

# 实验三：参数估计与非参数估计

Parameter estimation & Nonparameter estimation

# 参数估计 — Eager Learning

- 已知：
  - ✱ 样本的概率密度分布
- 求解：
  - ✱ 解析表达式的参数
- 主要方法有：

|      | 最大似然估计  | 最大后验概率估计  | 贝叶斯估计  |
|------|---|---|--|
| 学习过程 | 训练数据 $D$  | $D + \text{先验概率 } p(\theta)$                      | $D + \text{先验概率 } p(\theta)$                       |
| 估计过程 | $\hat{\theta} = \arg \max_{\theta} p(D   \theta)$ | $\hat{\theta} = \arg \max_{\theta} p(\theta   D)$ | $\hat{\theta} = \int \theta p(\theta   D) d\theta$ |
| 预测过程 | $y = p(x   \hat{\theta})$                         | $y = p(x   \hat{\theta})$                         | $y = p(x   \hat{\theta})$                          |

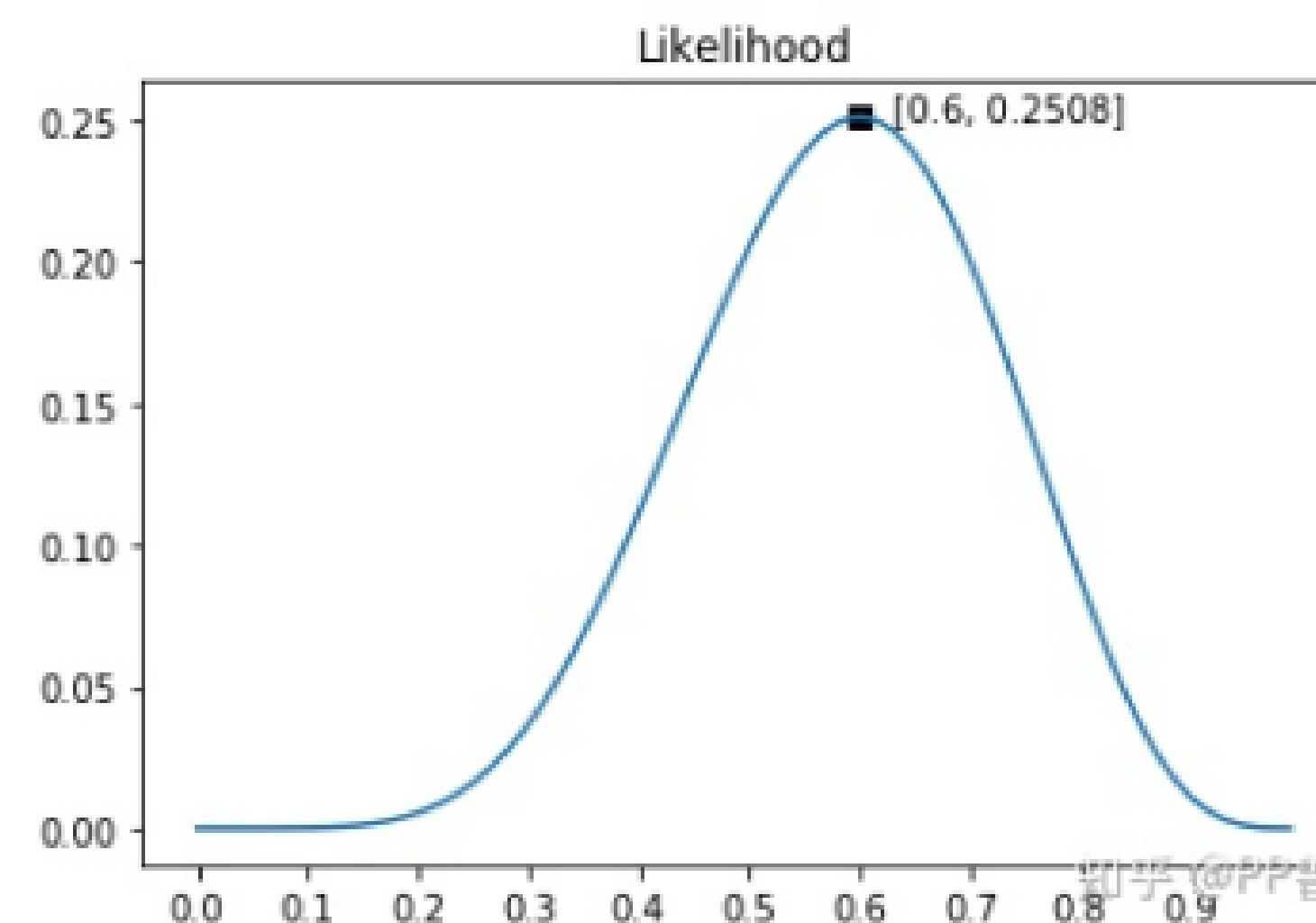
## ·似然估计

似然反映的是：**已知结果，反推原因**。似然函数表示的是基于观察的数据，取不同的参数 $\theta$ 时，统计模型以多大的可能性接近真实观察数据。例如：已经给你了一系列硬币正反情况，但你并不知道硬币的构造，下次下注时你要根据已有事实，反推硬币的构造。例如，当观察到硬币“10正0反”的事实，猜测硬币极有可能每次都是正面；当观察到硬币“6正4反”的事实，猜测硬币有可能不是正反均匀的，每次出现正面的可能性是0.6。

似然函数通常用L表示。观察到抛硬币“6正4反”的事实，硬币参数 $\theta$ 取不同值时，似然函数表示为：

$$L(\theta; 6\text{正}4\text{反}) = C_{10}^6 \times \theta^6 \times (1 - \theta)^4 \quad (2)$$

$$L(\theta; \mathbf{X}) = P_1(\theta; X_1) \times P_2(\theta; X_2) \dots \times P_n(\theta; X_n) = \prod P_i(\theta; X_i)$$



# 最大似然估计 (MLE)

---

- 给定随机样本  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} = \{\mathbf{x}_k\}_{k=1}^N$  来自概率密度  $p(\mathbf{x} | \theta)$

- 假设样本是独立同分布的, 则它们的联合概率分布为

$$p(\mathbf{X} | \theta) = p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N | \theta) = \prod_{k=1}^N p(\mathbf{x}_k | \theta)$$

- 估计使似然函数取最大值的参数  $\hat{\theta}$  :  $\hat{\theta} = \arg \max_{\theta} \prod_{k=1}^N p(\mathbf{x}_k | \theta)$

- 令似然函数对  $\theta$  的偏导数为零, 求解  $\hat{\theta}$ :  $\frac{\partial}{\partial \theta} \log \prod_{k=1}^N p(\mathbf{x}_k | \theta) = 0$

# 最大似然估计 (MLE)

---

- 定义对数似然函数:  $L(\theta) = \log \prod_{k=1}^N p(\mathbf{x}_k | \theta)$
- 令似然函数对  $\theta$  的偏导数为零:
$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} L(\theta) \\ &= \frac{\partial}{\partial \theta} \log \prod_{k=1}^N p(\mathbf{x}_k | \theta) \\ &= \sum_{k=1}^N \frac{\partial}{\partial \theta} \log p(\mathbf{x}_k | \theta) \\ &= \sum_{k=1}^N \frac{1}{p(\mathbf{x}_k | \theta)} \frac{\partial}{\partial \theta} p(\mathbf{x}_k | \theta) \end{aligned}$$
- 求得  $\hat{\theta}$ , 对于样本点进行预测  $y = p(x | \hat{\theta})$

