2.5.0

概率密度直方图方法：① 为估计在某个特定位置的概率密度
估计的
应考虑位于那个点的某个领域内的数据点

② 为获得好的结果，平滑参数的值既不能太大
也不能太小

2.5.1 假设有一未知的概率密度分布 $p(x)$，收集了 N 次观测，每个数据点落入 R 中的概率为

$$Bin(K,N,P)=\frac{N!}{K!(N-K)!}P^K(1-P)^{N-K}.$$

假定 N 较大，则 $P\simeq\frac{K}{N}$ 即 $K\simeq NP$ ①

假定 R 足够小，使在这个区域的概率密度 $p(x)$ 大致为常数

则 $P\simeq p(x)V$ 其中 V 为区域 R 的体积. ②

由①.②得 $p(x)=\frac{K}{NV}$

此式依赖于两个相互矛盾的假设 即区域 R 要足够小，使这个区域内的概率密度近似为常数，但也要足够大，使 K 足够让二项分布达到尖峰。

核方法：固定 V 然后从数据中确定 K.

取为以 x 为中心的边长小立方体 定义 $k(u)=\begin{cases}1 & |u_i|\le\frac{1}{2} \quad i=1,\dots,D\\ 0 & \text{其他.}\end{cases}$

∴当边长为 h 时 $k\left(\frac{x-x_n}{h}\right)=\begin{cases}1\\0\end{cases}$ ∴位于立方体内数据点总数为 $K=\sum_{n=1}^{N}k\left(\frac{x-x_n}{h}\right)$

即 $p(x)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{x-x_n}{h}\right)$ D 维边长为 h 的立方体 $V=h^D$

核密度估计 或 Parzen 估计

问题：人为带来的非连续性，且训练阶段只需存储数据.

2.5.1

一个以 x 为中心的小球体，且半径可以自由增长，直到他精确的包含 k 个数据点，利用公式 $P(x)=\frac{k}{NV}$ 计算概率分布

近邻方法：固定 k 的值然后确定 V

⇓

推广到分类.

一数据集 N_k 个数据属于类别 C_k，数据点总数为 N. $\sum N_k = N$.

现对一新的数据点进行分类. 画一个以 x 为中心的球且精确地包含 k 个数据点（无论属于哪个类别）设球体体积为 V. 且包含 C_k 类的 k_k 个数据点，则 $P(x|C_k)=\frac{k_k}{N_k V}$ ~ 每个类别关联的概率密度估计.

∵ 无条件概率密度为 $P(x)=\frac{k}{NV}$

先验为 $P(C_k)=\frac{N_k}{N}$

由贝叶斯定理得 $P(C_k|x)=\frac{P(x|C_k)P(C_k)}{P(x)}=\frac{k_k}{k}$

因为要最小化错误分类的概率，∴应把测试点 x 分配给有最大后验概率的类别. 对应最大的 $\frac{k_k}{k}$

最近邻 ($k=1$) 分类器在 $N\to\infty$ 时错误率不会超过最优分类器可达到的最小错误率的 2 倍

⇓

即有用真实概率分布的分类器.