華東師範大學
EAST CHINA NORMAL UNIVERSITY

数据科学与工程算法基础

Algorithm Foundations of Data Science and Engineering

第六章 EM算法

$$(1+x)^n = 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \cdots$$

课程提纲

Content

课程提纲

Content

参数估计（一）

- 抛一枚不均匀的硬币 10 次

H H H H T H H H H H

- 正面朝上的概率是多少?
- 这是一个典型的参数估计问题
  - 最大似然估计
  - 矩估计
  - $\hat{p} = \dfrac{\#\text{heads}}{\#\text{flips}}$
- 和我们遇到的很多情况一样，利用数据拟合模型参数

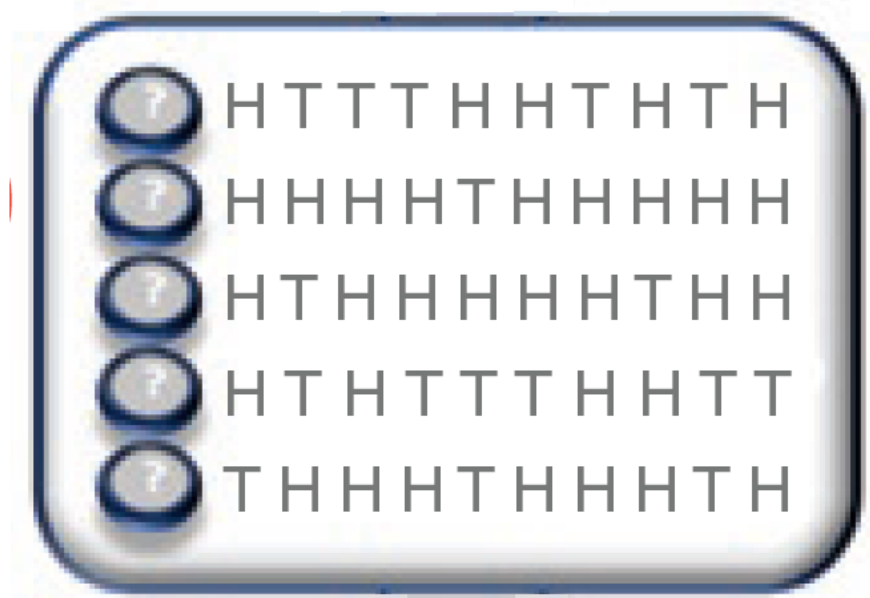
参数估计（二）

	Coin A	Coin B
H T T T H H T H T H		5 H, 5 T
H H H H T H H H H H	9 H, 1 T	
H T H H H H H T H H	8 H, 2 T	
H T H T T T H H T T		4 H, 6 T
T H H H T H H H T H	7 H, 3 T	
	24 H, 6 T	9 H, 11 T

- 罐子中装有两枚可分辨的硬币
  - 每次从罐子中挑其中一个硬币
  - 抛币 10 次
- 试估计两枚硬币正面朝上的概率分别是多少?

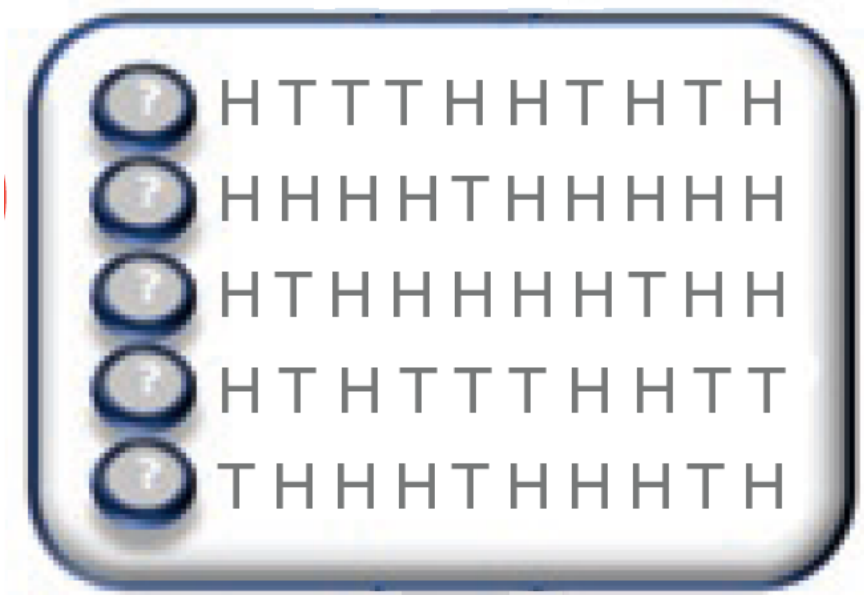
$$\widehat{p_A} = \frac{\text{#heads of A}}{\text{#flips of A}} \qquad \widehat{p_B} = \frac{\text{#heads of B}}{\text{#flips of B}}$$

