim}{A} = \{ (0.8, a) \}$ 时, 有 $\overline{\underset{\sim}{A}} = \{ (0.2, a) \}$, 则:

$$\underset{\sim}{A} \cup \overline{\underset{\sim}{A}} = \{ (0.8 \vee 0.2, a) \} = \{ (0.8, a) \}$$

$$\underset{\sim}{A} \cap \overline{\underset{\sim}{A}} = \{ (0.8 \wedge 0.2, a) \} = \{ (0.2, a) \}$$

互补律成立的特例:

$$\underset{\sim}{A} = \{ (0.0, a) \} \text{ 时, } \overline{\underset{\sim}{A}} = \{ (1.0, a) \};$$

$$\underset{\sim}{A} = \{ (1.0, a) \} \text{ 时, } \overline{\underset{\sim}{A}} = \{ (0.0, a) \}$$

此时模糊集合退化为经典集合。

1. 截集定义

设给定模糊集合 \tilde{A} ，论域 X ，对任意 $\lambda \in [0, 1]$ 称

普通集合 $(\tilde{A})_\lambda \equiv A_\lambda = \left\{ x \mid x \in X, \mu_{\tilde{A}}(x) \geq \lambda \right\}$

为 \tilde{A} 的 λ 截集（隶属度大于 λ 的成员集合）。

例： $\tilde{A} = 1/a + 0.9/b + 0.5/c + 0.2/d$ ，论域 $X = \{a, b, c, d\}$ ，

则： $A_1 = \{a\}$ ， $A_{0.9} = \{a, b\}$ ， $A_{0.8} = \{a, b\}$ ，

$A_{0.5} = \{a, b, c\}$ ， $A_{0.1} = \{a, b, c, d\}$ ， $A_0 = X$ 。

2. 截集的三个性质

$$\textcircled{1} \quad (\underline{A} \cup \underline{B})_{\lambda} = \underline{A}_{\lambda} \cup \underline{B}_{\lambda}$$

例： $\underline{A} = \{ (0.5, a), (0.8, b) \}$, $\underline{B} = \{ (0.9, a), (0.2, b) \}$, 则

左： $\underline{A} \cup \underline{B} = \{ (0.5 \vee 0.9, a), (0.8 \vee 0.2, b) \} = \{ (0.9, a), (0.8, b) \}$

$$(\underline{A} \cup \underline{B})_{0.5} = \{a, b\}$$

右： $\underline{A}_{0.5} = \{a, b\}$, $\underline{B}_{0.5} = \{a\}$, 有 $\underline{A}_{0.5} \cup \underline{B}_{0.5} = \{a, b\}$

所以：在 $\lambda = 0.5$ 时, $(\underline{A} \cup \underline{B})_{\lambda} = \underline{A}_{\lambda} \cup \underline{B}_{\lambda}$ 。在其他 λ 值时同样成立。

$$\textcircled{2} \quad (\underset{\sim}{A} \cap \underset{\sim}{B})_{\lambda} = A_{\lambda} \cap B_{\lambda}$$

$$\text{左: } \underset{\sim}{A} \cap \underset{\sim}{B} = \{ (0.5 \wedge 0.9, a), (0.8 \wedge 0.2, b) \} = \{ (0.5, a), (0.2, b) \}$$

$$(\underset{\sim}{A} \cap \underset{\sim}{B})_{0.3} = \{ a \}$$

$$\text{右: } A_{0.3} = \{ a, b \}, \quad B_{0.3} = \{ a \}, \quad \text{有 } A_{0.3} \cap B_{0.3} = \{ a \}$$

$$\text{所以: } \lambda = 0.3 \text{ 时, } (\underset{\sim}{A} \cap \underset{\sim}{B})_{\lambda} = A_{\lambda} \cap B_{\lambda}$$

$$\textcircled{3} \quad \text{若 } \lambda, \mu \in [0, 1], \text{ 且 } \lambda \leq \mu, \text{ 则 } A_{\lambda} \supseteq A_{\mu}.$$

$$\text{例: 对 } \underset{\sim}{A} = \{ (0.2, x_1), (0.5, x_2), (0.8, x_3), (1.0, x_4), (0.7, x_5) \}$$

$$\text{有 } A_{0.4} = \{ x_2, x_3, x_4, x_5 \}, \quad A_{0.5} = \{ x_2, x_3, x_4, x_5 \}, \quad A_{0.8} = \{ x_3, x_4 \}$$

$$\text{显然: } A_{0.5} \supset A_{0.8} \text{ (或 } A_{0.5} \supseteq A_{0.8} \text{)}$$

$$A_{0.4} = A_{0.5} \text{ (满足 } A_{0.4} \supseteq A_{0.5} \text{)}$$

七、神经网络模式识别

- 7.1 人工神经网络发展概况
- 7.2 神经网络基本概念
- 7.3 前馈神经网络
- 7.4 反馈网络模型Hopfield网络

7.1 人工神经网络发展概况

人工神经网络(Artificial Neural Networks, ANN):

