

机器学习初步

——贝叶斯分类器与EM算法

李爽

助理教授，特别副研究员

计算机学院 数据科学与知识工程研究所

E-mail: shuangli@bit.edu.cn

Homepage: shuangli.xyz



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

本章的主要内容

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 极大似然估计
- EM算法
- 小结

一、贝叶斯决策论：什么是贝叶斯决策论？

- 贝叶斯决策论 (Bayesian decision theory) 是概率框架下实施决策的基本方法。
- 贝叶斯决策的前提：各类别的总体概率分布已知、类别数已知。
- 对于分类任务来说，在所有相关概率已知的理想情况下，贝叶斯决策论考虑如何基于这些概率和误判损失来选择最优的类别标记。

一、贝叶斯决策论：什么是贝叶斯决策论？

- 多分类任务为例：

假设有 N 种可能的类别标记，即 $\mathcal{Y} = \{c_1, c_2, \dots, c_N\}$ 。 λ_{ij} 是将一个真实标记为 c_j 的样本误分类为 c_i 所产生的损失。基于后验概率 $P(c_i | x)$ 可获得将样本 x 分类为 c_i 所产生的期望损失 (expected loss)，即在样本 x 上的“条件风险” (conditional risk)：

$$\mathcal{R}(c_i | x) = \sum_{j=1}^N \lambda_{ij} P(c_j | x)$$

- 任务：寻找一个判定准则 $h: \mathcal{X} \mapsto \mathcal{Y}$ 以最小化总体风险：

$$\mathcal{R}(h) = \mathbb{E}_x[\mathcal{R}(h(x) | x)]$$

一、贝叶斯决策论：什么是贝叶斯决策论？

- 显然，对每个样本 x ，若 h 能最小化条件风险 $\mathcal{R}(h(x) | x)$ ，则总体风险 $\mathcal{R}(h)$ 也将最小化。

最小化贝叶斯风险

- 贝叶斯判定准则(Bayes decision rule):

为最小化总体风险，只需在每个样本选择上选择那个能使条件风险 $\mathcal{R}(c | x)$ 最小的类别标记，即

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

- 此时称 h^* 为贝叶斯最优分类器 (Bayes optimal classifier)，与之对应的总体风险 $\mathcal{R}(h^*)$ 称为贝叶斯风险 (Bayes risk)。
- $1 - \mathcal{R}(h^*)$ 反映了分类器所能达到的最好性能，即通过机器学习所能产生的模型精度的理论上限。

一、贝叶斯决策论：最小化错误率的贝叶斯决策

- 若目标是**最小化分类错误率**，则误判损失 λ_{ij} 可写为

$$\lambda_{ij} = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{otherwise} \end{cases}$$

此时**条件风险**为

$$\mathcal{R}(c | x) = 1 - P(c | x),$$

于是最小化分类**错误率**的**贝叶斯最优分类器**为

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

- 本质上，对每个样本 x ，选择能使**后验概率** $P(c | x)$ **最大**的类别标记。

条件风险：

$$\mathcal{R}(c_i | x) = \sum_{j=1}^N \lambda_{ij} P(c_j | x)$$

一、贝叶斯决策论：最小化风险的贝叶斯决策

- 最小化错误率的贝叶斯决策可以找到正确率最高的分类结果，但实际问题中，不同类别分类错误的代价可能不同。

例如，将有毒蘑菇分类为无毒蘑菇的代价远远大于将无毒蘑菇分类为有毒蘑菇的代价。

- 根据实际情况引入分类错误风险，使得贝叶斯决策更加科学。

- 以二分类任务为例：

将样本 x 划分到 ω_1 类别的风险为：

$$\mathcal{R}(\omega_1 | x) = \lambda_{11}P(\omega_1 | x) + \lambda_{12}P(\omega_2 | x)$$

将样本 x 划分到 ω_2 类别的风险为：

$$\mathcal{R}(\omega_2 | x) = \lambda_{21}P(\omega_1 | x) + \lambda_{22}P(\omega_2 | x)$$

决策依据： $\mathcal{R}(\omega_1 | x) < \mathcal{R}(\omega_2 | x)$ ，则判定 x 为 ω_1 类别。

$\mathcal{R}(\omega_1 | x) > \mathcal{R}(\omega_2 | x)$ ，则判定 x 为 ω_2 类别。

条件风险：

$$\mathcal{R}(c_i | x) = \sum_{j=1}^N \lambda_{ij} P(c_j | x)$$

一、贝叶斯决策论：估计后验概率的两种策略

- 不难看出，欲使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c | x)$ 。

现实任务中难以直接获得

- 如何估计后验概率 $P(c | x)$ ？

▣ 判别式模型 (discriminative models):

给定 x ，通过直接建模 $P(c | x)$ 来预测 c ；

e.g., 决策树、BP神经网络、支持向量机等；

▣ 生成式模型 (generative models):

