

2.4.1 基本原理

问题：已知样本集 $X = \{x_1, \dots, x_N\}$ ，其中样本均服从 $p(x)$ 的总体中独立抽取 (IID, Independently Identically Distribution)，求 $\hat{p}(x)$ 。

考虑随机向量 x 落入区域 \mathfrak{R} 的概率 $P = \int_{\mathfrak{R}} p(x) dx$

$$N : \text{样本总数} \quad k : \text{实际落到 } \mathfrak{R} \text{ 中的样本数} \quad \hat{P} = \frac{k}{N}$$

设 $p(x)$ 在 \mathfrak{R} 内连续，当 \mathfrak{R} 逐渐减小的时候，小到使 $p(x)$ 在其上几乎没有变化时：

$$P = \int_{\mathfrak{R}} p(x) dx = p(x) V \quad x \in \mathfrak{R}$$

V ：包含 x 的一个小区域 \mathfrak{R} 的体积，有 $V = \int_{\mathfrak{R}} dx$

$\hat{p}(x)$ ：为对 $p(x)$ 在小区域内的平均值的估计，即小区域内概率密度估计。

$$\hat{p}(x) = \frac{k}{NV}$$

2.4.2 直方图

直方图法（非参数概率密度估计的最简单方法） $\hat{p}(x) = \frac{k}{NV}$

(1) 把 x 的每个分量分成 s^d 个等间隔小窗（若 $x \in E^d$ ，则形成 s^d 个小船）

(2) 统计落入各个小船内的样本数 q_i

(3) 相应小船的概率密度为 $q_i / (NV)$ (N : 样本总数, V : 小船体积)

V 的选择：过大，估计粗糙；过小，可能某些区域中无样本。

设有样本总数为 n ，小船的体积为 V_n ，在 x 附近落入小船的

样本个数为 k_n 。当 n 趋近于无穷大时， $\hat{p}_n(x)$ 收敛于 $p(x)$ 的条件为：①③

$$(1) \lim_{n \rightarrow \infty} V_n = 0; \quad (2) \lim_{n \rightarrow \infty} k_n = \infty; \quad (3) \lim_{n \rightarrow \infty} \frac{k_n}{n} = 0$$

$$\hat{p}_n(x) = \frac{k_n}{nV_n}$$

* $\hat{p}_n(x) = \frac{k_n}{nV_n}$ 选择 V_n ，对 k_n 和 $\frac{k_n}{nV_n}$ 加限制以保证收敛 \Rightarrow Parzen 窗
区域选择策略 | 选择 k_n , V_n 包含 x 的 k_n 个近邻 $\Rightarrow k_n$ 近邻估计

2.4.3 k_n 近邻估计方法（最简单的分段线性分类器）
把各类型分为若干子类，以子类中心作为类别代表点，考查新样本到各代
表点的距离并把它分到最近的代表点所代表的类

极端情况 \Rightarrow 所有样本点均为代表点（最近邻）

样本集 $S_N = \{(x_1, \theta_1), (x_2, \theta_2), \dots, (x_N, \theta_N)\}$, x_i 为样本, θ_i 为类别标号，且 $\theta_i \in \{1, 2, \dots, c\}$

设样本 x_i 与 x_j 之间的距离为 $\delta(x_i, x_j)$ (比如欧氏距离 $\|x_i - x_j\|$)，对于未知样本 x ，设 S_N 中与之距离最近的样本为 x' (类别为 θ')， $\delta(x, x') = \min_{j=1, \dots, N} \delta(x, x_j)$

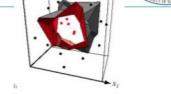
此时，将 x 分到 θ' 类，即 $\hat{\theta}(x) = \theta'$ 。
—— 最近邻决策

* 错误率渐近分析 $P^* \leq P_j \leq P^*(2 - \frac{c}{c-1} P^*) \leq 2P^*$
推导见 PPT 贝叶斯错误率 样本无穷多时最近邻错误率 两倍内

1. 最近邻

Voronoi 图：又名泰森多边形，是由一组连接两个
近点直线的垂直平分线组成的连续多边形组成。

Voronoi 网格：最近邻规则基于训练样本把特征空
间分成一个网格单元结构，称为 Voronoi 网格。



- V_i 是一个多边形，任何落在这个多边形的点
距离点都比其它已知样本点的距离更近。
- 每个网格包含一个训练样本点，该网格中
任何一个位置距离该训练样本点都距离
其它网格的训练样本点更远。
- 如果测试样本 x 落入该网格，则判别为该
网格样本点 S_i 所属的类别。

