Parzen窗估计

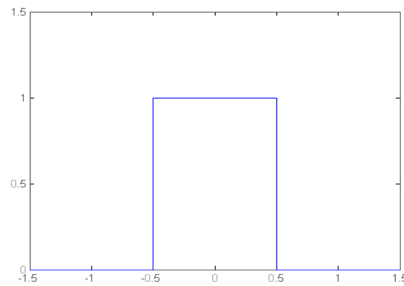
- **定义窗函数：假设R_n是一个d维的超立方体。令h_n为超立方体一条边的长度，则体积：**

$$V_n = h_n^d$$

立方体窗函数为:

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

中心在原点的
单位超立方体



1

Parzen窗估计

落入以X为中心的立方体区域的样本数为:

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

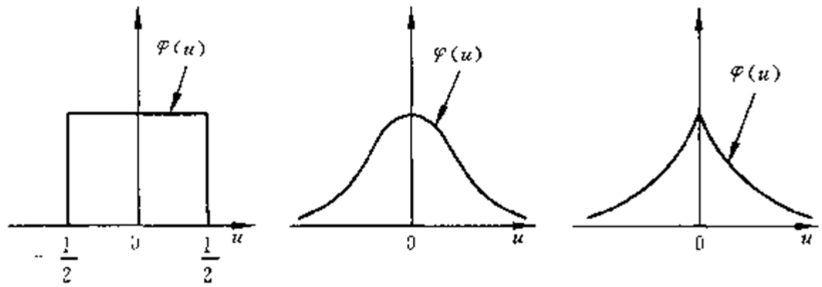
X处的密度估计为：

$$\hat{p}_n(\mathbf{x}) = \frac{k_n / n}{Vn} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

可以验证: $\hat{p}_n(\mathbf{x}) \ge 0$ $\int \hat{p}_n(\mathbf{x}) d\mathbf{x} = 1$

2

窗函数的形式



方窗函数

$$\varphi(u)=\begin{cases}1, |u|\le \frac{1}{2}\\ 0.\text{其他}\end{cases}$$

正态窗函数

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

指数窗函数

$$\varphi(u)=\exp\{-|u|\}$$

其中：$u=\frac{x-x_i}{h_n}$

3

K_n近邻估计

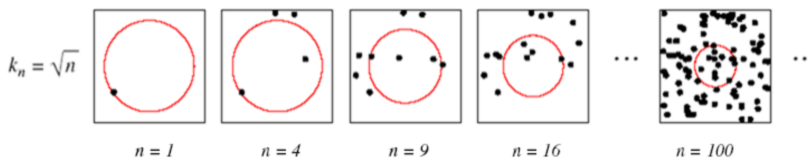
- **在Parzen窗估计中，存在一个问题：对h_n的选择。**
  - 若h_n选太小，则大部分体积将是空的（即不包含样本），从而使$P_n(x)$估计不稳定。
  - 若h_n选太大，则$P_n(x)$估计较平坦，反映不出总体分布的变化
- **K_n近邻法的思想：固定样本数量K_n ，调整区域体积大小V_n，直至有K_n个样本落入区域中**

4

K_n近邻估计

- **K_n近邻密度估计：**

固定样本数为 k_n ，在 $\mathbf{x}$ 附近选取与之最近的 k_n 个样本，计算该 k_n 个样本分布的最小体积 V_n



在X处的概率密度估计值为：$\hat{p}_n(\mathbf{x}) = \dfrac{k_n / n}{V_n}$

5

渐近收敛的条件

$\hat{p}_n(\mathbf{x})$ 渐近收敛的充要条件为：

$$\lim_{n\to\infty} k_n = \infty$$

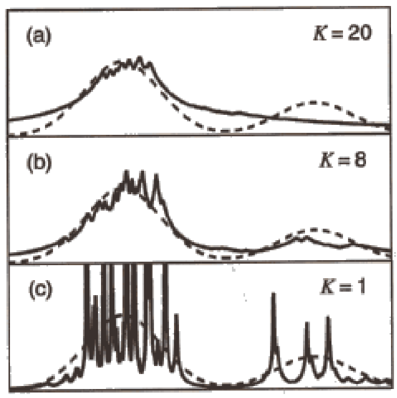
$$\lim_{n\to\infty} k_n / n = 0$$

通常选择：$k_n = \sqrt{n}$

6

K_n近邻估计

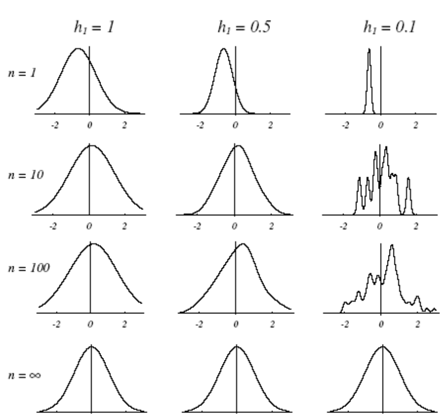
- 例子:



