（请大家预习）

第4章
非参数估计
Nonparametric Methods

张 燕 明

ymzhang@nlpr.ia.ac.cn

people.ucas.ac.cn/~ymzhang

模式分析与学习课题组(PAL)

多模态人工智能系统实验室 中科院自动化所

助教： 杨 奇 (yangqi2021@ia.ac.cn)

张 涛 (zhangtao2021@ia.ac.cn)

上讲主要内容回顾

- 数据属性不完整
  - 情形1：一些样本的属性缺失

$\mathbf{x}_1$=(22,0.58,4,53) $\mathbf{x}_2$=(20,0.46,3,*) $\mathbf{x}_3$=(19,*,*,*) $\mathbf{x}_4$=(23,*,3,98) $\mathbf{x}_5$=(27,0.65,2,47)

  - 情形2：一些隐藏属性(隐变量)不可观测

病人：（症状描述，血常规，体温，血压，心/肝/肺等器官的状态）

- EM算法
  - 用于数据属性不完整的概率密度参数估计

上讲主要内容回顾

- EM算法的基本概念
  - 观测数据： $X=\{\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_n\}$
  - 隐含数据： $Z=\{\mathbf{z}_1, \mathbf{z}_2,\ldots, \mathbf{z}_n\}$
  - 完全数据：$(\mathbf{x}_i, \mathbf{z}_i)$
  - 观测数据的对数似然函数： $\sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \boldsymbol{\theta})$
  - 完全数据的对数似然函数： $\sum_{i=1}^{n} \ln p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})$

上讲主要内容回顾

- EM算法步骤
  - 初始化 $\boldsymbol{\theta}^{old}$
  - Repeat
    - **E step:** 基于当前参数$\boldsymbol{\theta}^{old}$，估计每个样本隐变量的后验分布$p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})$
    - **M step:** 基于当前估计的$p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta}^{old})$更新参数$\boldsymbol{\theta}$：

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_i E_{p(\mathbf{z}_i|\mathbf{x}_i,\boldsymbol{\theta}^{old})}\left[\ln\left(p\left(x_i, z_i \mid \boldsymbol{\theta}\right)\right)\right]$$

$$= \sum_i \sum_{z_i} p(\mathbf{z}_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ln\left(p\left(x_i, z_i \mid \boldsymbol{\theta}\right)\right)$$

完全数据的对数似然

  - Until convergence

上讲主要内容回顾

- EM算法的两个说明
  - 为什么要最大化完全数据的似然$p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})$，而不直接最大化观测数据的似然$p(\mathbf{x}|\boldsymbol{\theta})$？

通常， $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$比$p(\mathbf{x}|\boldsymbol{\theta})$更简单。

  - EM算法的优化目标是什么？

EM通过坐标轮替法最大化$\ln p(\mathbf{x}|\boldsymbol{\theta})$的下界$L(q,\boldsymbol{\theta}) \equiv E_{q(\mathbf{z})} \ln \dfrac{p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})}$，

因此，也可以粗略地说EM在最大化$\ln p(\mathbf{x}|\boldsymbol{\theta})$。

EM for Gaussian mixture model

- 混合密度模型

$$p(\mathbf{x}\,|\,\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k\, p(\mathbf{x}\,|\,\boldsymbol{\theta}_k) \qquad \sum_{k=1}^{K} \pi_k = 1 \quad \forall k : \pi_k \geq 0$$

其中，$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K\}$，$\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_K\}$。称上述密度函数为混合密度；称概率密度函数 $p(\mathbf{x}|\boldsymbol{\theta}_k)$ 为成分密度；称π_k为混合参数(成分权重)；称K为成分个数。

- ✓ 目的：将简单密度函数线性组合，构造复杂概率密度函数，用以对复杂数据建模。
- ✓ 从样本生成的角度，可以先对分布$\boldsymbol{\pi}$采样选择一个成分z，再对相应的成分$p(\mathbf{x}|\omega_z, \boldsymbol{\theta}_z)$采样得到$\mathbf{x}$。（由此引入了隐变量z）

EM for Gaussian mixture model

- GMM

高斯密度函数

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \; p(\mathbf{x}\,|\,\boldsymbol{\theta}_k) \; = \sum_{k=1}^{K} \pi_k \; \mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- EM for GMM
  - E Step: 固定参数$\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$，对每个样本求$p(\mathbf{z}_i|\mathrm{x}_i)$

$$P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) = \frac{p(\mathbf{x}_i,z_i\,|\,\boldsymbol{\theta}^{old})}{p(\mathbf{x}_i\,|\,\boldsymbol{\theta}^{old})} = \frac{\pi_{z_i}\mathcal{N}\left(\mathbf{x}_i,\boldsymbol{\mu}_{z_i},\boldsymbol{\Sigma}_{z_i}\right)}{\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\mathbf{x}_i,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)}$$

  - M Step: 固定$\{\, p(\mathbf{z}_i|\mathbf{x}_i)\,\}$, 更新参数 $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} P(z_i = k\,|\,\mathbf{x}_i,\boldsymbol{\theta}^{old})$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} P(z_i = k\,|\,\mathbf{x}_i,\boldsymbol{\theta}^{old})\mathbf{x}_i}{\sum_{i=1}^{n} P(z_i = k\,|\,\mathbf{x}_i,\boldsymbol{\theta}^{old})} \qquad \hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{n} P(z_i = k\,|\,\mathbf{x}_i,\boldsymbol{\theta}^{old})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{i=1}^{n} P(z_i = k\,|\,\mathbf{x}_i,\boldsymbol{\theta}^{old})}$$

EM for hidden Markov model

- 序列数据：$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
  - n 为序列长度，$\mathbf{x}_t \in R^d$ 是 X 在第t 时刻的观测数据
- 马尔可夫性

$$\forall t \quad p\left(\mathbf{x}_t \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}\right) = p\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)$$

✓ 给定$\mathbf{x}_{t-1}$时，$\mathbf{x}_t$与$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-2}\}$无关
✓ 是一种很强的局部依赖假设，使条件分布大幅简化

- 一阶、静态、离散马氏链

$$p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) = p\left(\mathbf{x}_1\right)\prod_{t=2}^{n} p\left(\mathbf{x}_t \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}\right) = p(\mathbf{x}_1)\prod_{t=2}^{n} p\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)$$

✓ $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$称为状态转移概率，描述了系统状态转换的规则
✓ $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$可以由非负、行和为1的方阵表示

EM for hidden Markov model

- HMM
  - 观测序列由一个不可见的马尔可夫链生成。
  - HMM的随机变量可分为两组：
    - 状态变量$\{z_1, z_2, \ldots, z_n\}$：构成一阶、离散、静态马尔可夫链。描述系统内部的状态变化，通常是隐藏的，不可被观测的。其中，z_t表示第t时刻系统的状态。
    - 观测变量$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$：其中，$\mathbf{x}_t$表示第$t$时刻的观测变量，通过条件概率$p(\mathbf{x}_t \mid z_t)$由状态变量$z_t$生成。
  - HMM中的独立性：
    - 给定t 时刻的状态变量 z_t，t 时刻的观测变量 $\mathbf{x}_t$ 与其它状态无关：$p(\mathbf{x}_t \mid z_1, \ldots, z_t, \ldots, z_n) = p(\mathbf{x}_t \mid z_t)$
    - 给定t-1 时刻的状态变量z_{t-1}，t 时刻的状态变量 z_t 与之前 t-2 个状态无关：$p(z_t \mid z_1, \ldots, z_{t-1}) = p(z_t \mid z_{t-1})$

- HMM联合概率分布

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n) = p(\mathbf{z}_1)\prod_{t=2}^{n} p(\mathbf{z}_t \mid \mathbf{z}_{t-1})\prod_{t=1}^{n} p(\mathbf{x}_t \mid \mathbf{z}_t)$$

初始状态概率　状态转移概率　发射概率

- HMM的参数
  - 初始状态概率向量 **π**：

$$\pi_k = P(\mathrm{z}_1 = k), \quad 1 \le k \le K$$

  - 状态转移概率矩阵 **A**：

$$A_{i,j} = P(z_t = j \mid z_{t-1} = i), \quad 1 \le i, j \le K$$

  - 发射概率矩阵 **B**：　为简洁起见, 考虑离散的观测变量

$$B_{i,j} = P(x_t = j \mid z_t = i), \quad 1 \le i \le K, 1 \le j \le M$$

- Toy example

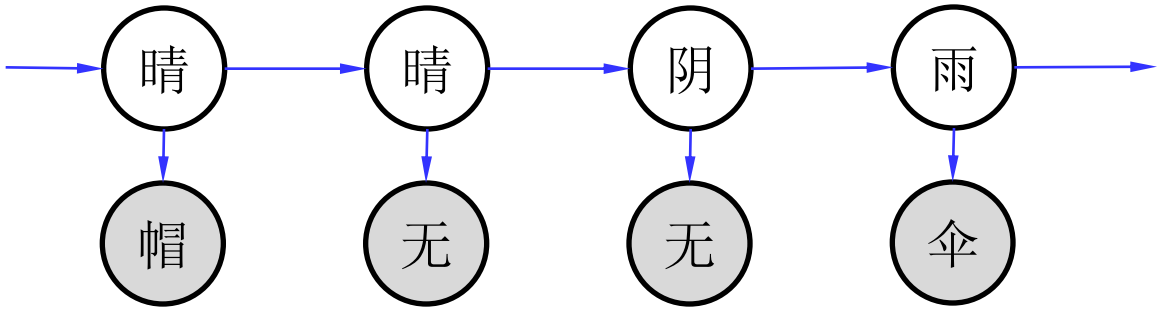
$z_t \in$ {“雨”，“晴”，“阴”}

$x_t \in$ {“打伞”，“戴帽”，“无伞无帽”}

今天 \ 明天	雨	晴	阴
雨	0.1	0.4	0.5
晴	0.1	0.6	0.3
阴	0.2	0.4	0.4

	打伞	戴帽	无伞无帽
雨	0.8	0.1	0.1
晴	0.1	0.4	0.5
阴	0	0.2	0.8

采样：



隐状态序列
（不可观测）

观测序列

EM for hidden Markov model

- HMM的三个基本问题
  - 给定观测序列$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$，如何调整HMM模型参数$\{\mathbf{A}, \mathbf{B}, \pi\}$使该序列出现的概率$P(\mathbf{x} \mid \mathbf{A},\mathbf{B},\pi)$最大？（参数估计）
    - ✓ 通过EM算法迭代求解。
    - ✓ E步，需通过前向-后向算法计算$\mathbf{z}_t$的后验分布和$\{\mathbf{z}_t, \mathbf{z}_{t+1}\}$的联合后验分布；M步，更新$\{\mathbf{A}, \mathbf{B}, \pi\}$
  - 给定HMM模型，如何有效地计算其产生观测序列$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$的概率$P(\mathbf{x}| \mathbf{A},\mathbf{B},\pi)$?
    - ✓ 通过前向算法或后向算法求解。
  - 给定HMM模型和观测序列$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$，如何找到与此观测序列相匹配的状态序列$\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$？（解码问题）
    - ✓ 搜索最优状态路径：$\mathbf{z}^* = \arg\max_{\mathbf{z}} p_\theta(\mathbf{z}_1, \ldots, \mathbf{z}_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)$
    - ✓ 通过维特比算法求解。

内容提要

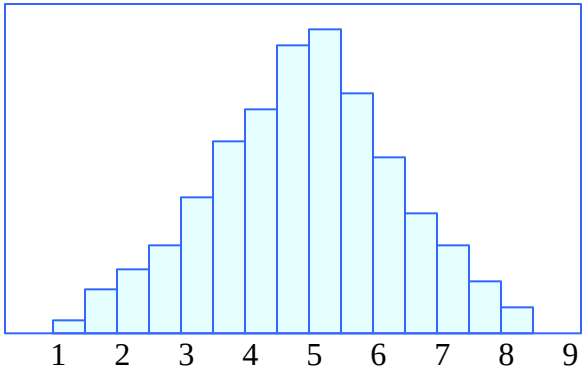
- 引言：非参数密度估计
- Parzen窗估计
- K近邻估计
- K近邻分类器
- 距离度量

4.1 非参数密度估计

- 引言
  - 最大似然方法和贝叶斯估计均属于参数估计方法：
    - 待估计的概率密度函数的形式已知，利用样本来估计函数中的某些参数。
  - 但是，在很多情形下，我们对样本的分布没有充分的了解，无法事先给出概率密度函数的具体形式，而且有些样本的分布也很难用简单的函数来描述。在这些情形下，就需要用非参数方法。
  - 非参数估计方法不需要对概率密度函数的形式作任何假设，而是直接用样本估计出整个函数。

4.1 非参数密度估计

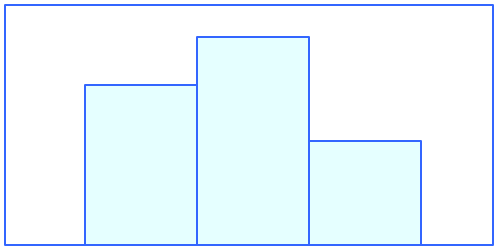
- 直方图方法
  - 最常用的对数据进行统计分析的方法，一种简单的、直观的非参数估计方法