# 最大似然估计 (MLE)

---

- 例：假设数据集  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ，服从高斯分布  $\mathcal{N}(\mu, \sigma^2)$ ，其中标准差  $\sigma$  已知，求均值  $\mu$  的最大似然估计。

$$\begin{aligned}\theta = \mu \Rightarrow \hat{\theta} &= \arg \max_{\theta} \sum_{k=1}^N \log p(x_k | \theta) \\ &= \arg \max_{\theta} \sum_{k=1}^N \log \left( \frac{1}{\sqrt{\pi}\sigma} \exp \left( -\frac{1}{2\sigma^2} (x_k - \mu)^2 \right) \right) \\ &= \arg \max_{\theta} \sum_{k=1}^N \left\{ \log \frac{1}{\sqrt{\pi}\sigma} - \frac{1}{2\sigma^2} (x_k - \mu)^2 \right\}\end{aligned}$$

$$\frac{\partial}{\partial \mu} \sum_{k=1}^N \left\{ \log \frac{1}{\sqrt{\pi}\sigma} - \frac{1}{2\sigma^2} (x_k - \mu)^2 \right\} = 0 \Rightarrow \hat{\theta} = \mu = \frac{1}{N} \sum_{k=1}^N x_k \quad \Rightarrow \quad \text{数据集的样本均值}$$

# 最大后验概率估计 (MAP)

---

- 已知：随机样本  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ，先验概率  $p(\theta)$
- 求解：后验概率  $p(\theta | \mathbf{X})$

- 根据贝叶斯规则：

$$p(\theta | \mathbf{X}) = \frac{\overset{\text{MLE}}{\boxed{p(\mathbf{X} | \theta)}} \overset{\text{Prior}}{\boxed{p(\theta)}}}{\cancel{p(\mathbf{X})} \text{ 常数项}}$$

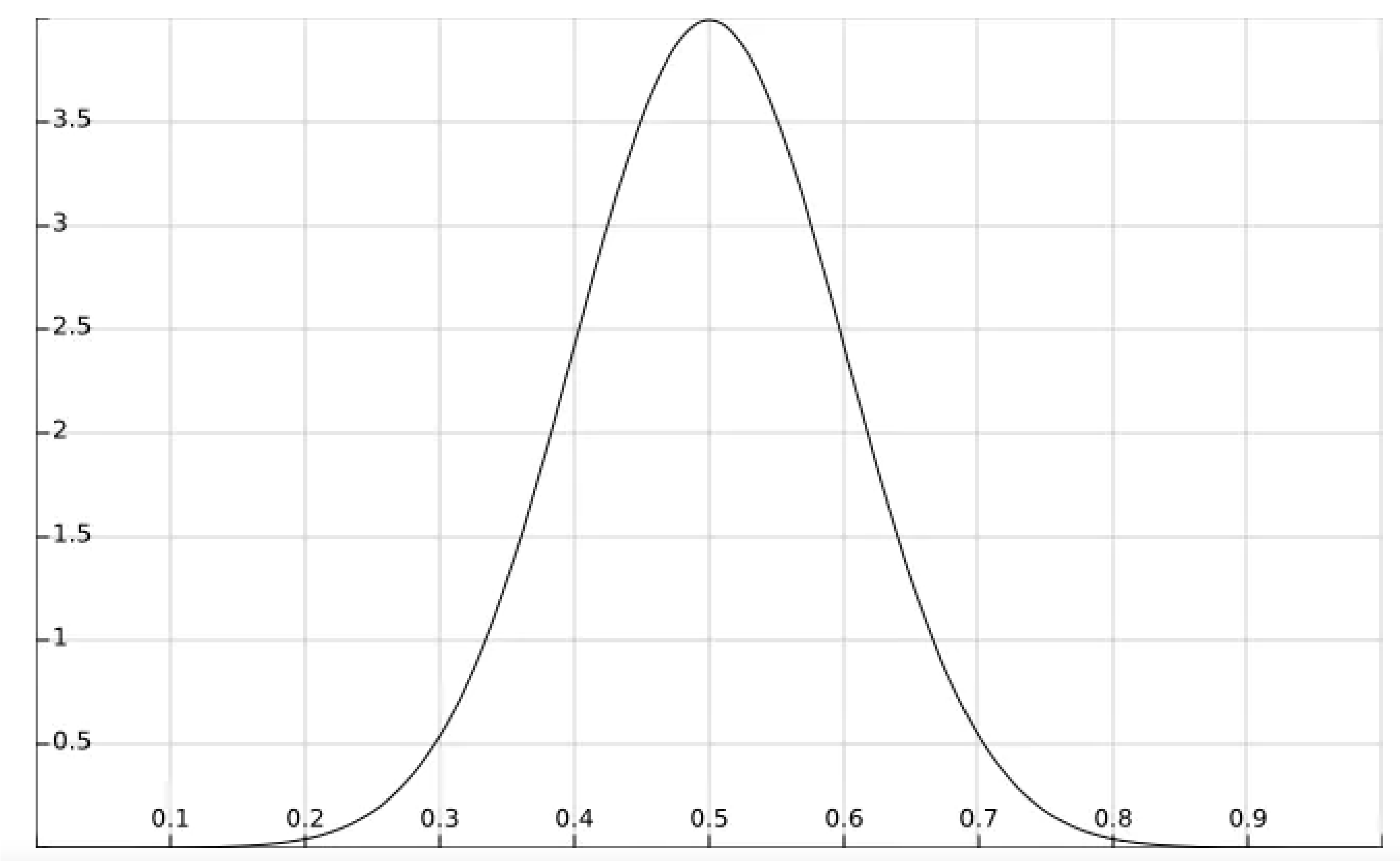
- 估计使最大后验概率取最大值的参数  $\hat{\theta}_{MAP}$ ：

$$\frac{\partial}{\partial \theta} p(\theta | \mathbf{X}) = 0 \text{ 或 } \frac{\partial}{\partial \theta} \left( p(\mathbf{X} | \theta) p(\theta) \right) = 0$$

在抛硬币的例子中，通常认为  $\theta = 0.5$  的可能性最大，因此我们用均值为 0.5，方差为 0.1 的高斯分布来描述  $\theta$  的先验分布，当然也可以使用其它的分布来描述  $\theta$  的先验分布。 $\theta$  的先验分布为：

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\theta-\mu)^2}{2\sigma^2}} = \frac{1}{10\sqrt{2\pi}}e^{-50(\theta-0.5)^2}$$

