Chapter 4: 非参数技术
Nonparametric Techniques

1

本章目录

中山大學

1. 概率密度的估计
2. Parzen窗方法
3. Kn-近邻估计
4. 最近邻规则
5. 距离度量和最近邻分类

2

4.1 引 言

中山大學

贝叶斯公式：

$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)}$$

- 参数估计方法的问题
  - 一般的给出的概率密度的形式很少符合实际情况
  - 所有经典密度函数的参数形式都是单模的
    - 而现实中，很多实际问题都是多模的密度函数
  - 高维密度函数通常表示成一些一维密度函数的乘积的假设通常不成立

3

3

中山大學

Rosenblatt和Parzen提出了非参数估计方法，即核密度估计方法．由于核密度估计方法不利用有关数据分布的先验知识，对数据分布不附加任何假定，是一种从数据样本本身出发研究数据分布特征的方法，因而，在统计学理论和应用领域均受到高度的重视。

4

4

4.1 引　言

中山大學

- 非参数方法
  - 能处理任意的概率分布
    - 而不必假设密度的参数形式已知
  - 基本方法
    - 从训练样本中估计概率密度函数$p(x|w_j)$
    - 直接估计后验密度概率$p(w_j|x)$
    - 直接进行判别函数的设计

5

5

中山大學

贝努利大数定理:

设n_A是n次独立重复试验中事件A发生的次数. p 是事件A在每次试验中发生的概率, 则对于任意正数$\varepsilon > 0$, 有

$$\lim_{n\leftarrow\infty} P\left\{\left|\frac{n_A}{n} - p\right| < \varepsilon\right\} = 1$$

或

$$\lim_{n\leftarrow\infty} P\left\{\left|\frac{n_A}{n} - p\right| \geq \varepsilon\right\} = 0$$

6

6

4.2 概率密度的估计

一个向量x落在区域R中的概率为

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

$\mathbf{x}_1, \cdots \mathbf{x}_n$ i.i.d.， 如果有k个样本落在R中，则其概率服从二项式定理：

$$P_k = \binom{n}{k} P^k (1-P)^{n-k},$$

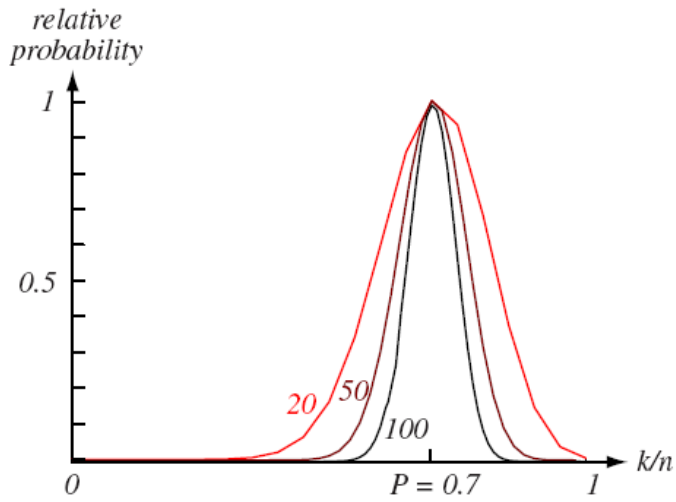
$\dot{P} \simeq k/n$ (伯努利大数定理)

假设区域R足够小时，$\int_R p(\mathbf{x}')d\mathbf{x}' \approx p(\mathbf{x})V$

$$p(\mathbf{x}) \approx \frac{k/n}{V}$$

7

7



8

8

中山大學

- **使用这种方法的问题**
  - **如果体积V固定，且能够获得越来越多的训练样本，那么所获得的也只是*p*(x)的空间平滑后的版本：**

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

  - **若样本个数n固定，且令V趋近于0，则区域R中可能不含任何样本，即*p*(x)趋向于0，估计结果就毫无意义了。**

9

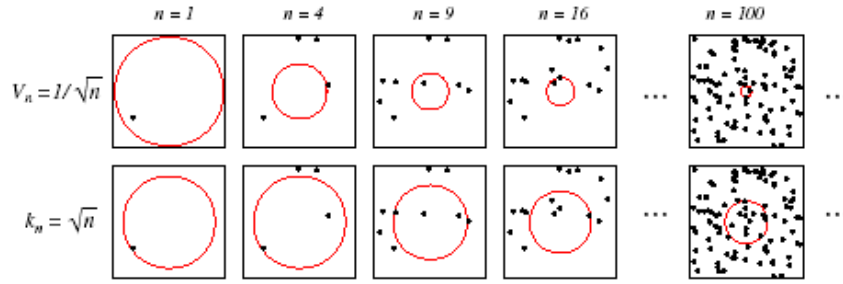
9

中山大學

- **较好的方法:**
  - 构造一系列包含x的区域：R_1, R_2 ……，V_n为R_n的体积，则$p_n(x)$表示对p(x)的第n次估计：

$$p_n(\mathbf{x}) = \frac{k_n / n}{V_n} \to p(\mathbf{x})$$

收敛的三个条件： $\lim_{n\to\infty} V_n = 0$, $\lim_{n\to\infty} k_n = \infty$, $\lim_{n\to\infty} k_n / n = 0$



10

10

4.3 Parzen窗方法

假设R_n是一个超立方体，所以$V_n = h_n^d$

定义窗函数：$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \le 1/2,\ j = 1,\cdots,d \\ 0 & \text{otherwise} \end{cases}$

则超立方体中的样本个数 $k_n = \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$

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

11

11

保证 $p_n(\mathbf{x})$ 是一个合理的概率密度函数，即，非负且积分为1

要求：$\varphi(\mathbf{x}) \ge 0$，且$\int \varphi(\mathbf{u})d\mathbf{u} = 1$，$V_n = h_n^d$

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

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

$= \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)d\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n}\int \varphi(u)du = \frac{1}{n}\cdot n = 1$

即分布是归一化的

12

12

中山大學

直接推导

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

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

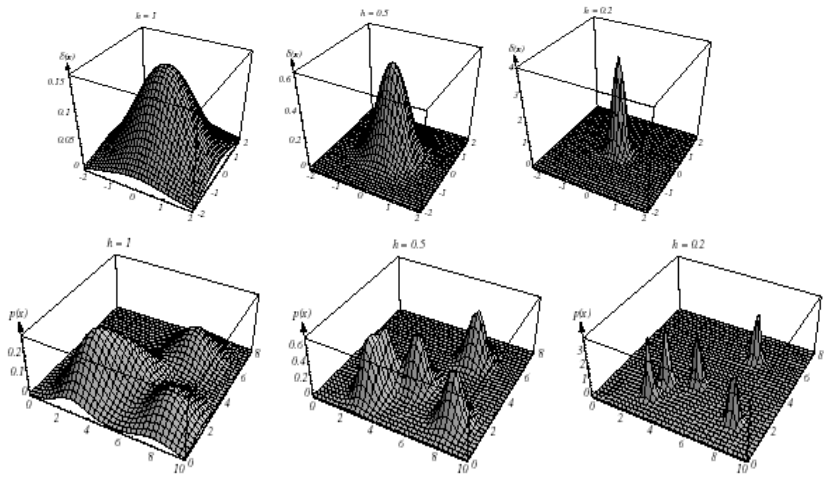
$$=\frac{1}{n}\cdot n=1$$

13

13

中山大學

- **Parzen窗的例子**



14

14

中山大學

■ 收敛性

$p_n(\mathbf{x})$ 依赖于样本 $\mathbf{x}_1,\cdots,\mathbf{x}_n$

$p_n(\mathbf{x})$有某种均值 $\bar{p}_n(\mathbf{x})$和方差$\sigma_n^2(\mathbf{x})$

若 $\lim_{n\to\infty}\bar{p}_n(\mathbf{x})=p(\mathbf{x}),\quad \lim_{n\to\infty}\sigma_n^2(\mathbf{x})=0,$