2. k -近邻（最近邻推广）

k -近邻法：找出 x 的 k 个近邻，看其中多数属于哪一类，则将 x 分到哪一类。

N 个样本，包含 c 个类别 ω_i ($i=1, \dots, c$)， k_i 为 x 的 k 个近邻中属于 ω_i 的样本数。

判别函数： $g_j(x) = k_j$ ($i=1, \dots, c$)

决策规则：If $g_j(x) = \max_{i=1, \dots, c} k_i$, then $x \in \omega_j$

x 的分类，是通过统计最近的 k 个样本的属性，用投票法将最常见的类别标记 x 。

渐近平均错误率的界：
 N 无穷大时， k 越大， P_k^* 的上限越低（越靠近下限）。但 k 应始终是 N 中的一小部分，保证 k 个近邻均充分接近 x 。否则这一关系不成立。

一般来说，总有

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

或者简化为

$$P^* \leq P_k \leq 2P^*$$

错误率分析 | $k > 1$ 时， k 近邻错误率低于最近邻
分析 | $k \rightarrow \infty$ 时， k 近邻错误率等于 Bayes 错误率

3. 快速最近邻

存储所有样本 与 将样本做比较
节省内存 / 计算量 \Rightarrow 减少计算 / 节省内存
最近邻有问题 | 噪声影响大 / 样本数接近时风险大 \Rightarrow 拒绝
有限样本性能不佳 \Rightarrow 各种实用加权投票法决策

加速方法

1. "部分距离" 计算

$$D(a, b) = \left(\sum_{k=1}^d (a_k + b_k)^2 \right)^{\frac{1}{2}} \quad r \leq d \Rightarrow D_r(a, b) = \left(\sum_{k=1}^r (a_k + b_k)^2 \right)^{\frac{1}{2}}$$

一旦其部分的距离大于目前最接近的样本的全欧式距离时终止计算

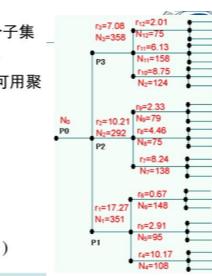
2. "预建立结构" 算法

改进的思路：

- 对样本集进行组织与整理，分群分层，尽可能将计算压缩到接近测试样本邻域的小范围内，避免盲目地与训练样本集中每个样本进行距离计算。
- 在原有样本集中挑选出对分类计算有效的样本，使样本总数合理地减少，以同时达到既减少计算量，又减少存储量的双重效果。

3. 分级分解构建搜索树

- 将整个样本集分成 l 个子集，每个子集又分为它的 l 个子集，如此进行若干次就能建立起一个样本集的树形结构。
- 分成子集的原则是该子集内的样本尽可能聚成堆，这可用聚类方法实现。
- 计算并存储 X_p 的 M_p , r_p 及 $D(x_i, M_p)$



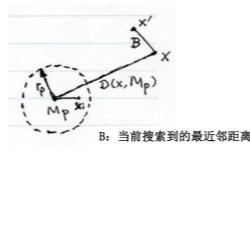
b. 搜索(分支定界算法)

搜索规则：

- 对新样本 x ，结点 X_p ，若 $D(x, M_p) > B + r_p$ 则 x 的近邻不可能在 X_p 中
- 对新样本 x ，结点 X_p 中的样本 $x_i \in X_p$ 若 $D(x, M_p) > B + D(x_i, M_p)$ ，则 x_i 不是 x 的最近邻。

其中 $r_p, D(x, M_p)$ 在训练（建树）过程中可以先计算保存，搜索过程只需计算 $D(x, M_p)$ 或更新 B 。

- 这种方法着眼于只解决减少计算量，但没有达到减少存储量的要求。
- 如果结构合理，可以降低计算时间。



4. 剪辑近邻法

基本理解：

处在两类交界处或分布重合区的样本可能误导近邻法决策。应将其从样本集中去掉。

- 考查样本是否为可能的误导样本
- 考查方法是通过试分类，认为错分样本为误导样本
- 若是则从样本集中去掉——剪辑

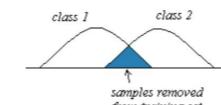
获得更准确的错误率

处在两类交界处或分布重合区的样本可能误导近邻法决策。应将其从样本集中去掉。

· 考查样本是否为可能的误导样本

· 考查方法是通过试分类，认为错分样本为误导样本

· 若是则从样本集中去掉——剪辑



基本做法：

将样本集分为考试集 X^{NT} 和参考集 X^{NR} : $X^N = X^{NT} \cup X^{NR}$, $X^{NT} \cap X^{NR} = \emptyset$

