

课程3： 贝叶斯分类

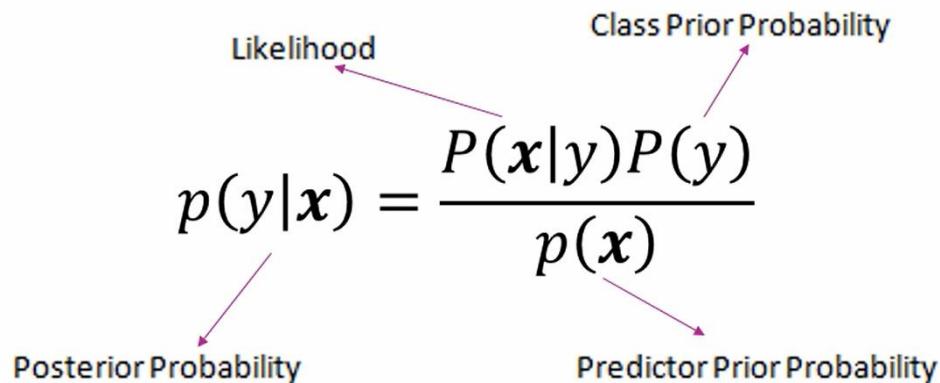
原理：

用贝叶斯公式根据某特征的先验概率计算出其后验概率, 然后选择具有最大后验概率的类作为该特征所属的类。

The diagram shows the Bayes' theorem formula with four labels and arrows pointing to the corresponding parts of the equation:

- Likelihood**: Points to $p(x | y)$ in the numerator.
- Class Prior Probability**: Points to $p(y)$ in the numerator.
- Posterior Probability**: Points to $p(y|x)$ on the left side of the equation.
- Predictor Prior Probability**: Points to $p(x)$ in the denominator.

$$p(y|x) = \frac{p(x | y) p(y)}{p(x)}$$



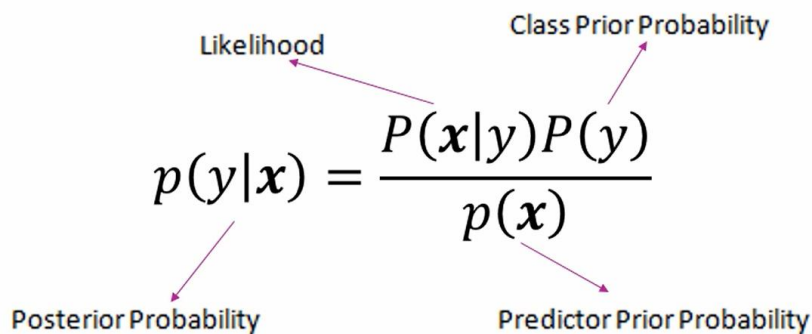
The diagram shows the formula $p(y|x) = \frac{P(x|y)P(y)}{p(x)}$ with four labels and arrows pointing to the corresponding terms: 'Likelihood' points to $P(x|y)$, 'Class Prior Probability' points to $P(y)$, 'Posterior Probability' points to $p(y|x)$, and 'Predictor Prior Probability' points to $p(x)$.

$$p(y|x) = \frac{P(x|y)P(y)}{p(x)}$$

先验概率 $p(y)$ ：反映了我们在实际观察之前对某种状态的**预期** --> 根据以往经验和分析得到的概率

举例：近一段时间以来接触到的一角的硬币比五角的硬币多，因此他觉得更可能是一角

证据因子 $p(x)$ ：使预估概率更接近真实概率 --> 起到归一化作用
若给定一个数据集，其中 $p(x)$ 的概率就被固定下来了，保持不变



A diagram showing the Bayes' theorem formula $p(y|x) = \frac{P(x|y)P(y)}{p(x)}$. Four purple arrows point from labels to parts of the formula: 'Likelihood' points to $P(x|y)$, 'Class Prior Probability' points to $P(y)$, 'Posterior Probability' points to $p(y|x)$, and 'Predictor Prior Probability' points to $p(x)$.

$$p(y|x) = \frac{P(x|y)P(y)}{p(x)}$$

似然函数 $p(x|y)$ ： 根据样本出现的频率来估计这个特征的概率

概率vs似然

- 概率：已知硬币的参数，就可以去推测抛硬币的各种情况的可能性
- 似然：我们对硬币的参数并不清楚，要通过抛硬币的情况去推测硬币的参数

后验概率 $p(y|x)$ ： 事件 x 发生后，得到结果 y 的概率

最小错误率贝叶斯决策

样本 x 上错误的概率 $P(e|x)=\begin{cases} P(w_2|x) & \text{如果决策 } x \in w_1 \\ P(w_1|x) & \text{如果决策 } x \in w_2 \end{cases}$

所有服从同样分布的独立样本上错误概率的期望 $P(e)=\int P(e|x)p(x)dx$

最小错误率贝叶斯决策 $\min P(e)=\int P(e|x)p(x)dx$

决策规则 如果 $P(w_1|x)>P(w_2|x)$, 则 $x \in w_1$; 反之, 则 $x \in w_2$

最小风险贝叶斯决策

条件风险（期望损失）：
$$R(c_i|x) = \sum_{j=1}^N k_{ij} P(c_j | x)$$

1. 错误的分类会带来损失

把病人误诊为健康 \longrightarrow 风险代价大

把正常人误诊为病人 \longrightarrow 风险代价小

2. 不同的错误带来的损失可能不同, 记作 k_{ij}

$k_{ij} \begin{cases} i \rightarrow \text{我们判断的类别} \\ j \rightarrow \text{原本真实的类别} \end{cases}$

例： $c = 1, 2, 3, 4$

$$R(c_1|x) = k_{1,1}P(c_1 | x) + k_{1,2}P(c_2 | x) + k_{1,3}P(c_3 | x) + k_{1,4}P(c_4 | x)$$

