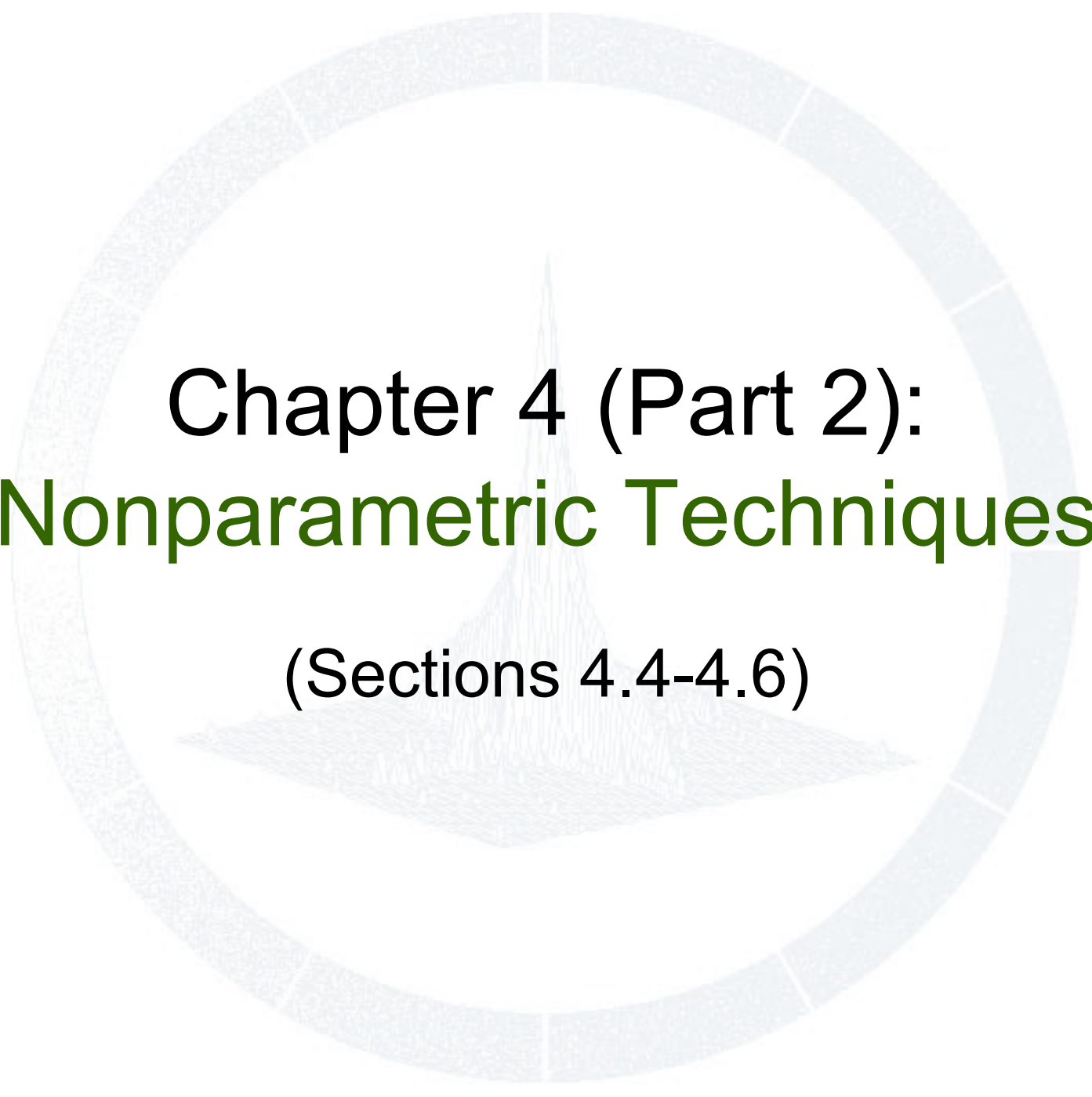




模式识别的理论与方法

Pattern Recognition

裴继红



Chapter 4 (Part 2): Nonparametric Techniques

(Sections 4.4-4.6)

本讲内容

- k -近邻估计
- 最近邻估计
- 距离度量和最近邻分类
- 衰减库仑能量（RCE）网络（选讲）



非参数估计回顾：估计 $p(x)$

- 假设在 n 个样本中，有 k 个落入区域 R 中。若样本的抽取过程是独立同分布的（*IID*），则 P (区域 R 的概率) 可以近似表示为 k/n .