剪辑：用 X^{NR} 中的样本对 X^{NT} 中的样本进行近邻法分类剪掉 X^{NT} 中被错分的样本;

X^{NT} 剩余样本构成剪辑样本集 X^{NTE} 。

分类：利用 X^{NTE} 和近邻法对未知样本 x 分类。

· 训练样本和测试样本没有独立性，会产生一个偏于乐观的估计。

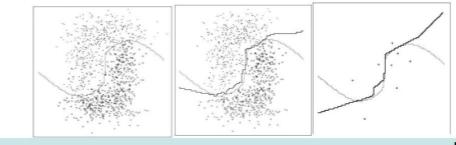
5. 压缩近邻法

主要用以减少计算量

将 X^N 分为 X_s 和 X_c ，开始时 X^N 中只有一个样本， X_c 中为其余样本。考查 X_c 中每个样本，若用 x 可正确分类则保留，否则移入 X_s ，……最后用 X_s 作分类的样本集。

可与剪辑法配合使用。

例：



6. 可拒绝决策近邻法

由于近邻法决策实际只取决于个别样本，因此有时风险较大，尤其是最近邻法和 k 近邻法当两类近邻数接近时，为此，可考虑引入拒绝决策。

方法：设某个 $k' > \frac{1}{2}(k+1)$ ($k' < k$)，

只有当 x 的 k 个近邻中有大于或等于 k' 个属于 ω_i 类时，才决策 $x \in \omega_i$ ，否则拒绝。

—— 简单多数 = 绝对多数

拒绝决策同样可引入改进的近邻法中，比如剪辑近邻法。

2.4.4 Parzen 窗法

概率密度估计： $\hat{p}(x) = \frac{1}{NV} = \frac{1}{N} \cdot \frac{1}{V}$

设 x 是 d 维特征向量，每个小船是一个超立方体，每一维的棱长是 h ，则小船体积为 $V = h^d$

核函数（窗函数）： $k(x, x_i) = \frac{1}{V} \exp\left(-\frac{\|x - x_i\|^2}{h^2}\right)$

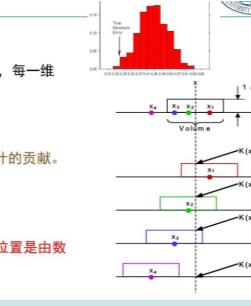
· 其反映了—个观测样本对 x 处的概率密度估计的贡献。

则概率密度估计为： $\hat{p}(x) = \frac{1}{N} \sum_{i=1}^N k(x, x_i)$

· 即在每一点上把所有观测样本的贡献平均。

注意到核函数估计和直方图法很相似，但窗的位置是由数据确定的。

窗函数条件： $k(x, x_i) \geq 0 \quad \int k(x, x_i) dx = 1$



常用窗函数

$$\text{① 超立方体窗 } k(x, x_i) = \begin{cases} \frac{1}{h^d}, & \text{if } \|x - x_i\| \leq \frac{h}{2}, j = 1, 2, \dots, d \\ 0, & \text{otherwise} \end{cases}$$

$$\text{② 正态窗 } k(x, x_i) = \frac{1}{\sqrt{(2\pi)^d h^d}} \exp\left\{-\frac{1}{2} \frac{(x - x_i)^T Q^{-1} (x - x_i)}{h^2}\right\} \quad (\Sigma = P^T Q)$$

$$\text{③ 超球窗 } k(x, x_i) = \begin{cases} V^{-1}, & \text{if } \|x - x_i\| \leq \frac{h}{2} \\ 0, & \text{otherwise} \end{cases}$$

窗长度 h 对概率密度估计 $p_{\hat{p}}(x)$ 的影响：

- 若 h 太大， $p_{\hat{p}}(x)$ 是 $p(x)$ 的一个平坦、分辨率低的估计，有平均误差。
- 若 h 太小， $p_{\hat{p}}(x)$ 是 $p(x)$ 的一个不稳定的起伏大的估计，有噪声误差。

$$\frac{1}{\sqrt{\pi}} e^{-\frac{(x-x_i)^2}{2h^2}} \text{ 高斯窗}$$