我们说$p_n(\mathbf{x})$收敛于$p(\mathbf{x})$

收敛的条件如下：

$$\sup_{\mathbf{u}}\varphi(\mathbf{u})<\infty,\quad \lim_{\|\mathbf{u}\|\to\infty}\varphi(\mathbf{u})\prod_{i=1}^{d}u_i=0$$

$$\lim_{n\to\infty}V_n=0,\quad \lim_{n\to\infty}nV_n=\infty$$

15

15

中山大學

4.3.1 均值的收敛性

$$\bar{p}_n(\mathbf{x})=E\left[p_n(\mathbf{x})\right]=\frac{1}{n}\sum_{i=1}^{n}E\left[\frac{1}{V_n}\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right]$$

$$=\int\frac{1}{V_n}\phi\left(\frac{\mathbf{x}-\mathbf{v}}{h_n}\right)p(\mathbf{v})d\mathbf{v}=\int\delta_n(\mathbf{x}-\mathbf{v})p(\mathbf{v})d\mathbf{v}$$

$$\lim_{n\to\infty}\delta_n(\mathbf{x}-\mathbf{v})=\delta(\mathbf{x}-\mathbf{v})$$

$$\therefore\lim_{n\to\infty}\bar{p}_n(\mathbf{x})=p(\mathbf{x})$$

16

16

4.3.2 方差的收敛性

$$\sigma_n^2(\mathbf{x}) = \sum_{i=1}^{n} E\left[\left(\frac{1}{nV_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) - \frac{1}{n}\bar{p}_n(\mathbf{x})\right)^2\right]$$

$$= nE\left[\frac{1}{n^2V_n^2}\varphi^2\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right] - \frac{1}{n}\bar{p}_n^2(\mathbf{x})$$

$$= \frac{1}{nV_n}\int \frac{1}{V_n}\varphi^2\left(\frac{\mathbf{x}-\mathbf{v}}{h_n}\right)p(\mathbf{v})d\mathbf{v} - \frac{1}{n}\bar{p}_n^2(\mathbf{x})$$

17

17

$$\lim_{n\to\infty}\frac{1}{n}\bar{p}_n^2(\mathbf{x}) = 0$$

$$\sigma_n^2(\mathbf{x}) \le \frac{\sup\left(\varphi(\bullet)\right)\bar{p}_n(\mathbf{x})}{nV_n}$$

因为$\lim_{n\to\infty} nV_n \to \infty$ (例如, $V_n = V_1/\sqrt{n}$ or $V_1/\ln n$)

$$\lim_{n\to\infty}\sigma_n^2(\mathbf{x}) = 0$$

18

18

4.3.3 举例说明

中山大學

- 例1. 一维的高斯分布

$$p(x) \sim N(0,1)$$

窗函数：$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$

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

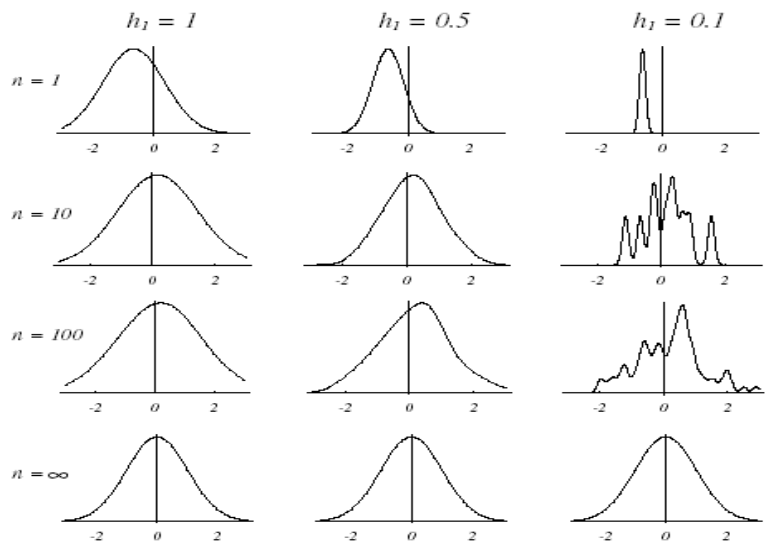
$$h_n = h_1 / \sqrt{n}$$

19

19

中山大學

- 一维的高斯分布

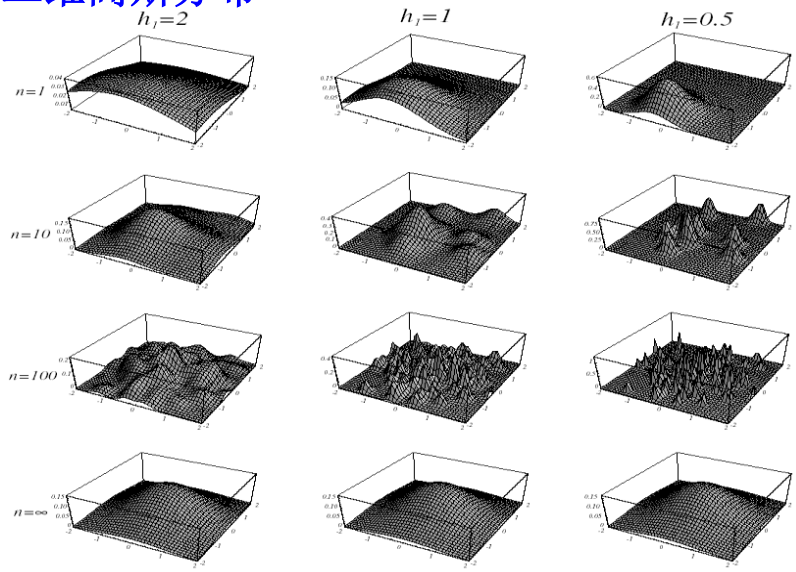


为了得到精确的估计，所需的样本个数将非常多.