先对联合概率分布 $P(x, c)$ 建模，然后再由此获得 $P(c | x)$ ；

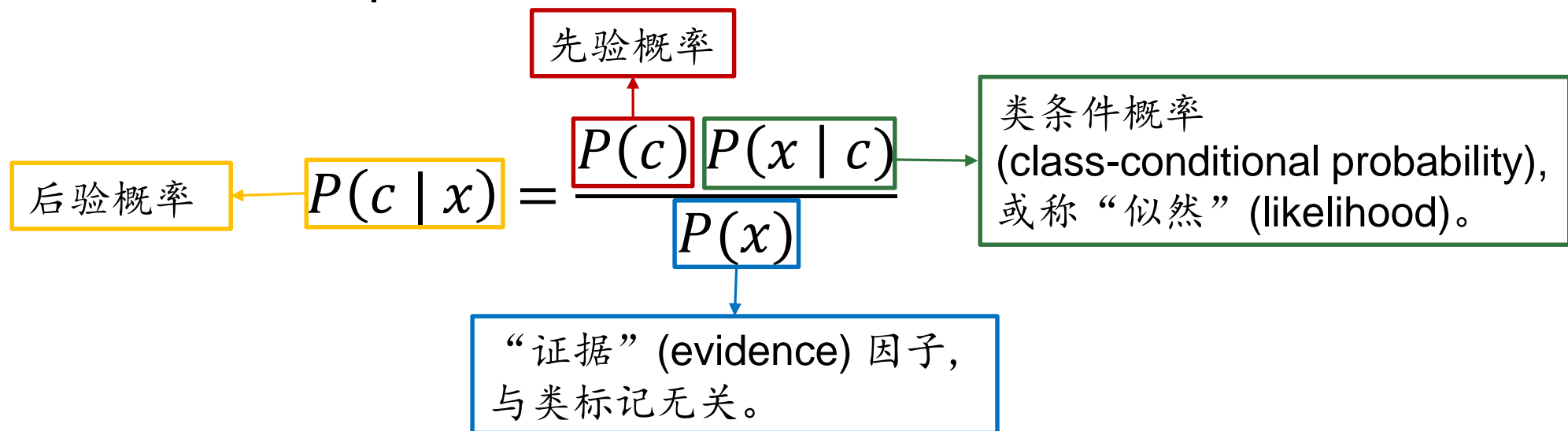
e.g., 朴素贝叶斯分类器、贝叶斯网络等；

一、贝叶斯决策论：生成式模型

- 对于生成式模型来说,

$$P(c | x) = \frac{P(x, c)}{P(x)}$$

基于贝叶斯定理, $P(c | x)$ 可写为



The diagram illustrates the components of Bayes' theorem for a generative model. The central equation is $P(c | x) = \frac{P(c)P(x | c)}{P(x)}$. Annotations include:

- 后验概率** (Posterior probability) pointing to $P(c | x)$.
- 先验概率** (Prior probability) pointing to $P(c)$.
- 类条件概率 (class-conditional probability), 或称“似然” (likelihood)** pointing to $P(x | c)$.
- “证据” (evidence) 因子, 与类标记无关。** (Evidence factor, independent of class label) pointing to $P(x)$.

估计 $P(c | x)$ 转化为估计先验 $P(c)$ 和似然 $P(x | c)$

一、贝叶斯决策论：估计类条件概率 $P(x | c)$

- 估计类条件概率 $P(x | c)$ 的常用策略：

先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布的参数进行估计。

- 参数估计的两个学派：

- 频率主义学派认为参数虽然未知，但却是客观存在的固定值，因此，可通过优化似然函数等准则来确定参数值；
- 贝叶斯学派认为参数是未观察到的随机变量，其本身也可有分布，因此，可假定参数服从一个先验分布，然后基于观测到的数据来计算参数的后验分布。

一、贝叶斯决策论：估计类条件概率 $P(x|c)$

- 贝叶斯学派认为，概率是一个人对于一件事的**信念强度**，概率是**主观**的。
- 频率主义学派所持的是不同的观念：他们认为**参数是客观存在的**，即使是未知的，但都是**固定值**，不会改变。
- 频率学派认为进行一定数量的**重复实验**后，如果出现某个现象的次数与总次数趋于某个值，那么这个比值就会倾向于固定。最简单的例子就是抛硬币了，在理想情况下，我们知道抛硬币正面朝上的概率会趋向于 $1/2$ 。
- 贝叶斯学派提出了一种截然不同的观念，他认为概率不应该这么简单地计算，而需要加入**先验概率**的考虑。

一、贝叶斯决策论：估计类条件概率 $P(x|c)$

- **先验概率**也就是说，我们先设定一个假设（或信念，belief）。然后通过一定的实验来证明/推翻这个假设，这就是后验。随后，旧的后验会成为一个新的先验，如此重复下去。
- **贝叶斯决策论**对于由**证据的积累**来推测一个**事物发生的概率**具有重大作用，它告诉我们当我们要预测一个事物，我们需要的是首先根据**已有的经验和知识**推断一个**先验概率**，然后在**新证据**不断积累的情况下**调整**这个概率。
- 用贝叶斯分析的方法，可以帮助我们解决生活中方方面面的问题。

一、贝叶斯决策论：估计类条件概率 $P(x | c)$

- 本节介绍如何利用频率主义学派的极大似然估计法 (Maximum Likelihood Estimation, 简称 MLE), 来估计概率分布 $P(x | c)$ 的参数。
- 任务描述:
具体地, 记类别 c 的类条件概率为 $P(x | \theta_c)$, 假设 $P(x | \theta_c)$ 具有确定的形式并且被参数向量 θ_c 唯一确定, 则我们的任务就是利用训练集 D 估计参数 θ_c 。
- 令 D_c 表示训练集 D 中第 c 类样本组成的集合, 假设这些样本是独立同分布的, 则参数 θ_c 对于数据集 D_c 的似然是

$$P(D_c | \theta_c) = \prod_{x \in D_c} P(x | \theta_c).$$

一、贝叶斯决策论：估计类条件概率 $P(x | c)$

- 对 θ_c 进行极大似然估计就是去寻找最大化似然 $P(D_c | \theta_c)$ 的参数值 $\hat{\theta}_c$ 。
- 由于 $P(D_c | \theta_c) = \prod_{x \in D_c} P(x | \theta_c)$ 中的连乘操作容易造成下溢，通常使用对数似然(log-likelihood)，此时参数 θ_c 的极大似然估计 $\hat{\theta}_c$ 为

$$\hat{\theta}_c = \arg \max_{\theta_c} LL(\theta_c),$$

其中 $LL(\theta_c) = \log P(D_c | \theta_c) = \sum_{x \in D_c} \log P(x | \theta_c)$.

一、贝叶斯决策论：估计类条件概率 $P(x | c)$

- 例如，在连续属性情形下，假设概率密度函数 $P(x | \theta_c) \sim N(\mu_c, \sigma_c^2)$ ，则参数 μ_c 和 σ_c^2 的极大似然估计为

$$\hat{\mu}_c = \frac{1}{|D_c|} \sum_{x \in D_c} x,$$
$$\hat{\sigma}_c^2 = \frac{1}{|D_c|} \sum_{x \in D_c} (x - \hat{\mu}_c)(x - \hat{\mu}_c)^T.$$

也就是说，通过极大似然法得到的正态分布均值就是样本均值，方差就是 $(x - \hat{\mu}_c)(x - \hat{\mu}_c)^T$ 的均值，这显然是一个符合直觉的结果。

- 需注意的是，这种参数化的方法虽能使类条件概率估计变得相对简单，但估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布。

本章的主要内容

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 极大似然估计
- EM算法
- 小结

二、朴素贝叶斯分类器：引入

- 基于贝叶斯公式来估计后验概率 $P(c | x)$ 的主要困难：

类条件概率 $P(x | c)$ 是所有属性上的联合概率，难以从有限的训练样本直接估计而得。

- 为避开这个障碍，朴素贝叶斯分类器 (Naive Bayes classifier) 采用了“属性条件独立性假设”：

对已知类别，假设所有属性相互独立，换言之，假设每个属性独立地对分类结果发生影响。

二、朴素贝叶斯分类器：引入

- 基于属性条件独立性假设，有

$$P(c | x) = \frac{P(c) P(x | c)}{P(x)} = \frac{P(c)}{P(x)} \prod_{i=1}^d P(x_i | c),$$