参数估计（三）

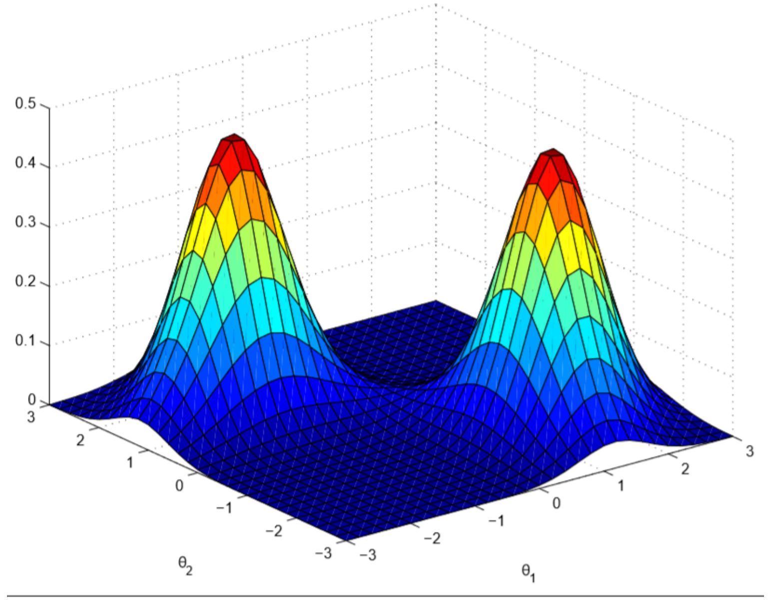


- 罐子中装有两枚不可分辨的硬币
  - 每次从罐子中挑其中一个硬币
  - 抛币 10 次
- 试估计两枚硬币正面朝上的概率分别是多少?

混合模型



混合贝努里模型



混合高斯模型

- 概率密度或者概率质量 $p(x)$ 可能是“多峰”的
  - 可将其建模为单峰分布的混合体
  - 不同单峰分布可以看作不同的总体（如男性和女性）
  - 因此，这类模型被称之为混合模型

隐变量

- 不可观测变量被成为**隐变量**
  - **虚构变量：** 旨在为数据生成过程提供一些简化，例如语音识别模型、混合模型
  - **无法观测变量：** 现实世界很多现象很难或无法测量，例如恒星的温度、疾病的原因、进化祖先
  - **缺失变量：** 由于某些原因导致数据无法获取，例如传感器故障、对象失联
- 离散隐变量可看作为数据标签，即混合物模型
- 如何解决包含隐变量的模型参数估计问题
  - 混合模型
  - 缺失数据模型

课程提纲

Content

什么是似然?

新手

拥有丰富经验的猎手

- 抛一枚**均匀**硬币 10 次，结果如下：

H H H H T H H H H H

你还相信这是一枚均匀的硬币吗？你觉得正面朝上的概率可能是 0.6 或者 0.7 吗?