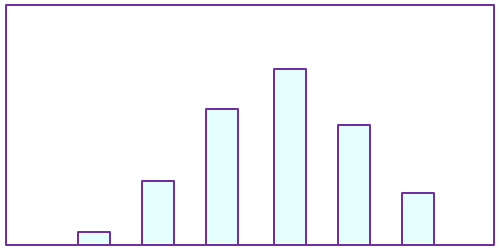
  - 直方图估计方法
    - 步1：将特征空间的每维特征在样本取值范围内分成 l 个等间隔的小窗。如果特征空间是 d 维，则得到 l^d 个小舱，每个小舱的体积记为V。
    - 步2：统计落入每个小舱的样本数目 k_i。
    - 步3：将小舱内的概率密度视为常数，并用 $k_i/(nV)$ 作为其估计值，其中 n 为总样本总数。

4.1 非参数密度估计

- 难点问题
  - 在直方图估计中，我们将特征空间在样本取值范围内进行等距划分。实际上选择多大尺寸的小舱对于密度估计的精度是至关重要的。
  - 如何选定这一尺寸是一个困难问题。如果小舱选择过大，由于假定小舱内$p(\mathbf{x})$ 为常数，则导致过于平均的估计结果。反之，如果小舱过小，落入小舱的样本将会很少，或者没有样本落入，从而导致对$p(\mathrm{x})$的估计不连续。



小窗过宽



小窗过窄

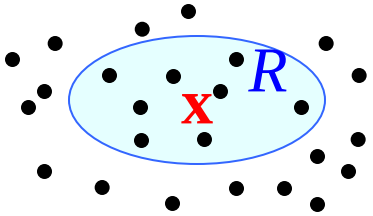
4.1 非参数密度估计

- 非参数估计的基本原理
  - 给定样本集 $D=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$，假定这些样本是从一个服从密度函数$p(\mathbf{x})$的总体分布中独立抽取出来的，目标是给出关于$p(\mathbf{x})$的估计
  - 考虑$\mathbf{x}$点处的一个小区域R，一个样本落入该区域概率是：

$$P = \int_R p(\mathbf{x}')d\mathbf{x}'$$

  - 若 n 个随机样本中有 k 个样本落入小区域 R，k/n 是对概率P的一个很好的估计（无偏估计、最大似然估计）
    - 落入 R 的样本数服从参数为P二项分布
    - 当样本数 n 越来越大时，该估计相当精确

4.1 非参数密度估计



- 非参数估计的基本原理
  - （接上页）k/n 是对 P 的一个很好的估计
  - 另一方面，假设$p(\mathbf{x})$连续，且小区域R的体积V足够小，可以假定在该小区域内 $p(\mathbf{x})$ 是常数，则

$$P = \int_R p(\mathbf{x}')d\mathbf{x}' \approx p(\mathbf{x})V$$

  - 于是有：

$$p(\mathbf{x})V \approx \frac{k}{n} \quad \Rightarrow \quad p(\mathbf{x}) \approx \frac{k}{nV}$$

$\mathbf{x}$点处的概率密度估计

如何决定局部区域的大小：随样本数 n 变化

4.1 非参数密度估计

- 收敛性
  - 对每个样本数n，都构造一个包含$\mathbf{x}$的小区域R_n，从而得到一个区域序列：$R_1, R_2, R_3, \ldots$。令V_n表示R_n的体积，k_n表示落入R_n的样本个数。则使用n个样本对$\mathbf{x}$点处的概率密度估计为：

$$p_n(\mathbf{x}) = \frac{k_n}{nV_n}$$

  - 为了使 $p_n(\mathbf{x})$ 收敛到 $p(\mathbf{x})$, 需要满足以下三个条件:

$$\lim_{n\to\infty} V_n = 0; \quad \lim_{n\to\infty} k_n = \infty; \quad \lim_{n\to\infty} \frac{k_n}{n} = 0$$

这几个条件说明：随着样本数目的增加，小舱的体积应该尽可能小，同时又必须保证小舱内有充分多的样本，但每个小舱内的样本又必须是总数的很小一部分。

4.1 非参数密度估计

- How to achieve the goal?
  - There are two common ways to obtain sequences of regions that satisfy these conditions:
    - One is to specify the volume V_n as some function of n, guaranteeing that $p_n(\mathbf{x})$ converges to $p(\mathbf{x})$. This is the Parzen-window method.（对给定的n，小舱体积在全局处处保持不变，小舱内的样本数可变）
    - The second method is to specify k_n as some function of n. Here the volume V_n is grown until it encloses k_n neighbors of $\mathbf{x}$. This is the k_n-nearest-neighbor estimation method. (对给定的n，小舱内的样本数不变，小舱大小可变)
  - Both of these methods do converge, although it is difficult to make meaningful statements about their finite-sample behavior.

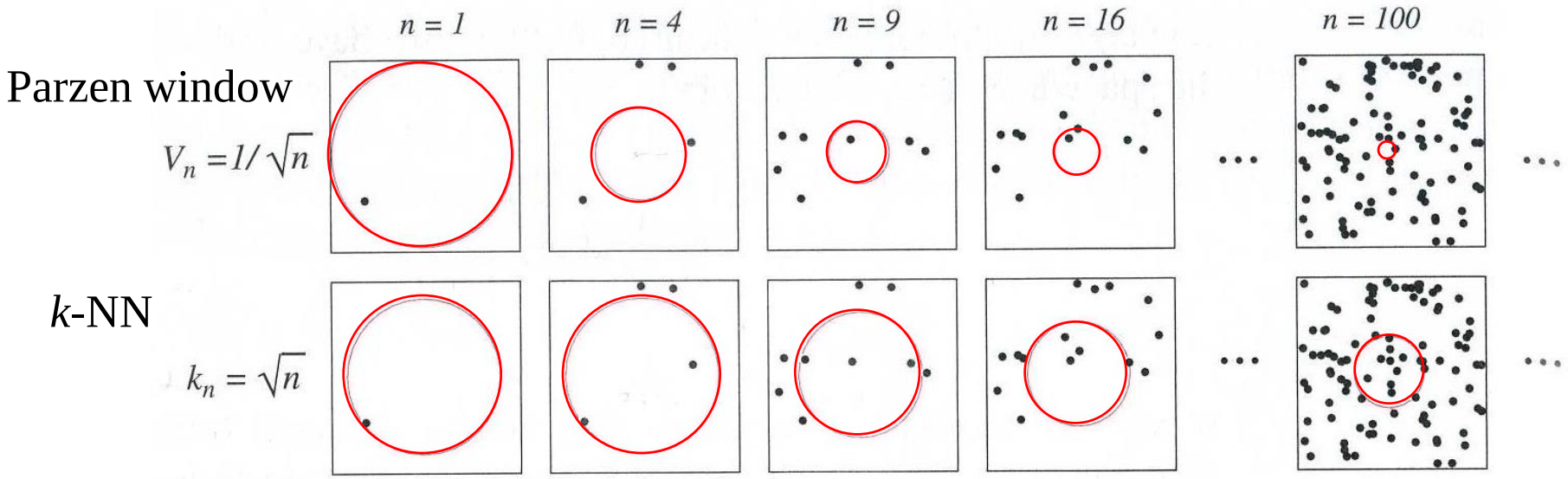
4.1 非参数密度估计

- How to achieve the goal?

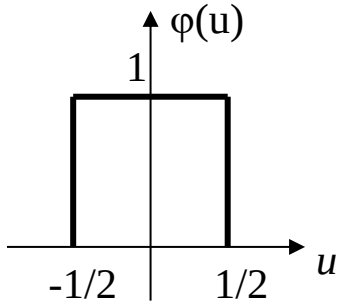
当n固定，**x**变化时：

- **Parzen window**: 固定局部区域体积V，k 变化
- **k-nearest neighbor**: 固定局部样本数k，V 变化

当**x**固定，n变化时：



4.2 Parzen窗方法



- 方法介绍
  - 假设 **x** 是 d 维空间中的点，并假设小舱是一个超立方体，其棱长为h_n，因此其体积为：

$$V_n = (h_n)^d$$

$$p_n(\mathbf{x}) = \frac{k_n}{nV_n}$$

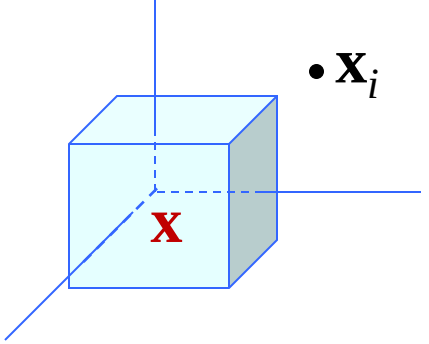
  - 现在的任务：要统计落入以**x**为中心的超立方体内样本的个数 k_n。为此，定义如下d 维单位方窗函数:

$$\varphi(\mathbf{u}) = \begin{cases} 1, & |u_j| \le \frac{1}{2}, \quad j = 1,2,...,d \\ 0, & \text{otherwise} \end{cases}$$

$$\left(\text{here } \mathbf{u} = [u_1, u_2, ..., u_d]^T \in R^d\right)$$

该函数在以原点为中心的 d 维单位超立方体内取值为1，而其它地方取值均为零。

4.2 Parzen窗方法



- 方法介绍
  - 对于点$\mathbf{x}$，考察样本$\mathbf{x}_i$是否在以$\mathbf{x}$ 为中心、h_n为棱长的超立方体内，可通过如下函数来判定：

$$\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

  - 若共有n 个样本 $D=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$，那么落入以$\mathbf{x}$为中心、$h_n$为棱长的超立方体内的样本总数为：

$$k_n=\varphi\left(\frac{\mathbf{x}-\mathbf{x}_1}{h_n}\right)+\varphi\left(\frac{\mathbf{x}-\mathbf{x}_2}{h_n}\right)+\cdots\varphi\left(\frac{\mathbf{x}-\mathbf{x}_n}{h_n}\right)=\sum_{i=1}^{n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

  - 从而有：

$$p_n(\mathbf{x})=\frac{k_n}{nV_n}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

4.2 Parzen窗方法

- $p_n(\mathbf{x})$是一个概率密度函数
  - 定义如下函数：

$$\delta_n(\mathbf{x}) = \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right) \Rightarrow p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta_n(\mathbf{x}-\mathbf{x}_i)$$

  - 引入如下积分变换函数：

$$\mathbf{u} = \frac{\mathbf{x}-\mathbf{x}_i}{h_n} \quad \Rightarrow \quad \mathbf{x} = h_n\mathbf{u} + \mathbf{x}_i$$

$$\int \delta_n(\mathbf{x}-\mathbf{x}_i)d\mathbf{x} = \int \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)d\mathbf{x} = \int \frac{1}{V_n}\left|\frac{\partial\mathbf{x}}{\partial\mathbf{u}}\right|\varphi(\mathbf{u})d\mathbf{u}$$

$$= \int \frac{h_n^d}{V_n}\varphi(\mathbf{u})d\mathbf{u}$$

$$= \int \varphi(\mathbf{u})d\mathbf{u}$$

$$= 1$$

$$\Rightarrow \int p_n(\mathbf{x})d\mathbf{x} = \frac{1}{n}n = 1$$

$p_n(\mathbf{x})$是一个概率密度函数

4.2 Parzen窗方法

- 窗函数的选择
  - 一般地，要保证 $p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta_n(\mathbf{x}-\mathbf{x}_i)$ 为一个概率密度函数，只需满足：

$$\delta_n(\mathbf{x}) \geq 0, \quad \int \delta_n(\mathbf{x})d\mathbf{x} = 1$$

  - $\delta_n(\mathbf{x} - \mathbf{x}_i)$反映了样本 $\mathbf{x}_i$ 对在 $\mathbf{x}$ 处概率密度估计贡献的大小，通常与 $\mathbf{x}_i$ 到 $\mathbf{x}$ 的距离有关
  - 概率密度估计$p_n(\mathbf{x})$就是把D中所有观测样本在$\mathbf{x}$点的贡献进行平均

4.2 Parzen窗方法

- 窗函数的选择
  - 除了以上方窗函数，还可定义其它函数：
    - 方窗（非单位长度）：

$$\delta(\mathbf{x}-\mathbf{x}_i)=\begin{cases}\dfrac{1}{h_n^d}, & |x_j-x_{i,j}|\le\dfrac{h_n}{2}, \quad j=1,2,...,d\\ 0, & \text{otherwise}\end{cases}$$

    - 正态窗：$\delta(\mathbf{x}-\mathbf{x}_i)=\dfrac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\dfrac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{x}_i)\right)$

    - 球窗：$\delta(\mathbf{x}-\mathbf{x}_i)=\begin{cases}\dfrac{1}{V}, & \|\mathbf{x}-\mathbf{x}_i\|_2\le r\\ 0, & \text{otherwise}\end{cases}$