简称神经网络。

模拟人脑神经细胞的工作特点:

- * 单元间的广泛连接;
- * 并行分布式的信息存贮与处理;
- * 自适应的学习能力等。

与目前按串行安排程序指令的计算机结构截然不同。

优点:

- (1) 较强的容错性;
- (2) 很强的自适应学习能力;
- (3) 可将识别和若干预处理融为一体进行;
- (4) 并行工作方式;
- (5) 对信息采用分布式记忆, 具有鲁棒性。

五个发展阶段：

第一阶段：启蒙期，始于1890年（Williams James）。

1943年：形式神经元的数学模型M-P模型提出；

1957年：感知器算法的此处（Frank Rosenblatt）

第二阶段：低潮期，始于1969年。

《感知器》(Perceptrons)一书出版，指出局限性。

第三阶段：复兴期，从1982年到1986年。

Hopfield的两篇论文提出新的神经网络模型；

《并行分布处理》出版，提出反向传播算法。

第四阶段：新兴期，1987年到2006年。

回顾性综述文章“神经网络与人工智能”。

第五阶段：深度学习热潮，2006年至今。

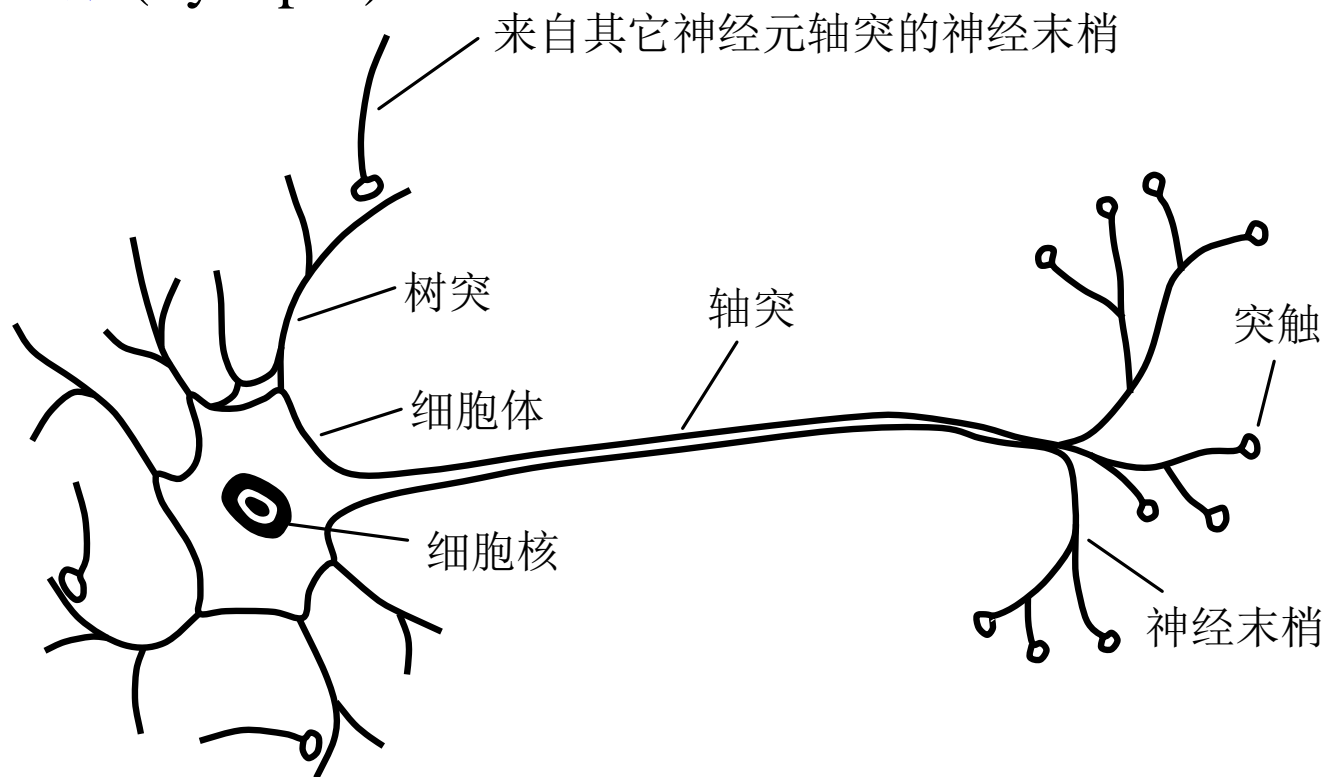
G. E. Hinton提出有效训练多层网络的随机NN算法。

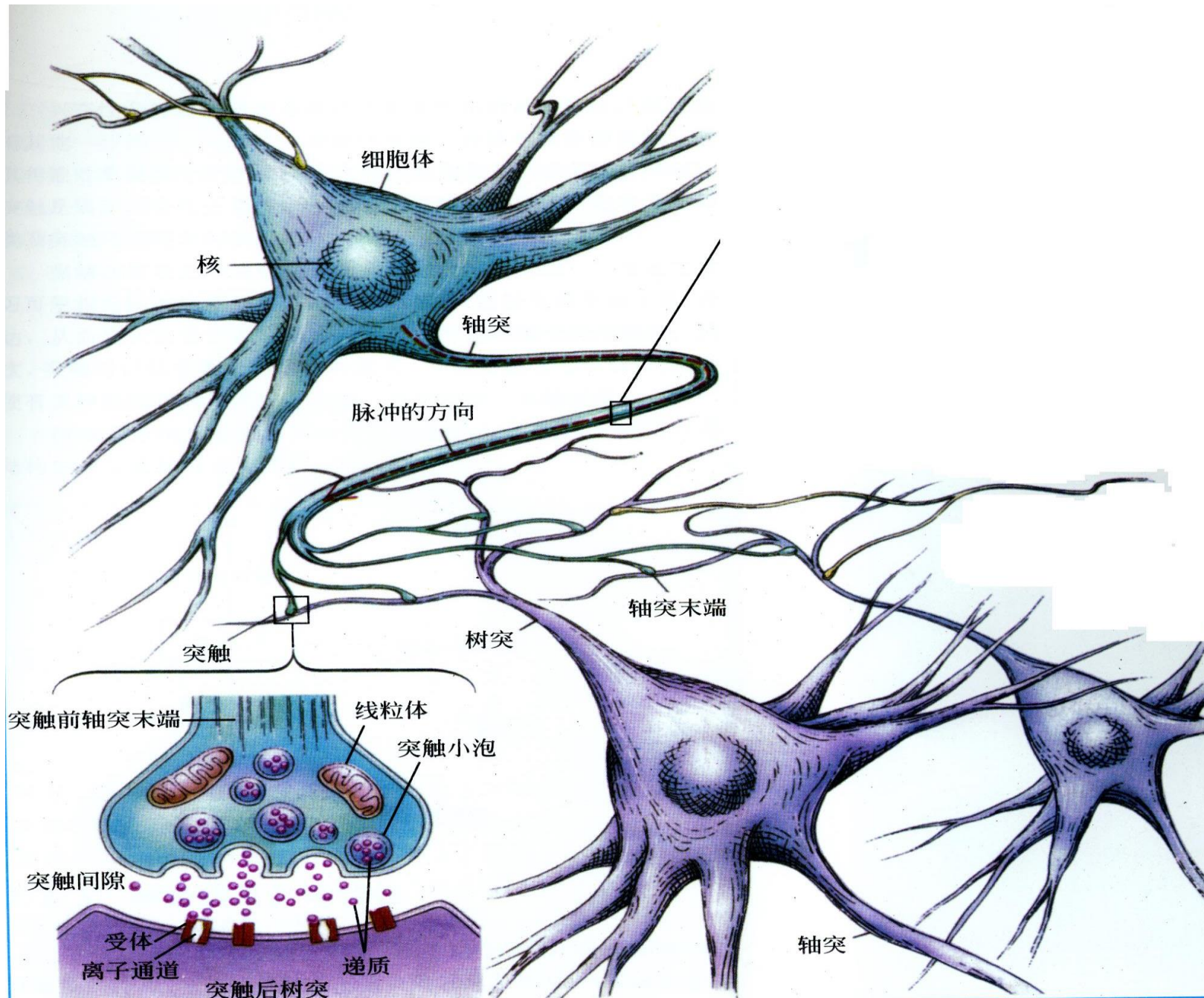
7.2 神经网络基本概念

7.2.1 生物神经元

1. 生物神经元的结构

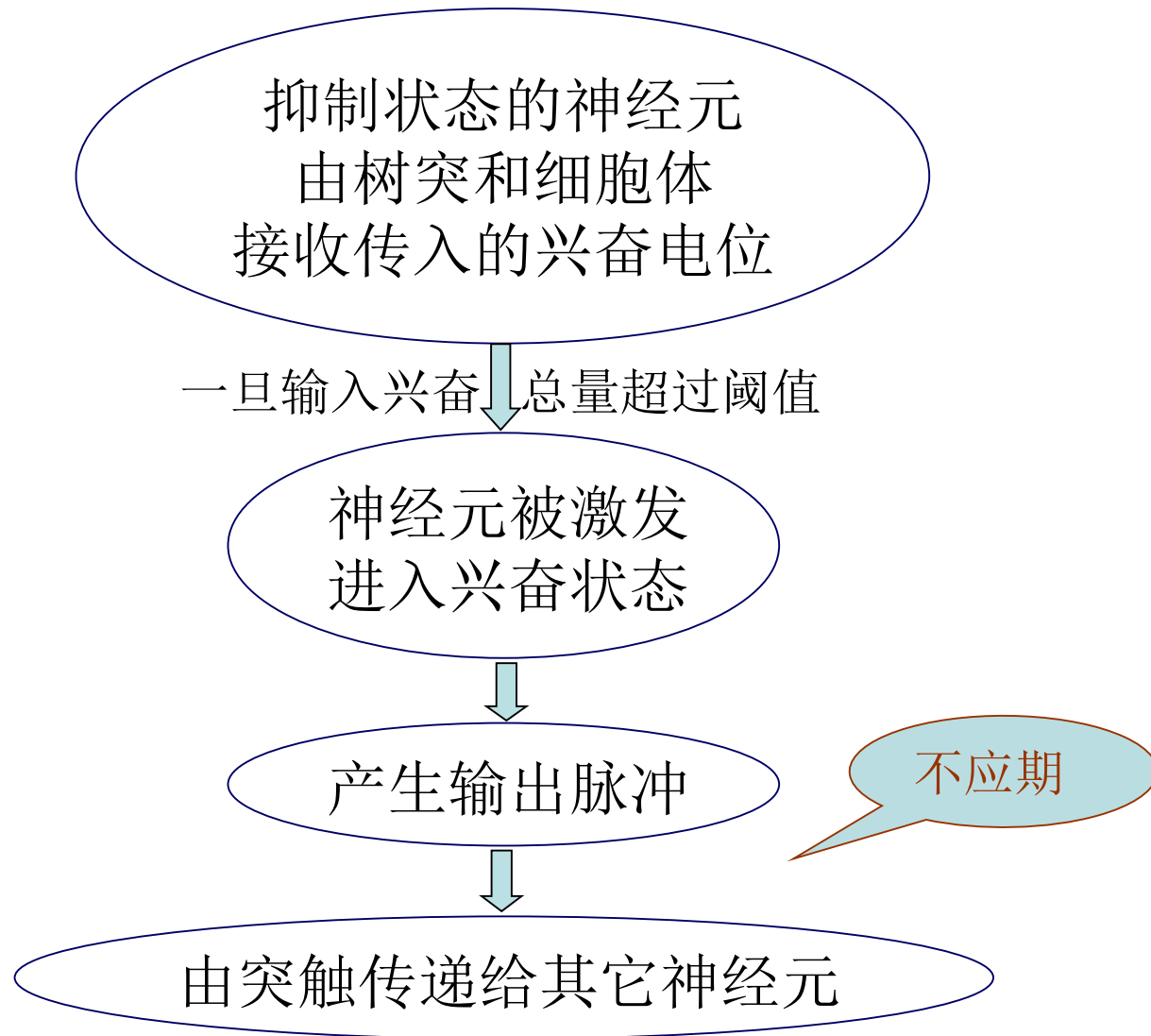
细胞体(Cell Body)、树突(Dendrite)、轴突(Axon)和突触(Synapse)。