由于 P 可以近似为 k/n
故有

$$p_n(x) \cong \frac{k_n/n}{V_n}$$



收敛条件回顾

- 为了求得真实的、而非平滑后的 $p(\mathbf{x})$ ，需要随着样本数量的增大而减小体积 V 的尺寸。

为此，给出了两种方法：

1) Parzen窗方法

将体积作为 n 的函数进行收缩，如： $V_n = 1/\sqrt{n}$

2) k 最近邻方法

将落入区域 R 的样本数 k 作为 n 的函数，如： $k_n = \sqrt{n}$

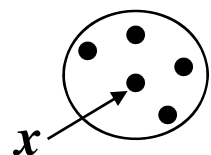


k -最近邻密度估计

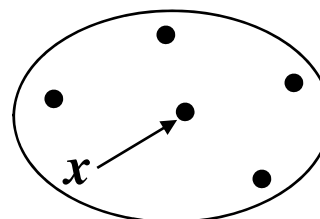
基本思想：将 x 置于元胞的中心，逐渐增大元胞，直到元胞中包含 k 个样本为止。这里， k 值是样本数 n 的函数

两种情况：

- 若 x 的邻域样本密度高，则元胞体积将比较小，
- 若 x 的邻域样本密度低，则元胞体积将比较大，



高密度区



低密度区



k -最近邻密度估计

不论在何种情况下，估计 $p(\mathbf{x})$ 的方法是一样的：

$$p_n(\mathbf{x}) \cong \frac{k_n/n}{V_n}$$

其中， k_n 的一种取法可以为：

$$k_n = \sqrt{n}$$

在这种情况下，体积 V_n 正比于

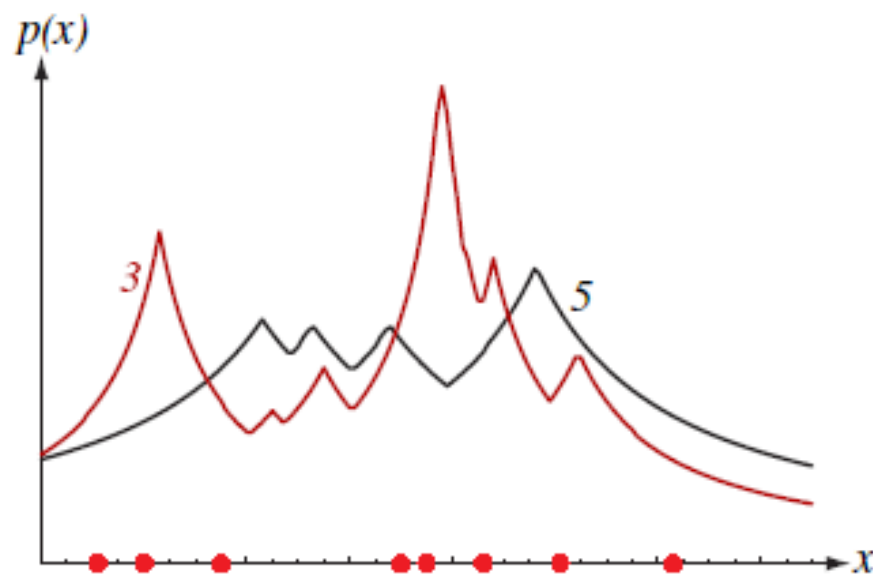
$$\frac{1}{\sqrt{n} \cdot p(\mathbf{x})}$$