20

20

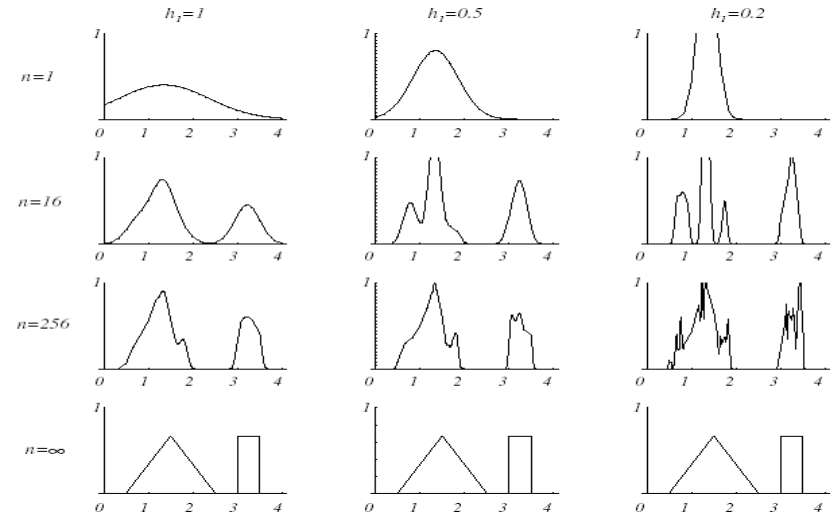
二维高斯分布



21

21

例2. 三角分布与均匀分布的混合

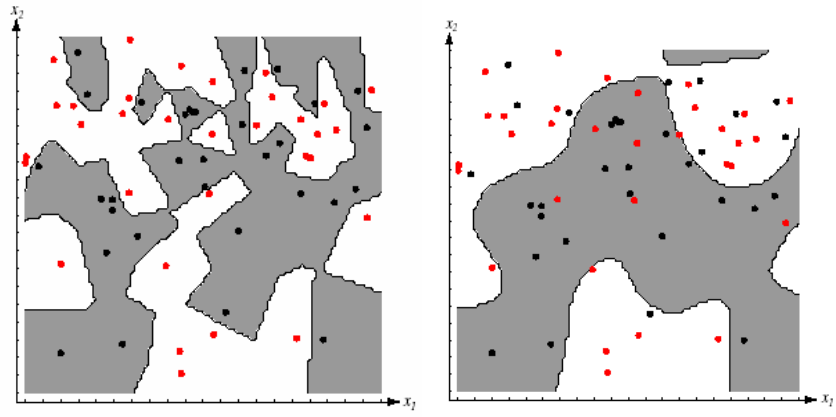


22

22

4.3.4 分类的例子

对每个类都独立估计概率密度，并根据最大后验概率的原则进行分类。



23

23

非参数方法的优点和局限性

- 优点在于通用性
- 实际需要的训练样本的个数大得非常惊人
  - 比知道分布的参数形式下进行估计所需要的样本个数多得多
- 维数灾难 (Curse of dimensionality)
  - 高维函数远比低维函数复杂，几乎无法进行有效的分析和掌握
  - 有效的处理方法是尽可能多的在处理问题时嵌入关于模式数据本身的可靠的先验知识

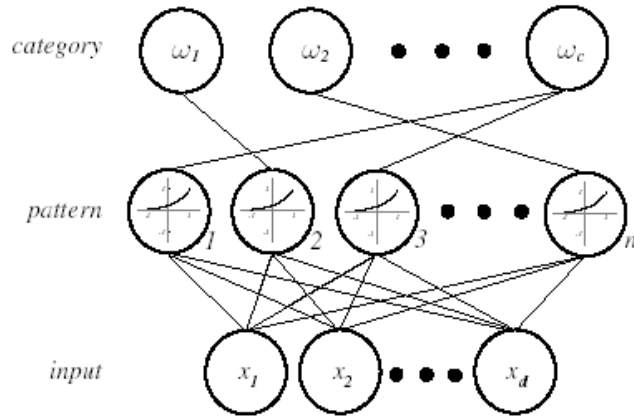
24

24

4.3.5 概率神经网络 (PNN)

中山大學

- 并行处理方式实现模式识别方法。
- Parzen窗方法的神经网络结构：
  假设从c类中随机抽取n个d维样本：



Training the network

中山大學

Algorithm

1. Normalize each pattern x of the training set to 1
2. Place the first training pattern on the input units
3. Set the weights linking the input units and the first pattern units such that: $w_1 = x_1$
4. Make a single connection from the first pattern unit to the category unit corresponding to the known class of that pattern
5. Repeat the process for all remaining training patterns by setting the weights such that $w_k = x_k$ $(k = 1, 2, \ldots, n)$

We finally obtain the following network

4.3.5 概率神经网络 (PNN)

中山大學

- **PNN训练**

initialize $j \leftarrow 0, n, a_{ji} \leftarrow 0$ for $j = 1, \cdots, n; i = 1, \cdots, c$

do $j \leftarrow j + 1$

$x_{jk} \leftarrow x_{jk} / \left(\sum_{i=1}^{d} x_{ji}^2 \right)^{1/2}$

$w_{jk} \leftarrow x_{jk}$

if $\mathbf{x}_j \in \omega_i$ then $a_{ji} \leftarrow 1$

until $j = n$

end

27

4.3.5 概率神经网络 (PNN)

中山大學

- 激活函数

$$net_k = \mathbf{w}_k^t \mathbf{x}$$

$$\varphi\left(\frac{\mathbf{x} - \mathbf{w}_k}{h_n} \right) \propto e^{-(\mathbf{x}-\mathbf{w}_k)^t (\mathbf{x}-\mathbf{w}_k)/2\sigma^2}$$

$$= e^{-(\mathbf{x}^t\mathbf{x} + \mathbf{w}_k^t\mathbf{w}_k - 2\mathbf{w}_k^t\mathbf{x})/2\sigma^2} = e^{(net_k - 1)/\sigma^2}$$

其中：$\mathbf{x}^t\mathbf{x}$和$\mathbf{w}_k^t\mathbf{w}_k$已被归一化。

28

Testing the network

Algorithm

1. Normalize the test pattern x and place it at the input units
2. Each pattern unit computes the inner product in order to yield the net activation

$$net_k = w_k^t . x$$

and emit a nonlinear function $f(net_k) = exp\left[\frac{net_k - 1}{\sigma^2}\right]$

3. Each output unit sums the contributions from all pattern units connected to it

$$P_n(x \mid \omega_j) = \sum_{i=1}^{n} \varphi_i \propto P(\omega_j \mid x)$$

4. Classify by selecting the maximum value of $P_n(x \mid \omega_j)$ $(j = 1, \ldots, c)$

29

4.3.5 概率神经网络 (PNN)

- PNN分类算法

initialize $k \leftarrow 0, \mathbf{x} \leftarrow$ test pattern, $g_i \leftarrow 0$

do $k \leftarrow k + 1$

$net_k \leftarrow \mathbf{w}_k^t \mathbf{x}$

if $a_{ki} = 1$ then $g_i \leftarrow g_i + \exp[(net_k - 1)/\sigma^2]$

until $k = n$

return $class \leftarrow \arg\max_i g_i(\mathbf{x})$

end

30

4.3.5 概率神经网络 (PNN)

- PNN学习速度快，存储空间要求高。
- 可以在线（增量）学习。
- 窗口体积序列的选择很重要。

通常选择$V_n = V_1/\sqrt{n}$，如果V_1非常小，则大多数体积内都是空的，估计$p_n(\mathbf{x})$将产生较大的误差；如果V_1非常大，则平滑效应大，概率密度的空间变化被掩盖了。

31

31

4.4 k_n-近邻估计

- 让体积成为训练样本的函数
- 原型样本
  - 训练样本
- 估计 $p(\mathbf{x})$
  - 以$\mathbf{x}$为中心
  - 让体积扩张直到包含进 k_n 个样本