似然函数

- 假设独立样本 $X_1, \ldots, X_n$ 的联合密度或概率质量函数为 $f(x \mid \theta)$，给定样本观测值 $x_1, \ldots, x_n$，参数 θ 的似然函数（likelihood function）定义为

$$L(\theta \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

  - 似然常常被用作“概率”的同义词，但又有不同
    - ✓ 概率描述了已知参数时的随机变量的输出结果
    - ✓ 似然则描述已知随机变量输出结果时，未知参数的可能取值

- 为了方便估计参数，常常采用对数似然函数

$$l(\theta \mid x_1, \ldots, x_n) = \log f(x_1, \ldots, x_n \mid \theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

  - 参数 θ 的似然和对数似然函数具有相同的最值点

最大似然估计

- 固定样本观测值 x, 令 $\hat{\theta}(x)$ 为似然函数 $L(\theta|x)$ 的最大值点，即 $\hat{\theta}(x) = \arg\max_{\theta} L(\theta|x)$，则称 $\hat{\theta}(x)$ 为参数 θ 的最大似然估计 (Maximum Likelihood Estimator, MLE)
  - MLE 是最有可能观察到固定样本点的参数取值
  - 最大似然估计广泛应用于统计、机器学习等领域，如参数估计和检验等
  - 最大似然估计有很好的理论性质
    - ✓ 相合性：当样本容量 n 充分大时，估计量可以以任意的精确程度逼近被估计参数的真值
    - ✓ 泛函不变性：$f(x)$ 的最大似然估计和 $g(f(x))$ 的最大似然估计结果相同

如何求解 MLE?

- 如果似然函数关于参数 θ 是可微的（θ 可能是一个参数向量），则 MLE 的可能候选者可以通过下式求解
$$\frac{\partial L(\theta \mid x)}{\partial \theta_i} = 0 \, , i = 1, \ldots, k$$
  - 因为一阶导数为 0 只是最大值的必要条件，而不是充分条件
  - 因此，还需要进一步确定 MLE
    - ✓ 一阶导数的零点仅位于函数域内部的极点
    - ✓ 如果最值点出现在边界上，则一阶导数可能不为 0
    - ✓ 为了求解最值点，还必须单独检查参数的边界

伯努利分布的 MLE

- 令 $X_1, \ldots, X_n$ 为独立同分布的伯努利分布 Bern(p) 样本，则对数似然函数为 $l(p\,|\,x) = \sum_{i=1}^{n} \log P(X_i = x_i) = \sum_{i=1}^{n} [x_i \log p + (1 - x_i)\log(1 - p)]$
- 当 $0 < \sum_{i=1}^{n} x_i < n$ 时，根据 $\frac{\mathrm{d}l(p\,|\,x)}{\mathrm{d}p} = 0$，得到 $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
- 当 $\sum_{i=1}^{n} x_i = 0$ 或者 $\sum_{i=1}^{n} x_i = n$ 时，则

$$l(p\,|\,x) = \begin{cases} n\log(1-p), & \text{if } \sum_{i=1}^{n} x_i = 0 \\ n\log p, & \text{if } \sum_{i=1}^{n} x_i = n \end{cases}$$

- 由于 $l(p\,|\,x)$ 为 p 的单调函数，并且容易验证此时 $\hat{p} = \bar{x}$
- 因此，$\hat{p}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$

泊松分布的 MLE

- 令 $X_1, \ldots, X_n$ 为独立同分布的泊松样本，令 $L(\lambda \mid x)$ 为参数 λ 的似然函数，则 $l(\lambda \mid x) = \log L(\lambda \mid x) = \sum_{i=1}^{n} (x_i \log \lambda - \log x_i! - \lambda)$
- 求解方程 $\frac{\mathrm{d} l(\lambda \mid x)}{\mathrm{d}\lambda} = \sum_{i=1}^{n} (x_i \frac{\mathrm{d} \log \lambda}{\mathrm{d}\lambda} - \frac{\mathrm{d}\lambda}{\mathrm{d}\lambda}) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$，得到 $\hat{\lambda} = \bar{x}$，因此得到似然函数的极值点
- 此外，容易得到 $\frac{\mathrm{d}^2 l(\lambda)}{\mathrm{d}\lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} < 0$，因此极值点即为最值点。所以，参数 λ 的最大似然估计为 $\hat{\lambda}_{\mathrm{MLE}} = \bar{x}$