一维样本密度的 k 最近邻估计: Figure 4.10

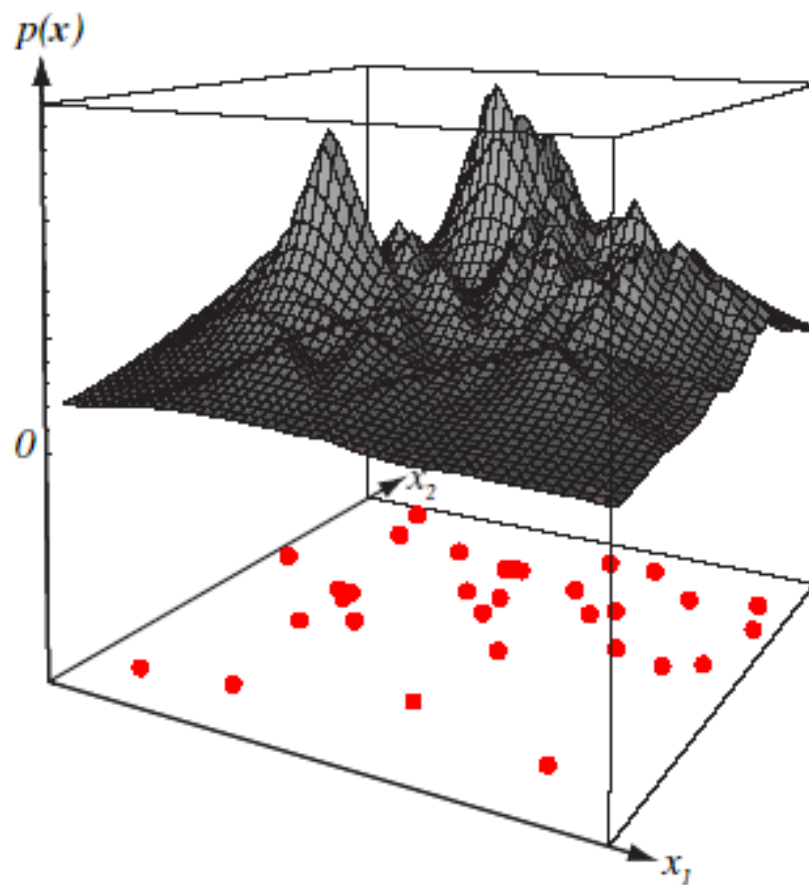
例: 8 个一维样本点及其 k 最近邻估计 ($k = 3, 5$)

- 注意, 在估计中存在斜坡不连续性, 这种不连续一般出现在离开样本位置的地方。



二维样本密度的 k 最近邻估计: Figure 4.11

- 在 $k = 5$ 时进行的二维 k -最近邻密度估计
- 注意, 由于样本数量 n 是有限的, 得到的估计结果具有明显的锯齿状。
- 另外, 注意到, 斜坡不连续性一般发生在离开样本点位置的地方。