其中 d 为属性数目， x_i 为 x 在第 i 个属性上的取值。

- 由于对所有类别来说， $P(x)$ 相同，所以基于最小化分类错误率的贝叶斯判定准则有 $h_{nb}(x) = \arg \max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^d P(x_i | c)$ ，这就是朴素贝叶斯分类器的表达式。

二、朴素贝叶斯分类器：训练过程

- 朴素贝叶斯分类器的训练过程就是基于训练集 D 来估计类先验概率 $P(c)$ ，并为每个属性估计条件概率 $P(x_i | c)$ 。
- 令 D_c 表示训练集 D 中第 c 类样本组成的集合，若有充足的独立同分布样本，则可容易地估计出类先验概率

$$P(c) = \frac{|D_c|}{|D|}.$$

对离散属性而言，令 D_{c,x_i} 表示 D_c 中在第 i 个属性上取值为 x_i 的样本组成的集合，则条件概率 $P(x_i | c)$ 可估计为

$$P(x_i | c) = \frac{|D_{c,x_i}|}{|D_c|}.$$

二、朴素贝叶斯分类器：训练过程

- 对连续属性，可考虑概率密度函数。

假定 $P(x_i | c) \sim N(\mu_{c,i}, \sigma_{c,i}^2)$ ，其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别是第 c 类样本在第 i 个属性上取值的均值和方差，则有

$$P(x_i | c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right).$$

二、朴素贝叶斯分类器：例子

- 用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试样例1进行分类：

西瓜数据集3.0									
编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

二、朴素贝叶斯分类器：例子

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测试样例1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

- 首先估计类先验概率 $P(c)$ ，显然有 $P(\text{好瓜} = \text{是}) = \frac{8}{17} \approx 0.471$ $P(\text{好瓜} = \text{否}) = \frac{9}{17} \approx 0.529$
- 然后，为每个属性估计条件概率 $P(x_i | c)$ ：

$$\begin{aligned}
 P_{\text{青绿}|\text{是}} &= P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375 & P_{\text{青绿}|\text{否}} &= P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333 \\
 P_{\text{蜷缩}|\text{是}} &= P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{是}) = \frac{5}{8} = 0.625 & P_{\text{蜷缩}|\text{否}} &= P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333 \\
 P_{\text{浊响}|\text{是}} &= P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750 & P_{\text{浊响}|\text{否}} &= P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{否}) = \frac{4}{9} \approx 0.444 \\
 P_{\text{清晰}|\text{是}} &= P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{是}) = \frac{7}{8} = 0.875 & P_{\text{清晰}|\text{否}} &= P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222 \\
 P_{\text{凹陷}|\text{是}} &= P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750 & P_{\text{凹陷}|\text{否}} &= P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.444 \\
 P_{\text{硬滑}|\text{是}} &= P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750 & P_{\text{硬滑}|\text{否}} &= P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{否}) = \frac{6}{9} \approx 0.667
 \end{aligned}$$

二、朴素贝叶斯分类器：例子

$$P_{\text{密度: } 0.697 | \text{是}} = P(\text{密度} = 0.697 | \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi} \cdot 0.129} \exp\left(-\frac{(0.697 - 0.574)^2}{2 \cdot 0.129^2}\right) \approx 1.959$$

$$P_{\text{密度: } 0.697 | \text{否}} = P(\text{密度} = 0.697 | \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi} \cdot 0.195} \exp\left(-\frac{(0.697 - 0.496)^2}{2 \cdot 0.195^2}\right) \approx 1.203$$

$$P_{\text{含糖: } 0.460 | \text{是}} = P(\text{含糖} = 0.460 | \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi} \cdot 0.101} \exp\left(-\frac{(0.460 - 0.279)^2}{2 \cdot 0.101^2}\right) \approx 0.788$$

$$P_{\text{含糖: } 0.460 | \text{否}} = P(\text{含糖} = 0.460 | \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi} \cdot 0.108} \exp\left(-\frac{(0.460 - 0.154)^2}{2 \cdot 0.108^2}\right) \approx 0.066$$

于是有，

$$P(\text{好瓜} = \text{是}) \times P_{\text{青绿} | \text{是}} \times P_{\text{蜷缩} | \text{是}} \times P_{\text{浊响} | \text{是}} \times P_{\text{凹陷} | \text{是}} \times P_{\text{硬滑} | \text{是}} \times P_{\text{密度: } 0.697 | \text{是}} \times P_{\text{含糖: } 0.460 | \text{是}} \approx 0.038$$

$$P(\text{好瓜} = \text{否}) \times P_{\text{青绿} | \text{否}} \times P_{\text{蜷缩} | \text{否}} \times P_{\text{浊响} | \text{否}} \times P_{\text{凹陷} | \text{否}} \times P_{\text{硬滑} | \text{否}} \times P_{\text{密度: } 0.697 | \text{否}} \times P_{\text{含糖: } 0.460 | \text{否}} \approx 6.8 \times 10^{-5}$$

由于 $0.038 > 6.8 \times 10^{-5}$ ，所以朴素贝叶斯分类器会将测试样本1判别为“好瓜”。

本章的主要内容

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 极大似然估计
- EM算法
- 小结

三、极大似然估计

- 极大似然估计 (maximum likelihood estimation, MLE) 提供了一种给定观察数据来估计模型参数的方法，即“**模型已定，参数未知**”。
- 极大似然估计的基本思想：
利用已知的样本结果信息，反推**最有可能（最大概率）**导致这些样本结果出现的模型参数值。
- 极大似然估计的步骤：
 - ① 写出似然函数
 - ② 对似然函数取对数
 - ③ 求导数，令导数为0，得到似然方程
 - ④ 解似然方程，得到的参数即为所求



三、极大似然估计

- 问题描述:

假设我们需要调查学校的男生和女生的身高分布。我们抽取100个男生和100个女生，将他们按照性别划分为两组。然后，统计抽样得到100个男生的身高数据和100个女生的身高数据。但是我们只知道他们的身高服从正态分布，不知道分布的均值 μ 和方差 σ^2 。所以我们需要估计参数 μ 和 σ^2 。