先验分布的函数图如下：



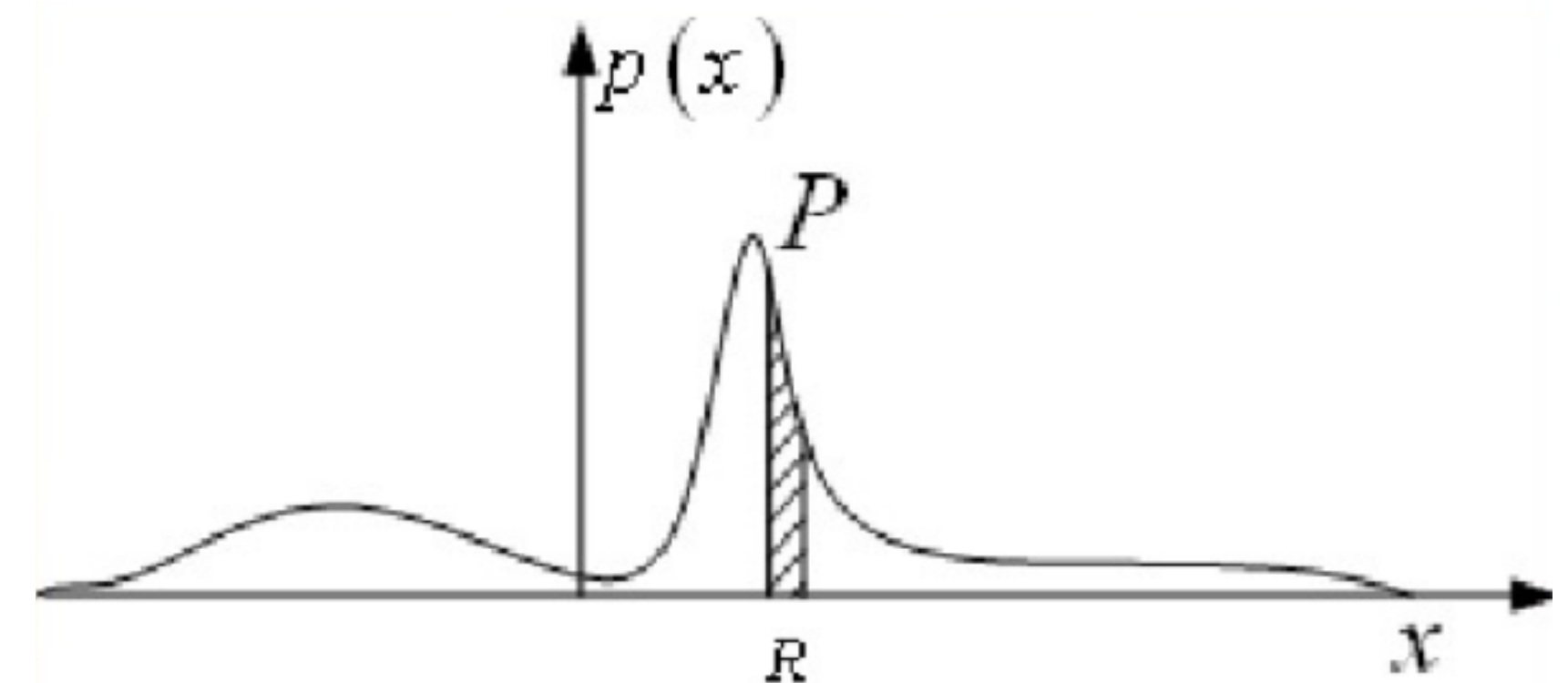
在最大似然估计中，已知似然函数为  $P(X|\theta) = \theta^6(1-\theta)^4$ ，因此：

$$P(X|\theta)P(\theta) = \theta^6 \times (1-\theta)^4 \times \frac{1}{10\sqrt{2\pi}} \times e^{-50(\theta-0.5)^2}$$



# 非参数估计 — Lazy Learning

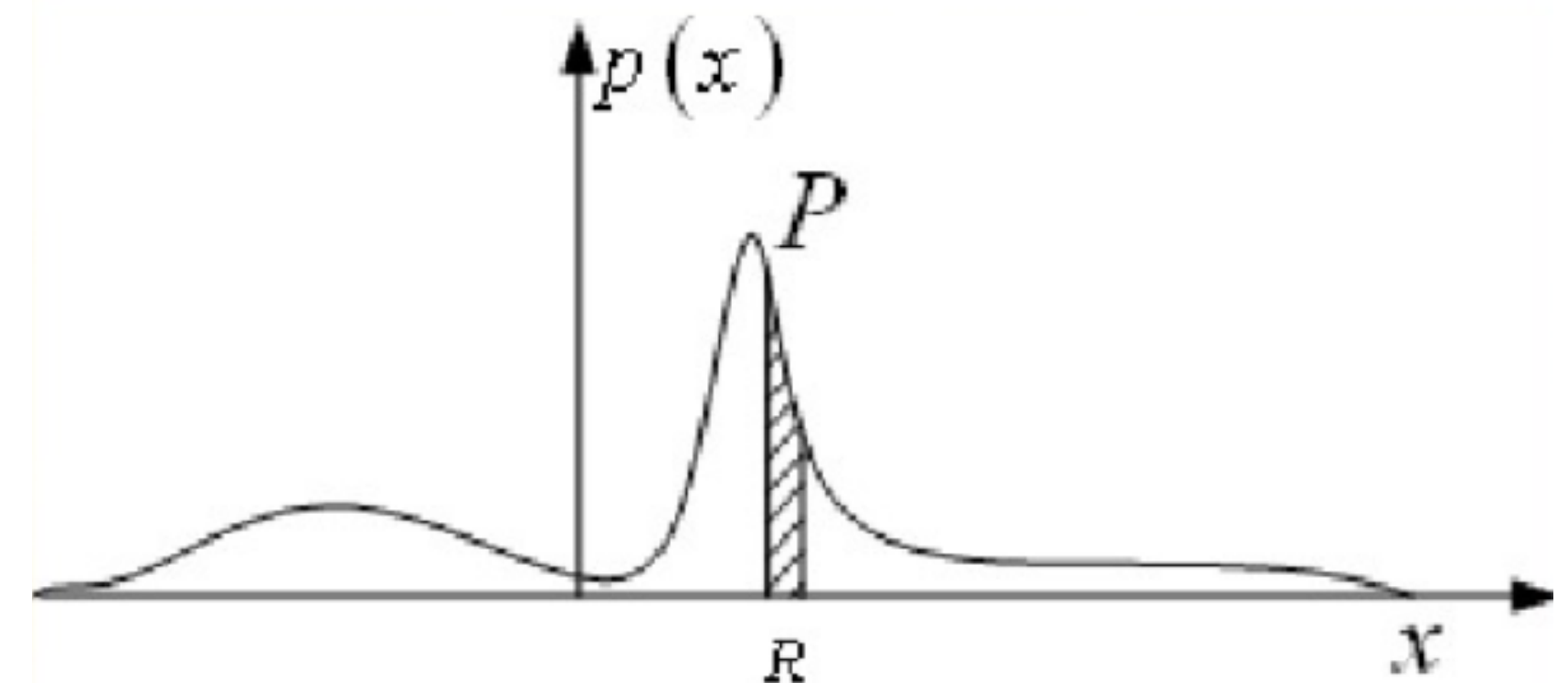
- 在非参数估计中，假定相似的输入具有相似的输出，不对基础密度假定任何形式的先验参数
- 非参数模型的复杂性依赖于训练集的大小，依赖于数据中问题的固有复杂性
- 当给定训练集时，并不计算模型，而将模型的计算推迟到给定一个检验实例时才进行，这会导致存储和计算量的增加。(比如：开卷考试)
- 核心思路：一个向量  $\mathbf{x}$  落入区间  $R$  中的概率为  $p = \int_R p(x)dx$
- 主要方法：
  - ✱ 直方图估计
  - ✱ 核估计
  - ✱ k最近邻估计

