

# 模式识别

推理和决策 (Inference and Decision)

密度估计 (Density Estimation)

吴建鑫

南京大学计算机系 & 人工智能学院, 2020

# 目标

- ✓ 理解并掌握各名词的含义，最低要求是使用这些名词的时候不能混淆
- ✓ 掌握基本的点估计方法，能在不查表（简单情况）或查表（复杂情况）下自主推导估计结果
- ✓ 掌握Bayesian相关方法的概念
- ✓ 理解、掌握非参数估计的方法
- ✓ 提高目标
  - **进一步**能通过独立阅读、了解Bayesian方法
  - **进一步**能通过独立阅读、了解非参数估计（如解决计算效率、核的选择等）

# 推理和决策

---

Inference and Decision

# 从统计学statistics的观点看

- ✓ 目的是得到映射:  $\mathcal{X} \mapsto \mathcal{Y}$ 
  - 数据分布  $p(\mathcal{X})$
  - 先验分布 prior distribution  $p(\mathcal{Y})$ 
    - *a priori*: Knowable without appeal to particular experience
    - a priori distribution: special meaning, do not misuse
  - 联合joint分布  $p(\mathcal{X}, \mathcal{Y})$
  - 类条件分布  $p(\mathcal{X} | \mathcal{Y} = i)$
  - 后验分布 posterior distribution  $p(\mathcal{Y} = i | \mathcal{X})$

# 如何表示/估计概率密度

- ✓ Parametric: 假设PDF服从某种函数形式  
functional form
  - 如高斯分布的函数形式，其包含若干参数
  - 当指定其所有参数值之后，PDF就完全确定
  - 不同的概率分布由不同的参数值决定
  - 估计PDF就是**估计参数**parameter estimation!
  - 所以叫参数估计

# 非参数估计

- ✓ Non-parametric: 不假设PDF是任何已知形式的函数
  - 从直观上更合理
  - 那么, 如何估计?
    - 使用训练数据直接估计空间中任意点的密度
    - $p(\mathbf{x}|D)$  - 后面具体讲
- ✓ 非参数不代表无参数!
  - 实际上是可以允许**无穷多**的参数
  - 而参数估计的参数个数是有限的
- ✓ 参数估计与非参数估计
  - Parametric and non-parametric estimate
  - 可能参数化估计与非参数化估计更切合其含义

# 推理与决策 inference & decision

## ✓ 生成模型和判别模型

- Generative (probabilistic) models: 估计 $p(\mathbf{x}|y = i)$ 和 $p(\mathbf{x})$ 
  - 然后用贝叶斯定理求 $p(y = i|\mathbf{x})$
- Discriminative (probabilistic) models: 直接估计 $p(y = i|\mathbf{x})$

## ✓ 这些模型分为两个步骤:

- 推理inference: 估计各种密度函数
- 决策decision: 根据估计得到的PDF对任意的 $\mathbf{x}$ 给出输出

# 参数估计

点估计 point estimation

贝叶斯估计 Bayesian estimation

KDE

# 以高斯分布为例

- ✓ 假设  $x \sim N(\mu, \sigma^2)$ , 从数据  $D = \{x_1, \dots, x_n\}$  估计
  - 数据独立同分布 i. i. d. (independently identically distributed)
- ✓ 参数记为  $\theta$ , 这里  $\theta = (\mu, \sigma)$ , 如何估计? 形式化?
- ✓ 一种直觉: 如果有两个不同的参数  $\theta_1$  和  $\theta_2$ 
  - 假设  $\theta$  是参数的真实值, 似然 (likelihood) 是
$$p(D|\theta) = \prod_i p(x_i|\theta) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
  - 若  $p(D|\theta_1) > p(D|\theta_2)$ , 该选择哪个?