k -最近邻估计的算法实现

➤ 如何有效地找出 k 个近邻？



k -近邻方法与Parzen窗方法的比较

k -Nearest Neighbor

$p(\mathbf{x})$ 的值永远不会等于0, 这是由于包围 \mathbf{x} 的元胞中的训练样本子集合永远不会为空

在样本数量低时, 估计结果趋向于变坏, 并且崎岖不平

Parzen Windows

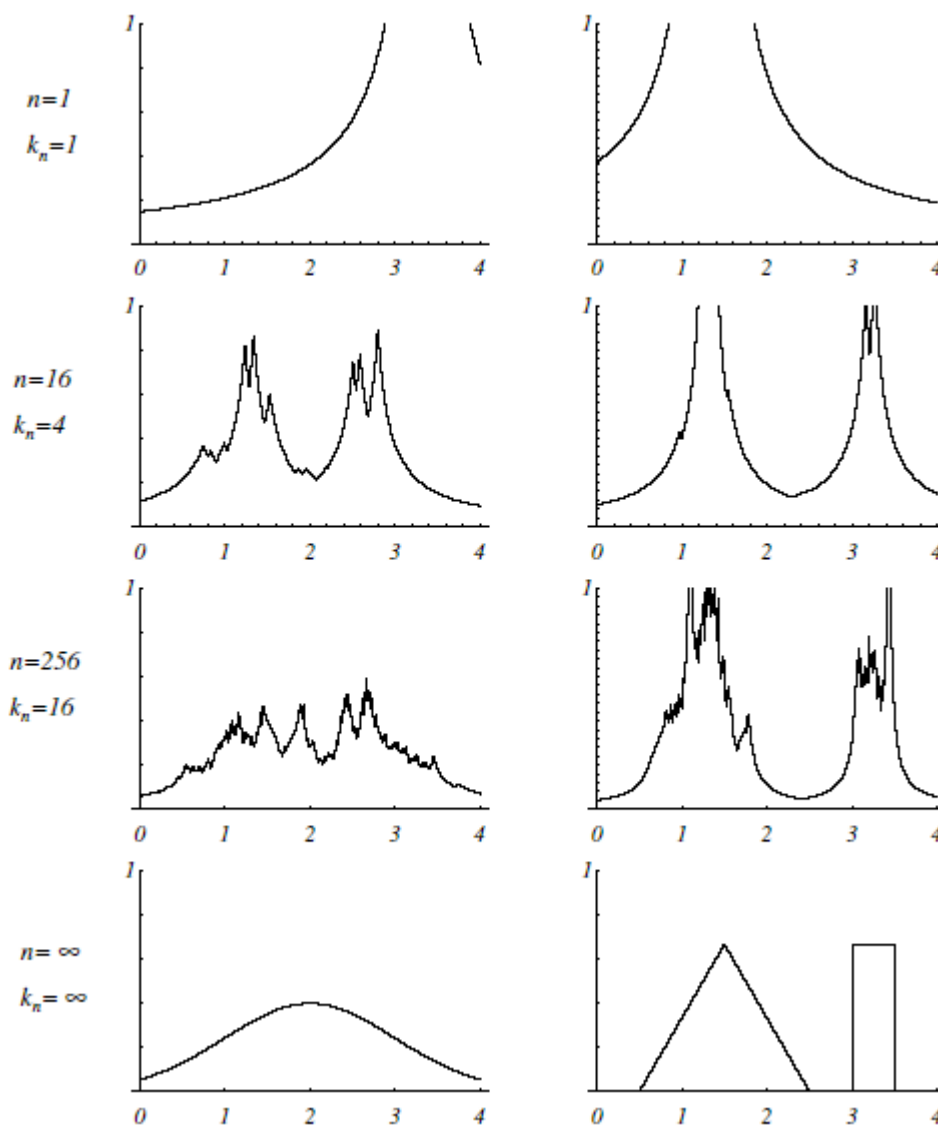
$p(\mathbf{x})$ 可能为0, 这是由于包围 \mathbf{x} 的元胞中的训练样本子集合可能为空