- 问题形式化:

以男生身高分布为例，已知

男生身高样本集： $X = \{x_1, x_2, \dots, x_N\}$, $N = 100$

且男生身高服从正态分布，即 $p(x_i | \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$, $\theta = \{\mu, \sigma^2\}$,

估计概率密度函数的参数 $\theta = \{\mu, \sigma^2\}$

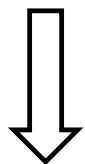
三、极大似然估计

- 问题求解:

似然函数: $L(\theta) = L(x_1, x_2, \dots, x_N | \theta) = \prod_{i=1}^N p(x_i | \theta), \theta \in \Theta$

对数似然: $LL(\theta) = LL(x_1, x_2, \dots, x_N | \theta) = \log \prod_{i=1}^N p(x_i | \theta) = \sum_{i=1}^N \log p(x_i | \theta), \theta \in \Theta$

求 $\hat{\theta} = \arg \max_{\theta} LL(\theta)$



μ 的估计值 $\hat{\mu}$: $\frac{\partial LL(\theta)}{\partial \mu} = 0 \Rightarrow \hat{\mu} = ?$

σ 的估计值 $\hat{\sigma}$: $\frac{\partial LL(\theta)}{\partial \sigma} = 0 \Rightarrow \hat{\sigma} = ?$

似然函数反映了在概率密度函数的参数是 θ 时, 得到 X 这组样本的概率

我们希望找到一个参数 θ , 使得抽到 X 这组的样本概率最大

为了防止连乘操作造成下溢, 通常采用对数似然

三、预备知识——极大似然估计

极大似然函数的应用一：回归问题中的极小化平方和（极小化代价函数）

- 假设线性回归模型具有如下形式: $h(x) = \sum_{i=1}^d \theta_i x_i + \varepsilon = \theta^T x + \varepsilon$, 其中 $x \in \mathbb{R}^d$, $\theta \in \mathbb{R}^d$, 误差 $\varepsilon \in \mathbb{R}$, $y \in \mathbb{R}^m$ 为对应的真值, 则如何求取最优的 θ ?
- 方法一: **最小二乘估计**。最合理的参数估计量应该使得模型能最好地拟合样本数据, 也就是估计值和观测值之差的平方和最小, 其推导过程如下所示:

$$J(\theta) = \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2 = \sum_{i=1}^m (\theta^T x_i - y_i)^2$$

- 求解方法是通过梯度下降算法, 训练数据不断迭代得到最终的值。
- 或者直接求出解析解

三、预备知识——极大似然估计

- 假设线性回归模型具有如下形式: $h(x) = \sum_{i=1}^d \theta_i x^i + \varepsilon = \theta^T x + \varepsilon$, 其中 $x \in \mathbb{R}^d$, $\theta \in \mathbb{R}^d$, 误差 $\varepsilon \in \mathbb{R}$, $y \in \mathbb{R}^m$ 为对应的真值, 则如何求取最优的 θ ?
- 方法二: 极大似然法。最合理的参数估计量应该使得从模型中抽取 m 组样本观测值的概率极大, 也就是似然函数极大。
- 假设误差项 $\varepsilon \in N(0, \sigma^2)$, 则 $y_i \in N(\theta^T x_i, \sigma^2)$

$$p(y_i | x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right)$$

$$\begin{aligned} L(\theta) &= \prod_{i=1}^m p(y_i | x_i; \theta) \\ &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right) \end{aligned}$$

三、预备知识——极大似然估计

- 假设线性回归模型具有如下形式: $h(x) = \sum_{i=1}^d \theta_i x^i + \varepsilon = \theta^T x + \varepsilon$, 其中 $x \in \mathbb{R}^d$, $\theta \in \mathbb{R}^d$, 误差 $\varepsilon \in \mathbb{R}$, $y \in \mathbb{R}^m$ 为对应的真值, 则如何求取最优的 θ ?
- 方法二: 极大似然法。

$$\begin{aligned} H(\theta) &= \log(L(\theta)) \\ &= \log \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right) \\ &= \sum_{i=1}^m \left(\log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right) \right) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^m (y_i - \theta^T x_i)^2 - m \ln \sigma \sqrt{2\pi} \end{aligned}$$

- 所以可得: $\arg \max_{\theta} H(\theta) \Leftrightarrow \arg \min_{\theta} J(\theta)$

极大似然函数等价于
最小二乘估计函数

三、预备知识——极大似然估计

极大似然函数的应用二：分类问题中极小化交叉熵（极小化代价函数）

- 在分类问题中，交叉熵的本质就是似然函数的极大化，逻辑回归的假设函数为：

$$h(x) = \hat{y} = \frac{1}{1 + e^{-\theta^T x + b}}$$

- 根据之前学过的内容我们知道 $\hat{y} = p(y = 1|x; \theta)$

当 $y = 1$ 时, $p_1 = p(y = 1|x; \theta) = \hat{y}$

当 $y = 0$ 时, $p_0 = p(y = 0|x; \theta) = 1 - \hat{y}$

$$p(y|x, \theta) = \hat{y}^y (1 - \hat{y})^{1-y}$$

- 合并上面两式子，可以得到

$$\begin{aligned} L(\theta) &= \prod_{i=1}^m p(y_i | x_i; \theta) \\ &= \prod_{i=1}^m \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i} \end{aligned}$$

三、极大似然估计

极大似然函数的应用二：分类问题中极小化交叉熵（极小化代价函数）

$$\begin{aligned} H(\theta) &= \log(L(\theta)) \\ &= \log \prod_{i=1}^m \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i} \\ &= \sum_{i=1}^m \log \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i} \\ &= \sum_{i=1}^m y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i) \\ J(\theta) &= -H(\theta) = - \sum_{i=1}^m y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i) \\ \arg \max_{\theta} H(\theta) &\Leftrightarrow \arg \min_{\theta} J(\theta) \end{aligned}$$

- 即将极大似然函数等价于极小化交叉熵函数损失。

本章的主要内容

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 极大似然估计
- EM算法
- 小结

四、EM算法：引入

- 前面的讨论中，一直**假设**训练样本所有属性变量的值都**已被观测到**，即训练样本是“完整”的。但现实中往往会遇到“**不完整**”的训练样本。
- 存在“**未观测**”变量的情形下，如何对模型参数进行估计呢？
- EM算法和极大似然估计的前提是一样的，都要**假设数据总体的分布**，如果不知道数据分布，是无法使用EM算法。