# 易混淆的表示法 notation

- ✓ 目前  $\theta$  不是随机变量，所以  $p(D|\theta)$  不是条件分布
  - $D$  固定， $\theta$  是变量， $p(D|\theta)$  是  $\theta$  的函数，不是一个 PDF！
  - $p(x_i|\theta)$  是一个 PDF，因为  $\theta$  不是随机变量，这不是一个条件分布，只是习惯上这么写，表明这个分布依赖于参数  $\theta$  的值， $x_i$  是 PDF 的变量
- ✓ 较好的表示法：定义似然函数 likelihood function
  - $\hat{l}(\theta) = p(D|\theta) = \prod_i p(x_i|\theta)$  （或者  $x_i$ ）
- ✓ 为了方便，定义对数似然函数 log-likelihood function
  - $l(\theta) = \ln p(D|\theta) = \sum_i \ln p(x_i|\theta)$

# 最大似然估计

- ✓ Maximum likelihood estimation, MLE

$$\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

- ✓ 高斯分布的最大似然估计

- 参数为  $(\boldsymbol{\mu}, \Sigma)$ , 数据为  $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$

- 练习: 通过对  $l(\boldsymbol{\theta})$  求导发现最佳的参数值, 可以查表,  
(猜一猜?)

$$\boldsymbol{\mu}^* = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

$$\boldsymbol{\Sigma}^* = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}^*)(\mathbf{x}_i - \boldsymbol{\mu}^*)^T$$

# 最大后验估计及其他

- ✓ Maximum a posteriori estimation, MAP
  - $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \hat{l}(\boldsymbol{\theta}) p(\boldsymbol{\theta})$
  - 即假设参数 $\boldsymbol{\theta}$ 也是随机变量，存在着先验分布
- ✓ 与MLE的关系
  - 假设我们对 $\boldsymbol{\theta}$ 一无所知，那么应该怎样设定 $p(\boldsymbol{\theta})$ ？
  - noninformative prior时，MLE等价于MAP
    - 若 $\boldsymbol{\theta}$ 是离散的随机变量，离散的均匀分布，  $p(\boldsymbol{\theta}) = \frac{1}{N}$
    - 若 $\boldsymbol{\theta}$ 是有限区间 $[a, b]$ 的连续随机变量，  $p(\boldsymbol{\theta}) = \frac{1}{b-a}$
    - 若 $\boldsymbol{\theta}$ 是 $(-\infty, +\infty)$ 上的连续随机变量， ?
    - 假设 $p(\boldsymbol{\theta}) = \text{const}$ ，称为improper prior

# 参数估计的一些性质

- ✓ 如果只有一个样例，参数估计会怎么样？
- ✓ 样例越多，估计越准！
- ✓ 渐进性质 asymptotic property：研究  $n \rightarrow \infty$  时的性质，如
  - 一致性 consistency：随样本容量增大收敛到参数真值的估计量
- ✓ 其他性质如
  - 无偏估计 unbiased estimate：指估计量的期望和被估计量的真值相等
- ✓ 进一步阅读：关于一致和无偏

# 贝叶斯参数估计

## ✓ Bayesian parameter estimation

- MLE: 视 $\theta$ 为固定的参数, 假设存在一个最佳的参数  
(或参数的真实值是存在的), 目的是找到这个值
  - MAP: 视 $\theta$ 为一个随机变量, 存在分布 $p(\theta)$ , 将其影响(先验分布)代入, 但仍然假设存在最优的参数
  - 以上均称为点估计point estimation
- ✓ 在贝叶斯观点中,  $\theta$ 是一个分布/随机变量, 所以估计应该是估计一个分布, 而不是一个值(点)!
- $p(\theta|D)$ : 这是贝叶斯参数估计的输出, 是一个完整的分布, 而不是一个点

# 高斯分布参数的贝叶斯估计

✓ 参数 $\theta$ 的先验分布 $p(\theta)$ , 数据 $D = \{x_1, \dots, x_n\}$ , 估计 $p(\theta|D)$ 。这里假设单变量, 只估计 $\mu$ , 方差 $\sigma^2$ 已知

- 第一步: 设定 $p(\mu)$ 的参数形式:  $p(\mu) = N(\mu_0, \sigma_0^2)$ , 目前假设参数 $\mu_0, \sigma_0^2$ 已知
- 第二步: 贝叶斯定理和独立性得到 $p(\mu|D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu} = \alpha p(D|\mu)p(\mu) = \alpha \prod_{i=1}^n p(x_i|\mu)p(\mu)$
- 第三步, 应用高斯的性质, 进一步得到其解析形式
  - 讲义第13章
  - 注意这里所有 $p(\cdot)$ 都是合法的密度函数