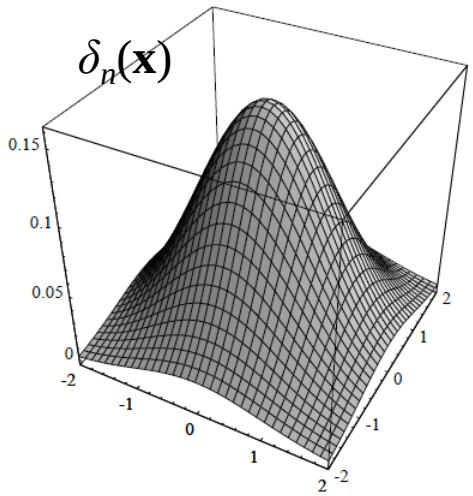
V: 超球体体积
r: 超球体半径

4.2 Parzen窗方法

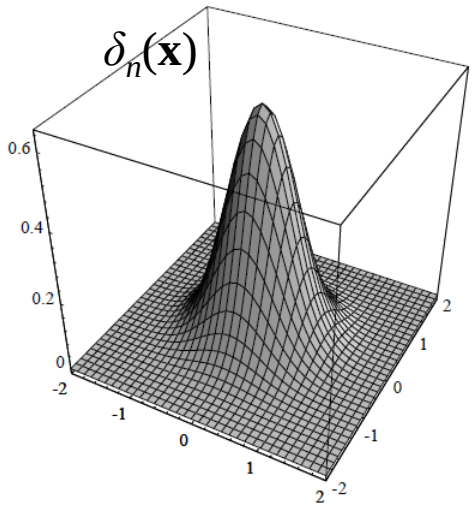
$$\delta_n(\mathbf{x}) = \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right)$$

$$\int \delta_n(\mathbf{x})d\mathbf{x} = 1$$

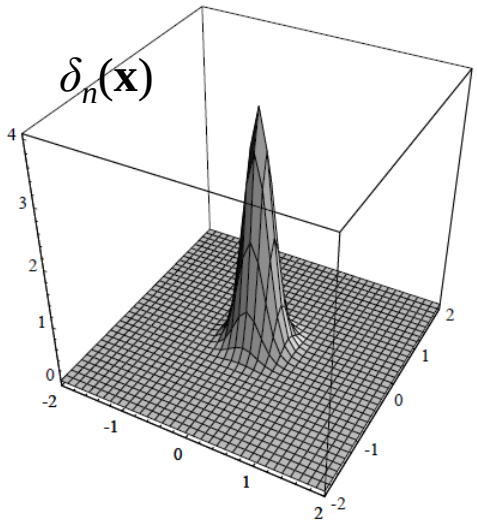
- 窗宽的影响
  - 当h_n较大时，$\delta_n(\mathbf{x})$ 低矮平滑；当h_n较小时，$\delta_n(\mathbf{x})$ 尖锐、变化剧烈；h_n趋近于0时，$\delta_n(\mathbf{x})$趋近于狄拉克delta函数



h_n=1

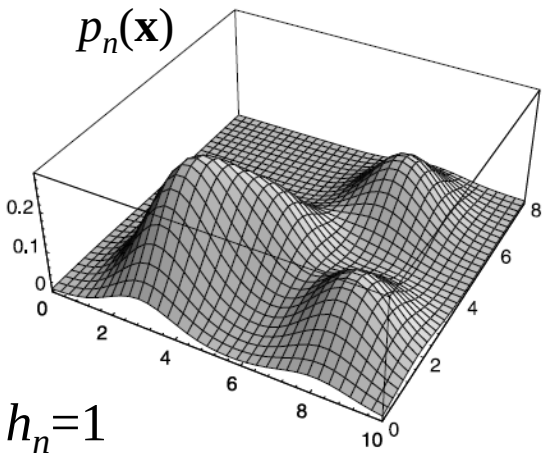


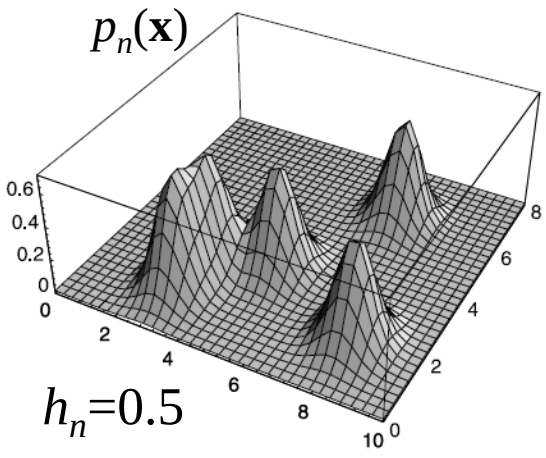
h_n=0.5

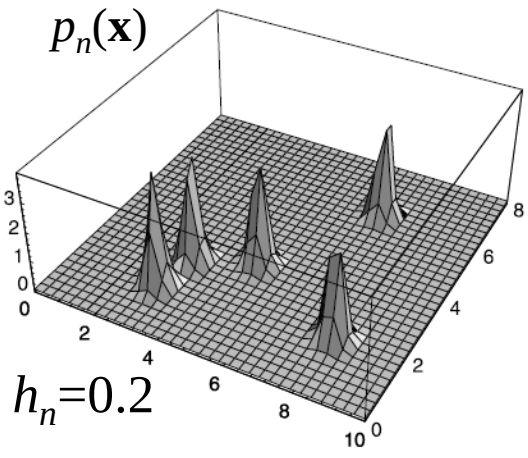


h_n=0.2

- 窗宽的影响







✓ 若 h_n 较大，则 $\delta_n(\mathbf{x}-\mathbf{x}_i)$ 的幅度将较小，宽度很大。此时，$p_n(\mathbf{x})$是 n 个幅度缓慢变化的函数的迭加，较平滑，难以跟上 $p(\mathbf{x})$ 的变化，分辨率较低。

✓ 若 h_n 较小，则 $\delta_n(\mathbf{x}-\mathbf{x}_i)$ 的幅度将较大，宽度很小，近似于以$\mathbf{x}_i$ 为中心的狄拉克delta 函数。此时，$p_n(\mathbf{x})$ 是 n 个以 $\mathbf{x}_i$ 为中心的脉冲在点 $\mathbf{x}$ 处的迭加，波动较大，不稳定，也可能不连续。