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

32

32

4.4 k_n-近邻估计

- $p_n(\mathbf{x})$ 收敛到 $p(\mathbf{x})$的充分必要条件

$$\lim_{n\to\infty} k_n = \infty, \quad \lim_{n\to\infty} k_n / n = 0$$

- 例

$$k_n = \sqrt{n}$$

$$V_n \approx \frac{1}{\sqrt{n}\,p(x)}, \quad V_n \approx \frac{V_1}{\sqrt{n}} \to 0 \text{ as } n \to \infty$$

33

33

4.4 k_n-近邻估计

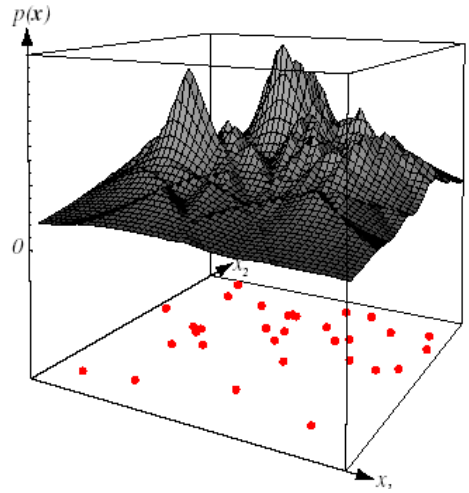


Figure 4.11: The k-nearest-neighbor estimate of a two-dimensional density for $k = 5$. Notice how such a finite n estimate can be quite "jagged," and that discontinuities in the slopes generally occur along lines away from the positions of the points themselves.

34

34

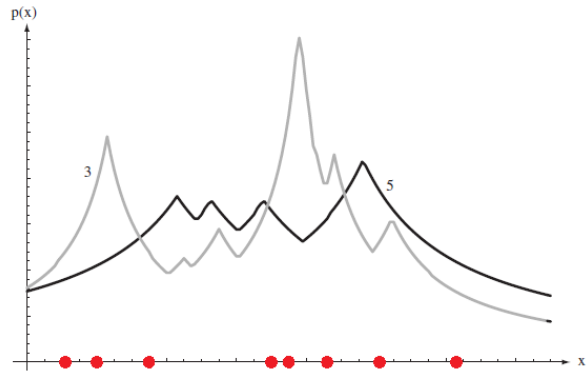
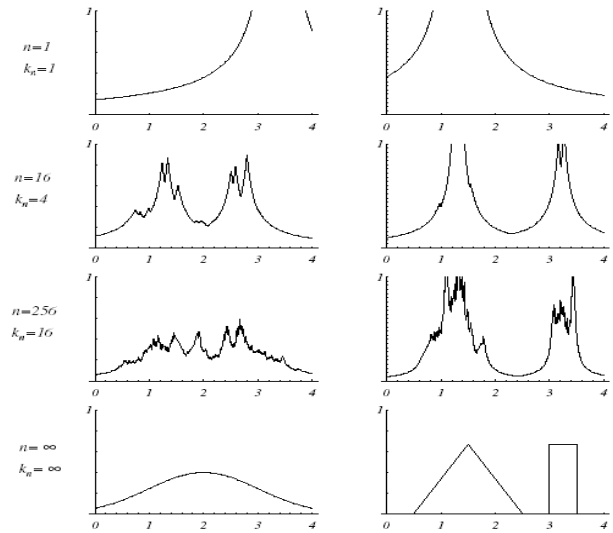


Figure 4.10: Eight points in one dimension and the k-nearest-neighbor density estimates, for $k = 3$ and 5. Note especially that the discontinuities in the slopes in the estimates generally occur *away* fom the positions of the points themselves.

35

35

4.4 k_n-近邻估计



36

36

4.4 k_n-近邻估计

- 后验概率的估计
  - 把一个体积放在**x**周围
  - 包含k 个样本
    - k_i 个属于类别 ω_i
  - 条件概率 $p(\mathbf{x},\omega_i)$的估计

$$p_n(\mathbf{x},\omega_i)=\frac{k_i/n}{V}$$

  - 后验概率 $p(\omega_i|\mathbf{x})$的估计

$$p_n(\omega_i\,|\,\mathbf{x})=\frac{p_n(\mathbf{x},\omega_i)}{\sum_{j=1}^{c}p_n(\mathbf{x},\omega_j)}=\frac{k_i}{k}$$

37

37

4.5 最近邻规则

- $D^n=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$表示一个由n个打上标记的类别组成的集合
- 对于测试样本点x,在集合D^n中距离它最近的点记为x'
- “最近邻规则”的分类方法就是把点x 分为 x'的类别。
- “最近邻规则”法是次优的方法。
  - 引致的误差率比贝叶斯误差率要大
  - 但不会超过贝叶斯误差率的两倍

38

38

4.5 最近邻规则

中山大學

- 启发式理解
  - 赋予最近邻点的标记 θ' 是一个随机变量
  - $P(\theta'=\omega_i|x')=P(\omega_i|x')$
  - 当样本的数目非常大的时候，有理由认为 x' 与 x 足够接近，使得 $P(\omega_i|x')$ 约等于 $P(\omega_i|x)$
  - $\omega_m(x)$ 定义为

$$P(\omega_m \mid \mathbf{x}) = \max_i P(\omega_i \mid \mathbf{x})$$

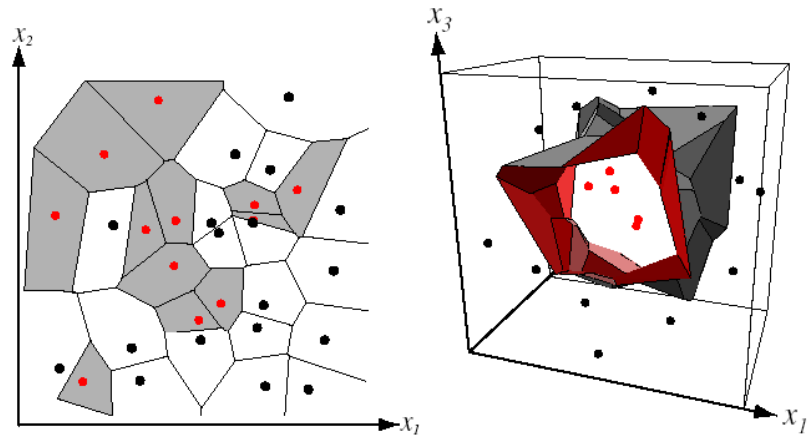
39

39

4.5 最近邻规则

中山大學

- Voronoi 网格



40

40

误差概率

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

$P(e\mid \mathbf{x})$ 的最小可能值记为 $P^*(e\mid \mathbf{x})$ 。

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

$P(e)$ 的最小可能值记为 P^* 。

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

$$P(e\mid \mathbf{x})=\int P(e\mid \mathbf{x},\mathbf{x}')\,p(\mathbf{x}'\mid \mathbf{x})d\mathbf{x}' \qquad (40)$$