最小风险贝叶斯决策

判定准则：

总体风险： 条件风险的期望

$$R(h) = E_x[R(h(x) | x)] \quad \longleftarrow \text{最小化}$$

最小化每一项 $R(c | x)$ 总_{min} = 样本1_{min} + 样本2_{min} + ...

$$h^*(x) = \arg \min_{c \in Y} R(c | x) = \sum k_{ij} P(c | x)$$

贝叶斯最优分类器

最小风险贝叶斯决策

0-1损失: $k_{ij} = \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise} \end{cases}$

定义 k_{ij}

$$h^*(x) = \arg \min_{c \in Y} R(c | x)$$

$$R(c | x) = 1 - p(c | x)$$

$$R(c_1|x) = k_{1,1}P(c_1 | x) + k_{1,2}P(c_2 | x) + k_{1,3}P(c_3 | x) + k_{1,4}P(c_4 | x)$$

$$h^*(x) = \arg \max_{c \in Y} P(c | x)$$

两类错误率

预测类别	真实类别	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

灵敏度 $S_n = \frac{TP}{TP+FN}$

特异度 $S_p = \frac{TN}{TN+FP}$

- P (Positive) 和 N (Negative) 表示模型的判断结果
- T (True) 和 F (False) 表示模型的判断结果是否正确
- FP: 假正例
- FN: 假负例 (第二类错误)
- TP: 真正例
- TN: 真负例 (第一类错误)

两类错误率

- 第一类错误率 α : 真实的阴性样本中被错误判断为阳性的比例
- 第二类错误率 β : 真实的阳性样本中被错误判断为阴性的比例

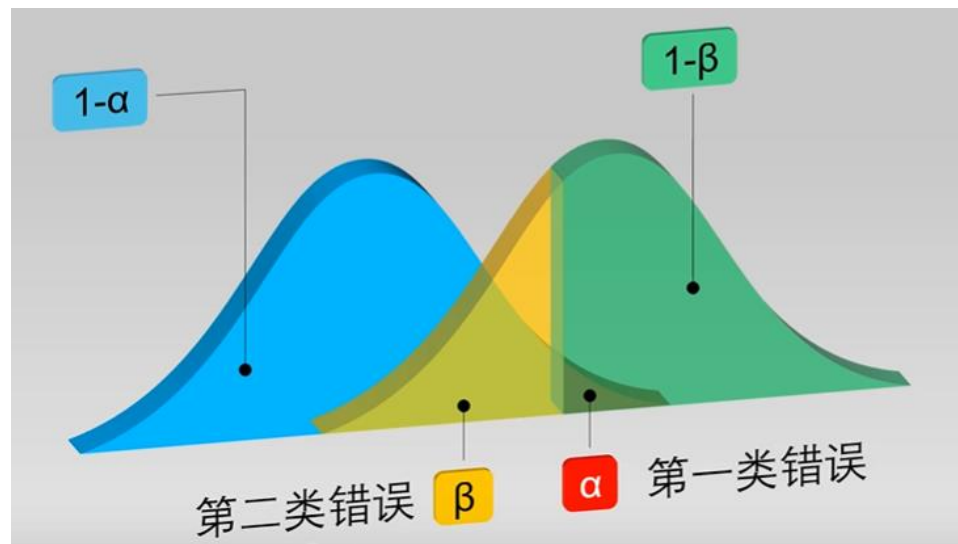
$$\text{第一类错误率: } P_1(e) = \int_{R_2} p(x | w_1) dx$$

$$\text{第二类错误率: } P_2(e) = \int_{R_1} p(x | w_2) dx$$

Neyman-Pearson决策

固定一类错误率、使另一类错误率尽可能小

$$\begin{aligned} &\min P_1(e) \\ &s.t. P_2(e) - \varepsilon_0 = 0 \end{aligned}$$