Large h_n: low variability, under-fitting

Small h_n: high variability, over-fitting

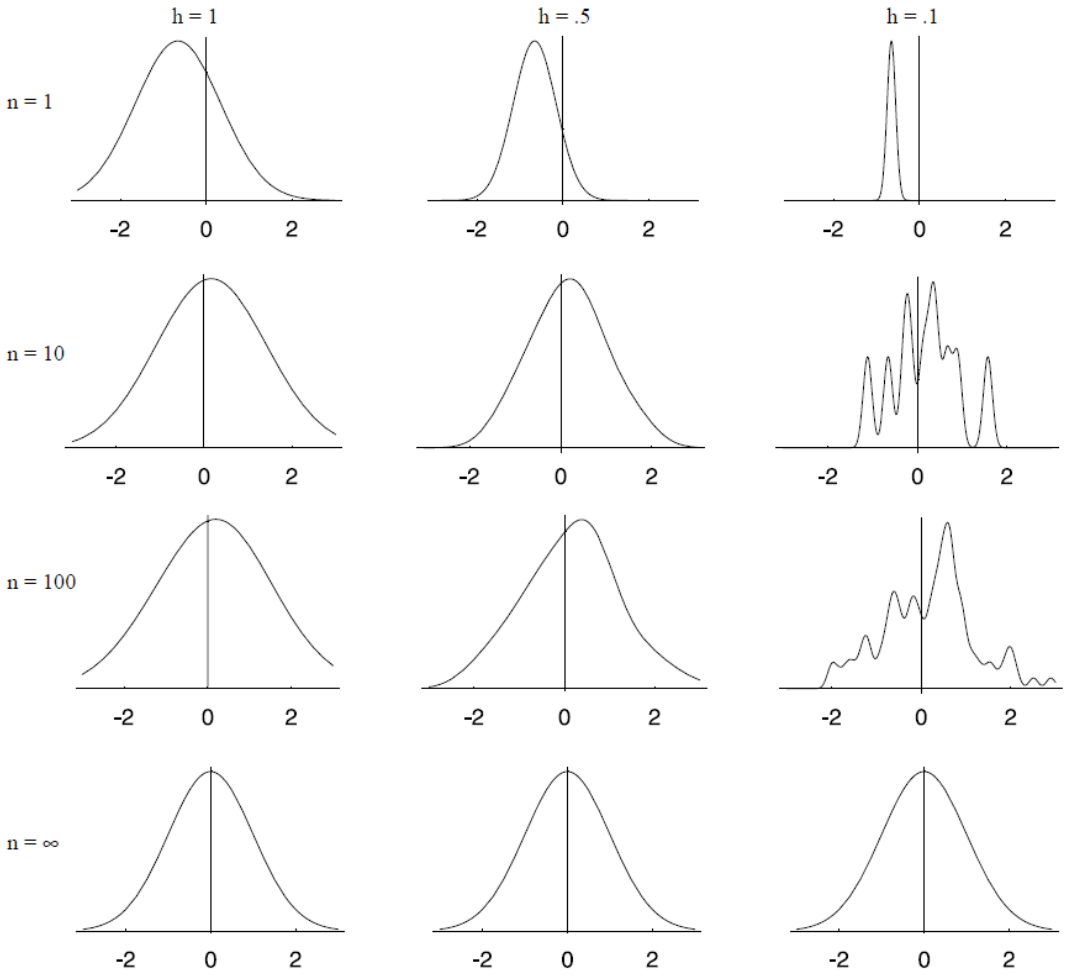
例子：用高斯窗估计高斯分布（一维）

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$$

$$p_n(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

$$h_n=\frac{h}{\sqrt{n}}$$

h: 初始窗宽，
一个可调参数



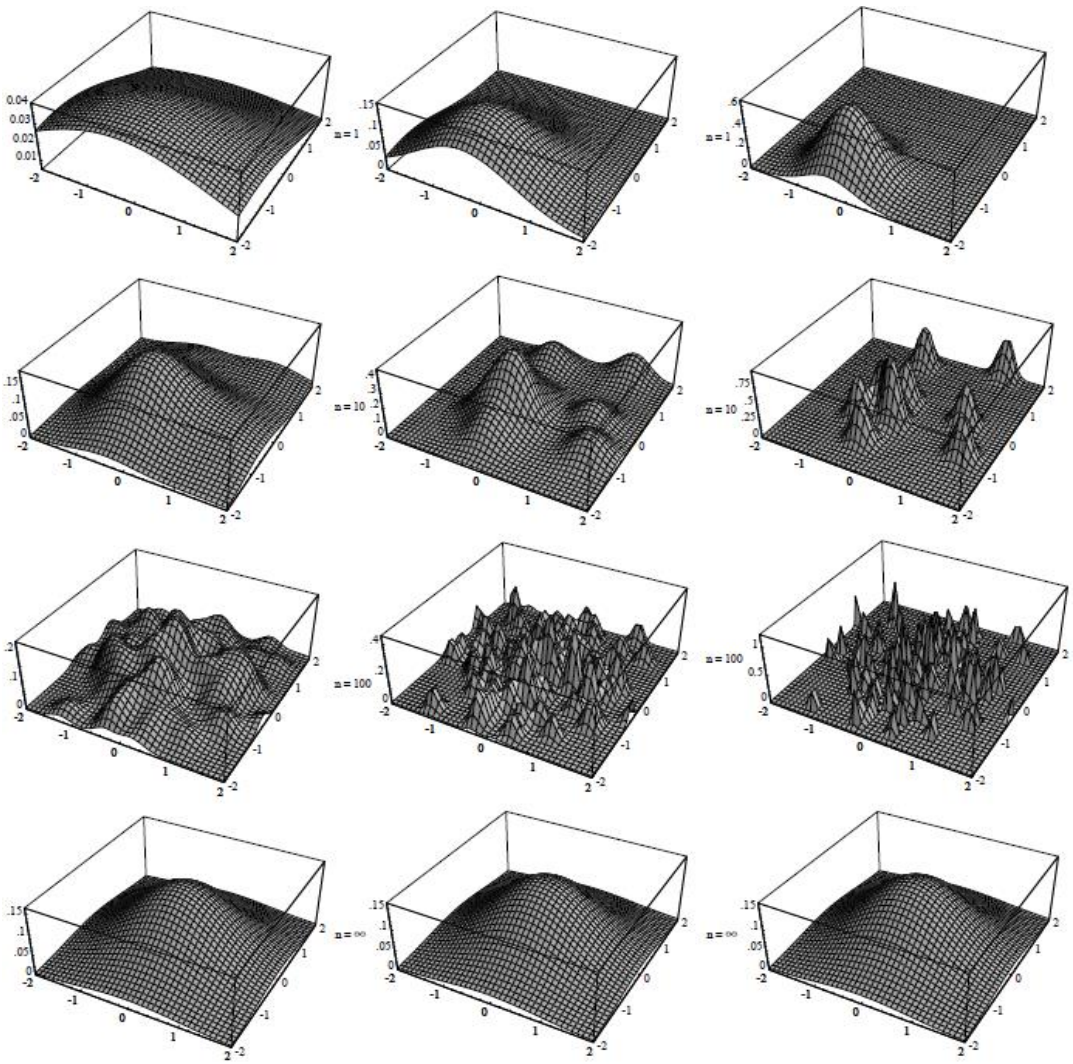
例子：用高斯窗估计高斯分布（二维）

n=1

n=10

n=100

n=∞

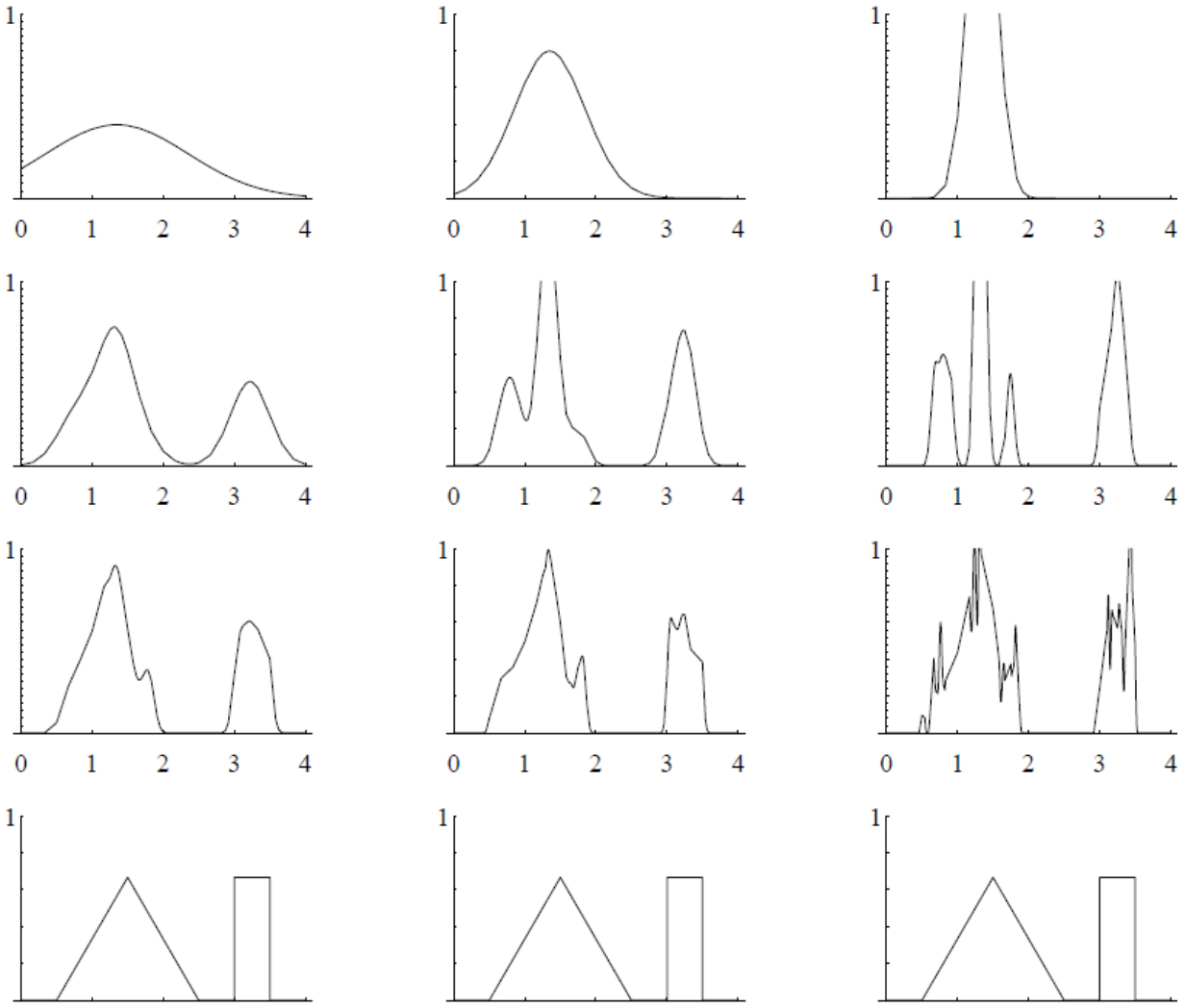


h: 初始窗宽

$h_n=h/\sqrt{n}$

h=2　　h=1　　h=0.5

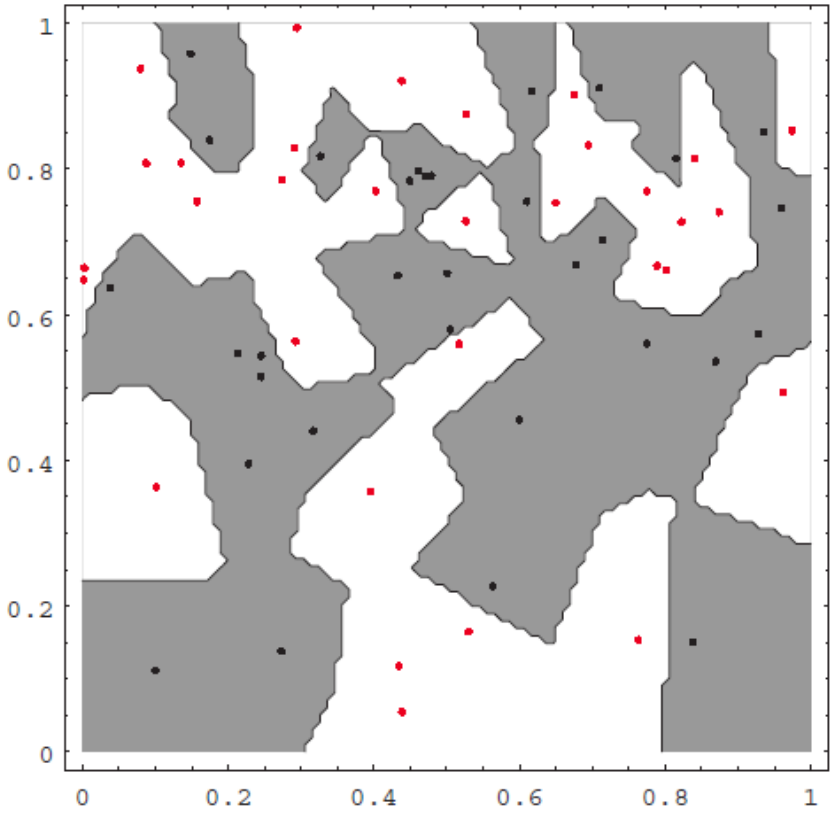
例子：用高斯窗估计多峰分布

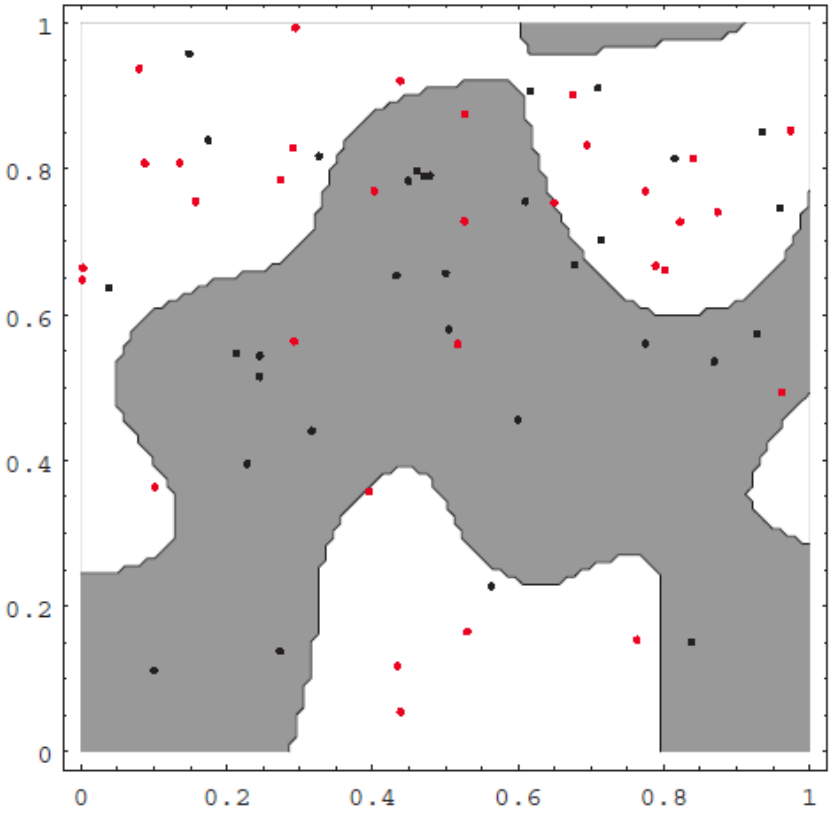


h: 初始窗宽

$h_n=h/\sqrt{n}$

使用Parzen窗方法进行分类：两类例子





- ✓ 对每个类别，用Parzen窗方法估计类条件概率密度$p(\mathbf{x}|\omega_i)$，用最大后验准则分类。
- ✓ 使用小窗宽得到的分类界面（左图）比大窗宽的分类界面（右图）更复杂。

4.2 Parzen窗方法

- 一个具体例子：
  - 给定一维空间中的五个点x_1=2, x_2=2.5, x_3=3, x_4=1.5, x_5=6。假定窗函数为σ=1的高斯函数。试计算x=3位置处的Parzen概率密度估计值。
  - 解：

$$\hat{p}(x)=\frac{1}{5}\sum_{i=1}^{5}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-x_i)^2}{2}\right)$$

将x=3代入得, x=3位置处的Parzen概率密度估计值0.225。进一步可将上述函数画出。

4.2 Parzen窗方法

- 算法特点
  - 适用范围广，无论概率函数是规则的或者是不规则的、单峰的还是多峰的；
  - 当样本数趋于无穷时，Parzen窗估计收敛于真实$p(x)$；
  - 但该方法要求样本的数量要大；
  - 选择合适的窗口函数将有利于提高估计的精度和减少样本的数量。
  - 与直方图仅仅在每个固定小窗内估计平均密度不同，Parzen窗用滑动的小窗来估计每个点上的概率密度。

4.3 K近邻估计

- Parzen窗估计中，小窗的体积V_n被视为样本个数n的函数，比如 $V_n = V_1/\sqrt{n}$
  - 当V_1选择得太小，导致大部分区域是空的，会使$p_n(\mathbf{x})$不稳定
  - 当V_1选择得太大，则$p_n(\mathbf{x})$会变得过于平坦，从而失去一些重要的空间变化
  - K近邻估计方法是克服这一问题的一种可能方法

4.3 K近邻估计

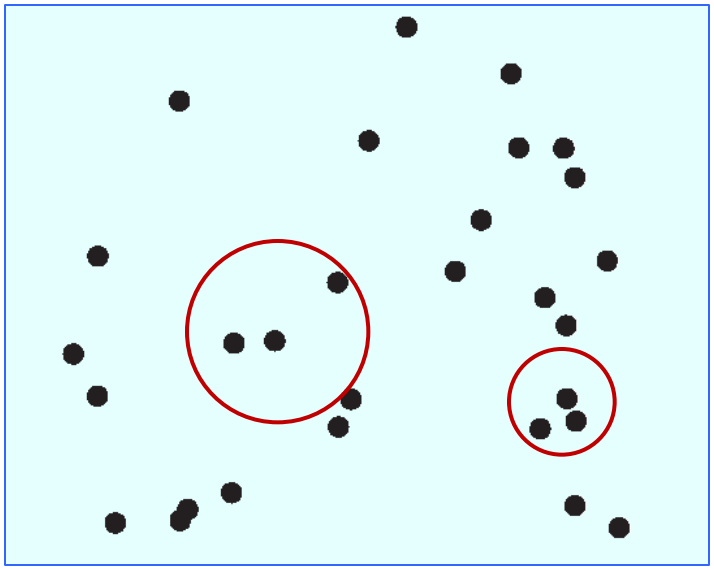
- 基本方法
  - K近邻估计是一种采用大小可变舱的密度估计方法。其基本做法是：根据总样本数确定一个参数k_n，要求每个小舱内拥有的样本数目是k_n。
  - 在估计$\mathbf{x}$点处的概率密度$p(\mathbf{x})$时，我们调整包含$\mathbf{x}$的小舱的体积，直到小舱内恰好落入k_n个样本，此时有：

$$p_n(\mathbf{x}) = \frac{k_n}{nV_n}$$

V_n不固定，与$\mathbf{x}$的位置相关

4.3 K近邻估计

- 基本方法
  - If the density is high near **x**, the cell will be relatively small, which leads to good resolution. If the density is low, the cell will grow large, but it will stop soon after it enters regions of higher density.