正态分布的 MLE

- 给定 $X_1, \ldots, X_n$ 为独立同分布的 $\mathrm{N}(\mu, \sigma^2)$ 样本，则

$$l(\mu, \sigma^2 \mid x) = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

- 分别令 $\dfrac{\partial l(\mu, \sigma^2)}{\partial \mu} = 0$ 和 $\dfrac{\partial l(\mu, \sigma^2)}{\partial \sigma} = 0$，得到

$$\begin{cases} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \\ -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \end{cases}$$

- 解得 $\hat{\mu} = \bar{x}$ 和 $\hat{\sigma}^2 = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

正态分布 MLE (续)

- 注意到对于任意 a，$\sum_{i=1}^{n}(x_i - \bar{x})^2 \le \sum_{i=1}^{n}(x_i - a)^2$
- 这意味着对于任意 θ，$e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2} \le e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
- 因此，最大似然函数变成一维情形，参考《Statistical Inference》一书例 7.2.11 和例 7.2.12
- 因此，$\hat{\mu}_{\mathrm{MLE}} = \bar{x}$ 和 $\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

George Casella, Roger L. Berger. Statistical Inference (Second Edition). Wadsworth Group, 2002.

课程提纲

Content

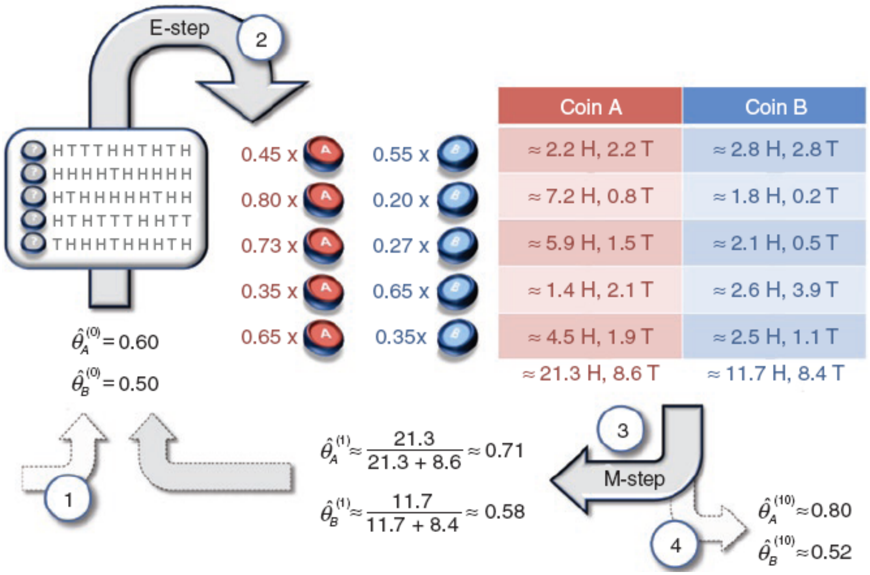
EM 算法引入

一次“猜”不准怎么办?