# 解的形式

$$p(\mu|D) = N(\mu_n, \sigma_n^2)$$

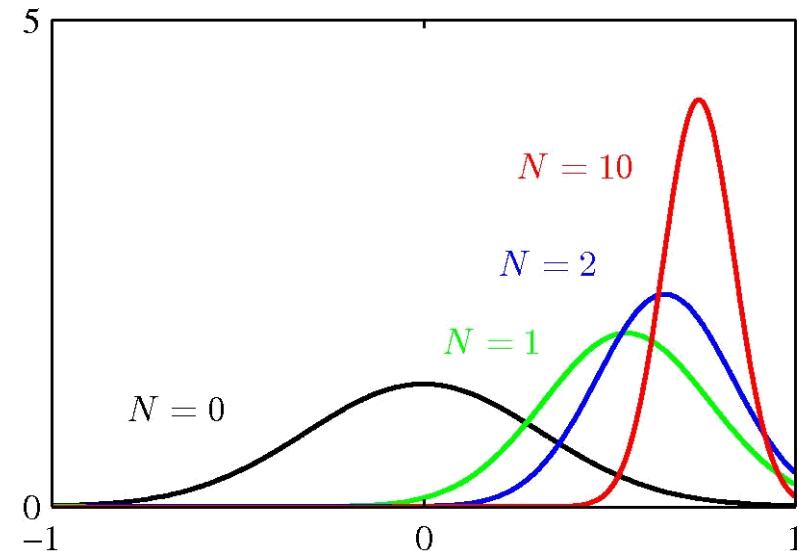
✓ 均值为  $\mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \mu_{\text{ML}}$

- 其中  $\mu_{\text{ML}}$  为MLE的估计值，即  $\mu_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n x_i$
- ✓ 方差为  $\sigma_n^2$ ，其值由如下公式确定：  $\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$ ,

或者为了便于记忆

$$(\sigma_n^2)^{-1} = (\sigma_0^2)^{-1} + \left(\frac{\sigma^2}{n}\right)^{-1}$$

✓ 先验和数据的综合！



图片来自PRML第2章

# 贝叶斯的进一步讨论

- ✓ 共轭先验conjugate prior
  - 若 $p(\mathbf{x}|\boldsymbol{\theta})$ , 存在先验 $p(\boldsymbol{\theta})$ , 使得 $p(\boldsymbol{\theta}|D)$ 和 $p(\boldsymbol{\theta})$ 有相同的函数形式, 从而简化推导和计算
  - 如高斯分布的共轭先验分布仍然是高斯分布
- ✓ 优缺点:
  - 理论上非常完备, 数学上很优美
  - 推导困难 (怎样求任意分布的共轭? 怎样用于决策?  
 $\mu_0$ 的prior) 、 计算量极大 (需要很多积分)
  - 在数据较多时, 学习效果常不如直接用discriminant function

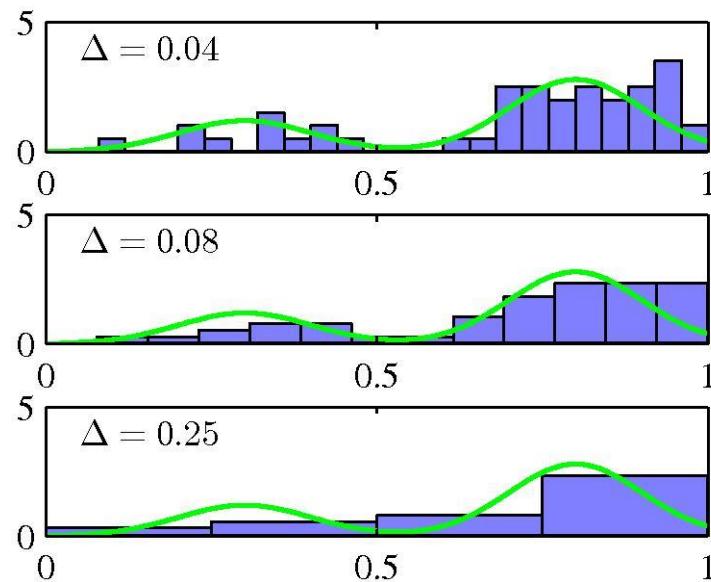
# 非参数估计：介绍

---

Non-parametric estimation: an introduction

# 非参数估计

- ✓ 常用的参数形式基本都是单模single modal的，不足以描述复杂的数据分布：即应该直接以训练数据自身来估计分布
  - 例如直方图histogram，基于计数counting