四、EM算法：引入

- 问题描述：

仍然是之前的身高调查问题，我们目前有100个男生和100个女生的身高，但是我们不知道这200个数据中哪个是男生的身高，哪个是女生的身高，即抽取得到的每个样本都不知道是从哪个分布中抽取的。此时要**分别估计**男生与女生分布的参数男生：均值 μ_1 和方差 σ_1^2 与女生：均值 μ_2 和方差 σ_2^2

这个时候，对于每个样本，就有两个未知量：

- (1) 这个身高数据是来自男生数据集还是女生数据集？
- (2) 男生和女生身高数据的正态分布的参数分别是多少？

已知

- (1) 样本服从的分布模型
- (2) 随机抽取的样本

如何估计？

未知

- (1) 每个样本来自哪个分布
- (2) 模型的参数

四、EM算法：引入

- 这个时候，对于每一个样本或者你抽取到的人，就有两个问题需要估计，一是这个人是男的还是女的，二是男生和女生**对应的身高的正态分布的参数**是多少。这两个问题是相互依赖的：
- 当我们知道了每个人是男生还是女生，我们可以很容易**利用极大似然**对男女各自的身高的分布进行估计。
- 反过来，当我们知道了**男女身高的分布参数**我们才能知道每一个人更有可能是男生还是女生。
- 例如我们已知男生的身高分布为 $N(\mu_1 = 172, \sigma_1^2 = 5^2)$ ，女生的身高分布为 $N(\mu_2 = 162, \sigma_1^2 = 5^2)$ ，一个学生的身高为180，我们可以推断出这个学生为男生的可能性更大。

四、EM算法：引入

- 现在我们既不知道每个学生是男生还是女生，也不知道男生和女生的身高分布。这就成了一个先有鸡还是先有蛋的问题了。为了解决这个你依赖我，我依赖你的循环依赖问题，总得**有一方要先打破僵局**，可以先随便整一个值出来，看你怎么变，然后再根据你的变化调整我的变化，然后如此**迭代着不断互相推导**，最终就会收敛到一个解，这就是EM算法的基本思想了。
- EM的意思是“Expectation Maximization”，**具体方法**为：
 1. 先设定男生和女生的身高分布参数(初始值)，例如男生的身高分布为 $N(\mu_1 = 172, \sigma_1^2 = 5^2)$ ，女生的身高分布为 $N(\mu_2 = 162, \sigma_1^2 = 5^2)$ 作为初值，当然了，刚开始肯定没那么准；
 2. 然后计算出每个人更可能属于第一个还是第二个正态分布中的（例如，这个人的身高是180，那很明显，他极大可能属于男生），这个是属于Expectation一步；

四、EM算法：引入

- 3. 我们已经大概地按上面的方法将这 200 个人分为男生和女生两部分，我们就可以根据之前说的极大似然估计分别对男生和女生的身高分布参数进行估计（这不变成了极大似然估计了吗？极大即为Maximization）这一步称为 Maximization；
- 4. 然后，当我们更新这两个分布的时候，每一个学生属于女生还是男生的概率又变了，那么我们就再需要调整E步；
- 5.如此往复，直到参数基本不再发生变化或满足结束条件为止。
- 上面的学生属于男生还是女生我们称之为隐含参数，女生和男生的身高分布参数称为模型参数

四、EM算法：引入

- 令 Y 表示已观测变量集， Z 表示隐变量集， Θ 表示模型参数。若欲对 Θ 做极大似然估计，则应最大化对数似然

$$LL(\Theta | Y, Z) = \log P(Y, Z | \Theta).$$

- 然而，由于 Z 是隐变量，上式无法直接求解。
- 此时我们可以通过对 Z 计算期望，来最大化已观测数据的对数“边际似然” (marginal likelihood):

$$LL(\Theta | Y) = \log P(Y | \Theta) = \log \sum_Z P(Y, Z | \Theta).$$

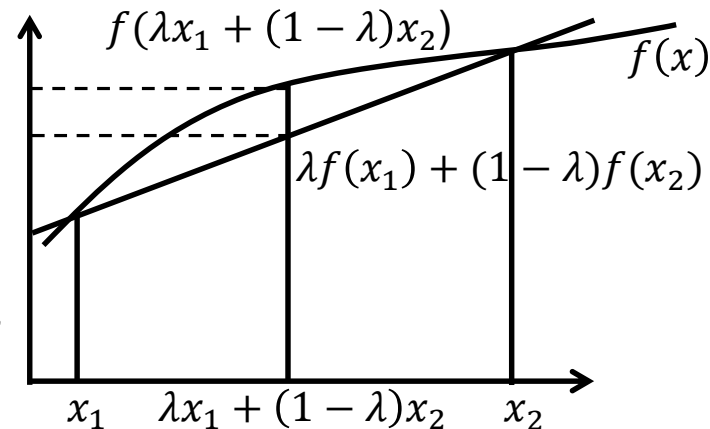
四、EM算法：预备知识——Jensen不等式

- **凹函数定义**：设 f 是定义在区间 $I = [a, b]$ 上的实值函数，如果对于任意的 x_1 和 $x_2 \in I, \lambda \in [0, 1]$ ，下列式子成立，则称 f 是 I 上的**凹函数**。

$$f(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

- **凹函数的判定**：函数的二阶导数 ≤ 0 。
- **Jensen不等式**：设 f 是定义在区间 $I = [a, b]$ 上的凹函数，对于任意的 $x_i \in I, \lambda_i \in [0, 1], i = 1, 2, \dots, n, \sum_{i=1}^n \lambda_i = 1$ ，下列不等式成立：

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \geq \sum_{i=1}^n \lambda_i f(x_i)$$



四、EM算法：预备知识——Jensen不等式

- Jensen不等式证明：归纳法，

当 $n=1$ 和 $n=2$ 时，明显成立。假设 $f(\sum_{i=1}^n \lambda_i x_i) \geq \sum_{i=1}^n \lambda_i f(x_i)$ 对 $n > 2$ 成立，则对于 $n+1$ 来说，

$$f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) = f\left(\lambda_{n+1} x_{n+1} + \sum_{i=1}^n \lambda_i x_i\right) = f\left(\lambda_{n+1} x_{n+1} + (1 - \lambda_{n+1}) \frac{1}{1 - \lambda_{n+1}} \sum_{i=1}^n \lambda_i x_i\right)$$

$$\geq \lambda_{n+1} f(x_{n+1}) + (1 - \lambda_{n+1}) f\left(\frac{1}{1 - \lambda_{n+1}} \sum_{i=1}^n \lambda_i x_i\right)$$

$$= \lambda_{n+1} f(x_{n+1}) + (1 - \lambda_{n+1}) f\left(\sum_{i=1}^n \frac{\lambda_i}{1 - \lambda_{n+1}} x_i\right)$$

$$\sum_{i=1}^n \frac{\lambda_i}{1 - \lambda_{n+1}} = 1$$

$$\geq \lambda_{n+1} f(x_{n+1}) + (1 - \lambda_{n+1}) \sum_{i=1}^n \frac{\lambda_i}{1 - \lambda_{n+1}} f(x_i)$$

$$= \lambda_{n+1} f(x_{n+1}) + \sum_{i=1}^n \lambda_i f(x_i) = \sum_{i=1}^{n+1} \lambda_i f(x_i)$$