可以证明: $P^*\le P\le P^*(2-\frac{c}{c-1}P^*)$

41

41

最近邻规则的收敛性

■S为以x为中心的超球体，任何样本落在S中的概率为：

$$P_s=\int_{\mathbf{x}'\in S} p(\mathbf{x}')d\mathbf{x}'\neq 0$$

■所有n个样本都落在S之外的概率：

$$(1-P_s)^n\to 0 \text{ as } n\to\infty$$

$$\therefore \mathbf{x}'\to\mathbf{x}$$

$$p(\mathbf{x}'\mid \mathbf{x})\to\delta(\mathbf{x}'-\mathbf{x})$$

当n无穷大时，$p(\mathbf{x}'\mid \mathbf{x})$ 趋近于以x为中心的狄拉克函数

42

42

最近邻规则的误差率

测试样本x的标签 θ 待定。

$\mathbf{x}_n^{'}$ 是n个训练样本中距x最近的向量，其类别为 $\theta_n^{'}$

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

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

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

注意，当 $\theta\neq\theta_n^{'}$ 为分类错误。

43

43

$$P_n(e\,|\,\mathbf{x}) = \int P_n(e\,|\,\mathbf{x},\mathbf{x}_n^{'})\,p(\mathbf{x}_n^{'}\,|\,\mathbf{x})d\mathbf{x}_n^{'}$$

$$\lim_{n\to\infty} P_n(e\,|\,\mathbf{x})$$

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

$$=1-\sum_{i=1}^{c} P^2(\omega_i\,|\,\mathbf{x})$$

$$P=\lim_{n\to\infty} P_n(e)=\lim_{n\to\infty}\int P_n(e\,|\,\mathbf{x})\,p(\mathbf{x})d\mathbf{x}$$

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

44

44

误差界

假定贝叶斯误差率 $P^* = \int P^*(e|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ 比较低。

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

$$1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x}) \approx 1 - P^2(\omega_m|\mathbf{x}) \approx 2\left(1 - P(\omega_m|\mathbf{x})\right)$$

$$= 2P^*(e|\mathbf{x})$$

$$P \approx \int 2P^*(e|\mathbf{x})p(\mathbf{x})d\mathbf{x} = 2P^*$$

45

45

更精确的误差上界计算

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

$$P(\omega_i|\mathbf{x}) \geq 0, \quad \sum_{i \neq m} P(\omega_i|\mathbf{x}) = 1 - P(\omega_m|\mathbf{x})$$

$\sum_{i \neq m} P^2(\omega_i|\mathbf{x})$ 达到最小值当所有 $P(\omega_i|\mathbf{x})$（$i \neq m$）都相等。

46

46

中山大學

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

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

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

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

47

中山大學

$$\mathrm{Var}\left[P^*(e \mid \mathbf{x})\right] = \int \left[P^*(e \mid \mathbf{x}) - P^*\right]^2 p(\mathbf{x}) d\mathbf{x}$$

$$= \int P^{*2}(e \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} - P^{*2} \geq 0$$

$$\int P^{*2}(e \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \geq P^{*2}$$

$$P = \int \left[2P^*(e \mid \mathbf{x}) - \frac{c}{c-1} P^{*2}(e \mid \mathbf{x})\right] p(\mathbf{x}) d\mathbf{x}$$

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

48

中山大學

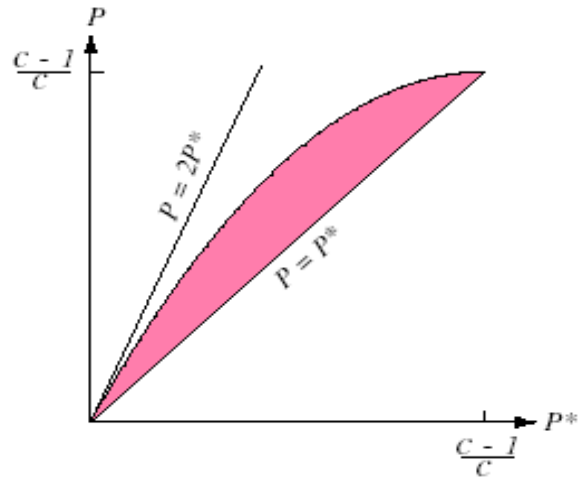
- 误差上界在“零信息”情况下取到。
  - $p(x|w_i)$ 都相等
  - $P(w_i|x) = P(w_i)$
  - $P^*(e|x)$ 与 x互相独立
- P^* 在 0 到$(c-1)/c$之间。

49

49

中山大學

最近邻规则的误差率P的边界



50

50

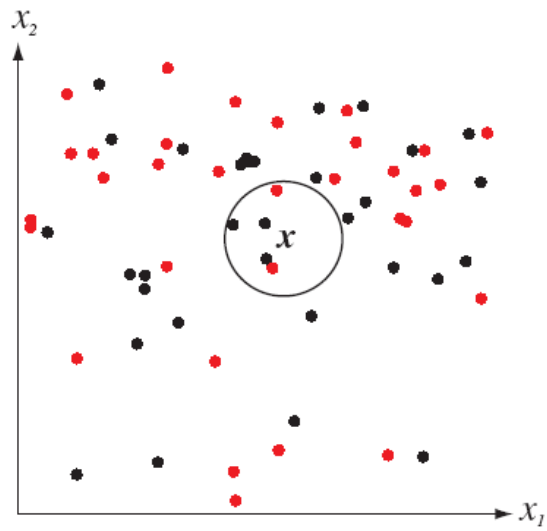
收敛速度

- 收敛速度可能会任意的慢
- $P_n(e)$ 未必会随着n的增加而单调递减

51

51

4.5.4 k-近邻规则



52

52

4.5.4 k-近邻规则

- 简化分析
  - 考虑一个两类问题，取k为奇数
  - k个近邻的标记都是随机变量， $P(\omega_i|\mathbf{x})$，i=1，2都是相互独立的
  - 当k个最近邻中的大多数的标记为ω_m，才判决为类别为ω_m，做出这样的选择的概率为

$$\sum_{i=(k+1)/2}^{k}\binom{k}{i}P(\omega_m\mid\mathbf{x})^i\left[1-P(\omega_m\mid\mathbf{x})\right]^{k-i}$$

当k个越大,选择类别w_m概率越大。

53

53

4.5.4 k-近邻规则

- 简化分析
  - 大样本个数时的k-近邻规则的二类误差率的上界为函数$C_k(P^*)$，其中$C_k(P^*)$为大于下式的最小的凹函数

$$\sum_{i=0}^{(k-1)/2}\binom{k}{i}\left[\left(P^*\right)^{i+1}(1-P^*)^{k-i}+\left(P^*\right)^{k-i}(1-P^*)^{i+1}\right]$$

习题18，证明。

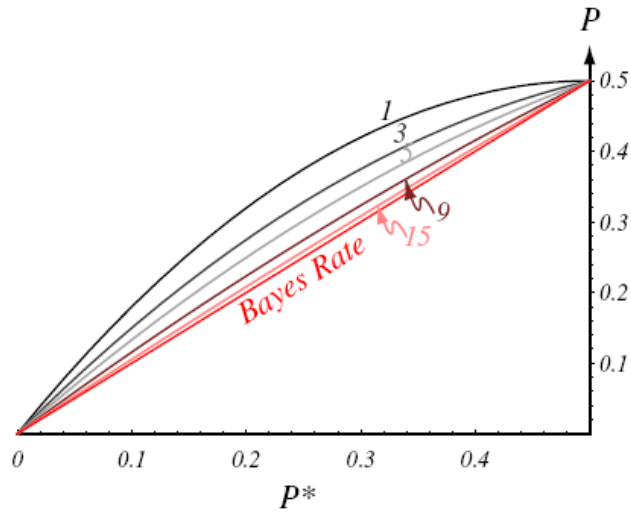
54

54

4.5.4 k-近邻规则

中山大學

- 对于一个两类问题的误差率上界$C_k(P^*)$，如下图



4.5.4 k-近邻规则

中山大學

- **k-近邻规则的进一步讨论**
  - **k-近邻规则可以被看作是另一种从样本中估计后验概率$P(\omega_i|x)$的方法**
    - **为了得到可靠的估计，k越多越好**
  - **另外，希望x的k个近邻x'距离x越近越好，因为这样能保证$P(\omega_i|x')$尽可能逼近$P(\omega_i|x)$**
  - **只有当n趋近于无穷大时，我们才能保证k-近邻规则几乎是最优的分类规则**