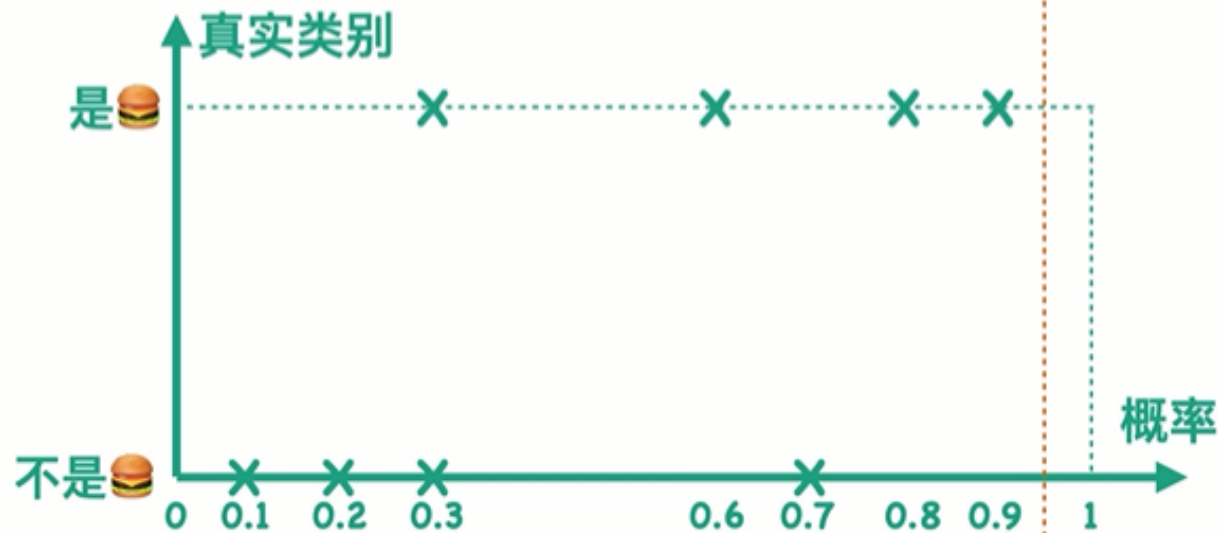
ROC曲线

分类问题: 判断是否为汉堡🍔的图片

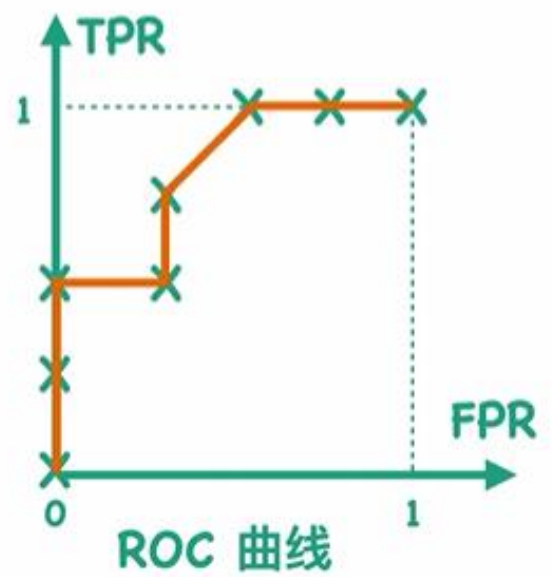
$$TPR = \frac{TP}{TP + FN} = \frac{0}{0 + 4} = 0$$
$$FPR = \frac{FP}{FP + TN} = \frac{0}{0 + 4} = 0$$

预测类别

	真实类别	
	是🍔	不是🍔
是🍔	0	0
不是🍔	4	4



阈值: (0.9, 1]



练习

(1) 假设在某个局部地区细胞识别中正常 (w_1) 和异常 (w_2) 两类的先验概率分别为：正常状态 $P(w_1)=0.9$ ，异常状态 $P(w_2)=0.1$

现有一待识别的细胞，其观察值为 x ，从类条件概率密度曲线上分别查得 $p(x|w_1)=0.2$ ， $p(x|w_2)=0.4$ ，试对该细胞 x 进行分类

(2) 在 (1) 给出条件的基础上, 利用表2-2的决策表, 按最小风险贝叶斯决策进行分类

表 2-2 例 2.2 的决策表

决策	状 态	
	ω_1	ω_2
α_1	0	6
α_2	1	0

解: 利用贝叶斯公式, 分别计算出 w_1 及 w_2 的后验概率

$$P(w_1 | x) = \frac{p(x | w_1)P(w_1)}{\sum_{j=1}^2 p(x | w_j)P(w_j)} = \frac{0.2 * 0.9}{0.2 * 0.9 + 0.4 * 0.1} = 0.818$$

$$P(w_2 | x) = 1 - P(w_1 | x) = 0.182$$

根据贝叶斯决策规则式(2-8), 因为

$$P(w_1 | x) = 0.818 > P(w_2 | x) = 0.182$$

所以合理的决策是把 x 归类于正常状态

解: 已知条件为 $P(w_1)=0.9, P(w_2)=0.1$
 $P(x|w_1)=0.2, P(x|w_2)=0.4$
 $\lambda_{11} = 0, \lambda_{12} = 6$
 $\lambda_{21} = 0, \lambda_{22} = 0$

根据例2.1的计算结果可知后验概率为

$$P(w_1 | x) = 0.818, P(w_2 | x) = 0.182$$

再按式(2-26)计算出条件风险

$$R(\alpha_1 | x) = \sum_{j=1}^2 \lambda_{1j} P(w_j | x) = \lambda_{12} P(w_2 | x) = 1.092$$