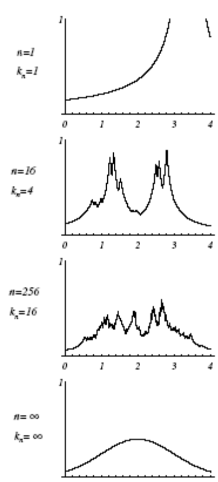
7

- 例子:

Parzen windows



k_n-nearest-neighbor



$k_n = \sqrt{n}$

斜率不连续

当n值为有限值时K_n近邻估计十分粗糙

8

K_n近邻估计

- **K_n近邻后验概率估计：**

给定i.i.d.样本集 $X=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}$，共 c 类。把一个体积V放在x周围，能够包含进k个样本，其中有 k_i 个样本属于第i类。那么联合概率密度的估计为：

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

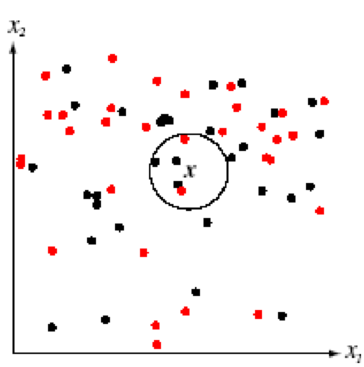
- **后验概率：**

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

9

K_n近邻估计

- 例子



X属于第i类的后验概率就是体积中标记为第i类的样本个数与体积中全部样本点个数的比值。

为了达到最小误差率，选择比值最大的那个类别作为判决结果。

如果样本足够多、体积足够小，这样的方法得到的结果是比较准确的！

10

最近邻分类器(NN)

- 假设i.i.d.样本集 $X=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_l,y_l)\}$
- 对于样本 $\mathbf{x}$ ，NN采用如下的决策：

$$if\quad i=\arg\min_{\mathbf{x}_i\in X} d\left(\mathbf{x}_i,\mathbf{x}\right)$$
$$then\ y=y_i$$

- 相当于采用 $k=1$ 近邻方法估计后验概率，然后采用最大后验概率决策。
- 分类一个样本的计算复杂度：$O\left(ld\right)$（采用欧氏距离）

11

最近邻分类器

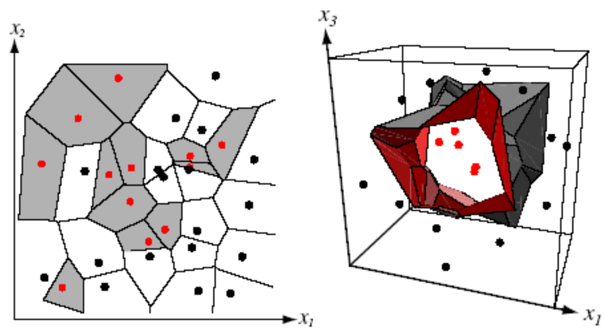
- **样本 *x = (0.10, 0.25)* 的类别?**

Training Examples	Labels	Distance
(0.15, 0.35)	ω_1	0.118
(0.10, 0.28)	ω_2	0.030
(0.09, 0.30)	ω_5	0.051
(0.12, 0.20)	ω_2	0.054

12

最近邻分类器

- 决策边界： Voronoi网格



NN分类规则将特征空间分成许多Voronoi网格

（ Voronoi网格：由一组由连接两邻点直线的垂直平分线组成的连续多边形组成 ）

13

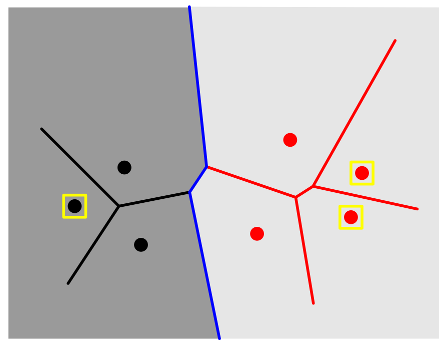
最近邻分类器

- 决策边界
  - 在一个Voronoi网格中，每一个点到该 Voronoi网格原型的距离小于到其它所有训练样本点的距离。
  - NN分类器将该Voronoi网格中的点标识为与该原型同类。