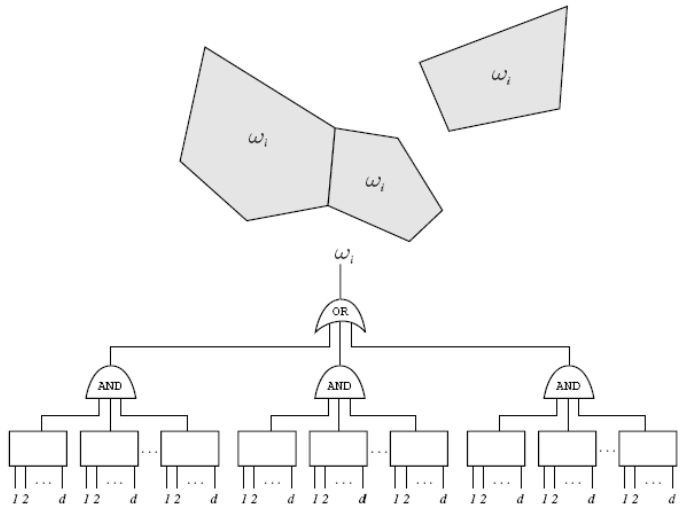
4.5.5 k-近邻规则的计算复杂度

- 搜索每一个训练样本点的复杂度为$O(n^2)$
- 每一个距离（欧氏距离）的计算复杂度为$O(d)$
- 找出距离最近的那一个，总的计算复杂度为$O(dn^2)$

57

57

一个并行的最近邻算法的硬件电路实现



58

58

4.5.5 k-近邻规则的计算复杂度

- 降低最近邻规则搜索的复杂度的方法
  - 1、计算部分距离

$$D_r(\mathbf{a},\mathbf{b}) = \left(\sum_{k=1}^{r} (a_k - b_k)^2 \right)^{1/2}, \quad r < d$$

  - 2、预建立结构方法
    - 建立某种形式的搜索树，在这个搜索树上，各个原型样本点都被有选择的相互连接
    - 但是，不能保证找到的结果是真正的最近邻
  - 3、在训练过程中有选择的消去那些对于问题“无用”的训练样本

59

59

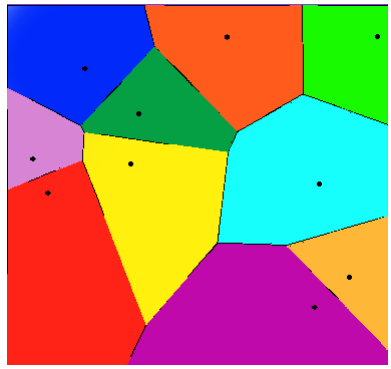
最近邻剪辑算法

1. Initialize $j \leftarrow 0$, $D \leftarrow$ data set, $n \leftarrow$ 原型点个数
2. 构造D的全部Voronoi图
3. do $j \leftarrow j+1$; 对每一个原型点 $\mathbf{x}'_j$
4. 找到 $\mathbf{x}'_j$ 的所有Voronoi近邻
5. if 这些近邻中存在不是和$\mathbf{x}'_j$同一类别的点， then 标记 $\mathbf{x}'_j$
6. until $j=n$
7. 删除所有没有被标记的点
8. 构造剩余点的Voronoi图
9. end

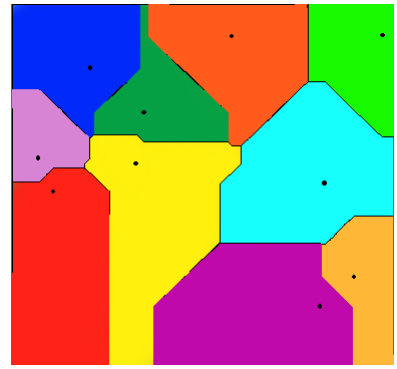
60

60

Voronoi Diagram (Voronoi Tessellation)



The Voronoi diagram of 10 point sites in the plane under the euclidean distance



The Voronoi cells of 10 point sites in the plane, under the Manhattan distance

61

61

4.5.5 k-近邻规则的计算复杂度

- 这一剪辑算法的计算复杂度为 $O(d^3 n^{\lfloor d/2 \rfloor} \ln n)$
- 这个算法不能保证找到最少需要的原型样本点集
- 在不影响精度的前提下，这个算法能显著降低计算复杂度
- 这个算法可以与部分距离法和预建立结构法结合使用

62

62

4.6 度量距离和最邻近分类

- **4.6.1** 度量的性质

一个度量，必须满足4个性质，对于任意的向量a, b, c, 有：

非负性： $D(a,a) \geq 0$

自反性： $D(a,b) = 0$ 当且仅当 $a = b$

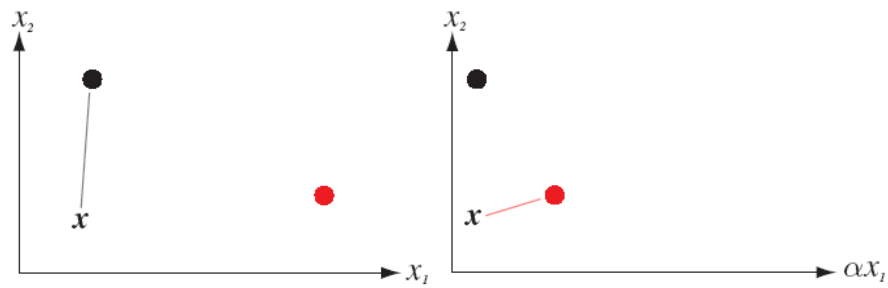
对称性： $D(a,b) = D(b,a)$

三角不等式： $D(a,b) + D(b,c) \geq D(a,c)$

63

63

尺度变换对欧几里德距离度量的影响



64

64

更加广义的d维空间中的度量为Minkowski距离度量

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

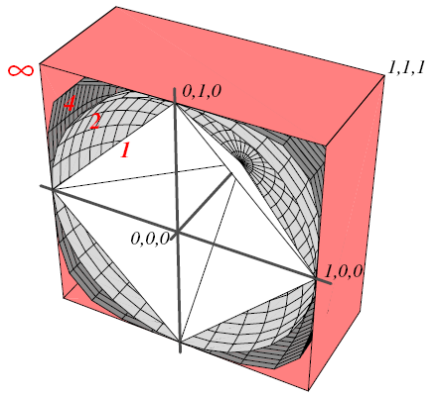


图4-19

每一个彩色的平面距离原点为1.0（使用不同的k值的Minkowski距离）的点所形成。这样，白色的表面对应于L_1范数（Manhattan距离）。浅灰色的球体对应于L_2范数（欧几里德距离），暗灰色表面对应于L_4范数，而粉红色的立方体对应于L_∞范数