$k_n=3$

4.3 K近邻估计

- Convergence
  - We want k_n to go to infinity as n goes to infinity, since this assures us that k_n/n will be a good estimate of the probability that a point will fall in the cell of volume V_n.
  - However, we also want k_n to grow sufficiently slowly that the size of the cell needed to capture k_n training samples will shrink to zero. Thus, it is clear from the estimation about $p_n(\mathbf{x})$ that the ratio k_n/n must go to zero.

$$\lim_{n\to\infty} k_n = \infty; \quad \lim_{n\to\infty} \frac{k_n}{n} = 0$$

4.3 K近邻估计

- Convergence
  - **x**点处，概率密度函数$p(\mathbf{x})$的估计值为 $p_n(\mathbf{x}) = \dfrac{k_n}{nV_n}$
  - 如果取 $k_n = \sqrt{n}$

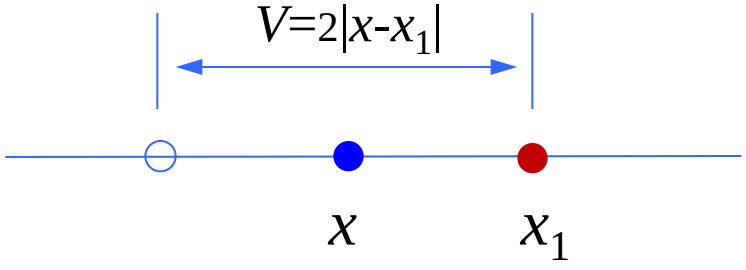
$$p_n(\mathbf{x}) = \frac{\sqrt{n}}{nV_n} \quad \Rightarrow \quad p(\mathbf{x}) \approx \frac{1}{\sqrt{n}V_n} \quad \Rightarrow \quad V_n \approx \frac{1}{\sqrt{n}\,p(\mathbf{x})} = \frac{V_1}{\sqrt{n}}$$

$$V_1 = \frac{1}{p(\mathbf{x})}$$

(与Parzen窗方法不同， V_1不再是一个固定的值，而是与密度相关)

4.3 K近邻估计

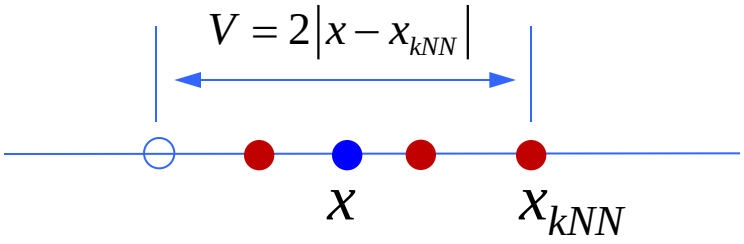
- 密度函数估计的解析表达式
  - 在**x**点处，概率密度函数的估计值为 $p_n(\mathbf{x})=\dfrac{k_n}{nV_n}$
  - 关键是如何计算V_n
  - 对于一维来讲，考虑最近邻情形，即k_n=1。假定样本x_1是x点的最近邻样本，则x点处的最近邻估计为：



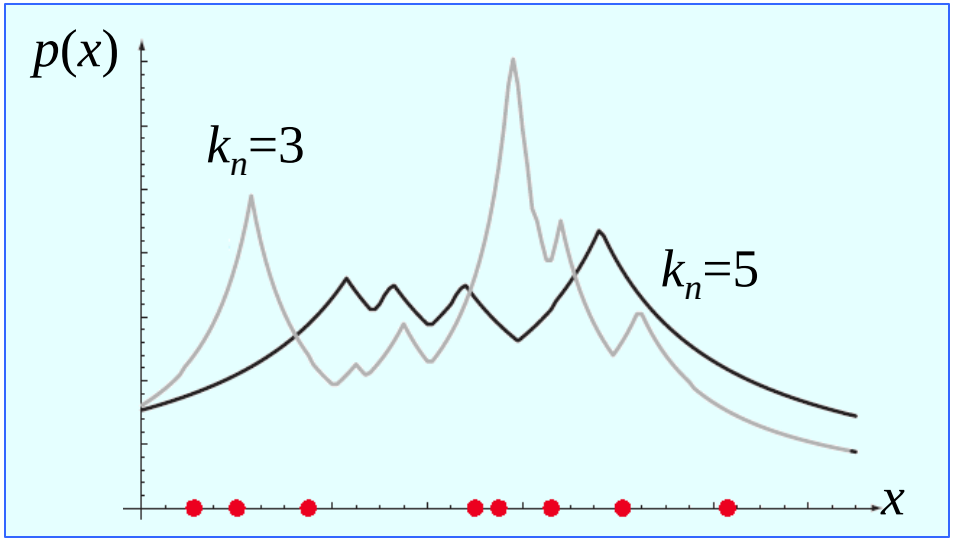
$$p_n(\mathbf{x})=\frac{k_n}{nV_n}=\frac{1}{2n|x-x_1|}$$

此时，*K*近邻估计在特征空间积分为正无穷，不是严格意义的概率密度函数

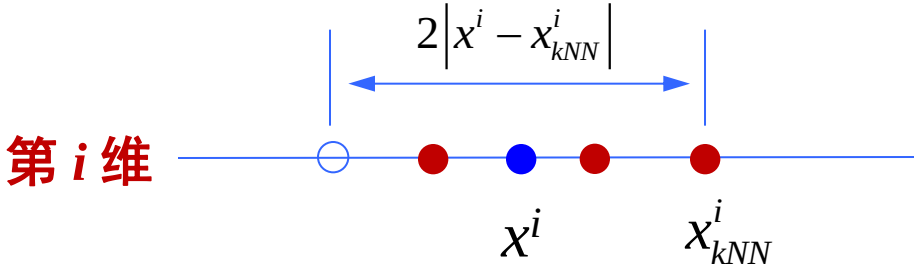
- 密度函数估计的解析表达式
  - 对于一维来讲，考虑k_n近邻情形，此时：



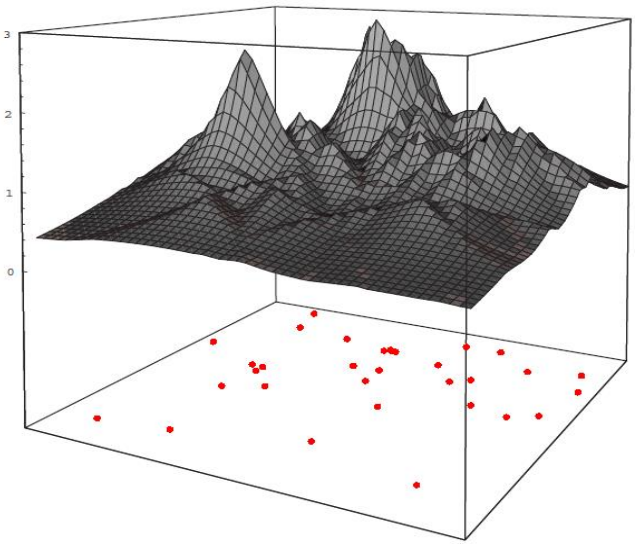
$$p_n(\mathbf{x}) = \frac{k_n}{nV_n} = \frac{k_n}{2n|x - x_{kNN}|}$$



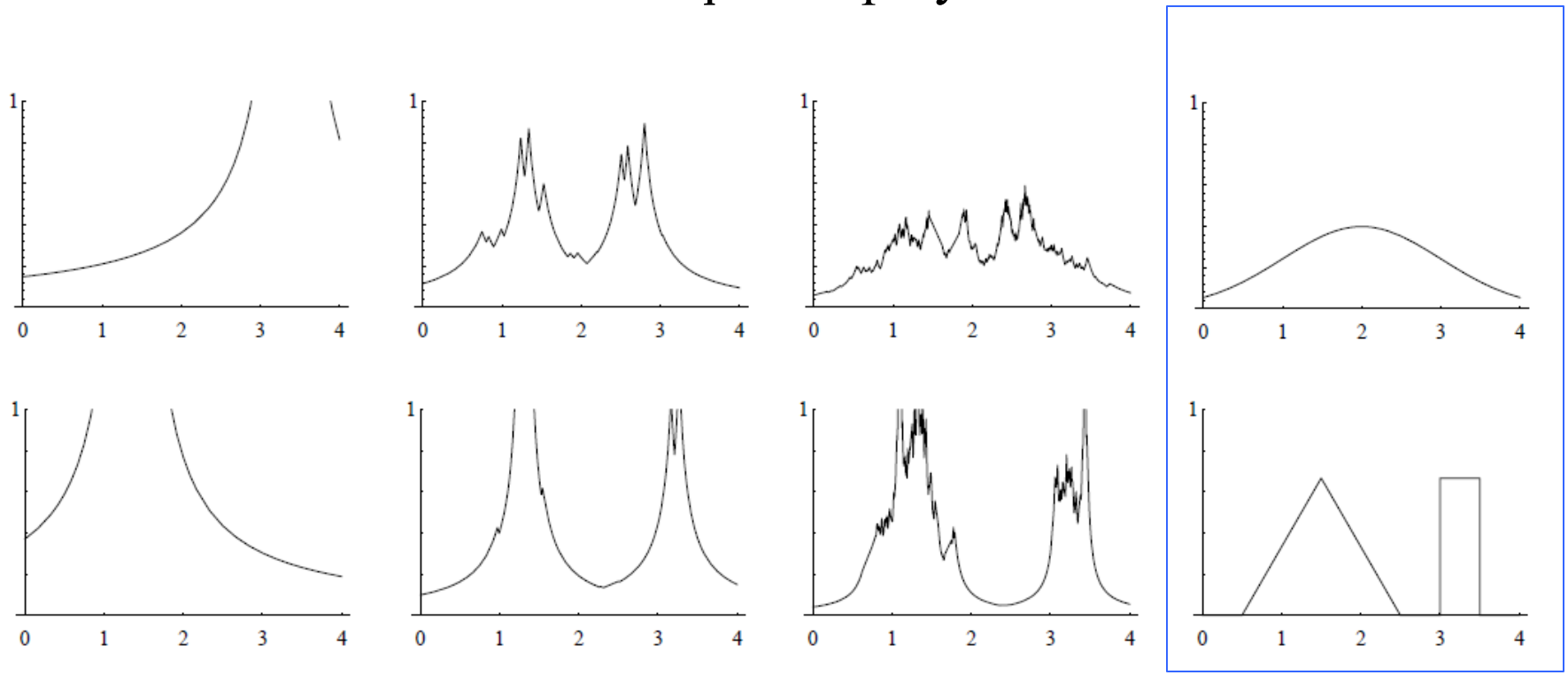
- 密度函数估计的解析表达式
  - 对于多维情形，可以用球来度量K近邻小舱，也可以用立方体包围盒来度量K近邻小舱。
  - 比如，采用立方体包围盒：



$$p_n(\mathbf{x}) = \frac{k_n}{nV_n} = \frac{k_n}{2^d n \prod_{i=1}^{d} \left|x^i - x^i_{kNN}\right|}$$



- Several *k*-nearest-neighbor estimates of two unidimensional densities: a Gaussian and a bimodal distribution. Notice how the finite *n* estimates can be quite "spiky."



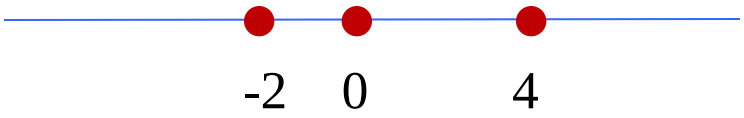
n=1, k_n=1 n=16, k_n=4 n=256, k_n=16 n=∞, k_n=∞

$$k_n = \sqrt{n}$$

4.3 K近邻估计

- 一个具体的例子
  - 给定一维空间三个样本点{-2,0,4}，请写出概率密度函数$p(x)$的最近邻（1-NN）估计，并画出概率密度函数曲线图。

$$p_n(x)=\frac{k_n}{nV_n}=\begin{cases}\dfrac{1}{6|x+2|}, & \text{if } x<-1\\ \dfrac{1}{6|x|}, & \text{if } -1<x<2\\ \dfrac{1}{6|x-4|}, & \text{if } x>2\end{cases}$$



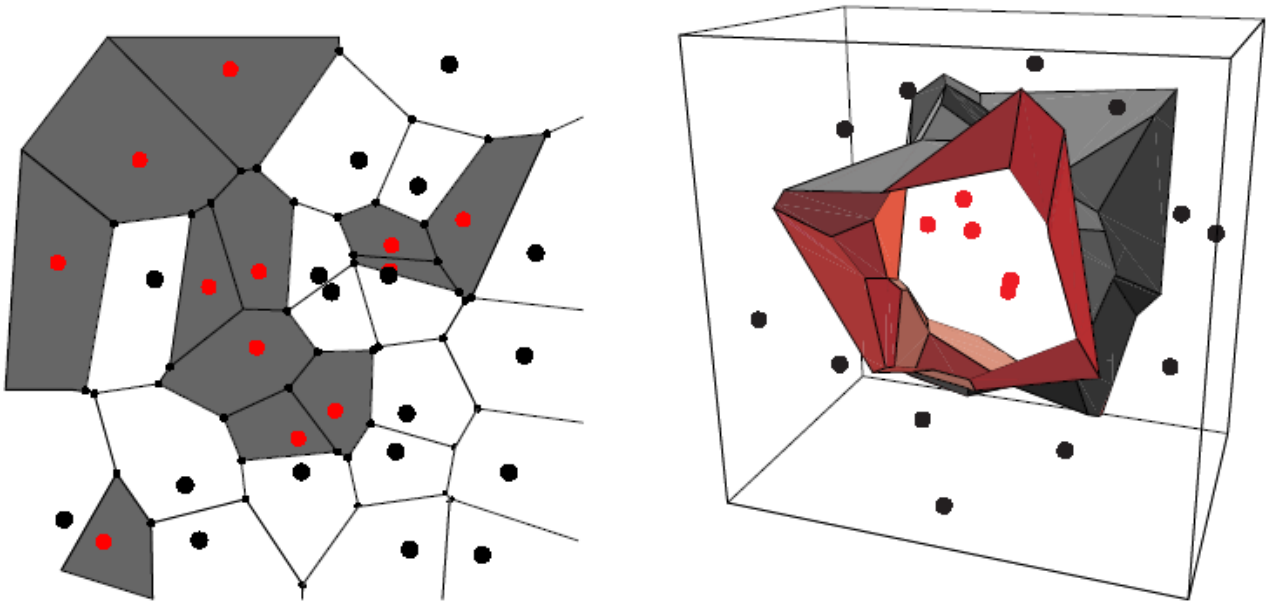
4.4 K近邻分类器

- 最近邻分类器
  - 方法：对测试样本$\mathbf{x}$，将其与训练样本逐一进行比较，找出距离$\mathbf{x}$最近的训练样本，以该样本的类别作为$\mathbf{x}$的类别。
  - 直观理解：给定训练样本集D，对于测试样本$\mathbf{x}$，假设$\mathbf{x}' \in D$为其最近邻样本，ω_m是$\mathbf{x}'$的类别标签。如果$\mathbf{x}'$与$\mathbf{x}$充分接近，有理由相信它们的类别相同，或者类后验概率相同，即对任意类别ω_i，$P(\omega_i \mid \mathbf{x}') \approx P(\omega_i \mid \mathbf{x})$。