$$R(\alpha_2 | x) = \lambda_{21} P(w_1 | x) = 0$$

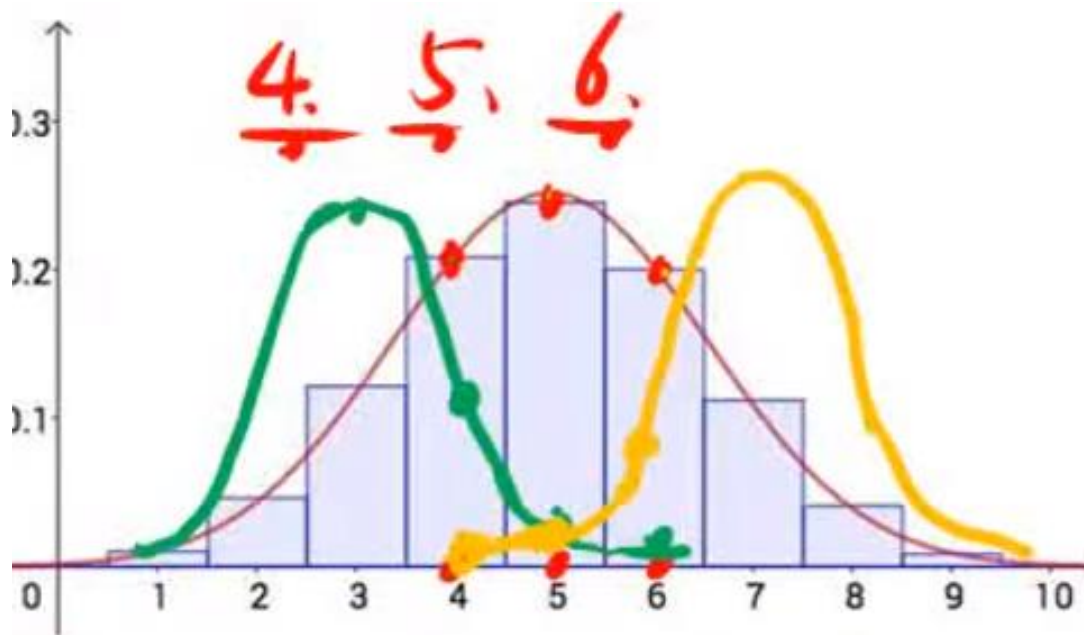
$$R(\alpha_1 | x) > R(\alpha_2 | x)$$

即决策为 w_2 的条件风险小于决策为 w_1 的条件风险, 因此我们采取决策行动 α_2 , 即判断待识别的细胞 x 为 w_2 类—异常细胞。

最大似然估计

原理：通过事实数据，猜测模型参数情况

- 相信试验/数据反映的客观规律
- 试验结果/数据既然已经发生，就说明该事件发生的可能性是很大的，是个大概率事件



最大似然估计求解

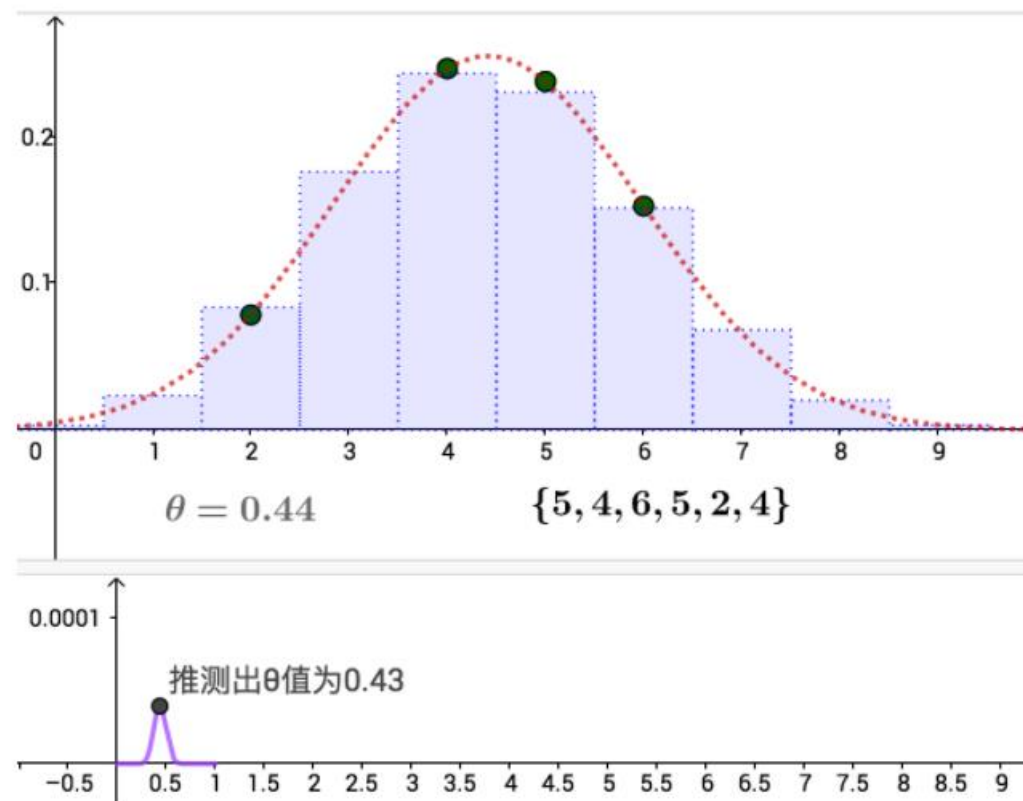
$$x_1, x_2, \dots, x_6 = \{5, 4, 6, 5, 2, 4\}$$

$$L(\theta) = [C_{10}^5 \theta^5 (1-\theta)^5] [C_{10}^4 \theta^4 (1-\theta)^6] \dots$$

最大似然函数步骤：

- 写似然函数 $L(\theta)$
- 取对数 $\ln L(\theta)$
- 求偏导 $\frac{\partial \ln L(\theta)}{\partial \theta} = 0$

$$\hat{\theta} \longrightarrow \max L(\theta)$$



正态分布下的最大似然估计

概率密度函数 $f(x) = \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(x-\mu)^2}{2\delta^2}}$

写似然函数 $L(\mu, \delta^2) = \prod_{i=1}^N f(x_i | \theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(x_i-\mu)^2}{2\delta^2}} = (2\pi\delta^2)^{-\frac{n}{2}} e^{-\frac{1}{2\delta^2} \sum_{i=1}^n (x_i-\mu)^2}$

取对数 $\ln L(u, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - u)^2$

求偏导
$$\begin{cases} \frac{\partial \ln L(u, \sigma^2)}{\partial u} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - u) = 0 \\ \frac{\partial \ln L(u, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - u)^2 = 0 \end{cases} \longrightarrow \begin{cases} u^* = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ \sigma^{*2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