65

65

描述两个集合之间的Tanimoto度量距离在分类学（taxonomy）中得到广泛的应用。其定义为：

$$D_{Tanimoto}(S_1, S_2) = \frac{(n_1 - n_{12}) + (n_2 - n_{12})}{n_1 + n_2 - n_{12}}$$

其中n_1和 n_2分别是s_1 和s_2的元素个数，而n_{12}是这两个集合的交集中的元素个数。

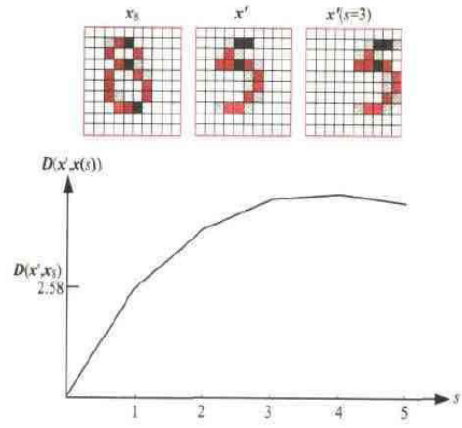
66

66

4.6.2 切空间距离

- 在最邻近规则中，如果不加考虑的任意选择距离度量，会有很多问题。

图 4 20 因为忽略平移不变性问题而不加分辨地使用欧几里德距离有时候会带来严重的误差。上图中的模式 x' 代表一个手写体字符"5"，而 $x'(s=3)$ 代表同一个形状，但是经过了向右的 3 个像素的平移。这样，欧几里德距离度量的结果 $D(x', x'(s=3))$ 要比 $D(x', x_8)$ 大得多，其中的 x_8 表示一个手写体字符"8"。这样，使用欧几里德距离度量的最近邻规则分类器就会导致很大的分类误差。所以，为了解决这个问题，我们必须寻找一个对一些已知的变换（比如平移，旋转，尺度变换等）不敏感的距离度量



67

67

中山大學

在理想情况下，除非我们已经把两个模式变换地尽可能相似，否则不会过早地计算这两个模式之间的距离。

而这样的预变换的计算复杂度通常是非常大的。而在通常情况下，我们甚至不知道应该需要旋转多少角度，因此必须进行不同角度的尝试，而每一次尝试都需要进行一次距离的计算，来检验这时候是否达到了最佳的效果。

如果在分类时，对每一个训练样本都进行这样尝试的话，这样做的计算复杂度几乎是无法忍受的。

68

68

中山大學

- 切空间距离分类器使用一个全新的距离的度量和一个可以近似任意变换的线性逼近。
- 假设已经知道所需处理的问题会设计r种变换，比如水平平移，垂直平移，剪切，旋转，尺度变换，线条的细化等，在设计分类器时，我们对每一个原型样本点**x**',都进行每一种变换操作$F_i(\mathbf{x}';\alpha_i)$，这样$F_i(\mathbf{x}';\alpha_i)$就能够代表图像**x**',经过角度$\alpha_i$的旋转得到的新的图像。然后，对于每一种操作，我们都构造一个切向量 TV_i ：

$$\mathbf{TV}_i = F_i(\mathbf{x}';\alpha_i) - \mathbf{x}'$$

69

69

中山大學

切向量的线性组合（图4-21）

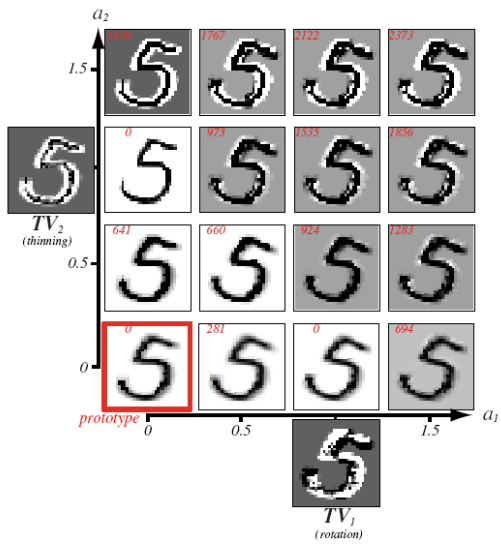


Figure 4.21: The pixel image of the handwritten 5 prototype at the lower left was subjected to two transformations, rotation, and line thinning, to obtain the tangent vectors **TV**1 and **TV**2

70

70

中山大學

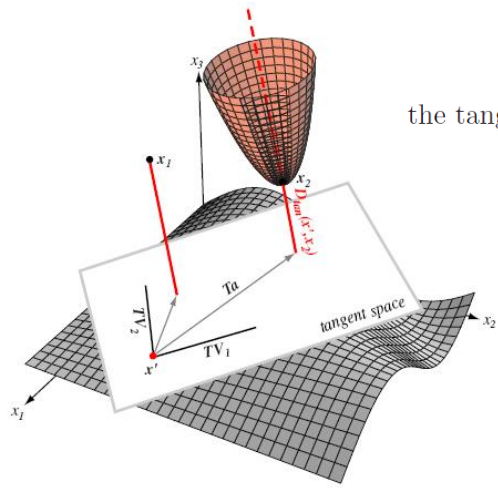
切空间距离：
对每一个原型样本点，构造r×d的矩阵T，矩阵T由x’,
处的切向量组成。
如果矩阵T由x’ 处的r个切向量组成，那么测试点x’, 到原型样本点x的距离为：

$$D_{tan}(\mathbf{x}',\mathbf{x}) = \min_{\mathbf{a}} \left[\left\| (\mathbf{x}'+\mathbf{T}\mathbf{a}) - \mathbf{x} \right\| \right]$$

71

71

中山大學



the tangent distance $D_{tan}(\mathbf{x}', \mathbf{x}_2)$.

Thus although the Euclidean distance from $\mathbf{x}'$ to $\mathbf{x}_1$ is less than to $\mathbf{x}_2$

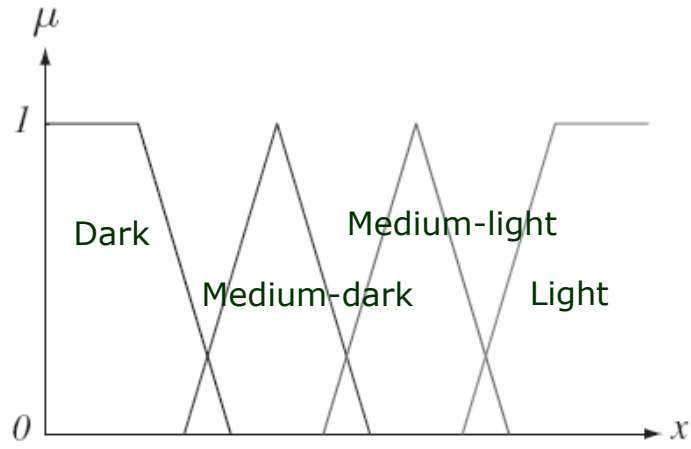
Euclidean distance from $\mathbf{x}_2$ to the tangent space of $\mathbf{x}'$ is a quadratic function of the parameter vector $\mathbf{a}$,