不断“猜”，直到准为止！

- 在混合贝努里模型中联合概率分布中的未知变量
  - 模型参数：θ_A 和 θ_B
  - 硬币标签：A 或者 B
- 如何估计模型参数?
  - 固定一个未知变量，推断另一个
    - ✓利用抛币结果“猜测”是 A 还是 B：固定参数，“猜”隐变量
    - ✓根据猜测结果推断参数估计：固定隐变量，推断参数

EM 算法示例

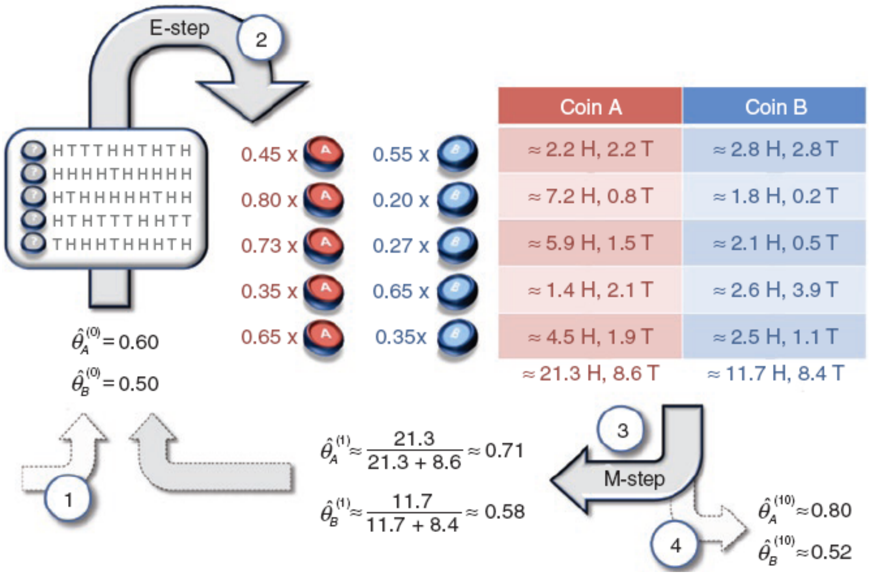


$$
\begin{aligned}
P(A \mid X_1) &= \frac{P(X_1 \mid A)P(A)}{P(X_1 \mid A)P(A) + P(X_1 \mid B)P(B)} \\
&= \frac{(0.4^5 \cdot 0.6^5) \cdot 0.5}{(0.4^5 \cdot 0.6^5) \cdot 0.5 + (0.5^5 \cdot 0.5^5) \cdot 0.5} \\
&\approx 0.45
\end{aligned}
$$

$$
P(B \mid X_1) = 1 - P(A \mid X_1) \approx 0.55
$$

- 迭代式更新
  - 依据后验概率“猜测”样本来自哪枚硬币?
  - 更新模型参数
  - 直到前后两次参数不再变化为止

EM 算法示例



$$\hat{\theta}_A^{(1)} = \frac{\sum_{i=1}^{n} x_i P(A \mid X_i)}{\sum_{i=1}^{n} P(A \mid X_i)} \approx \frac{2.124}{2.98} \approx 0.71$$

$$\hat{\theta}_B^{(1)} = \frac{\sum_{i=1}^{n} x_i P(B \mid X_i)}{\sum_{i=1}^{n} P(B \mid X_i)} \approx \frac{1.176}{2.02} \approx 0.58$$