有很多问题：

- 多维怎么办？
- 怎么确定bin的个数？
- 连续？
- 需要保存数据吗？

# 维数灾难

## ✓ Curse of dimensionality

- 以直方图为例，需要保存的参数是什么？
  - 如果每维 $n$ 个参数，那么 $d$ 维应该保存多少个参数？
  - 如果 $n = 4$ ,  $d = 100$ , 那么应该保存多少个参数？
- 
- $4^{100} = 2^{200} \approx 10^{60}$ ! 那么，需要多少样例来学习?
    - $1G = 10^9$

## ✓ 不仅局限于直方图、非参数估计，在参数估计、以及很多其他统计学习方法中都是如此

# Kernel density estimation

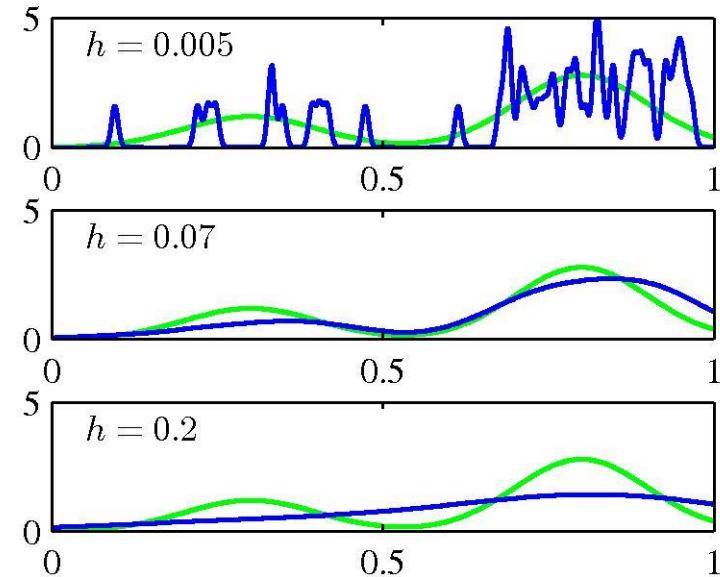
- ✓ KDE，注意这里的kernel与上一章中SVM中的核含义不完全一致，其要求的条件也不完全一致
- ✓ 举例：Parzen window（一维，使用高斯核）