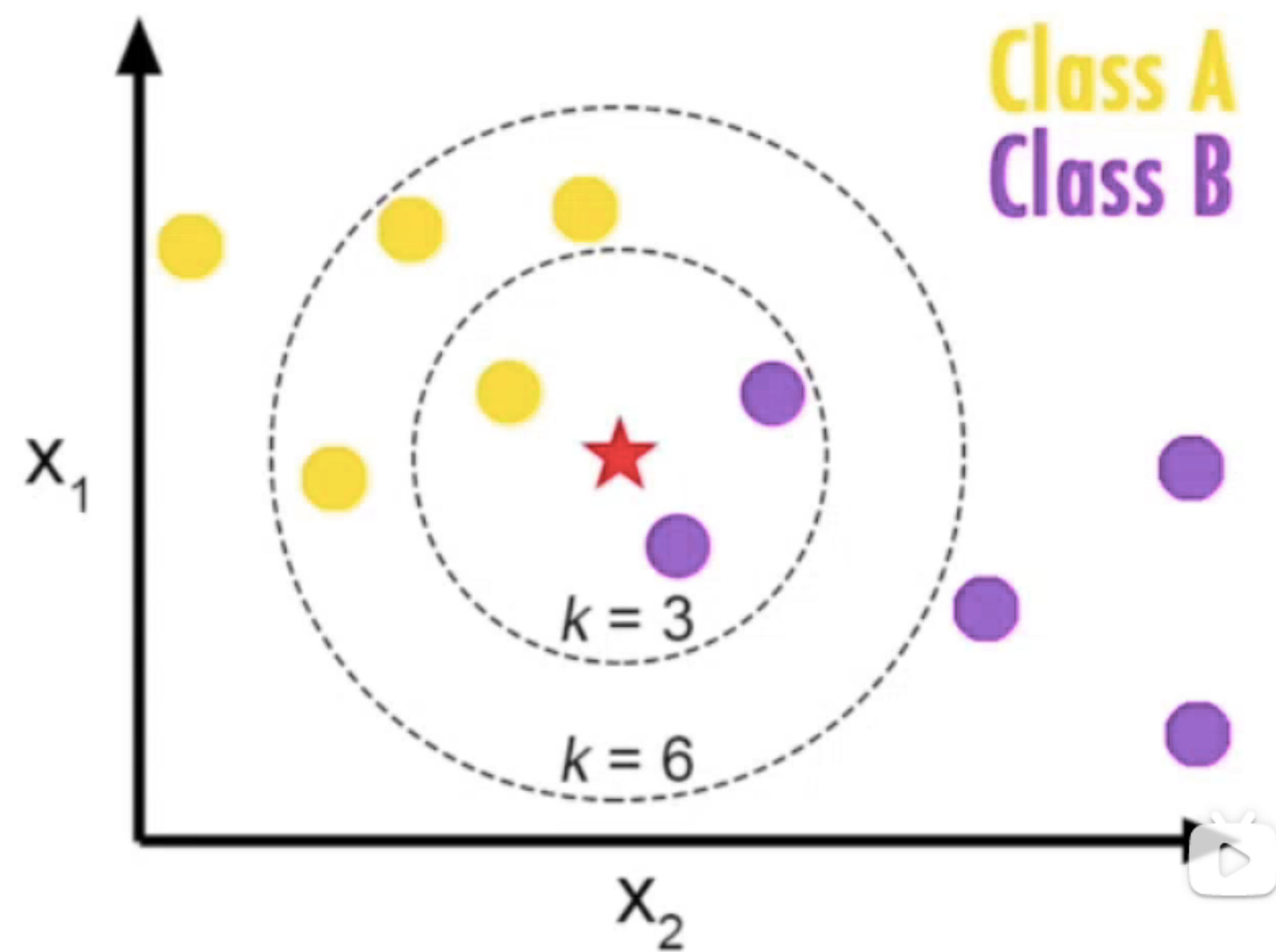
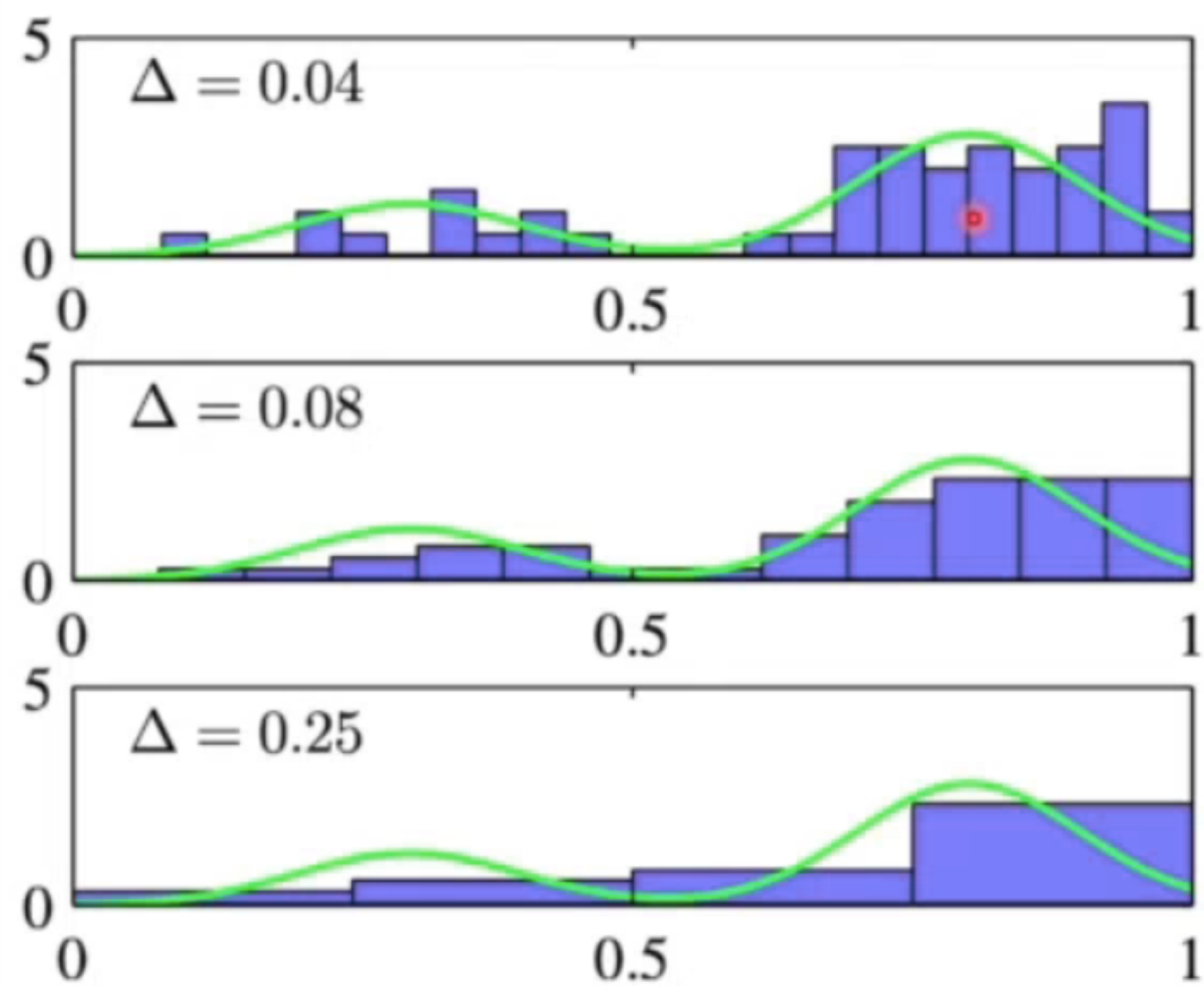