最大后验估计MAP

最大后验估计与最大似然估计相似

不同点：在于估计 θ 的函数中允许加入一个先验 $p(\theta)$ ，此时不是要求似然函数最大，而是要求由贝叶斯公式计算出的整个后验概率最大

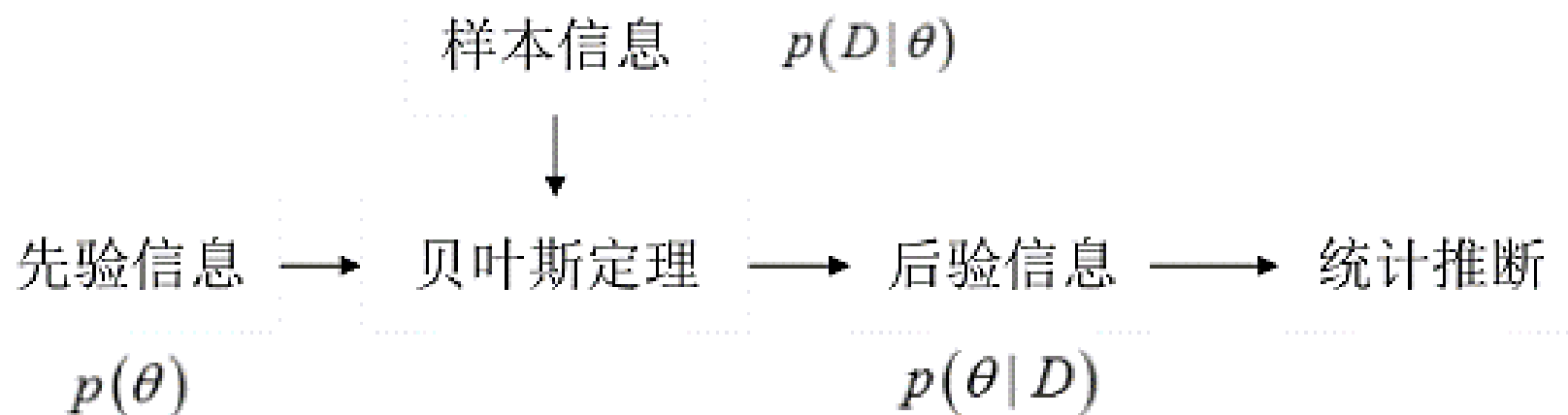
$$\begin{aligned}\hat{\theta}_{MAP} &= \arg \max \frac{p(X | \theta) p(\theta)}{p(X)} \\ &= \arg \max p(X | \theta) p(\theta) \\ &= \arg \max \{L(\theta | X) + \log p(\theta)\} \\ &= \arg \max \left\{ \sum_{x \in X} p(\theta | x) + \log p(\theta) \right\}\end{aligned}$$

贝叶斯估计

本质：贝叶斯估计的本质是通过贝叶斯决策得到参数 θ 的最优估计，使得总期望风险最小

区别：不直接估计参数的值，而是允许参数服从一定概率分布

基本原理：



贝叶斯估计

$P(\theta)$ 是待估计参数 θ 的先验概率, θ 取值与样本集 $D = \{x_1, x_2, \dots, x_N\}$ 有关, $\lambda(\hat{\theta}, \theta)$ 是 $\hat{\theta}$ 作为 θ 的估计量时的损失函数。

定义样本 x 下的条件风险为 $R(\hat{\theta} | x) = \int \lambda(\hat{\theta}, \theta) p(\theta | x) d\theta$

则有
$$R = \int_{E^d} R(\hat{\theta} | x) p(x) dx$$

又 $R(\hat{\theta} | x)$ 非负, 则由贝叶斯决策知, 求 R 最小即求 $R(\hat{\theta} | x)$ 最小。

即: $\theta^* = \arg \min_{\hat{\theta}} R(\hat{\theta} | x)$

可得最优估计: $\theta^* = \int \theta p(\theta | x) d\theta$

贝叶斯估计步骤

贝叶斯估计的基本步骤(基于平方误差损失函数)

1. 确定参数 θ 的先验分布 $p(\theta)$;

2、由样本集 $D = \{x_1, x_2, \dots, x_N\}$ 求出样本联合分布 $P(D | \theta)$, 它

是 θ 的函数:
$$P(D | \theta) = \prod_{n=1}^N p(x_n | \theta)$$

3、利用贝叶斯公式, 求 θ 的后验分布
$$P(\theta | D) = \frac{P(D | \theta)p(\theta)}{\int_{\theta} P(D | \theta)p(\theta)d\theta}$$