四、EM算法：算法简介

- EM算法是一种迭代优化策略，常用于含有隐变量的概率模型参数的极大似然估计或极大后验概率估计。
- EM算法的每次迭代由两步组成：
 - E步，求期望 (Expectation)，
 - M步，求极大 (Maximization)。
- EM算法的基本思想：
 - 若参数 Θ 已知，则可根据训练数据推断出最优隐变量 Z 的值 (E步)。
 - 反之，若隐变量 Z 的值已知，则可方便地对参数 Θ 做极大似然估计 (M步)。

四、EM算法：算法步骤

- EM算法：

- 输入：观察变量数据 Y ，隐变量数据 Z ，联合分布 $P(Y, Z | \theta)$ ，条件分布 $P(Z | Y, \theta)$

- 输出：模型参数 θ

- ① 选择参数初始值 $\theta^{(0)}$ ，开始迭代；

- ② E步：记 $\theta^{(i)}$ 为第 i 次迭代参数 θ 的估计值，在第 $i + 1$ 次迭代的E步，计算

$$Q(\theta, \theta^{(i)}) = \mathbb{E}_Z[\log P(Y, Z | \theta) | Y, \theta^{(i)}] = \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta)$$

- ③ M步：求使 $Q(\theta, \theta^{(i)})$ 极大化的 θ ，确定第 $i + 1$ 次迭代的参数的估计值 $\theta^{(i+1)}$

$$\theta^{(i+1)} = \arg \max_{\theta} Q(\theta, \theta^{(i)})$$

- ④ 重复步骤②-③，直至收敛；

算法的
核心，
需要自
己构造

四、EM算法：算法的导出

- 面对一个含有隐变量的概率模型，我们的目标是极大化观测数据 Y 关于参数 θ 的对数似然函数，即最大化：

$$LL(\theta) = \log P(Y | \theta) = \log \sum_{\mathbf{Z}} P(Y, \mathbf{Z} | \theta) = \log (\sum_{\mathbf{Z}} P(Y | \mathbf{Z}, \theta) P(\mathbf{Z} | \theta))$$

- 最大化这个式子的困难在于式中含有隐变量 \mathbf{Z} ，并且含有和（或者积分）的对数。
- 事实上，EM算法是通过迭代逐步近似极大化 $LL(\theta)$ 的。
- 假设在第 i 次迭代后， θ 的估计值是 $\theta^{(i)}$ 。我们希望新估计值 θ 能使 $LL(\theta)$ 增加，即 $LL(\theta) > LL(\theta^{(i)})$ ，并逐步达到极大值。为此，考虑二者的差：

$$LL(\theta) - LL(\theta^{(i)}) = \log (\sum_{\mathbf{Z}} P(Y | \mathbf{Z}, \theta) P(\mathbf{Z} | \theta)) - \log P(Y | \theta^{(i)})$$

四、EM算法：算法的导出

- 利用Jensen不等式得到其下界：

Jensen不等式： $\log \sum_j \lambda_j y_j \geq \sum_j \lambda_j \log y_j$ ，其中 $\lambda_j \geq 0$ ， $\sum_j \lambda_j = 1$ 。

$$LL(\theta) - LL(\theta^{(i)})$$

$$\begin{aligned} &= \log \left(\sum_Z P(Z | Y, \theta^{(i)}) \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)})} \right) - \log P(Y | \theta^{(i)}) \\ &\geq \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)})} - \log P(Y | \theta^{(i)}) \\ &= \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)})} - \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y | \theta^{(i)}) \\ &= \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)}) P(Y | \theta^{(i)})} \end{aligned}$$

四、EM算法：算法的导出

- 令 $B(\theta, \theta^{(i)}) = LL(\theta^{(i)}) + \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)}) P(Y | \theta^{(i)})}$ ，则可以得到

$$LL(\theta) \geq B(\theta, \theta^{(i)})$$

- 可以知道 $B(\theta, \theta^{(i)})$ 是 $LL(\theta)$ 的一个下界，且 $LL(\theta^{(i)}) = B(\theta^{(i)}, \theta^{(i)})$ 。
- 因此任何可以使得 $B(\theta, \theta^{(i)})$ 增大的 θ ，也可以使得 $LL(\theta)$ 增大。
- 为了使得 $LL(\theta)$ 有尽可能大的增长，选择 $\theta^{(i+1)}$ 使得 $B(\theta, \theta^{(i)})$ 达到极大，即

$$\theta^{(i+1)} = \arg \max_{\theta} B(\theta, \theta^{(i)})$$

四、EM算法：算法的导出

- 现在求 $\theta^{(i+1)}$ ，省略对 θ 极大化而言是常数的项：

$$\begin{aligned}\theta^{(i+1)} &= \arg \max_{\theta} B(\theta, \theta^{(i)}) \\ &= \arg \max_{\theta} (LL(\theta^{(i)}) + \sum_Z P(Z | Y, \theta^{(i)}) \log \frac{P(Y | Z, \theta) P(Z | \theta)}{P(Z | Y, \theta^{(i)}) P(Y | \theta^{(i)})}) \\ &= \arg \max_{\theta} (\sum_Z P(Z | Y, \theta^{(i)}) \log [P(Y | Z, \theta) P(Z | \theta)]) \\ &= \arg \max_{\theta} \left(\sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta) \right) \\ &= \arg \max_{\theta} Q(\theta, \theta^{(i)})\end{aligned}$$