# 非参数估计

- 核心思路：一个向量  $\mathbf{x}$  落入区间  $R$  中的概率为  $p = \int_R p(x)dx$
- 当样本数  $n$  足够大时，可以近似地认为  $P \approx \frac{k}{n}$ ，其中  $k$  是出现该特征的频数。
- 假设密度函数  $p(x)$  是连续的，那么在区域  $R$  足够小时，我们可以近似地认为  $p(x)$  是一个常值函数，因此  $P \approx p(x)V$ ，其中  $V$  在多维情况下是区域  $R$  的体积。

$$\frac{k}{n} \approx P \approx p(x)V \Rightarrow p(x) \cong \frac{k}{nV}$$





# 非参数估计

---

$$p(x) \cong \frac{k}{nV}$$

积分近似:  $\lim V_n = 0$

固定  $V$  值, 使用样本确定  $K$  值——核函数密度估计

频率近似:  $\lim k_n = \infty$  &  $\lim \frac{k_n}{n} = 0$

固定  $K$  值, 使用样本确定  $V$  值——最近邻密度估计

- 当  $n \rightarrow \infty$  时,  $V$  值适当减小,  $k$  值适当增加, 上述两种方法均收敛于真实的概率密度函数。
- 优点: 不需要任何对分布形式的假设, 能根据数据特性自适应地估计出相应地密度函数

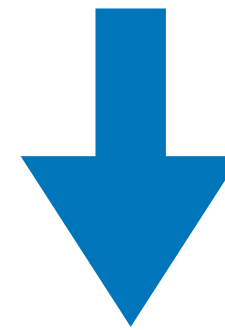
# 核函数密度估计

✱ 方窗函数

$$\varphi(u) = \begin{cases} 1, |u| \leq 1/2 \\ 0, otherwise \end{cases}$$

✱ 高斯窗口函数

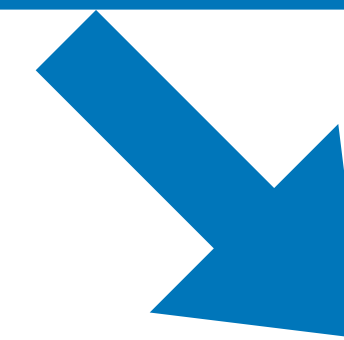
$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \|u\|^2\right\}$$



|                       |  |           |
|-----------------------|--|-----------|
| $p(x) = \frac{k}{NV}$ | $K = \sum_{n=1}^N \varphi\left(\frac{x - x_n}{h}\right)$ | $V = h^D$ |
|-----------------------|--|-----------|



$$p(x) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h^D} \varphi\left(\frac{x - x_n}{h}\right)$$

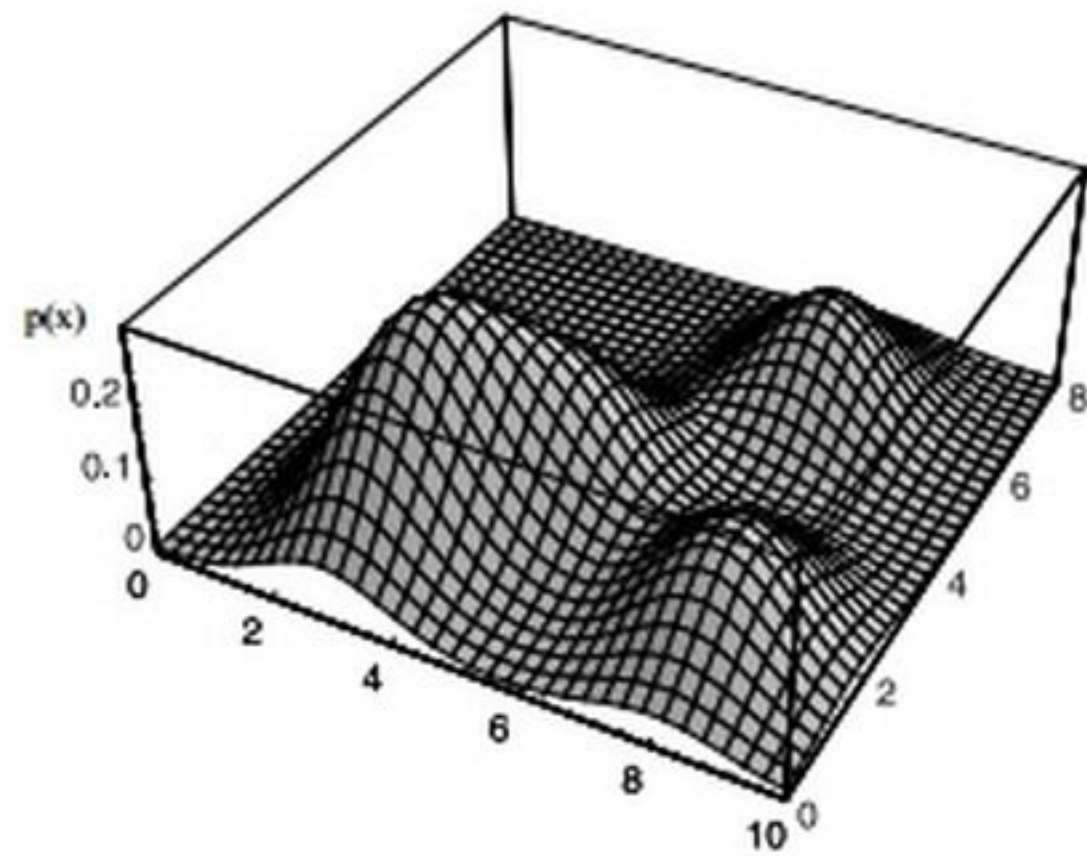


$$p(x) = \frac{1}{N} \sum_{n=1}^N \frac{1}{\sqrt{2\pi h^2}} \exp\left\{-\frac{\|x - x_n\|^2}{2h^2}\right\}$$