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

问题：

- 连续吗？
- 多维：多个维度乘积
- 需要保存数据吗？
  - 存储和计算实际代价大
  - 无穷多的参数
- 怎么确定  $h$ ？

图片来自PRML第2章



① h和bin的个数是超参数

② h决定有一定理论，常用交叉验证

③ 多维计算量大

## 决策

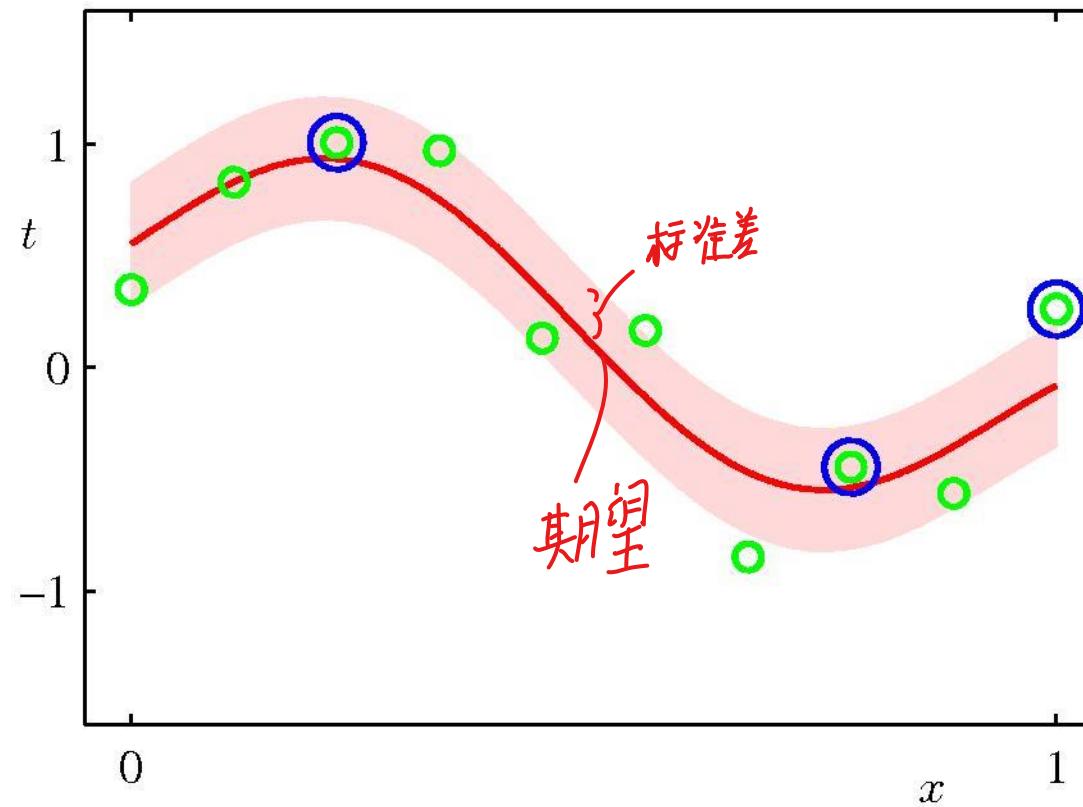
---

Decision, 或预测 prediction

# 决策、预测

- ✓ 当inference完成之后，如果给定输入 $x$ 
  - 应当给出什么样的输出？
  - 怎么给出？
- ✓ 点估计：
  - 根据参数得到后验概率 $p(y|x; \theta)$
  - 根据其给出结果（如分类，如何输出？）
- ✓ Bayesian decision
  - 输出，也是一个随机变量，称为预测分布predictive distribution
  - 结果通常根据其期望决定，同时还可以给出方差

# Bayesian prediction例子



图片来自PRML第7章

# 点估计下的例子

- ✓ 在0-1风险时，选择后验概率最大的那个类别

$$\operatorname{argmax}_i p(y = i | \boldsymbol{x}; \boldsymbol{\theta})$$

- ✓ 在discriminant function观点下，可以定义函数

$$g_i(\boldsymbol{x}) = p(y = i | \boldsymbol{x}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{x} | y = i; \boldsymbol{\theta}) p(y = i)}{p(\boldsymbol{x}; \boldsymbol{\theta})}$$

- ✓ 或者定义为  $g_i(\boldsymbol{x}) = p(\boldsymbol{x} | y = i; \boldsymbol{\theta}) p(y = i)$ ，为什么？

- ✓ 或者定义为

$$g_i(\boldsymbol{x}) = \ln(p(\boldsymbol{x} | y = i; \boldsymbol{\theta})) + \ln(p(y = i))$$

# 高斯分布条件下的判别函数

✓ 作业：假设 $p(\mathbf{x}|y = i) = N(\boldsymbol{\mu}_i, \Sigma)$ ，即在一个2分类问题中，各类条件分布都是高斯分布，虽期望不同，但协方差矩阵是一样的。若同时假设两类的先验概率均为0.5，那么

- $g_1$ 和 $g_2$ 的最简化的表达式是什么？
- 在两类分类问题中，可以使用单个判别函数而不是两个来进行分类。对此问题，其单个判别函数是？
- 和FLD的关系？

# 进一步的阅读

✓ 如果对本章的内容感兴趣，可以参考如下文献

- All of statistics, All of Nonparametric Statistics: 两本书
- PRML—这本书非常Bayesian!
- ESL
- EM算法和GMM
  - 参数估计的常用优化算法
  - 讲义第十四章
- 共轭分布见第十三章讲义的习题