该式子等价于EM算法的一次迭代，即求Q函数及其极大化。

- 所以EM算法是通过不断求解下界极大化逼近求解对数似然函数极大化的算法。

四、EM算法：算法的导出

- EM算法的直观解释：

- 函数 $LL(\theta)$ 和 $B(\theta, \theta^{(i)})$ 在点 $\theta = \theta^{(i)}$ 处相等
EM算法找到下一个点 $\theta^{(i+1)}$ 使函数 $B(\theta, \theta^{(i)})$ 极大化，也使函数 $Q(\theta, \theta^{(i)})$ 极大化。这时，由于 $LL(\theta) \geq B(\theta, \theta^{(i)})$ ，函数 $B(\theta, \theta^{(i)})$ 的增加，保证对数似然函数 $LL(\theta)$ 在每次迭代中也是增加的。
- EM算法在点 $\theta^{(i+1)}$ 重新结算Q函数值，进行下一次迭代。在这个过程中，对数似然函数 $LL(\theta)$ 不断增大。
- 从图可以推断EM算法不能保证找到全局最优值。

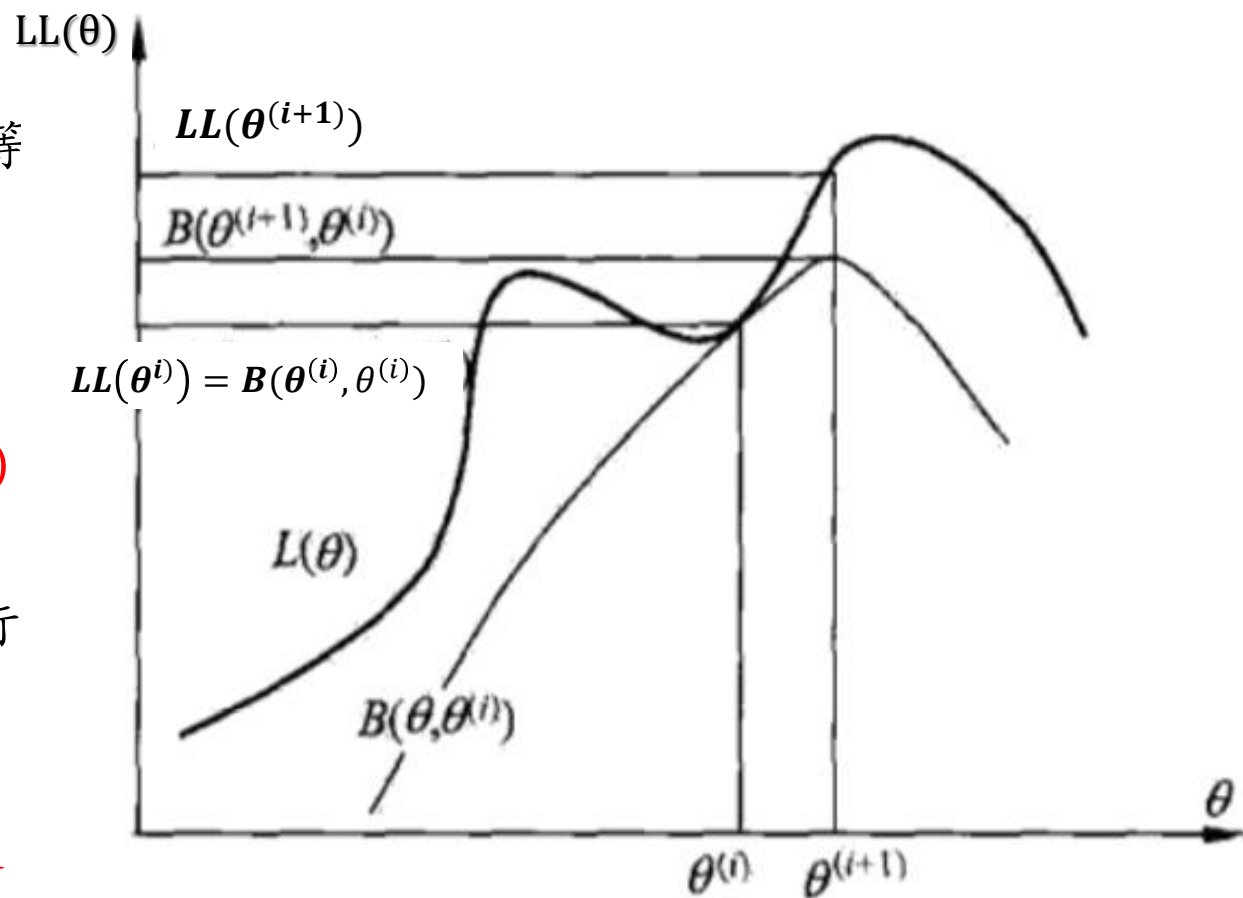


图 9.1 EM 算法的解释

四、EM算法：收敛性证明

- 要证EM算法的收敛性，即证 $P(Y | \theta^{(i+1)}) \geq P(Y | \theta^{(i)})$.
- 因为 $P(Y | \theta) = \frac{P(Y, Z | \theta)}{P(Z | Y, \theta)}$ ，两边同时取对数，有
$$\log P(Y | \theta) = \log P(Y, Z | \theta) - \log P(Z | Y, \theta).$$

左右两边同时求期望，则有

$$\begin{aligned} \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y | \theta) &= \log P(Y | \theta) \\ &= \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta) - \sum_Z P(Z | Y, \theta^{(i)}) \log P(Z | Y, \theta), \end{aligned}$$

其中 $\sum_Z P(Z | Y, \theta^{(i)}) = 1$ ， $\log P(Y | \theta)$ 是与 Z 无关的变量，因此积分等于它本身。

四、EM算法：收敛性证明

• 根据Q函数的定义： $Q(\theta, \theta^{(i)}) = \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta)$,

令 $H(\theta, \theta^{(i)}) = \sum_Z P(Z | Y, \theta^{(i)}) \log P(Z | Y, \theta)$, 则有

$$\begin{aligned} \log P(Y | \theta) &= \sum_Z P(Z | Y, \theta^{(i)}) \log P(Y, Z | \theta) - \sum_Z P(Z | Y, \theta^{(i)}) \log P(Z | Y, \theta) \\ &= Q(\theta, \theta^{(i)}) - H(\theta, \theta^{(i)}) \end{aligned}$$

$$\log P(Y | \theta^{(i+1)}) - \log P(Y | \theta^{(i)})$$

$$= Q(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i+1)}, \theta^{(i)}) - (Q(\theta^{(i)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}))$$

$$= Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)}) - (H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}))$$

□ 因为 $\theta^{(i+1)}$ 使得 $Q(\theta, \theta^{(i)})$ 增加, 所以 $Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)}) \geq 0$ 。