4. 求出贝叶斯估计值
$$\hat{\theta} = \int_{\theta} \theta P(\theta | D)d\theta$$

非参数估计

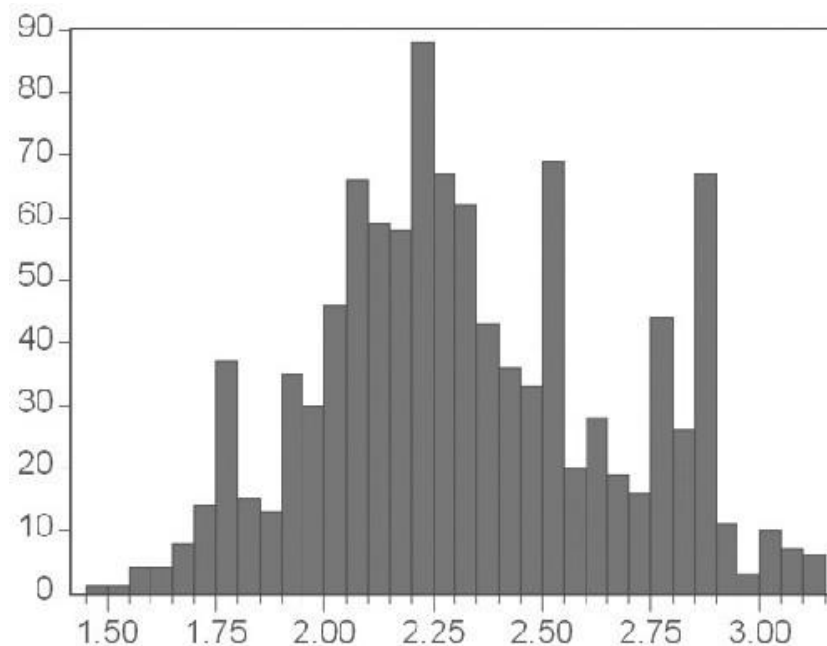
参数估计：已知样本类别和函数模型（假设一个模型），根据样本估计模型中的未知参数

非参数估计：已知样本类别，未知函数模型（不假设模型），直接从样本中学习估计模型

非参数估计 —— 直方图

直方图方法

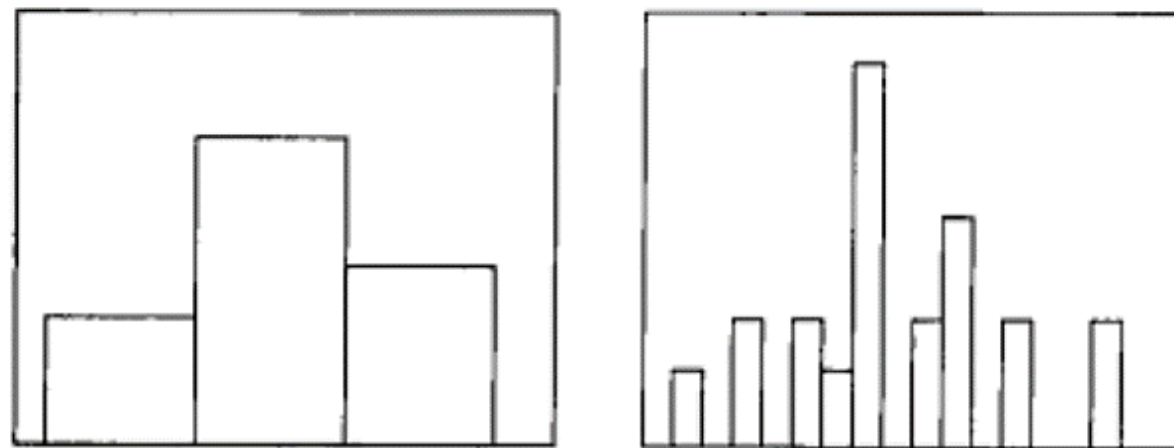
- 把 x 的每个分量分成 k 个等间隔小窗,
($x \in E^d$, 则形成 k^d 个小舱)
- 统计落入各个小舱内的样本数 q_i
- 相应小舱的概率密度为: $q_i / (NV)$
(N : 样本总数, V : 小舱体积)



直方图

小区间的大小选择与估计的效果是密切相连的。

- 如果区域选择过大，会导致最终估计出来的概率密度函数非常粗糙；
- 如果区域的选择过小，可能会导致有些区域内根本没有样本或者样本非常少，这样会导致估计出来的概率密度函数很不连续



(a) 小窗过宽

(b) 小窗过窄

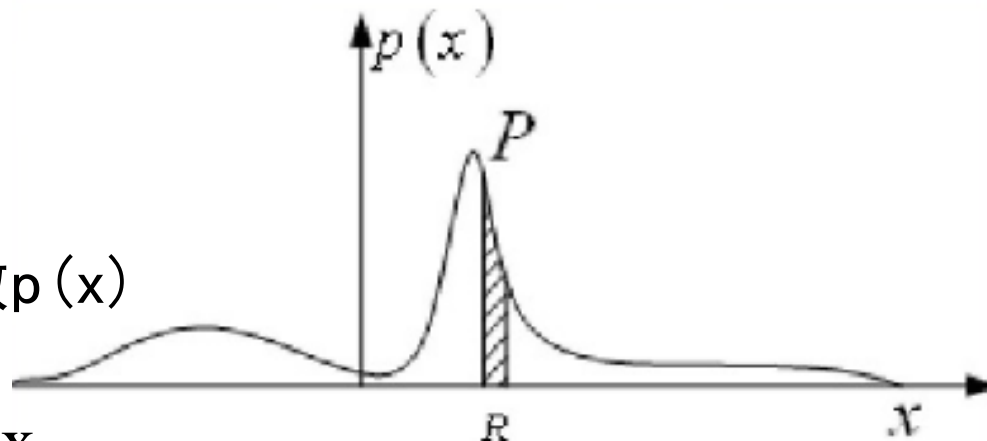
小窗宽度对直方图估计的影响示意

非参数估计思路

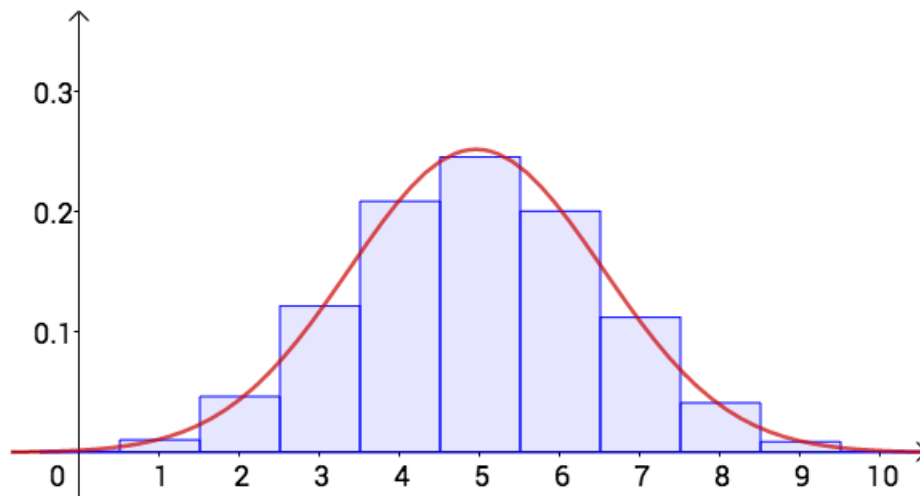
非参数概率密度估计的核心思路：

因此, 可以通过统计概率 P 来估计概率密度函数 $p(x)$

一个向量 x 落在区域 R 中的概率 P 为 $P = \int_R p(x) dx$



用直方图逼近概率密度函数



非参数估计思路

假设 N 个样本的集合 $X = \{x_1 \cdots x_N\}$ 是根据概率密度函数为 $p(x)$ 的分布独立抽取得到的
那么, 有 k 个样本落在区域 R 中的概率服从二项式 $P_k = C_N^k p^k (1-p)^{N-k}$

k 的期望值为 $E[k] = NP$

对 P 的估计 $\hat{P} = \frac{k}{N}$ 当 $N \rightarrow \infty$ 时, 估计是非常精确的

非参数估计

- 假设 $p(x)$ 是连续的, 且 R 足够小使得 $p(x)$ 在 R 内几乎没有变化
- 令 R 是包含样本点 x 的一个区域 a , 其体积为 V , 设有 N 个训练样本, 其中有 k 落在区域 R 中, 则可对概率密度作出一个估计:

$$P = \int_R p(x) dx = p(x)V \quad \hat{P} = \frac{k}{N}$$

$$\hat{P}(x) = \frac{k}{NV} \quad \text{对 } p(x) \text{ 在小区域内的平均值的估计}$$

非参数估计收敛问题

收敛性问题: 样本数量 N 无穷大时, 估计的概率函数是否收敛到真实值?