对于合适的 h 值, 估计可以取得好的结果



两个一维密度的几种 k -最近邻估计 **Figure 4.12**

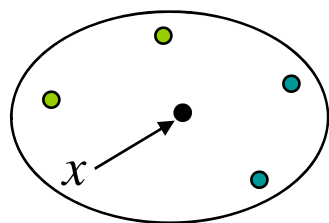
- 高斯分布、双模分布，
- 注意，有限 n 带来的估计结果的崎岖不平性



后验概率估计

可以用下面的方法直接估计后验概率 $p(\omega_j | x)$

➤ 以 x 为中心，构造胞形，使其包含 k 个样本



对 $p(\omega_j | x)$ 的一个合理的估计是

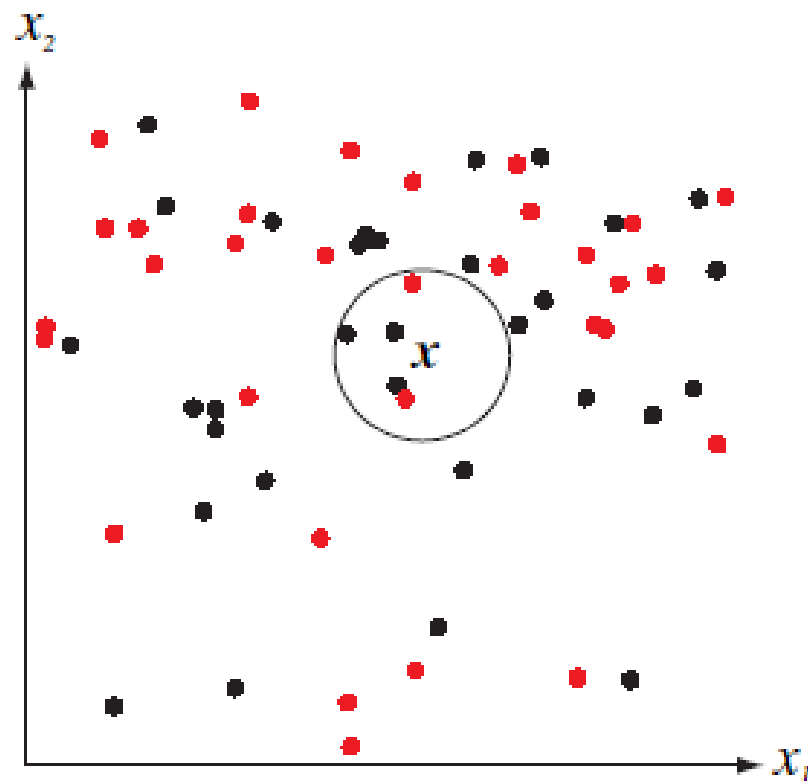
$$p(\omega_j | x) = \frac{k_j}{k}$$

其中, k_j 是落在胞形内部的 k 个样本中, 属于类 ω_j 的样本的数量



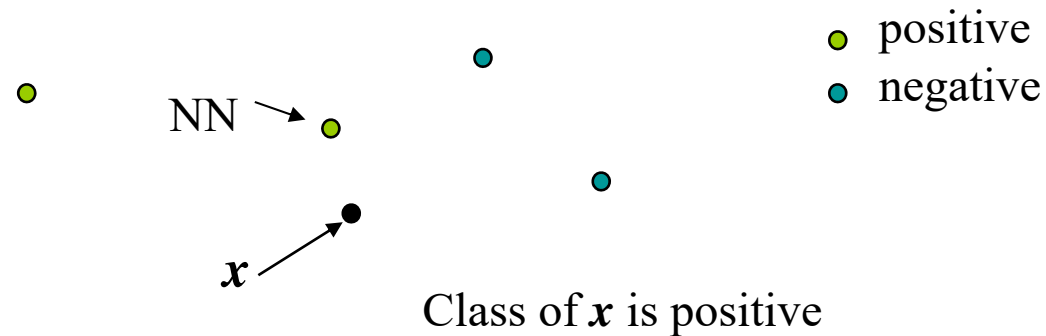
k -最近邻分类: Figure 4.15

- 在测试样本 x 处采用 k -最近邻进行决策
- 逐渐地放大以 x 为中心的球形区域直到包含了 k 个训练样本为止。而对测试点 x 的分类依据这些样本点的投票，以多数原则决定。
- 在如图所示， $k=5$ 的情况下，测试样本点 x 的类别标记为黑色的训练样本所在的类别



