1. 密度估计

设概率密度为 $p(x)$，则 x' 落在一特定区域 R 中的概率为：

$$P = \int_R p(x')dx'$$

现假设区域 R 内等概率密度，R 的体积为 V，n 个样本落在 V 中的数目为 k，则用频率代替概率：

$$\frac{k}{n} \approx \int_R p'(x)dx' = p(x)\cdot V \Rightarrow p_n(x) = \frac{\frac{k_n}{n}}{V_n} \text{ 估计 } p(x).$$

$p_n(x) \longrightarrow p(x)$ 的条件为：

① $\lim_{n\to\infty} V_n = 0$　（从特征空间 R 平均概率 $\longrightarrow$ x 点概率）

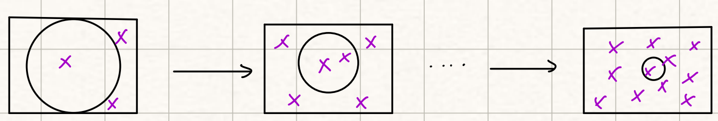
② $\lim_{n\to\infty} k_n = \infty$　（有足够多样本确保频率 $\to$ 概率，且 $n\to\infty \Rightarrow k_n\to\infty$）

③ $\lim_{n\to\infty} \frac{k_n}{n} = 0$　（由 $V_n\to 0$ 迫使 $\frac{k_n}{n}\to 0$，不然不收敛）。

* Parzen window：固定局部体积 V，k 在变化。

基本 Parzen 窗思想：每一时刻 n，给定特征空间 V_n（如 $V_n = \frac{1}{\sqrt{n}}$），计算给定特征空间中频率 $\frac{k}{n}$，

则 $p_n = \frac{\frac{k_n}{n}}{V_n}$ 估计 P



核化 Parzen window：

对于基本的 Parzen window，我们定义一个窗函数。

$$\varphi_n(u) = \begin{cases} 1 & u \in V_n \\ 0 & u \notin V_n \end{cases}$$

则有 $p_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(x_i)$.

现在我们用概率密度来代替这种硬分类，即核函数 φ 满足：

$$\varphi(x) \geq 0,\quad \int \varphi(u)\,du = 1.$$

考虑到尺度变换与位置变换，此时区域内的样本数量

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

故概率密度估计 $p_n = \frac{\frac{k_n}{n}}{V_n} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h_n}\right)$

注意：h 取值过大：估计较平稳，但容易训练不足（$h(x)$ 效应明显）。

h 取值过小：估计不平稳，容易过拟合（样本点效应明显）

Parzen window 估计方法收敛性：

$$\bar{p}_n(x) = E[p_n(x)]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h_n}\right)\right] \qquad (x_i \text{ 是 i.i.d 样本})$$

$$= \int \frac{1}{V_n}\varphi\left(\frac{x-v}{h_n}\right)\cdot p(v)\,dv \qquad (x_i \sim p(x))$$

$$= \int \delta_n(x-v)\,p(v)\,dv. \qquad \left(\delta_n(x) = \frac{1}{V_n}\varphi\left(\frac{x}{h_n}\right)\right)$$

$\bar{p}_n(x)$ 是 $p(x)$ 与 $\delta_n(x)$ 的卷积.

当 $n \to \infty$，$\delta_n(x-v) \longrightarrow \delta(x) = \begin{cases} 1 & x\text{点处} \\ 0 & \text{otherwise} \end{cases}$

$$\bar{p}_n(x) \longrightarrow p(x).$$

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

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

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

去掉第二项. 代入上式. 我们有

$$\sigma^2(x) \leq \frac{\sup\{\varphi(x)\}}{nV_n}\,\bar{p}_n(x).$$

窗宽 h_n 选择： * 一般而言 n 越大. 或密度很大时. h_n 越小. 相应 V_n 也减小

* 可以用交叉验证进行选择.