- 迭代式更新
  - 依据后验概率“猜测”样本来自哪枚硬币?
  - 更新模型参数
  - 直到前后两次参数不再变化为止

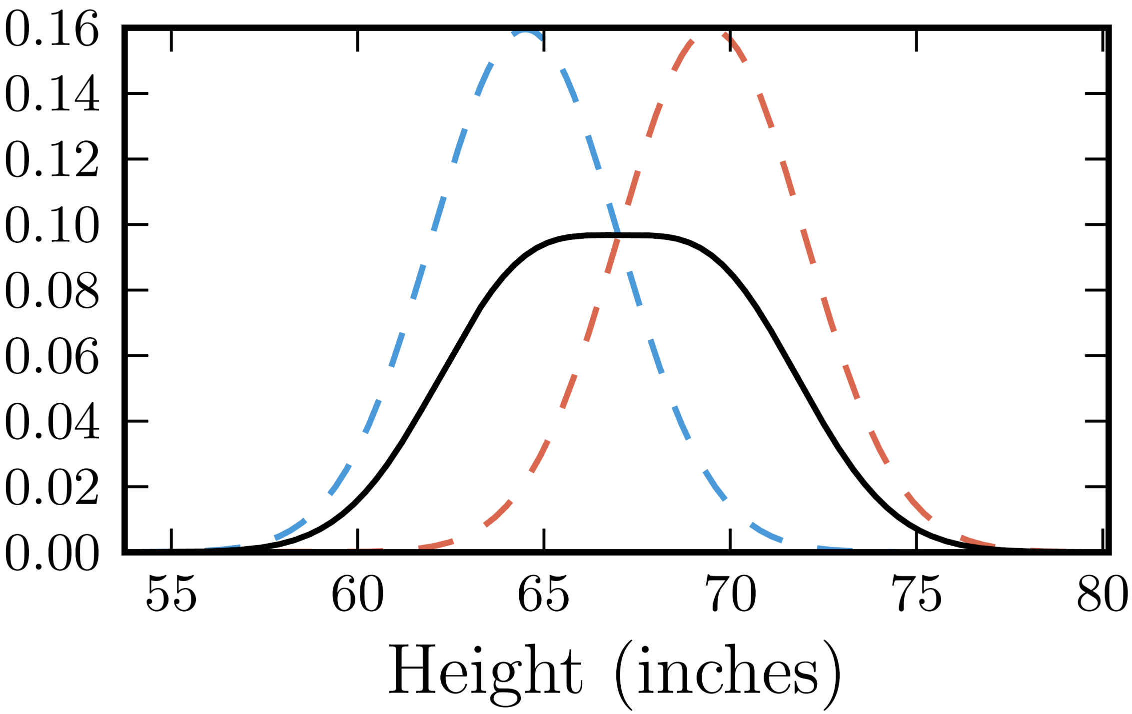
EM 算法迭代过程

- 令 $C_i \in \{A, B\}$ 标记硬币标签，联合概率表示为
  $P(X_i, C_i) = \left(P(X_i \mid A)P(A)\right)^{I_{C_i=A}}\left(P(X_i \mid B)P(B)\right)^{I_{C_i=B}}$
- 对数似然函数为 $l(\theta_1, \theta_2 \mid X_i, C_i) = \sum_{i=1}^{n} \left(I_{C_i=A} \log P(X_i \mid A)P(A) + I_{C_i=B} \log P(X_i \mid B)P(B)\right)$

- 给定参数初始值 $\hat{\theta}^{(0)} = (\hat{\theta}_1^{(0)}, \hat{\theta}_2^{(0)})$
- EM 算法执行流程
  - E 步：求条件期望
    - ✓ 为每一组样本确定属于硬币 A 或硬币 B 的后验概率
    - ✓ 计算 $E_{C|\hat{\theta}^{(t)}}[l(\theta_1, \theta_2 \mid X, C)]$
  - M 步：求最值点
    - ✓ 估计新一轮参数值 $\hat{\theta}^{(t+1)} = \arg\max_{\theta} E_{C|\hat{\theta}^{(t)}}[l(\theta_1, \theta_2 \mid X, C)]$

混合高斯模型 (GMM)

- 男生和女生的身高分别服从 $\mathrm{N}(\mu_M, \sigma^2)$ 和 $\mathrm{N}(\mu_F, \sigma^2)$ 的正态分布（已知方差 σ^2），随机选择 n 个人（未知男女），能否估计该混合模型的参数?