- 最近邻分类器
  - 如果将最近邻决策写成判别函数的形式，ω_i类的判别函数可以写作：

$$g_i(\mathbf{x}) = \max_{\mathbf{x}_j \in \omega_i} -d(\mathbf{x}, \mathbf{x}_j), \quad i = 1, 2, ..., c$$

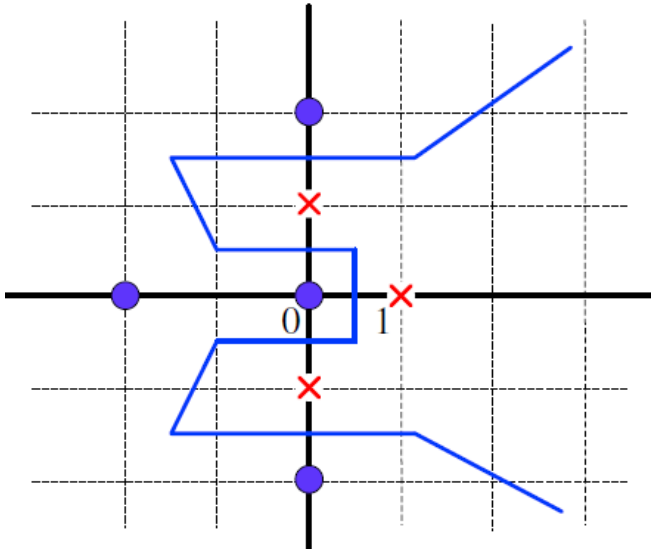
  - 决策规则：$\arg\max_{i \in \{1,...,c\}} g_i(\mathbf{x})$



Voronoi tessellation of the space

4.4 K近邻分类器

- 一个具体的例子
  - 现有7个二维向量：$\mathbf{x}_1 = (1, 0)^T$ ，$\mathbf{x}_2 = (0, 1)^T$ ，$\mathbf{x}_3 = (0,-1)^T$, $\mathbf{x}_4 = (0, 0)^T$, $\mathbf{x}_5 =(0, 2)^T$, $\mathbf{x}_6 = (0,-2)^T$, $\mathbf{x}_7 = (-2, 0)^T$ 。上标T表示向量转置。假定前三个为 ω_1类，后四个为ω_2类。画出最近邻法决策面。



4.4 K近邻分类器

- 最近邻规则的错误率分析
  - 研究表明，在训练样本足够多时，最近邻决策可以取得很好的效果。
  - Our analysis of the behavior of the nearest-neighbor rule will be directed at obtaining the infinite-sample conditional average probability of error $P(e|x)$, where the averaging is with respect to the training samples.
  - Generally, the unconditional average probability of error will be found by averaging $P(e|\mathbf{x})$ over all $\mathbf{x}$:

$$P(e) = \int P(e \mid \mathbf{x})\, p(\mathbf{x}) d\mathbf{x}$$

4.4 K近邻分类器

- 最近邻规则的错误率分析
  - We should recall that the Bayes decision rule minimizes $P(e)$ by minimizing $P(e|\mathbf{x})$ for every $\mathbf{x}$.
  - Recall again that if we let $P^*(e|\mathbf{x})$ be the minimum possible value of $P(e|\mathrm{x})$, and P^* be the minimum possible value of $P(e)$, then

贝叶斯错误率：

$$P^*(e\,|\,\mathbf{x}) = 1 - P(\omega_m\,|\,\mathbf{x}), \qquad P^* = \int P^*(e\,|\,\mathbf{x})\,p(\mathbf{x})d\mathbf{x}$$

where $\omega_m = \arg\max_i P(\omega_i\,|\,\mathbf{x}_i)$

4.4 K近邻分类器

- 最近邻规则的错误率分析
  - We now wish to evaluate the average probability of error for the nearest-neighbor rule.
  - In particular, if $P_n(e)$ is the n-sample error rate, and if we denote

$$P(e) = \lim_{n \to \infty} P_n(e)$$

then, we have

$$P^* \le P(e) \le P^*(2 - \frac{c}{c-1} P^*)$$

- 最近邻规则的错误率[略]

$$P(e) = \int P(e|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$$

$$P(e|\mathbf{x}) = \int P(e|\mathbf{x},\mathbf{x}')p(\mathbf{x}'|\mathbf{x})\,d\mathbf{x}' \qquad \mathbf{x}'\text{: NN of } \mathbf{x}$$

  - 当n→∞, $p(\mathbf{x}'|\mathbf{x})$ (x'为x的最近邻的概率)趋近以**x** 为中心的delta函数
  - 对$P(e|\mathbf{x},\mathbf{x}')$, 假设 **x** 和 $\mathbf{x}_j'$ (最近训练样本，与**x** 独立)的类别标号分别为 θ 和 θ_n'

$$P(\theta,\theta_n'|\mathbf{x},\mathbf{x}_n') = P(\theta|\mathbf{x})P(\theta_n'|\mathbf{x}_n')$$

$$P_n(e|\mathbf{x},\mathbf{x}_n') = 1-\sum_{i=1}^{c} P(\theta=\omega_i,\theta_n'=\omega_i|\mathbf{x},\mathbf{x}_n') = 1-\sum_{i=1}^{c} P(\omega_i|\mathbf{x})P(\omega_i|\mathbf{x}_n')$$

$$\lim_{n\to\infty} P_n(e|\mathbf{x}) = \int[1-\sum_{i=1}^{c} P(\omega_i|\mathbf{x})P(\omega_i|\mathbf{x}_n')]\delta(\mathbf{x}_n'-\mathbf{x})d\mathbf{x}_n' = 1-\sum_{i=1}^{c} P^2(\omega_i|\mathbf{x})$$

- 最近邻规则的错误率[略]
  - Asymptotic error rate $\lim_{n\to\infty} P_n(e|\mathbf{x}) = 1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x})$

$$\begin{aligned} P &= \lim_{n\to\infty} P_n(e) \\ &= \lim_{n\to\infty} \int P_n(e|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \\ &= \int \left[1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x})\right] p(\mathbf{x})\, d\mathbf{x} \end{aligned}$$

- Error bound of 1-NN rule

$$\sum_{i=1}^{c} P^2(\omega_i|\mathbf{x}) = P^2(\omega_m|\mathbf{x}) + \sum_{i\neq m} P^2(\omega_i|\mathbf{x})$$

Minimized when P_i ($i \neq m$) are equal

$$P(\omega_i|\mathbf{x}) = \begin{cases} \frac{P^*(e|\mathbf{x})}{c-1} & i \neq m \\ 1 - P^*(e|\mathbf{x}) & i = m \end{cases}$$

$P^*(e|\mathbf{x}) = 1 - P(\omega_m|\mathbf{x})$

(Bayes error)

$$\sum_{i=1}^{c} P^2(\omega_i|\mathbf{x}) \geq (1 - P^*(e|\mathbf{x}))^2 + \frac{P^{*2}(e|\mathbf{x})}{c-1}$$

- Error bound of 1-NN rule [略]

$$\sum_{i=1}^{c} P^2(\omega_i|\mathbf{x}) \geq (1 - P^*(e|\mathbf{x}))^2 + \frac{P^{*2}(e|\mathbf{x})}{c-1}$$

$$1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x}) \leq 2P^*(e|\mathbf{x}) - \frac{c}{c-1} P^{*2}(e|\mathbf{x})$$

  - Error rate

$$P = \int \left[1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x})\right] p(\mathbf{x})\, d\mathbf{x} \longrightarrow P \leq 2P^*$$

$$\begin{aligned} \mathrm{Var}[P^*(e|\mathbf{x})] &= \int [P^*(e|\mathbf{x}) - P^*]^2 p(\mathbf{x})\, d\mathbf{x} \\ &= \int P^{*2}(e|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} - P^{*2} \geq 0 \longrightarrow \int P^{*2}(e|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \geq P^{*2} \end{aligned}$$

  - Error bound

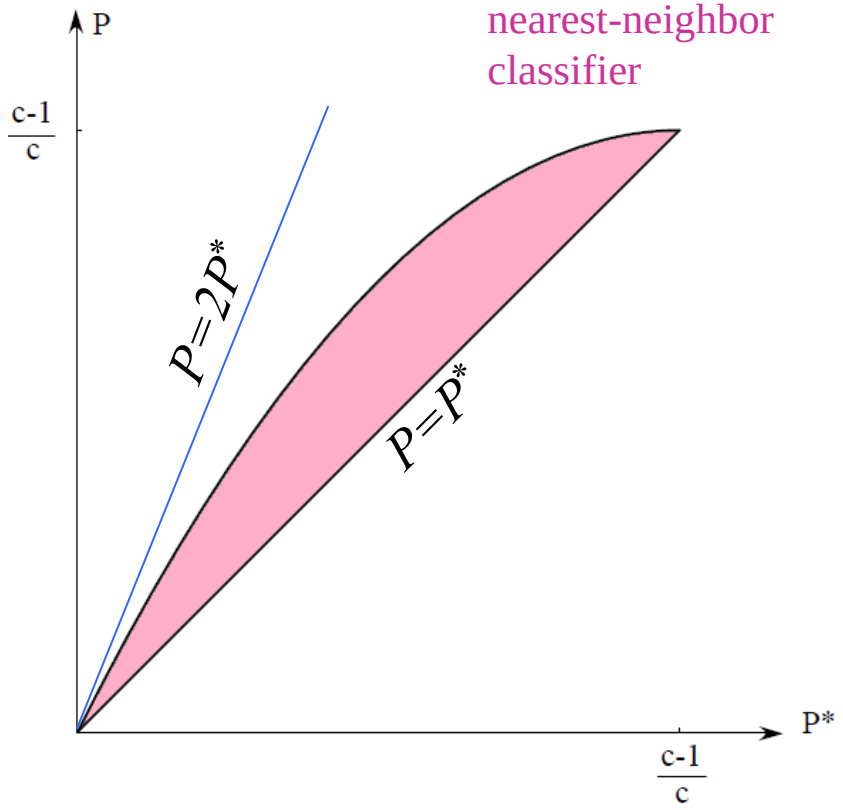
$$P^* \leq P \leq P^*\left(2 - \frac{c}{c-1} P^*\right)$$

详见R.O. Duda etc., Pattern Classification (chapter 4.5), 2000.

4.4 K近邻分类器

- 最近邻错误率分析
  - 这个结论告诉我们：最近邻法的渐近错误率最坏不会超过两倍的贝叶斯错误率，而最好则有可能接近或达到贝叶斯错误率。
  - 最近邻法的渐近错误率总落入在如下阴影之中。

possible asymptotic error rates of the *c*-category nearest-neighbor classifier



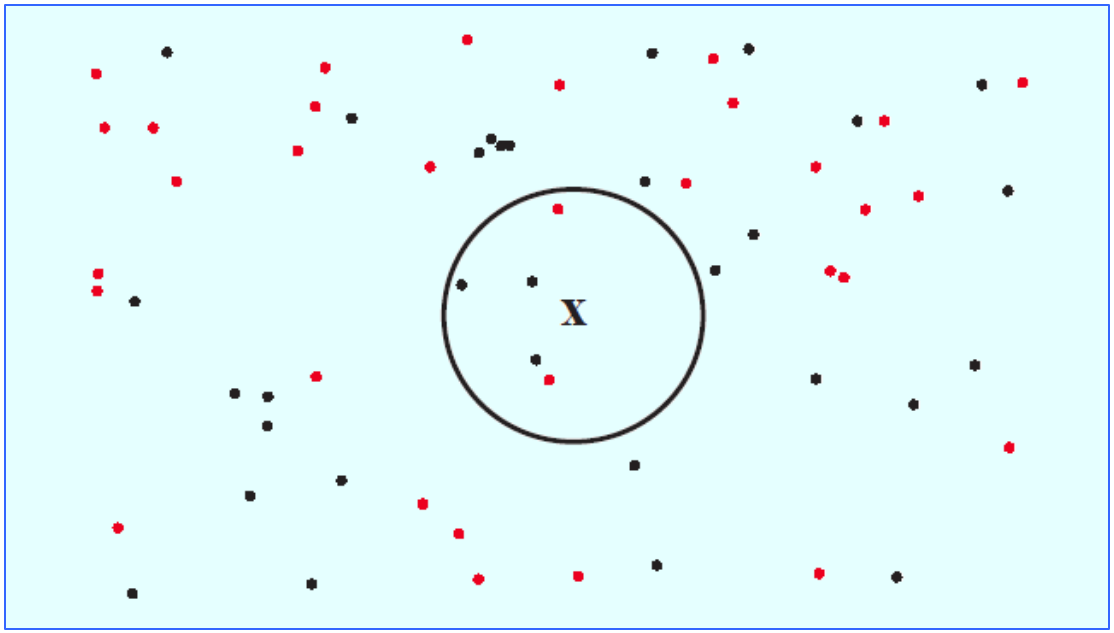
记：$P = P(e)$

4.4 K近邻分类器

- K近邻分类器
  - 在很多情况下，将决策建立在最近邻样本上有一定的风险。（数据分布复杂，存在噪声）
  - 一种自然的改进就是引入投票机制，选择前 k 个距离测试样本最近的训练样本，用它们的类别投票来决定新样本的类别。
  - 给定训练集D，对测试样本$\mathbf{x}$，设$X_{k\mathrm{NN}}$ 是$\mathbf{x}$在D中的 k 个最近邻样本。设ω_i 类的判别函数 $g_i(\mathbf{x}) = k_i$，k_i是$X_{k\mathrm{NN}}$中ω_i 类样本的个数。
  - 决策规则：$\arg\max_{i\in\{1,\ldots,c\}} g_i(\mathbf{x})$