实际中, $\hat{p}(x)$ 越精确, 要求: $R \rightarrow 0$ $\lim_{N \rightarrow \infty} \hat{p}_N(x) = p(x)$

实际中, N 是有限的

当 $R \rightarrow > 0$ 时, 绝大部分区间没有样本: $\hat{p}(x) = 0$

如果侥幸存在一个样本, 则: $\hat{p}(x) = \infty$

非参数估计理论结果

设有一系列包含X的区域 $R_1, R_2, \dots, R_n, \dots$,

对 R_1 采用1个样本进行估计, 对 R_2 用2个, R_n 包含 k_n 个样本。 V_n 为 R_n 的体积。

$$P_n(x) = \frac{k_n / N}{V_n} \quad \text{为 } P(x) \text{ 的第 } n \text{ 次估计}$$

如果要求 $P_n(x)$ 能够收敛到 $P(x)$, 那么必须满足:

$$\lim_{n \rightarrow \infty} V_n = 0$$

$$\lim_{n \rightarrow \infty} K_n = \infty$$

$$\lim_{n \rightarrow \infty} \frac{K_n}{n} = 0$$

$$\lim_{n \rightarrow \infty} V_n = 0$$

$$\lim_{n \rightarrow \infty} K_n = \infty$$

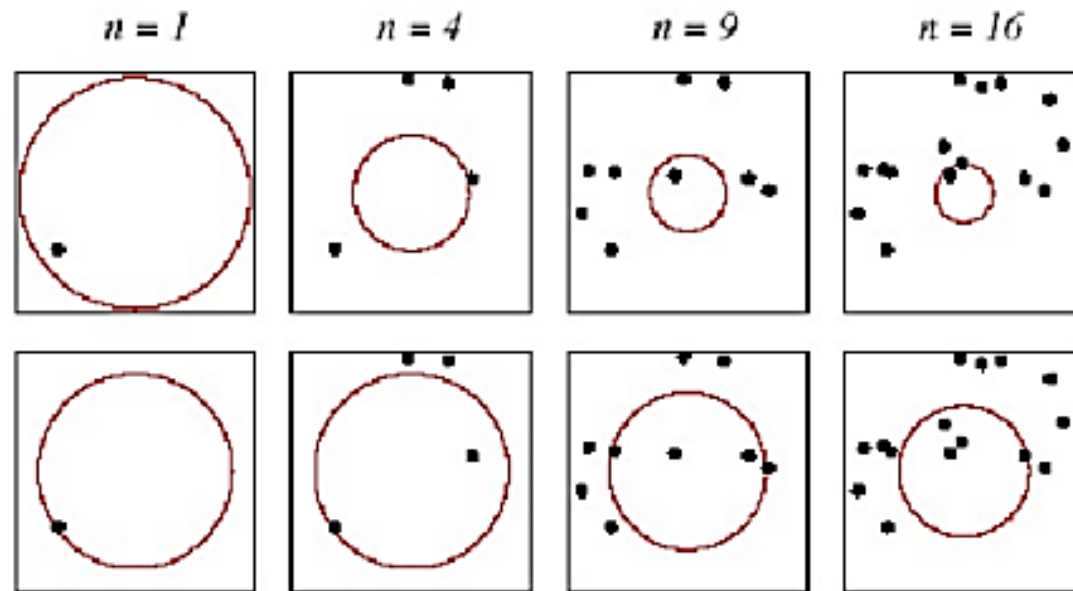
$$\lim_{n \rightarrow \infty} \frac{K_n}{n} = 0$$

选择 V_n

$$V_n = 1/\sqrt{n}$$

选择 K_n

$$k_n = \sqrt{n}$$



有两种途径获得区域:

1. 选择 V_n , (比如 $V_n = \frac{1}{\sqrt{n}}$), 同时对 k_n 和 $\frac{k_n}{n}$ 加限制以保证收敛

——Parzen 窗法

2. 选择 k_n , (比如 $V_n = \frac{1}{\sqrt{n}}$), V_n 为正好包含 x 的 k_n 个近邻。

—— k_N 近邻估计



Kn近邻估计

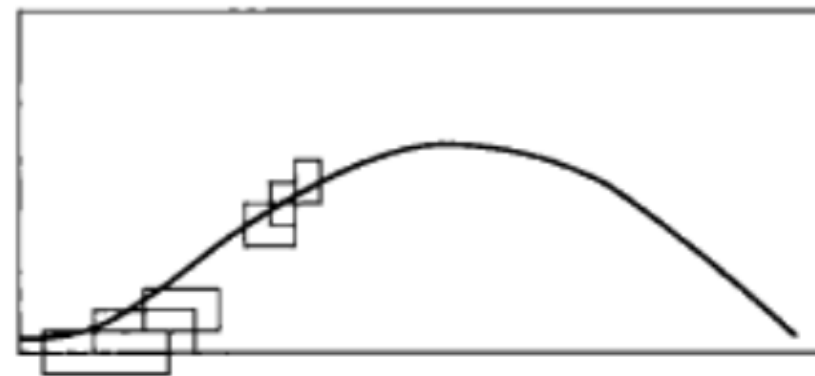
N近邻估计就是一种采用可变大小的小舱的密度估计方法, 基本做法是:

根据总样本确定一个参数 K_N , 即在总样本数为N时我们要求每个小舱内拥有的样本个数。

- 在样本密度比较高的区域小舱的体积就会比较小
- 而在密度低的区域则小舱体积自动增大



这样就能够比较好地兼顾在高密度区域估计的分辨率和在密度区域估计的连续性。



窗口宽度和样本密度关系示意图

Parzen窗法

思想：将核函数看作一个窗，统计样本落在窗内的个数来估计概率密度函数

假定区域 R_n 是一个 d 维超立方体， h_n 是超立方体一条边的长度，则有体积

$$V_n = h_n^d$$

窗(核)函数定义为

$$k_N = \sum_{i=1}^N \psi\left(\frac{x - x_i}{h}\right)$$

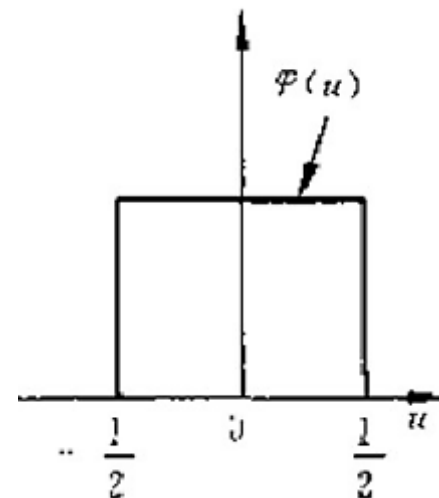
估计的密度函数为

$$\hat{p}(x) = \frac{1}{NV} \sum_{i=1}^N \psi\left(\frac{x - x_i}{h}\right)$$

Parzen窗法的形式

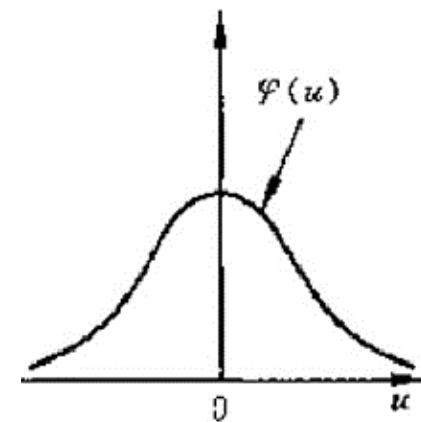
方窗

$$k(x, x_i) = \begin{cases} \frac{1}{h_d}, & \text{若 } |x^j - x_i^j| \leq h/2, j=1, 2, \dots, d \\ 0, & \text{其他} \end{cases}$$



正态窗

$$k(x, x_i) = \frac{1}{\sqrt{(2\pi)^d \rho^{2d} |Q|}} e^{-\frac{1}{2} \frac{((x-x_i)^T Q^{-1} (x-x_i))}{\rho}}$$



超球窗

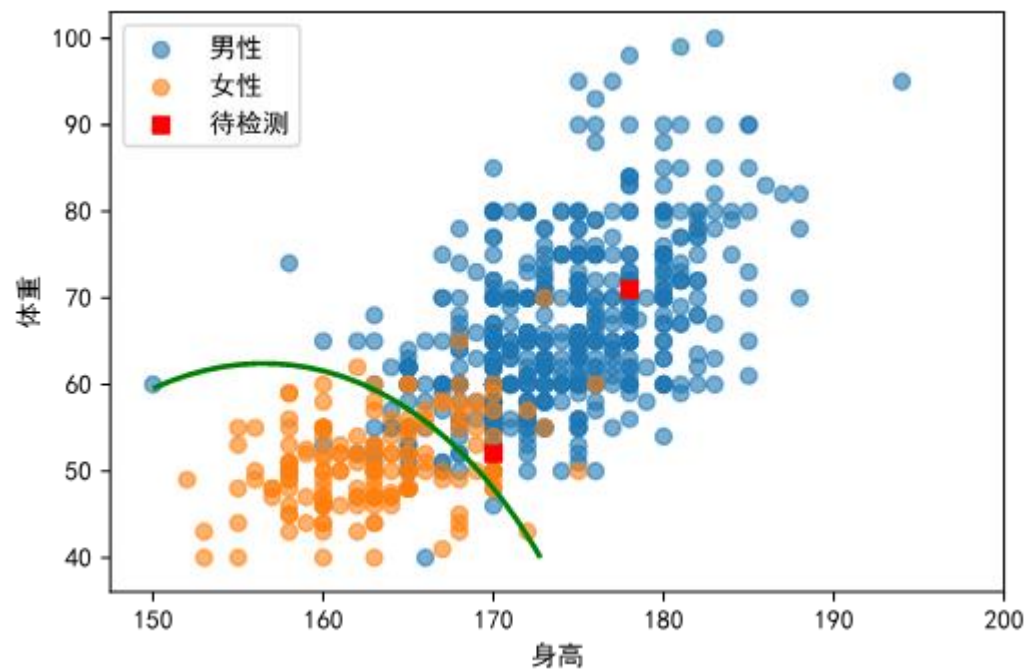
$$k(x, x_i) = \begin{cases} V^{-1}, & \text{若 } |x - x_i| \leq \rho \\ 0, & \text{其他} \end{cases}$$

作业二：利用贝叶斯分类器实现基于身高体重的性别分类

- 推荐编程环境: Anaconda+Jupyter notebook+pytorch

安装教程: [点这](#)

- 要求可视化决策面以及分类结果



参考文献

1. D. M. Chickering, D. Heckerman, C. Meek. Large-Sample Learning of Bayesian Networks is NP-Hard[J]. Journal of Machine learning Research, 2004:1287-1330
2. N. Friedman, D. Geiger and M. Goldszmidt. Bayesian Network Classifiers[J]. Journal of Machine learning Research, 1997:919-931
3. D. Grossman, P. M. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood[C]. Proceedings of the International Conference on Machine Learning (ICML), 2004.

相关论文会放到课程网页中，如有需要请自行下载。