- 假设有样本点 $x_1, \ldots, x_n$，每个样本都有一个未观测的性别标签 $C_i \in \{M, F\}$，标记该样本点是男生或女生，因此类别标签 C_i 服从伯努利分布

GMM 建模

- 令 $C_i \in \{M, F\}$ 标记为男生或女生，样本点概率表示为
$$P(x_i \mid C_i) = \big(P(x_i \mid M)P(M)\big)^{I_{C_i=M}}\big(P(x_i \mid F)P(F)\big)^{I_{C_i=F}}$$
- 对数似然函数为
$$l(\mu_M, \mu_f \mid x, C) = \sum_{i=1}^{n} \big(I_{C_i=M} \log P(x_i \mid M)P(M) + I_{C_i=F} \log P(x_i \mid F)P(F)\big)$$
- 由于 $I_{C_i=M}$ 是伯努利随机变量，可以计算后验概率 $P(C_i = M \mid x_i)$：
$$\begin{aligned} P(C_i = M \mid x_i) &= \frac{P(x_i \mid C_i = M)P(C_i = M)}{P(x_i \mid C_i = M)P(C_i = M) + P(x_i \mid C_i = F)P(C_i = F)} \\ &= \frac{\pi_M N(x_i \mid \mu_M, \sigma^2)}{\pi_M N(x_i \mid \mu_M, \sigma^2) + \pi_F N(x_i \mid \mu_F, \sigma^2)} \\ &:= q(M) \end{aligned}$$
$$P(C_i = F \mid x_i) = 1 - q(M)$$

求解 GMM

- 初始化参数为 $\hat{\mu}^{(0)} = (\hat{\mu}_M^{(0)}, \hat{\mu}_F^{(0)})$， $\pi_M^{(0)} = \pi_F^{(0)} = 0.5$
- E 步骤：计算后验概率

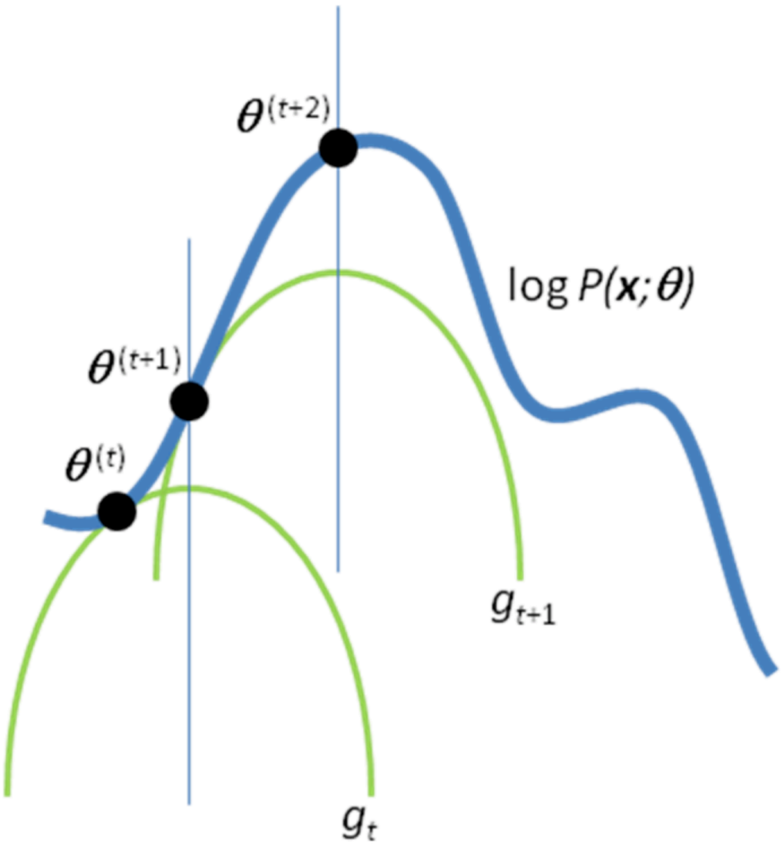
$$q_i(M)^{(t+1)} = P(C_i = M \mid x_i, \mu^{(t)}) = \frac{\pi_M^{(t)} N(x_i \mid \mu_M^{(t)}, \sigma^2)}{\pi_M^{(t)} N(x_i \mid \mu_M^{(t)}, \sigma^2) + \pi_F^{(t)} N(x_i \mid \mu_F^{(t)}, \sigma^2)}$$