72

72

4.7 模糊分类

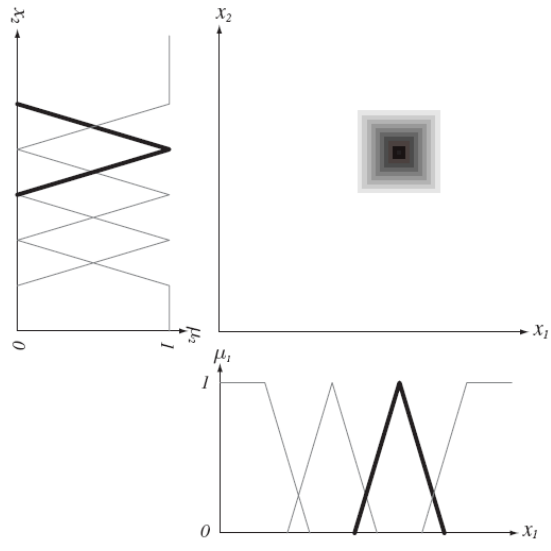
- 模糊类别隶属度函数



73

73

合取规则和分类函数



74

74

4.7 模糊分类

- 类别隶属度函数的Cox-Jaynes公理

$$P(a\,|\,d) > P(b\,|\,d) \text{ and } P(b\,|\,d) > P(c\,|\,d) \;\Rightarrow\; P(a\,|\,d) > P(c\,|\,d)$$

$$P(\text{not } a\,|\,d) = F_1\left[P(a\,|\,d)\right]$$

$$P(a,b\,|\,d) = F_2\left[P(a\,|\,d), P(b\,|\,d)\right]$$

75

75

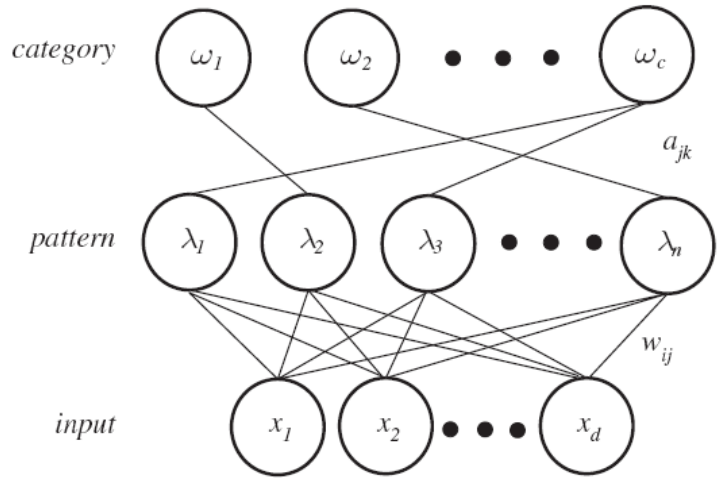
贡献和局限性

- 贡献：指引人们如何把一种语言形式的知识转化为确定的分类函数
- 局限：纯粹模糊技术不依赖训练样本

76

76

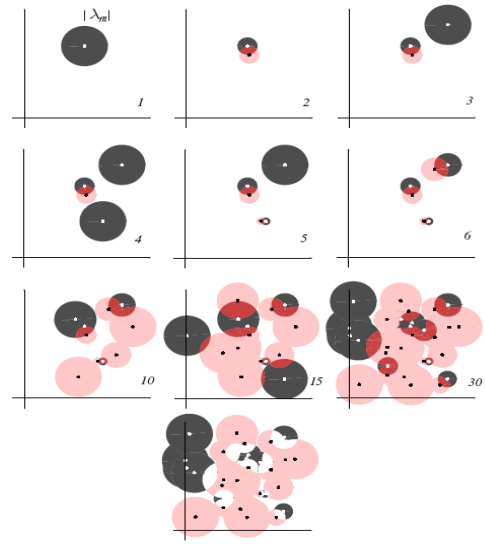
4.8 RCE网络（衰减库仑势法）



77

77

RCE训练



78

78

REC网络的训练算法

initialize $j \leftarrow 0, \varepsilon \leftarrow$ 小模式, $\lambda_m \leftarrow$ 最大半径

do $j \leftarrow j+1$, x $\leftarrow \mathbf{x}_j^{'}, \omega_k$ 是 $\mathbf{x}_j^{'}$的类别

$w_{ij} \leftarrow x_i$（训练权重）

$\hat{\mathbf{x}} \leftarrow \arg\min_{\mathbf{x}' \notin \omega_k} D(\mathbf{x}, \mathbf{x}')$（找到不属于$\omega_k$的最近邻点）

$\lambda_j \leftarrow \max\left[\min\left[D(\hat{\mathbf{x}}, \mathbf{x}), \lambda_m\right], \varepsilon\right]$

$a_{jk} \leftarrow 1$

until $j = n$

end

79

79

RCE网络分类算法

initialize $j \leftarrow 0, \mathbf{x} \leftarrow$ 测试模式, $D_t \leftarrow \{\ \}$

do $j \leftarrow j+1$

if $D(\mathbf{x}, \mathbf{x}_j^{'}) < \lambda_j$ then $D_t \leftarrow D_t \cup \mathbf{x}_j^{'}$

until $j = n$

if 所有 $\mathbf{x}_j^{'} \in D_t$ 的标记相同,

then return 所有 $\mathbf{x}_k \in D_t$的标记

else return "模糊"标记

end

80

80

4.9 级数展开逼近

$$\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\approx\sum_{j=1}^{m}a_j\psi_j(\mathbf{x})\chi_j(\mathbf{x}_i)$$

$$\sum_{i=1}^{n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)=\sum_{j=1}^{m}a_j\psi_j(\mathbf{x})\sum_{i=1}^{n}\chi_j(\mathbf{x}_i)$$

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

$$b_j=\frac{a_j}{nV_n}\sum_{i=1}^{n}\chi_j(\mathbf{x}_i)$$

81

81

一维例子

$$\sqrt{\pi}\varphi(u)=e^{-u^2}\approx\sum_{j=0}^{m}\frac{(-1)^j u^{2j}}{j!}$$

$$\sqrt{\pi}\varphi\left(\frac{x-x_i}{h}\right)\approx 1-\left(\frac{x-x_i}{h}\right)^2=1+\frac{2}{h^2}xx_i-\frac{1}{h^2}x^2-\frac{1}{h^2}x_i^2$$

$$\sqrt{\pi}\,p_n(x)\approx b_0+b_1x+b_2x^2$$

$$b_0=\frac{1}{h}-\frac{1}{h^3}\frac{1}{n}\sum_{i=1}^{n}x_i^2,\ b_1=\frac{2}{h^3}\frac{1}{n}\sum_{i=1}^{n}x_i,\ b_2=-\frac{1}{h^3}$$

$|x-x_i|\le h$ is required

82

82

本章小结

中山大學

1. 概率密度的估计
2. Parzen窗方法
3. Kn-近邻估计
4. 最近邻规则
5. 距离度量和最近邻分类

本章小结

中山大學

- 非参数估计方法有两种基本途径：
  - 第一种途径，概率密度函数被估计，并且被用于后面的分类中。如Parzen窗方法，及其硬件实现方式PNN；
  - 第二种途径，不估计概率密度函数，直接根据样本进行分类，如k-近邻方法和几种松弛网络
- 松弛方法（如势函数）建立包围在原型样本点周围的“吸引盆”。RCE算法就是一种，调整吸引盆，以包含进周围尽可能多的同一类别的训练样本点。