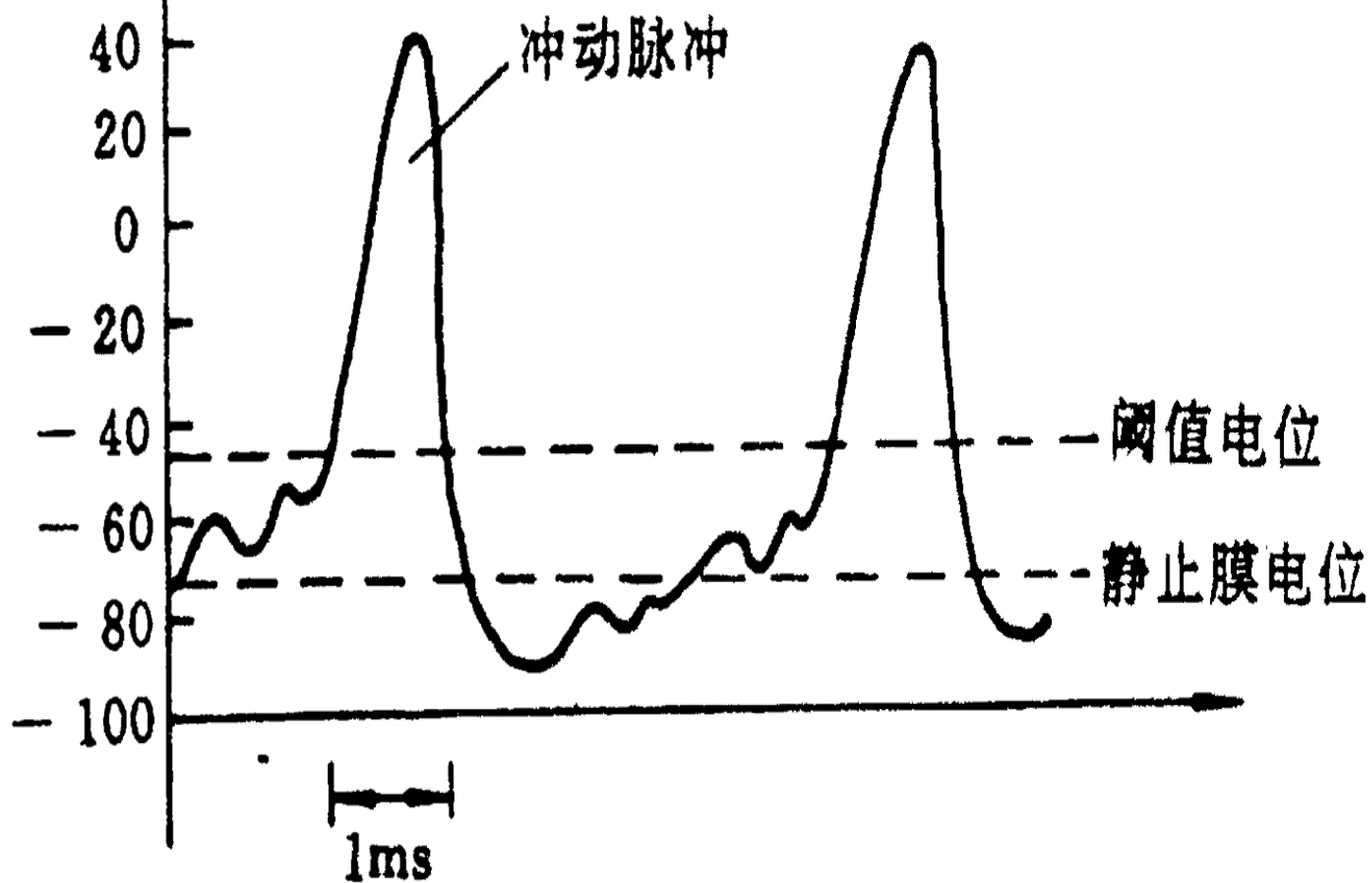


2. 生物神经元的工作机制

兴奋和抑制两种状态。



膜电位 (mV)