4.4 K近邻分类器

- K近邻分类器
  - 决策规则： $\arg\max_{i\in\{1,\ldots,c\}} g_i(\mathbf{x})$



k=5

4.4 K近邻分类器

- K近邻分类器与K近邻估计的关系
  - 假设训练集D包含n个样本，其中ω_i类样本n_i个，$n=\sum_{i=1}^{c} n_i$。
  - 测试样本$\mathbf{x}$处，包含k个最近邻样本的区域体积为V，其中ω_i类样本k_i个，$k=\sum_{i=1}^{c} k_i$.
  - 则类条件概率密度$p(\mathbf{x}\mid\omega_i)$和类先验概率$p(\omega_i)$的估计为：

$$p_n(\mathbf{x}\mid\omega_i)=\frac{k_i}{n_i V},\qquad p(\omega_i)=\frac{n_i}{n}$$

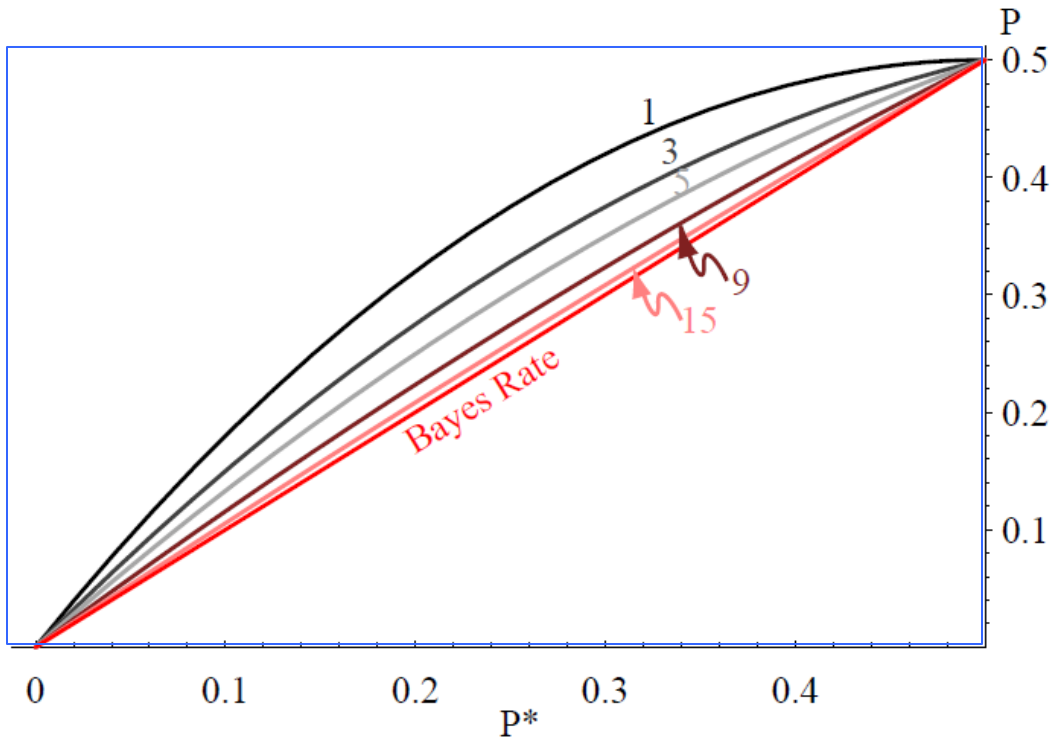
样本$\mathbf{x}$的类后验概率为：

$$p_n(\omega_i\mid\mathbf{x})=\frac{p_n(\mathbf{x}\mid\omega_i)p(\omega_i)}{\sum_{i=1}^{c} p_n(\mathbf{x}\mid\omega_i)p(\omega_i)}=\frac{k_i}{k},$$

$$\omega_m=\arg\max_i\left\{p_n(\omega_i\mid\mathbf{x})\right\}$$ 这就是*K*近邻分类器！

4.4 K近邻分类器

- K近邻分类器错误率分析
  - For the error rates will be further decreased, compared with the 1-NN case.



4.5 K近邻的快速计算

- 搜索*K*近邻的计算复杂度
  - O(*dnK*)
- 快速搜索策略
  - 部分距离(partial distance)
  - 搜索树
  - 剪辑/压缩算法(pruning, condensing)
- 部分距离

Partial square distance (r<d): $D_r^2(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{r}(a_i - b_i)^2$

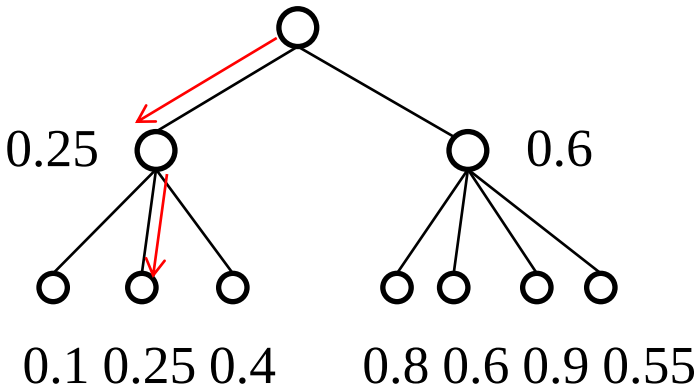
Full distance to the current closest prototype $D^2(\mathbf{x}, \mathbf{x}')$
Terminate computing if the partial square distance is greater than $D^2(\mathbf{x}, \mathbf{x}')$

4.5 K近邻的快速计算

- 搜索树
  - 将所有训练样本以树的形式组织起来，树中每个节点代表一个样本（树的构造过程省略，可参见kd tree）
  - 通常，在特征空间距离相近的样本，在树中对应节点的路径长度较短
  - 查找**x**的近邻时，先与根节点的子节点比较，选择最优的子树，递归向下搜索。

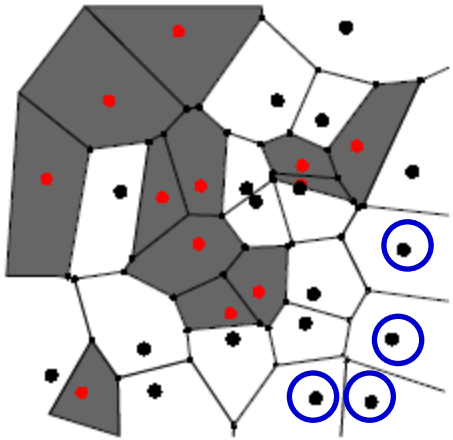
设样本服从(0,1)上的均匀分布，D={0.1, 0.8, 0.25, 0.6, 0.9, 0.4, 0.55}

x=0.3



4.5 K近邻的快速计算

- 压缩近邻法
  - Remove prototypes that are surrounded by samples of same class
  - 考虑近邻法的分类原理，远离分类边界的样本对于分类决策没有贡献。只要能够找出各类样本中最有利于与其它类相互区分的代表性样本，则可将很多训练样本去掉，简化决策过程中的计算。



4.5 K近邻的快速计算

- 压缩近邻法
  - 将样本集分为两个活动的子集： X_S和X_G，前者称为储存集（Storage）, 后者称为备选集（GrabBag）。
    - **首先，**在算法开始时，X_S中只有一个样本，其余样本均在X_G中。
    - **然后，**考查X_G中的每一个样本，如果采用X_S中的样本能够对其正确分类，则该样本仍然保留在X_G中，否则移动到X_S中，从而扩大代表集合。依次重复进行上述操作，直到没有样本需要搬移为止。
    - **最后，**用X_S中的样本作为代表样本，对新来的样本进行分类。

4.6 距离度量

- 距离度量是模式分类的基础
  - 设有 d 维空间的三个样本$\mathbf{x}$，$\mathbf{y}$ 和 $\mathbf{z}$，记 $d(.,.)$为一个$\mathrm{R}^d \times \mathrm{R}^d \rightarrow \mathrm{R}$的映射，如满足如下几个条件则称$d(.,.)$为一个距离:
    - $d(\mathbf{x}, \mathbf{y}) \geq 0$ 非负性
    - $d(\mathbf{x}, \mathbf{x}) = 0$ 自相似性
    - $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ 对称性
    - $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ 三角不等式
  - 距离可以描述点对间的相异程度，距离越大，两个点越不相似；距离越小，两个点越相似。

4.6 距离度量

- 设$\mathbf{x}, \mathbf{y} \in \mathrm{R}^d$，Minkowski 距离度量定义如下：

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{d} |x_i - y_i|^q \right)^{\frac{1}{q}}$$

$q = 1$	$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{d} \|x_i - y_i\|$	城区距离 曼哈顿距离
$q = 2$	$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{d} \|x_i - y_i\|^2}$	欧氏距离
$q = \infty$	$d(\mathbf{x},\mathbf{y}) = \max_{1 \le i \le d} \|x_i - y_i\|$	切比雪夫距离

4.6 距离度量

- 欧氏距离：

$$\mathbf{x} = [\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare]^T \in \mathbf{R}^d$$
$$\mathbf{y} = [\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare]^T \in \mathbf{R}^d$$

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} = \left((\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})\right)^{\frac{1}{2}}$$

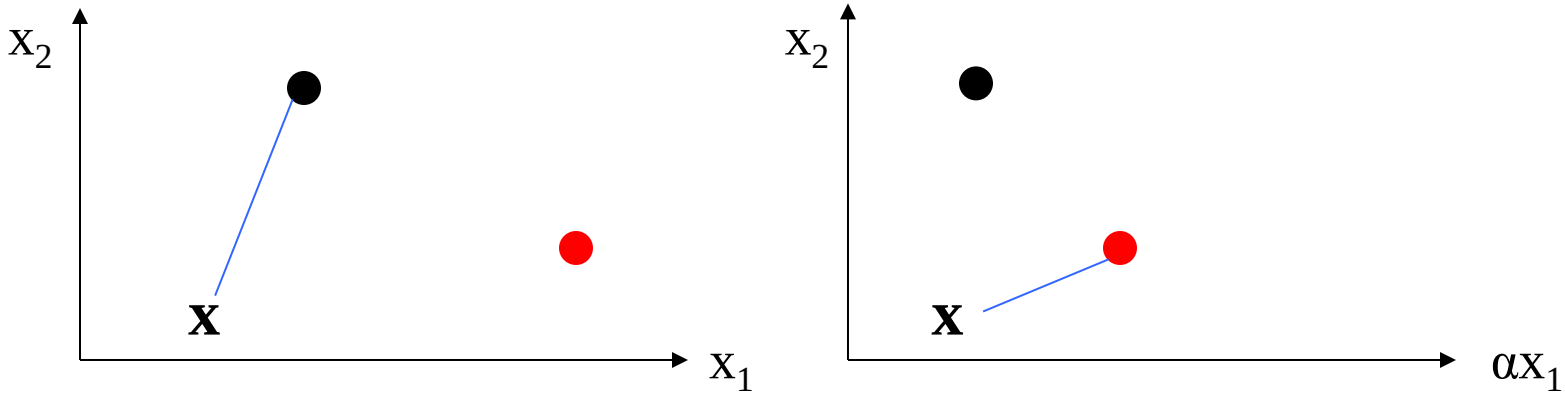
- **Mahalanobis (马氏)距离：** 设$\mathbf{x}, \mathbf{y} \in R^d$，定义如下：

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{M}(\mathbf{x}-\mathbf{y})}$$ （其中，**M**是半正定矩阵）

  - **M**为单位矩阵时，退化为欧氏距离
  - **M**为对角矩阵时，退化为特征加权欧氏距离
  - 由$\mathbf{M}=\mathbf{Q}^T\mathbf{Q}$，$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{M}(\mathbf{x}-\mathbf{y})} = \sqrt{(Q\mathbf{x}-Q\mathbf{y}))^T(Q\mathbf{x}-Q\mathbf{y}))}$，可看作是在新特征空间中的欧氏距离。