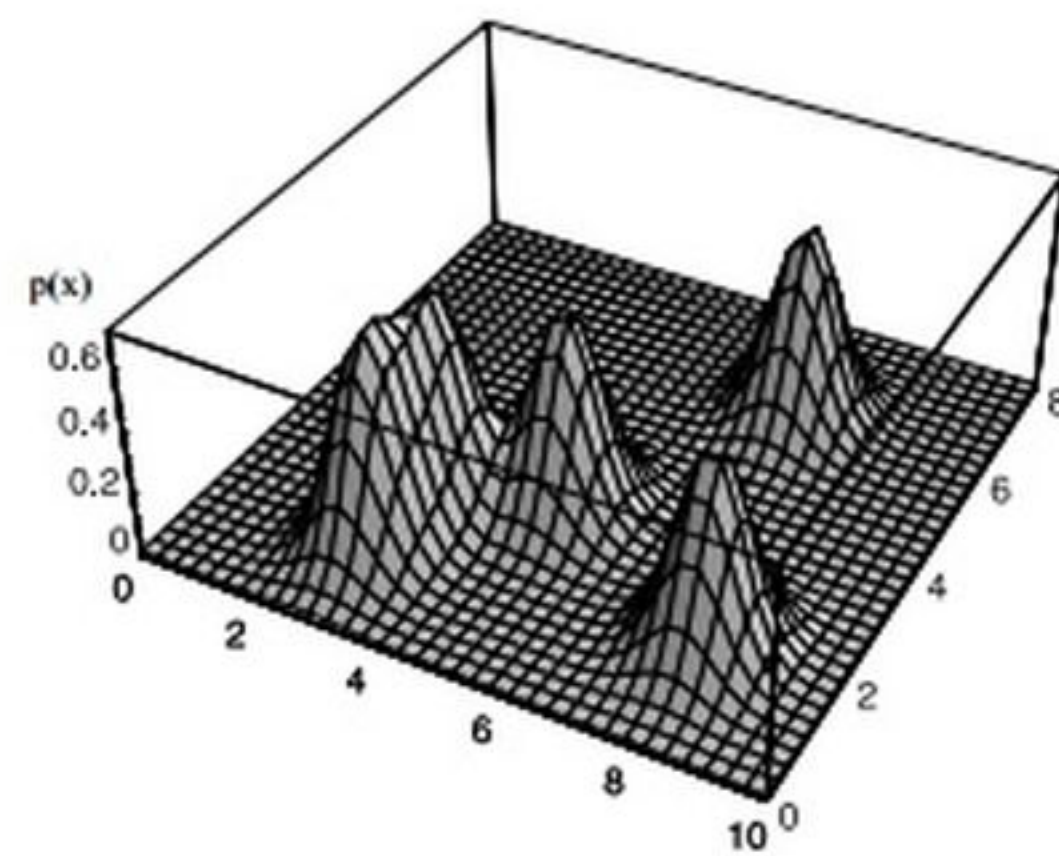
密度函数的估计



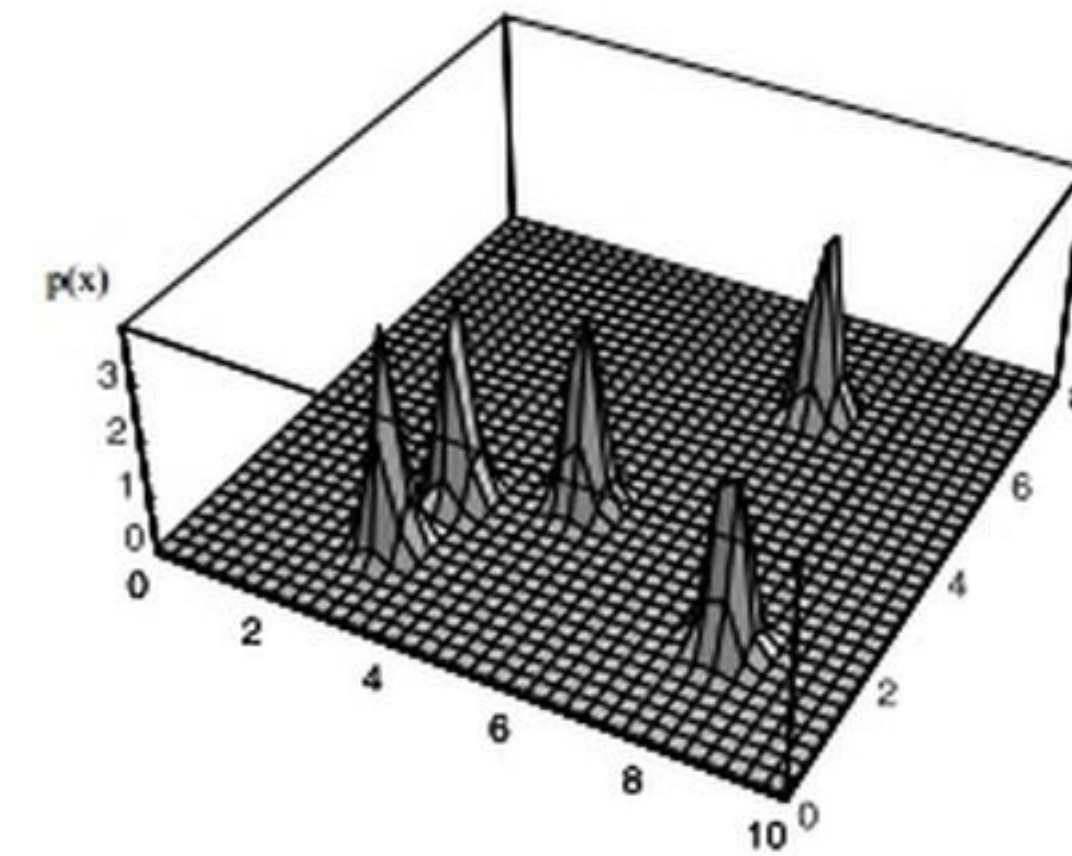
# 核函数密度估计



$h=1$



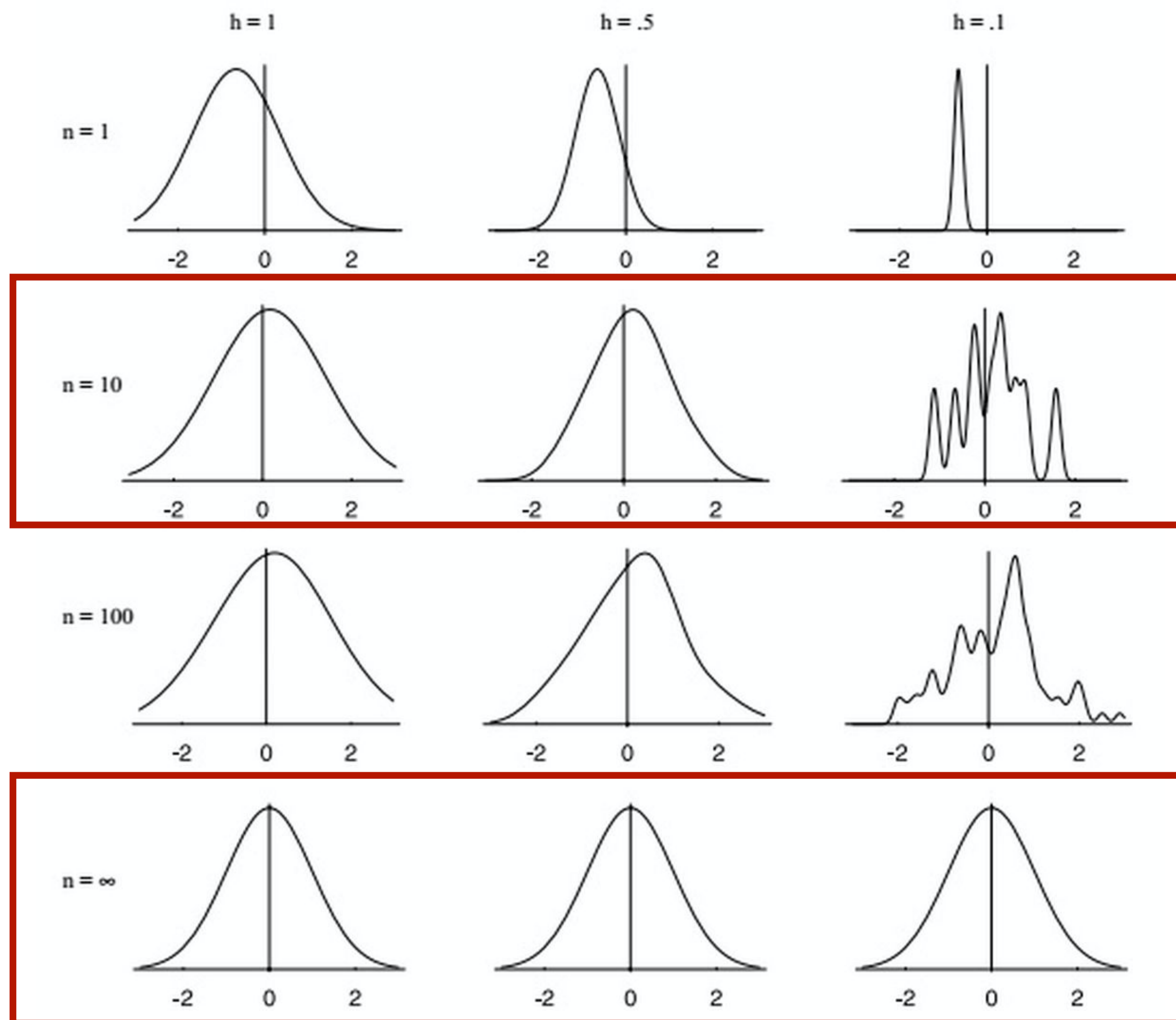
$h=0.5$



$h=0.2$

- 如何选择合适的窗口  $h$  是合理估计密度函数的关键问题
  - ❖ 较大取值将产生过度平滑的密度估计，模糊了数据的空间结构。
  - ❖ 较小取值将产生又长又尖的密度估计，解释比较困难。

# 核函数密度估计



## 交叉检验确定最优 $h$ 值

在样本量有限的情况下

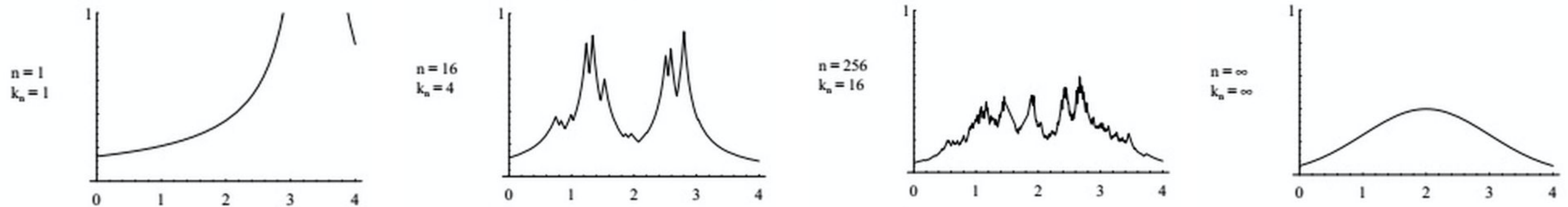
不同的  $h$  对非参估计的影响很大

当样本量趋于无穷时

非参估计都和真实的生成密度函数相匹配

# k最近邻密度估计

- 在  $k$  近邻法中，区域大小  $V$  不再是样本数量  $N$  的函数，而是这  $k$  个训练数据的函数
- 窗口函数集中于  $V$  or  $h$  的选取，而  $k$  近邻法集中于  $k$  的选取
- 最近邻法可以看做是  $k$  近邻法的一种平凡情况（ $n=1, k=1$ ）
- 当样本量  $n$  足够大时， $k$  近邻法能有效地估计出真实的密度函数





# k最近邻密度估计

---

用于分类时，非参估计可以直接用于估计类条件密度

- 已知：一个体积为  $V_n$  的区域  $R_n$ ，区域内共有  $k_n$  个样本，其中  $k_n^i$  个样本属于第  $i$  个类别  $w_i$
- 求解： $w_i$  类条件密度  $p(x | w_i)$