神经元的整合

空间整合：同一时刻产生的刺激所引起的膜电位变化，大致等于各单独刺激引起的膜电位变化的代数。

时间整合：各输入脉冲抵达神经元的时间先后不一样。总的突触后膜电位为一段时间内的累积。

神经元及其突触是神经网络的基本器件。因此，模拟生物神经网络应首先模拟生物神经元人工神经元（节点）。

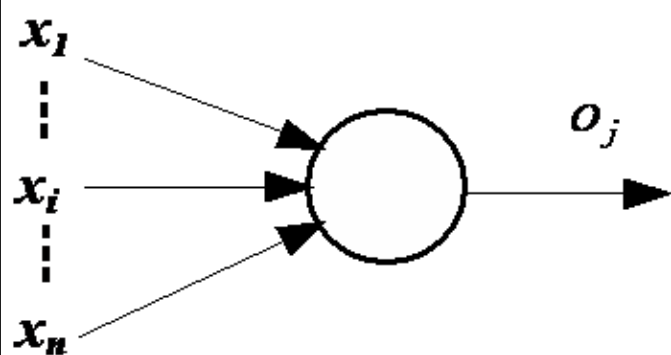
从三个方面进行模拟：

- 节点的信息处理能力（神经元模型）
- 节点与节点之间连接（网络拓扑结构）
- 节点相互连接的强度（通过学习调整）

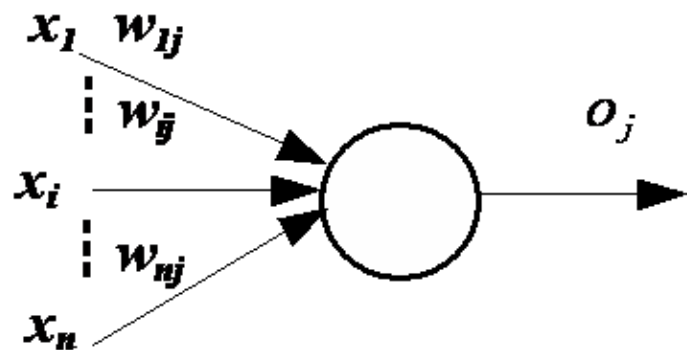
神经元模型的六点假设：

- (1) 每个神经元都是一个多输入单输出的信息处理单元；
- (2) 神经元输入分兴奋性输入和抑制性输入两种类型；
- (3) 神经元具有空间整合特性和阈值激活特性；
- (4) 神经元输入与输出间有固定时滞, 取决于突触延搁；
- (5) 忽略时间整合作用和不应期；
- (6) 神经元本身是非时变的, 即其突触时延和突触强度均为常数。

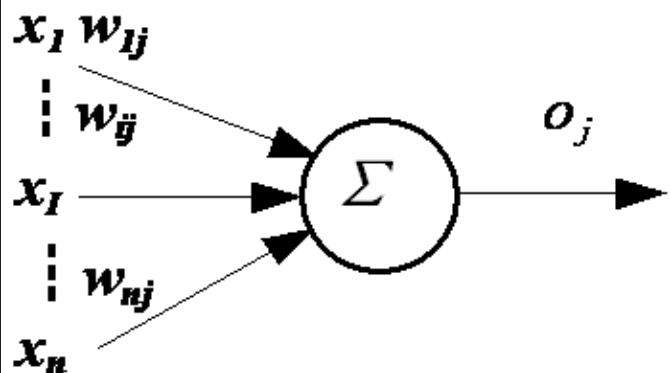
神经元模型示意图



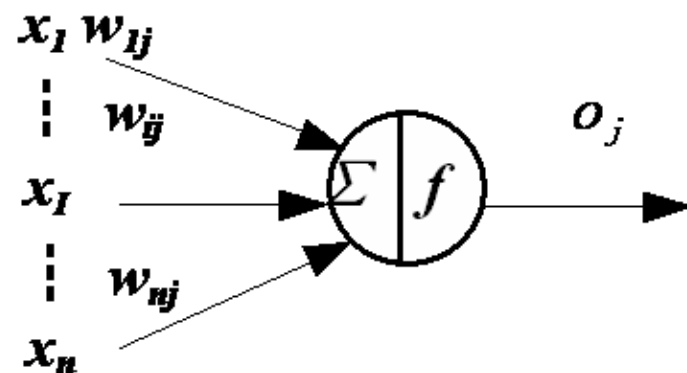
(a)多输入单输出



(b)输入加权



(c)输入加权求和



(d)输入-输出函数

End of This Part