其中 $\pi_M^{(t)} = 1 - \pi_F^{(t)} = \frac{1}{n}\sum_{i=1}^{n} q_i(M)^{(t)}$

- M 步骤：更新参数估计

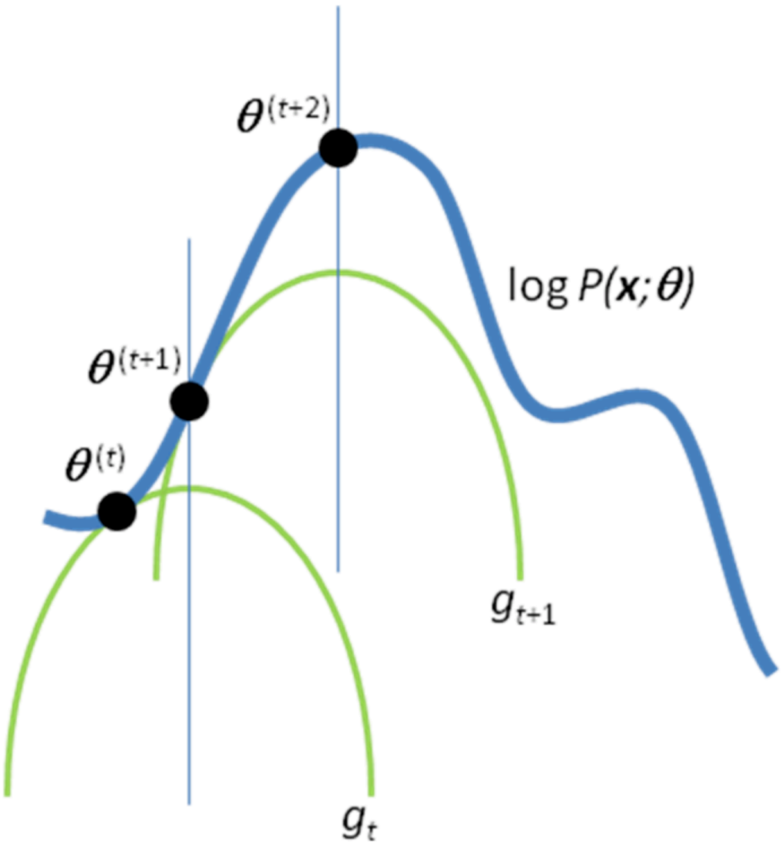
$$\mu_M^{(t+1)} = \frac{\sum_{i=1}^{n} x_i q_i(M)^{(t+1)}}{\sum_{i=1}^{n} q_i(M)^{(t+1)}}, \quad \mu_F^{(t+1)} = \frac{\sum_{i=1}^{n} x_i q_i(F)^{(t+1)}}{\sum_{i=1}^{n} q_i(F)^{(t+1)}}$$

EM 算法的收敛性



- E 步骤
  - 固定 $\theta^{(t)}$，构建目标函数 $\log P(\mathbf{x};\theta)$ 的下界函数 g_t，使得下界函数 g_t 与目标函数 $\log P(\mathbf{x};\theta)$ 在 $\theta^{(t)}$ 处相等
- M 步骤
  - 找到下界函数 g_t 的最大值，并更新 $\theta^{(t)}$ 到最大值 $\theta^{(t+1)}$
- 重复以上过程，直到收敛到目标函数的某个极大值点

EM 算法不能保证最优解



- EM 算法是否会收敛到全局最优解?
  - 答案是否定的，因为对数似然函数可能有多个极值点
  - EM 算法仅能保证收敛到其中一个极值点，即局部最优解
- 解决方案
  - 选择不同的初始值运行多次 EM 算法，选择其中最好（对数似然函数值最大）的结果

总结

- MLE vs. EM 算法
  - 解决的问题不同
  - 都为最好地“拟合”参数，但方法差别很大
- EM 算法应用
  - 缺失值处理
  - 噪音去除
  - 指数族混合模型
  - 主题模型：pLSA 和 LDA