密度估计为：

$$p(x | w_i) = \frac{k_n^i}{nV_n}$$

先验密度的最大似然估计为：

$$p(w_i) = \frac{n_i}{N}$$

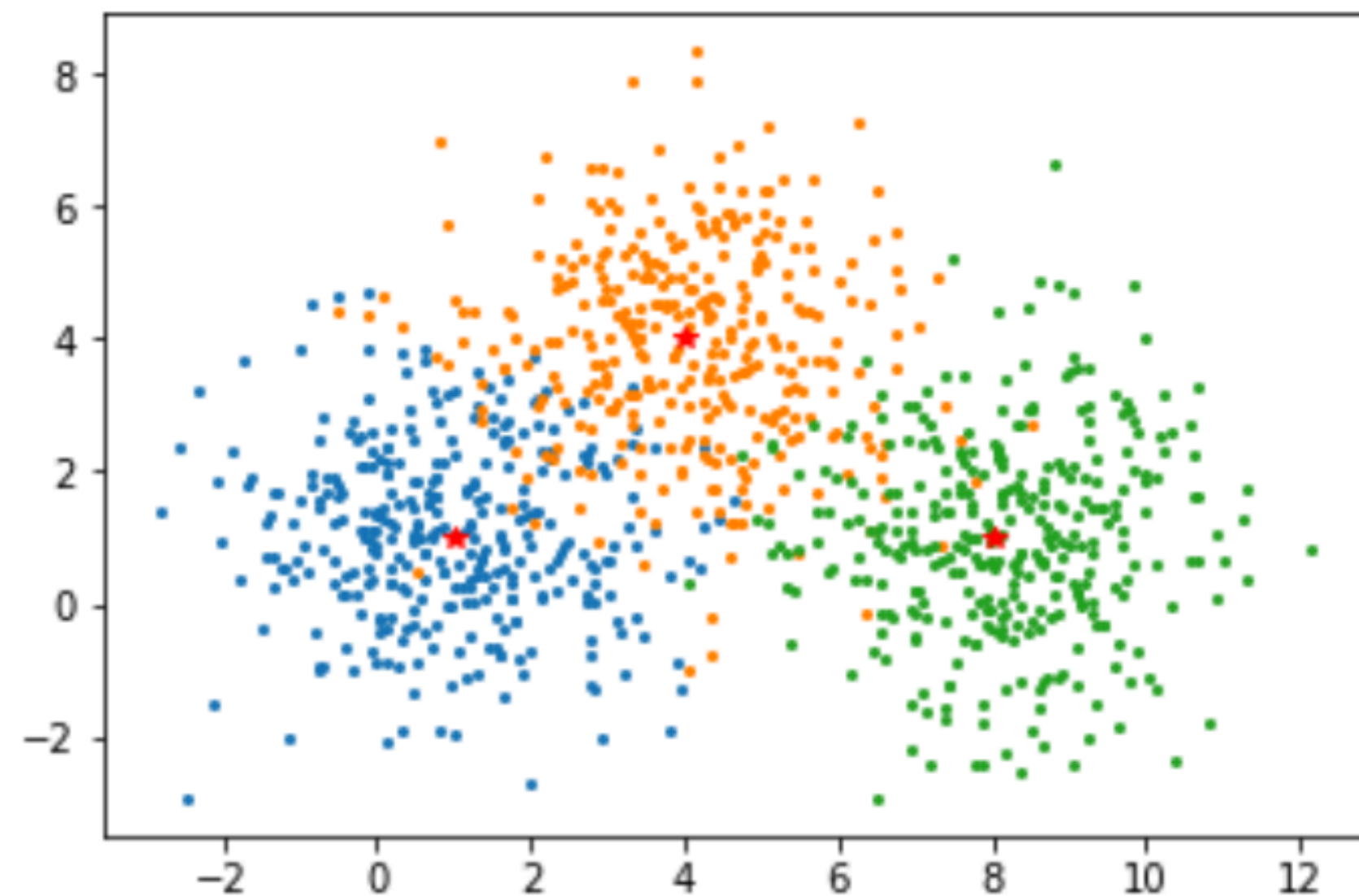
后验密度为：

$$p(w_i | x) = \frac{p(x | w_i)p(w_i)}{p(x)} = \frac{\frac{k_i}{n_i V} \cdot \frac{n_i}{N}}{\frac{k}{NV}} = \frac{k_i}{k}$$

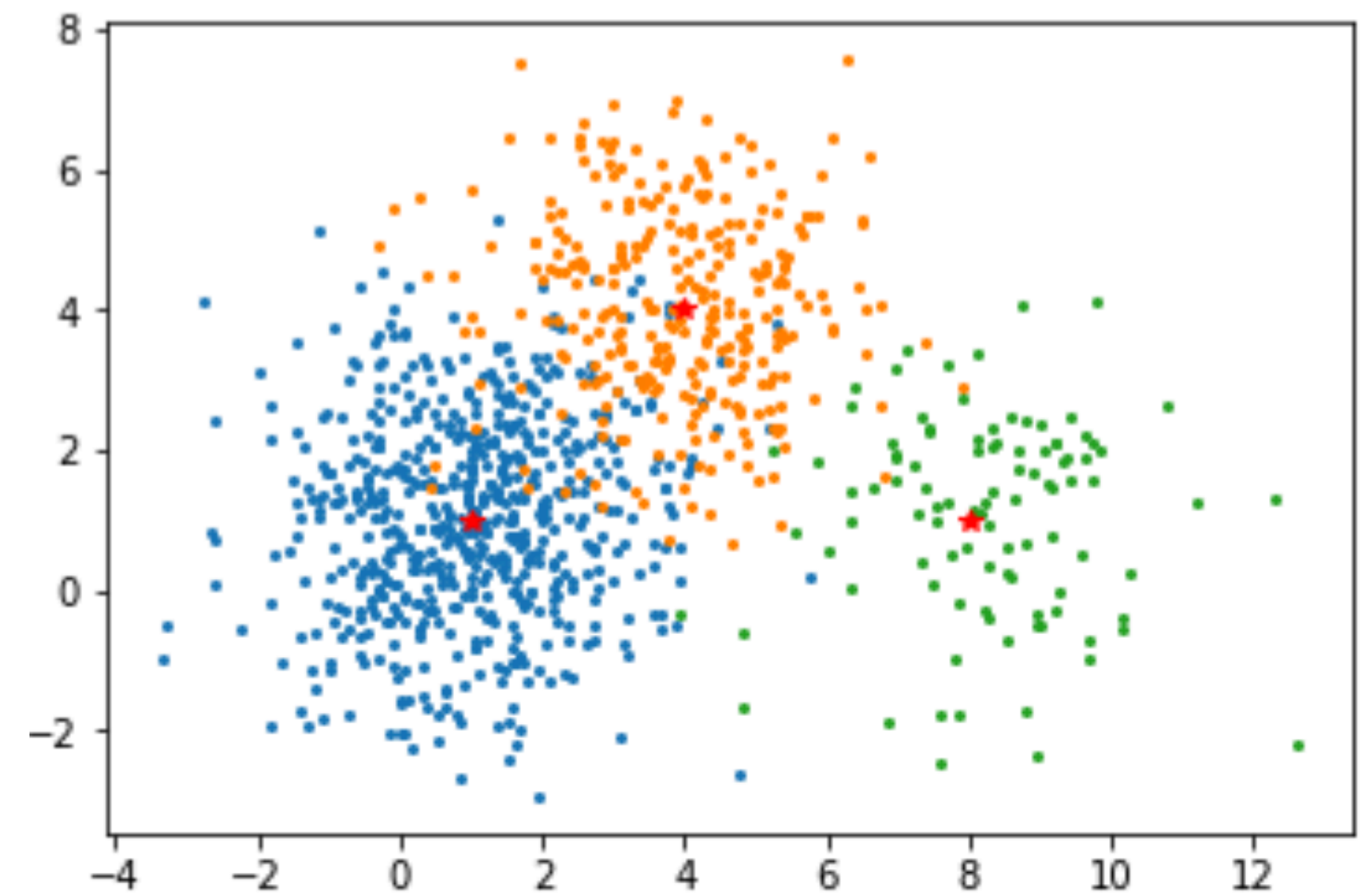
# 实验要求

`np.random.multivariate_normal(mean, cov, temp_num)`

- 数据: 生成两个各包含  $N=1000$  个二维随机矢量的数据集  $X_1$  和  $X_2$ , 数据集中随机矢量来自于三个分布模型, 分别满足均值矢量  $\mathbf{m}_1 = [1, 1]^T, \mathbf{m}_2 = [4, 4]^T, \mathbf{m}_3 = [8, 1]^T$  和协方差矩阵  $\mathbf{S}_1 = \mathbf{S}_2 = \mathbf{S}_3 = 2\mathbf{I}$ , 其中  $\mathbf{I}$  是  $2 \times 2$  的单位矩阵。在生成数据集  $X$  时, 假设来自三个分布模型的先验概率相同  $p(w_1) = p(w_2) = p(w_3) = 1/3$ ; 而在生成数据集  $X'$  时, 先验概率分别为  $p(w_1) = 0.6, p(w_2) = 0.3, p(w_3) = 0.1$ 。



$X_1$



$X_2$

# 实验要求

---

- 基本要求 (3')

在两个数据集合上应用“最大后验概率规则”进行分类实验，计算分类错误率，分析实验结果。

- 中级要求 (2')

在两个数据集合上使用高斯核函数估计方法，应用“似然率测试规则”分类，在  $[0.1, 0.5, 1, 1.5, 2]$  范围内交叉验证找到最优  $h$  值，分析实验结果。

- 提高要求

在两个数据集合上使用进行k-近邻概率密度估计，计算并分析  $k=1, 3, 5$  时的概率密度估计结果。

- 拓展要求

在两个数据集合上应用“贝叶斯规则”进行分类实验，计算分类错误率，分析实验结果。