最近邻划分规则

- 对前面方法（后验概率估计）的一个简化是只考虑与 x 最靠近的单个样本的情况，用该样本的类去预测 x 的类



- 值得指出的是：这种方法非常简单。并且，虽然误差一般比Bayes误差要大，但是在训练样本数量无限的情况下，其误差不会超过Bayes误差的2倍。



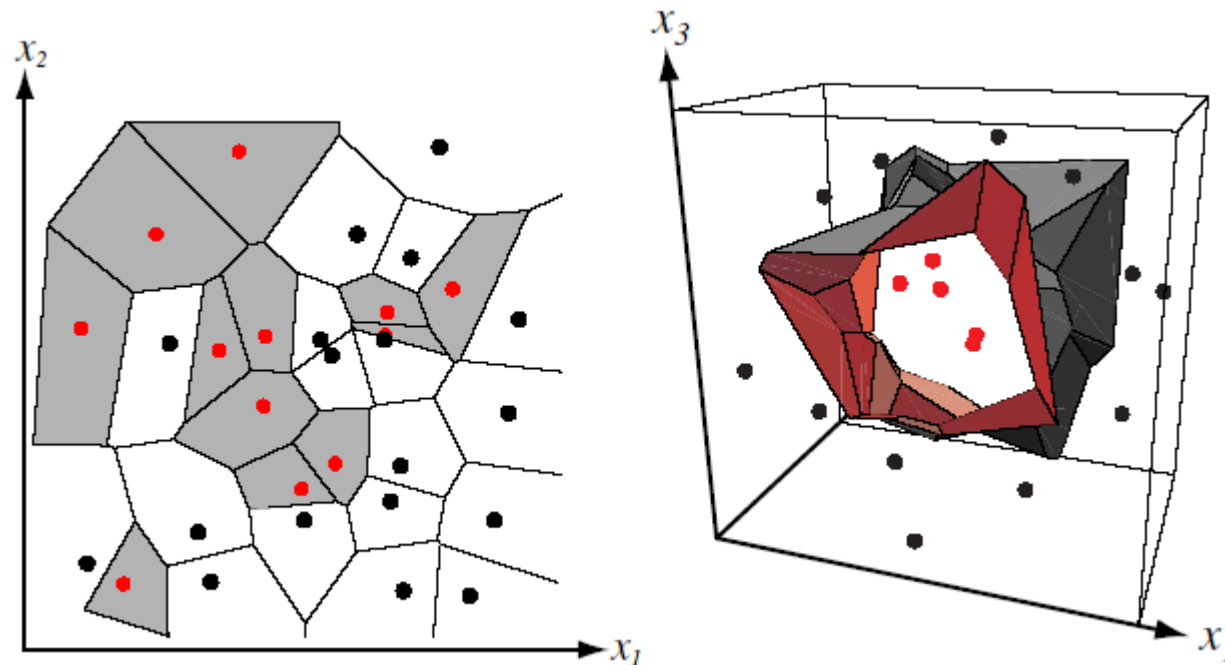
Voronoi-Tessellation图

- 最近邻规则将特征空间划分为胞形结构。
- 在胞形内部的所有点都标记为胞形中心的训练样本点所在的类。
- 该胞形结构称为Voronoi-Tessellation图



最近邻估计: Figure 4.13

- 在二维空间中，最近邻算法将空间划分为一系列Voronoi元胞（**cell**），每一个元胞由包含在其中的训练样本点标记其类别。
- 在三维空间中的元胞，其决策面类似于水晶表面结构



k -最近邻划分规则

- 将只包含单个样本的最近邻划分规则推广到在胞形中包含 k 个样本的情况
- 主要思想是通过检验 k 个最近邻样本的类别标记，通过投票的原则来作出决策



k -最近邻划分方法的计算复杂度

假设有 n 个 d 维空间的样本，寻找与 \mathbf{x} 最近的单个样本点。算法如下：

算法：

- ① 对每一个点 \mathbf{x}
- ② 对每一个点 \mathbf{x}'
- ③ 计算 \mathbf{x}' 与 \mathbf{x} 之间的距离
- ④ 将已搜索过的最靠近的点保存在内存中
- ⑤ 内层循环结束
- ⑥ 存储最靠近 \mathbf{x} 的点
- ⑦ 结束

每一个距离计算的复杂度为 $O(d)$. 总的搜索的复杂度为 $O(d n^2)$.
另外，存在时间复杂度为 $O(1)$ ，空间复杂度为 $O(n)$ 的并行算法。



减小计算代价的方法

□ 有几种改进的减小计算代价的方法:

➤ **部分距离法**。只使用样本 d 维分量中的其中 r 维计算距离，当计算的距离值已经较大时，则停止计算:

$$D_r(a, b) = \sqrt{\sum_{k=1}^r (a_k - b_k)^2}$$

➤ **搜索树法**。建立一个将模式原型连接在一起的搜索树，计算测试样本到一个或几个原型以及与这些原型相连接的距离。



度量问题

➤ 最近邻方法首先需要定义一个距离度量。

➤ 一个距离度量所具备的特性:

非负性: $D(a,b) \geq 0$

反射性: $D(a,b) = 0$ 当且仅当 $a = b$

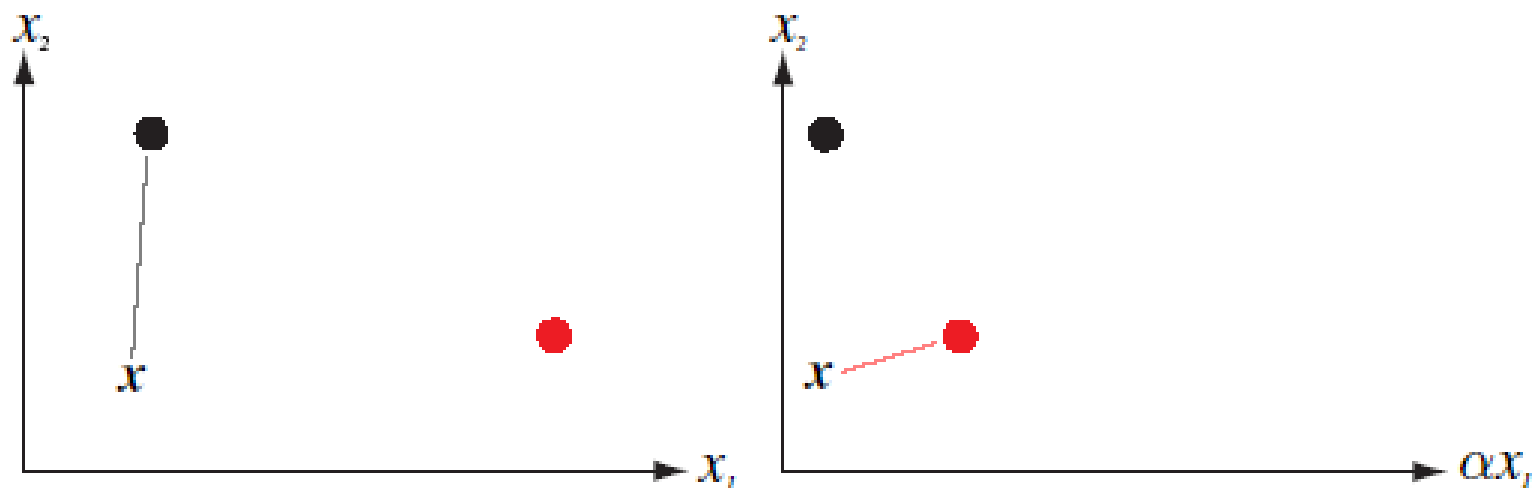
对称性: $D(a,b) = D(b,a)$

三角不等式: $D(a,b) + D(b,c) \geq D(a,c)$

□ 通常的欧式距离具有上述三个特性



分量的尺度变换对度量的影响: **Figure 4.18**



一些距离度量

➤ Minkowski 度量:

$$L_k(a, b) = \left(\sum_{i=1}^d |a_i - b_i|^k \right)^{1/k}$$

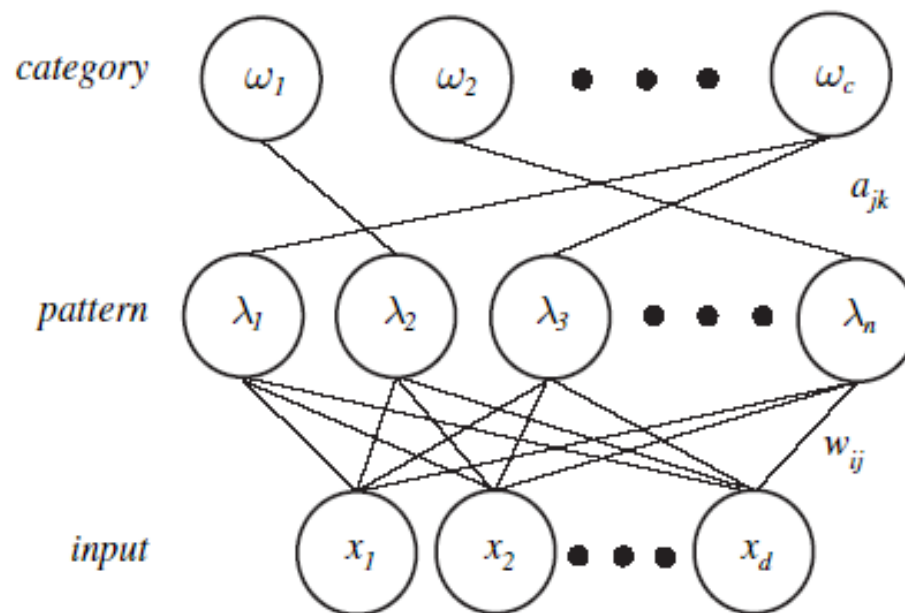
➤ Manhattan 距离

L_1 范数 ($k = 1$ 时的Minkowski距离).