14

最近邻分类器

- **决策边界：**
  - 在NN分类器中，分类边界对于分类新样本是足够的。
  - 但是计算或者存储分类边界是非常困难的
  - 目前已经提出许多算法来存储简化后的样本集，而不是整个样本集，使得分类边界不变。



15

NN分类器的渐近误差界

若 $P_n(error)$ 是***n***个样本时的误差率，并且：

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

P^* 为最小**Bayesian**错误率，***c***为类别数。

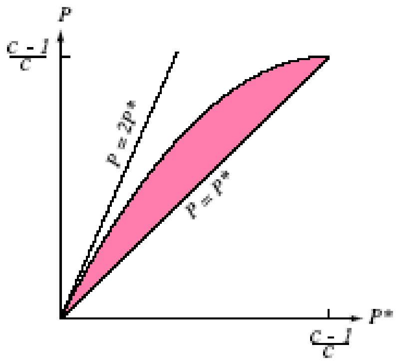
$$P^*(error|\mathbf{x}) = 1 - P(\omega_j|\mathbf{x}),\ \ j = \arg\max_{j=1,\cdots,c} P(\omega_j|\mathbf{x})$$

$$P^* = \int p(error|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

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

16

NN分类器的渐近误差界



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

假设能够得到无限多的训练样本和使用任意复杂的分量规则，我们至多只能使误差率降低一半。

也就是说，分类信息中的一半信息是由最邻近点提供的！

17

最近邻分类器

- 当样本有限的情况下，最近邻分类器的分类效果如何?
  - 不理想！
- 随着样本数量的增加，分类器收敛到渐近值的速度如何?
  - 可能会任意慢，而且误差未必会随着n的增加单调递减！

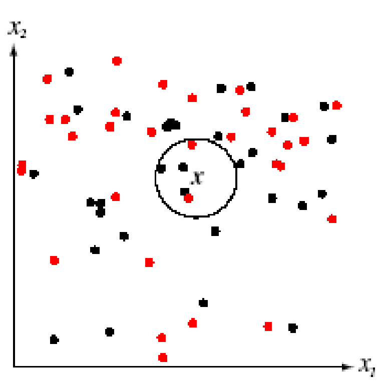
18

k-近邻分类器（k-NN）

- 假设i.i.d.样本集 $X=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_l,y_l)\}$
- 对于样本 $\mathbf{x}$，k-NN采用如下的决策：
- 搜索与 $\mathbf{x}$ 最近的 k 个近邻，如果 k 个近邻中属于 ω_j 类的样本最多，则判决 $\mathbf{x}$ 属于 ω_j
- **原理：**相当于采用 k 近邻方法估计后验概率，然后采用最大后验概率决策。
- 分类一个样本的计算复杂度：$O(kld)$（采用欧氏距离）

19

k-近邻分类器



从测试样本x开始生长，不断扩大区域，直至包含进k个训练样本；

把测试样本x的类别归为与之最近的k个训练样本中出现频率最大的类别。

20

例：

$k = 3$ (odd value)

and $x = (0.10, 0.25)^t$

Prototypes	Labels
(0.15, 0.35)	ω_1
(0.10, 0.28)	ω_2
(0.09, 0.30)	ω_5
(0.12, 0.20)	ω_2

- 选择 k-NN to x

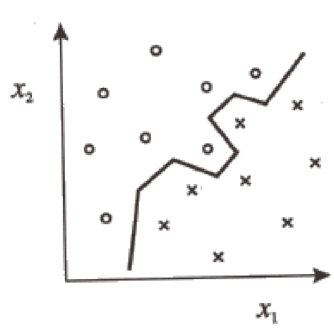
  { $(0.10, 0.28, \omega_2)$;
  $(0.12, 0.20, \omega_2)$;
  $(0.09, 0.30, \omega_5)$ }

- X属于 ω_2。

21

k-近邻分类器

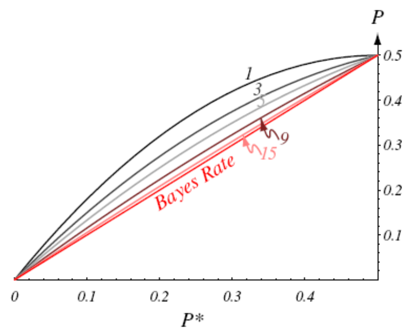
- **决策面：**
  - 分段线性超平面
  - 每一个超平面对应着最近两点的中垂面。



22

k-近邻分类器

- **k-NN分类器的误差率在样本数无穷大时趋向于Bayesian最小错误率！**



23