4.6 距离度量

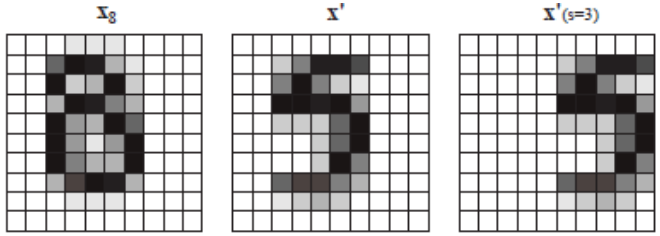
- 距离对坐标变换敏感

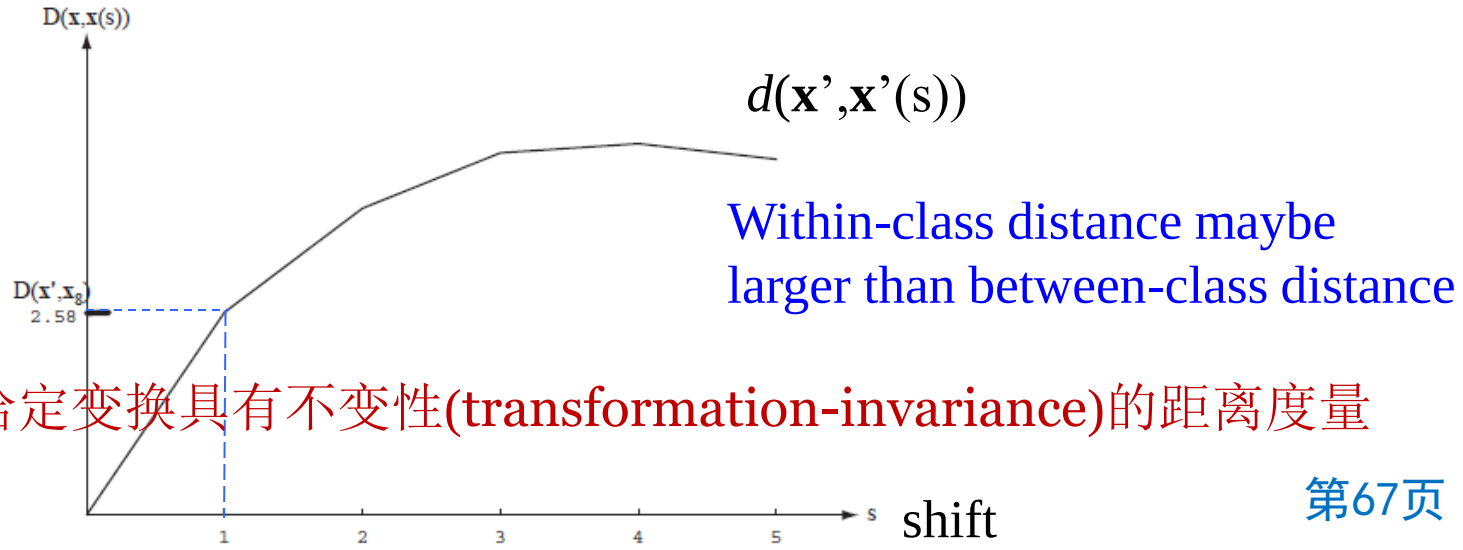


- 向量间距离的扩展
  - 分布间距离：KL散度、交叉熵
  - 集合间距离：见chapter 9（聚类）
- 距离通常是无监督概念
  - 如何利用先验知识、监督信息？

4.6 距离度量

- Tangent Distance
  - 图像变换对距离的影响
    - Shift (translation), rotation, scaling, distortion
    - Distance sensitive to transformation



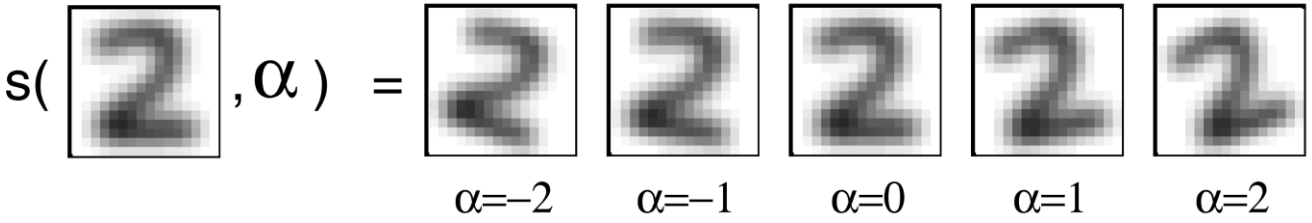


Within-class distance maybe larger than between-class distance

希望找到对给定变换具有不变性(transformation-invariance)的距离度量

4.6 距离度量

- Tangent Distance
  - 变换的参数化表示： $s(\mathbf{x}, \alpha)$，α是变换s的参数
    - 例：s表示旋转变换，α表示角度



  - 变换s下样本x的表示

$$S_{\mathbf{x}} = \{\mathbf{x}' \mid \exists \alpha \text{ for which } \mathbf{x}' = s(\mathbf{x}, \alpha)\}$$ 样本空间中的一维流形

  - 样本y到x的距离

$$d(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x}' \in S_{\mathbf{x}}} \|\mathbf{y} - \mathbf{x}'\| = \min_{\alpha} \|\mathbf{y} - s(\mathbf{x}, \alpha)\|$$ 在$S_{\mathbf{x}}$中找与y距离最近的点

该距离对变换s具有不变性

4.6 距离度量

- Tangent Distance
  - 对变换s的线性近似

$$s(\mathbf{x},\alpha) = s(\mathbf{x},0) + \alpha\frac{\partial s(\mathbf{x},0)}{\partial \alpha} + O(\alpha^2) \approx \mathbf{x} + \alpha T$$

s在α=0点的泰勒展开

其中 $T = \left.\frac{\partial s(\mathbf{x},\alpha)}{\partial \alpha}\right|_{\alpha=0}$ 称为切向量(tangent vector)

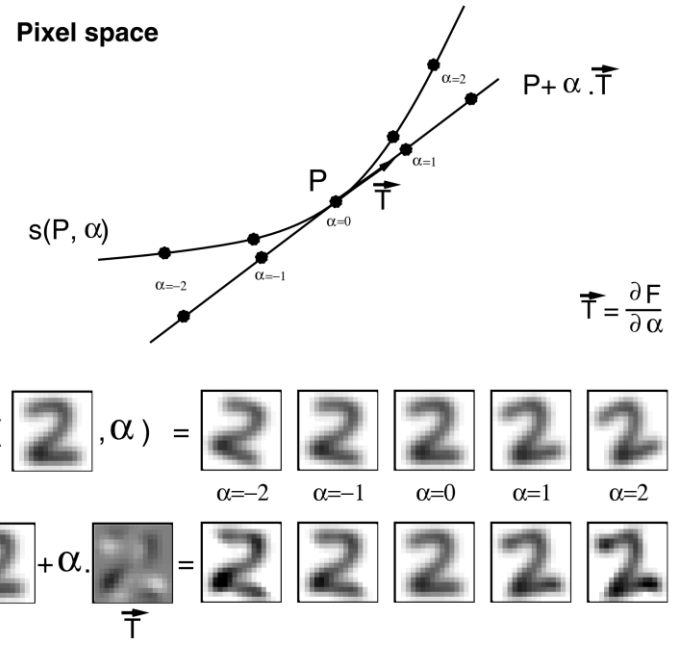
  - 用线性子空间近似S_x

$$\{\mathbf{x}' \mid \exists \alpha \text{ for which } \mathbf{x}' = \mathbf{x} + \alpha T\}$$

  - 样本y到x的切距离

$$\min_{\alpha} \|\mathbf{y} - (\mathbf{x} + \alpha T)\|$$

y到子空间的最小距离



4.6 距离度量

- Tangent Distance
  - 一般地，假设变换s 包含m 个变换参数：$\alpha=(\alpha_1, \alpha_2, \ldots, \alpha_m)$
    - 例如：$\alpha_1, \alpha_2, \alpha_3$分别表示旋转角度、水平位移、垂直位移
  - 此时，变换s 下样本**x** 的表示

    $S_{\mathbf{x}} = \{\mathbf{x}' \mid \exists \alpha \in R^m \text{ for which } \mathbf{x}' = s(\mathbf{x}, \alpha)\}$　样本空间中的m维流形
  - 用切平面（m维线性子空间）近似$S_{\mathbf{x}}$

    $$\{\mathbf{x}' \mid \exists \alpha \in R^m \text{ for which } \mathbf{x}' = \mathbf{x} + \sum_{i=1}^{m} \alpha_i T_i\} \qquad T_i = \left.\frac{\partial s(\mathbf{x}, \alpha)}{\partial \alpha_i}\right|_{\alpha=0}$$

  - 样本**y**到**x**的切距离

    $$\min_{\boldsymbol{\alpha} \in R^m} \left\| \mathbf{y} - (\mathbf{x} + \sum_{i=1}^{m} \alpha_i T_i) \right\|$$

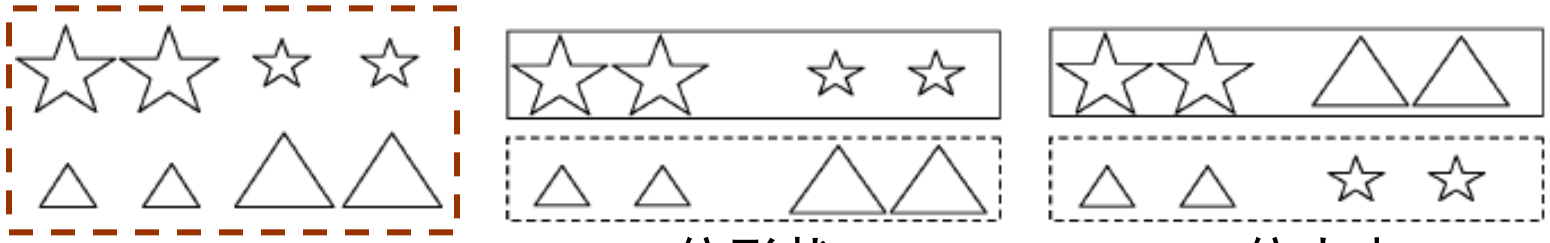
P.Y. Simard etc. Efficient pattern recognition using a new transformation distance. NIPS, 1992.

4.6 距离度量

- 距离度量学习

能否让距离度量反映用户偏好或某些先验知识？

形状检索



图像检索

Can we utilize **the prior knowledge** to learn a distance metric to meet the needs of the users?

4.6 距离度量

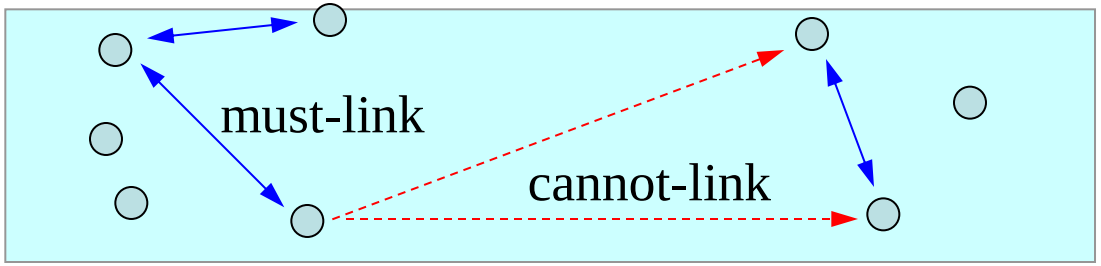
- 距离度量学习
  - 为了反映特征之间的关联关系
  - 为了充分利用先验知识
  - 反映特定问题的特定需求

图像检索

Can we utilize **the prior knowledge** to learn a distance metric to meet the needs of the users?

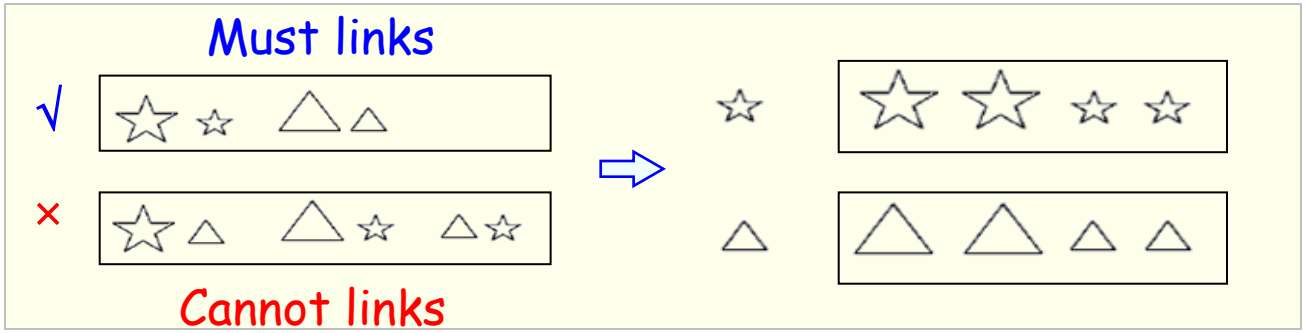
4.6 距离度量

- 一种典型的距离度量问题描述
  - Assume data points in $D = \{\mathbf{x}_i\} \subset R^d$. Given:
    - a must-link set: $S = \{ (\mathbf{x}_i, \mathbf{x}_j)$ is a must-link $\}$
    - a cannot-link set: $D = \{ (\mathbf{x}_i, \mathbf{x}_j)$ is a cannot-link $\}$



侧信息 (side information)

(举例)



4.6 距离度量

- 一个经典的学习模型 (xing et al.)

不相似点对马氏距离尽可能大

$$d_{\mathbf{A}}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{A}(\mathbf{x}-\mathbf{y})}$$

$$\max_{\mathbf{A}} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D} d_{\mathbf{A}}(\mathbf{x}_i,\mathbf{x}_j)$$

$$\text{s.t.} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S} [d_{\mathbf{A}}(\mathbf{x}_i,\mathbf{x}_j)]^2 \le 1$$

$$\mathbf{A} \in R^{m\times m} \succeq 0$$

半正定

相似点对马氏距离尽可能小

E. Xing, M. Jordan, S. J. Russell, A. Ng. Distance metric learning with application to clustering with side-information, Advances in neural information processing systems, 2002

总结

- 非参数法的基本思想
  - 没有给定概率密度函数形式
  - 基于概率和密度的原始定义，以训练样本的局部分布近似**x**的局部密度
- Parzen window
- K-nearest neighbor (k-NN)
  - 1-nearest neighbor (1-NN), Error bound
  - 快速搜索
- 距离度量
  - Tangent distance
  - Metric learning

下次课内容

- 线性判别函数与决策面
- 广义线性判别函数
- 线性感知器准则及松驰方法
- 线性最小二乘方法

致谢

- PPT由向世明老师提供

Thank All of You!

(Questions?)

张燕明

ymzhang@nlpr.ia.ac.cn

people.ucas.ac.cn/~ymzhang

模式分析与学习课题组(PAL)

中科院自动化研究所· 模式识别国家重点实验室