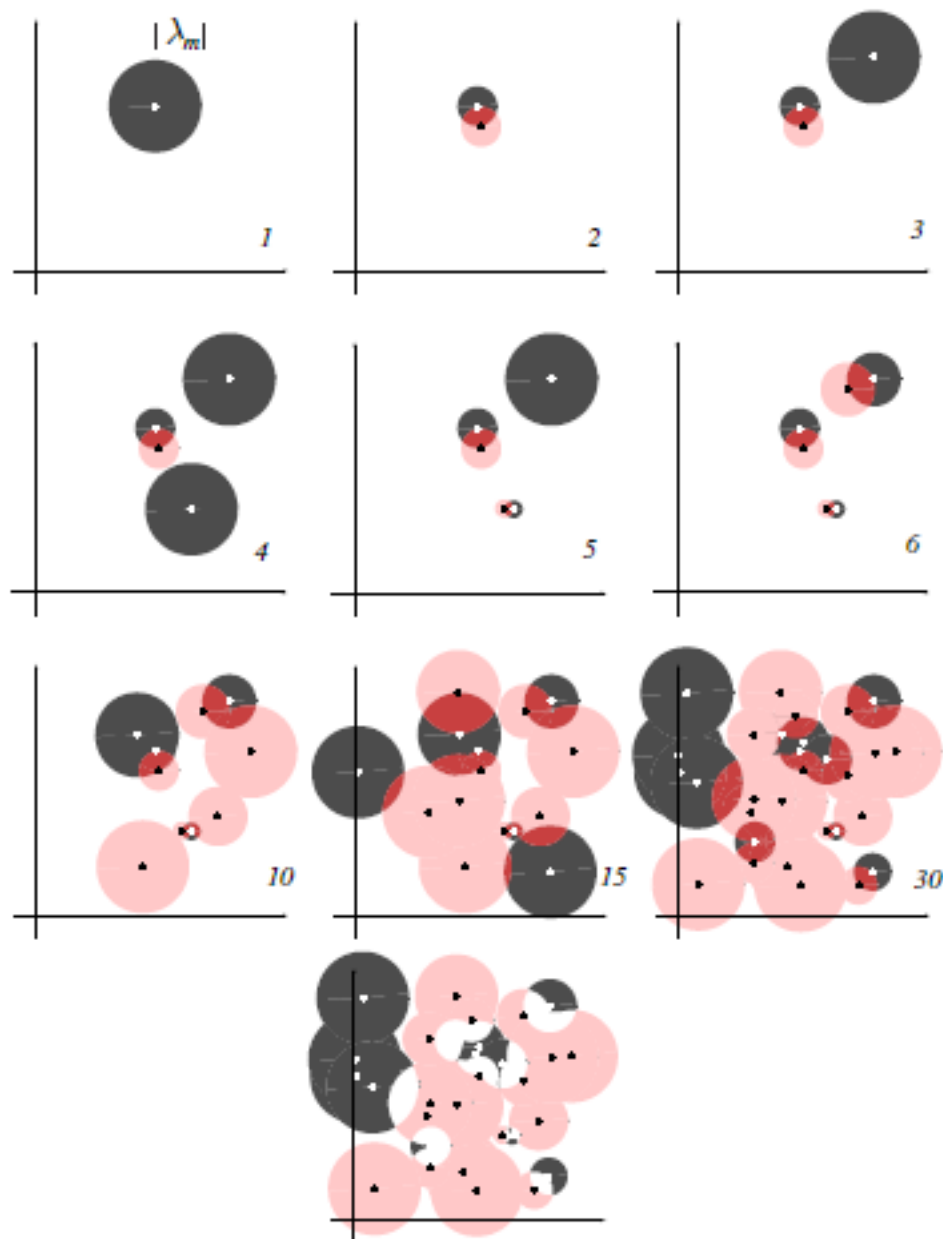
RCE网络

- **RCE网络的学习算法：**
对每一个已标记类别的训练样本，确定只包含同类样本的最大超球体的半径 λ
- **RCE网络的分类算法：**
对未知样本，寻找出其落入到的所有超球体的已标记样本



□ RCE网络 训练示例

- 红色：类别1
- 灰色：类别2
- 深红色：模糊区域



基于排序的 k -近邻算法

- 目的寻找与 x 最接近的 k 个样本模式
- 算法设计:



一些编程问题