□ 所以, 我们的目标是证明 $H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}) \leq 0$ 。

四、EM算法：收敛性证明

- $H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}) \leq 0$ 的证明如下：

$$H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)})$$

$$= \sum_Z P(Z | Y, \theta^{(i)}) \left(\log \frac{P(Z | Y, \theta^{(i+1)})}{P(Z | Y, \theta^{(i)})} \right)$$

$$\leq \log \left(\sum_Z P(Z | Y, \theta^{(i)}) \frac{P(Z | Y, \theta^{(i+1)})}{P(Z | Y, \theta^{(i)})} \right) \quad \text{----- Jensen不等式}$$

$$= \log (\sum_Z P(Z | Y, \theta^{(i+1)})) = \log 1 = 0$$

所以,

$$\log P(Y | \theta^{(i+1)}) - \log P(Y | \theta^{(i)}) \geq 0$$

$$P(Y | \theta^{(i+1)}) - P(Y | \theta^{(i)}) \geq 0$$

$$P(Y | \theta^{(i+1)}) \geq P(Y | \theta^{(i)}), \text{ 得证。}$$

$$H(\theta, \theta^{(i)}) = \sum_Z P(Z | Y, \theta^{(i)}) \log P(Z | Y, \theta)$$

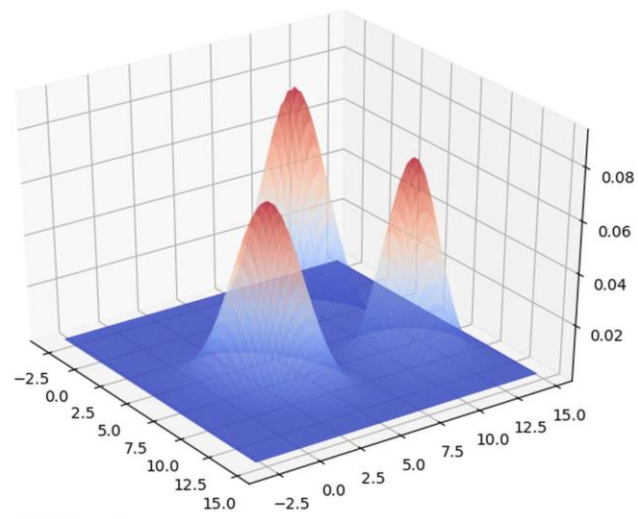
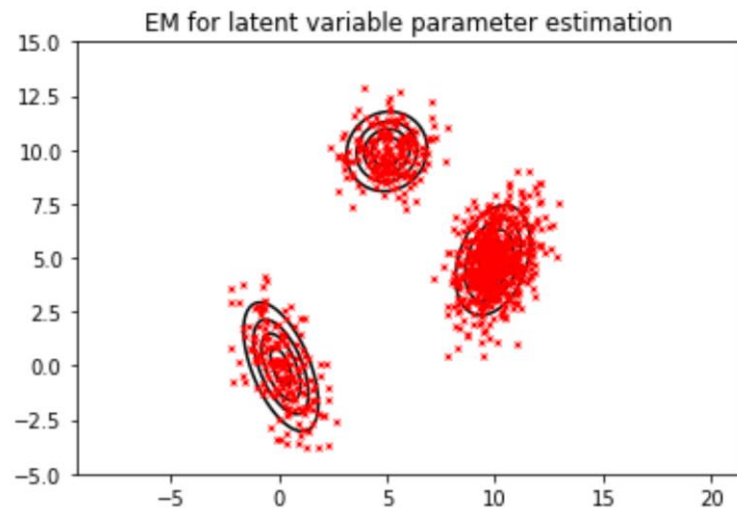
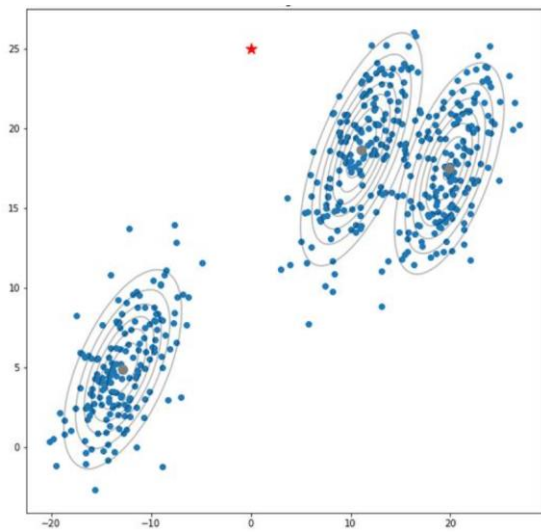
四、EM算法：在高斯混合模型学习中的应用

- 高斯混合模型：

指具有如下形式的概率分布模型，

$$P(y | \theta) = \sum_{k=1}^K \alpha_k \phi(y | \varphi_k),$$

其中 α_k 是系数， $\alpha_k \geq 0$, $\sum_{k=1}^K \alpha_k = 1$, $\phi(y | \varphi_k) = \frac{1}{\sqrt{2\pi} \sigma_k} \exp(-\frac{(y-\mu_k)^2}{2\sigma_k^2})$ 是第 k 个高斯分布模型， $\varphi_k = (\mu_k, \sigma_k^2)$.



四、EM算法：在高斯混合模型学习中的应用

- 高斯混合模型：
- 任务描述：

假设观测数据 y_1, y_2, \dots, y_N 是由高斯混合模型生成，即

$$P(y | \theta) = \sum_{k=1}^K \alpha_k \phi(y | \varphi_k),$$

其中 $\theta = (\alpha_1, \alpha_2, \dots, \alpha_k; \varphi_1, \varphi_2, \dots, \varphi_k)$,

利用EM算法估计高斯混合模型的参数 θ 。

四、EM算法：在高斯混合模型学习中的应用

- 1. 明确隐变量，写出完全数据的对数似然函数

可以设想观测数据 $y_j, j = 1, 2, \dots, N$ ，是这样产生的：

- ✓ 首先依赖概率 α_k 选择第 k 个高斯分布模型 $\phi(y | \varphi_k)$ ；然后依第 k 个高斯分布模型的概率分布生成观测数据 $y_j, j = 1, 2, \dots, N$ 。
- ✓ 这时，观测数据 $y_j, j = 1, 2, \dots, N$ 是已知的；反映观测数据 y_j 来自第 k 个高斯分布模型的数据是未知的，以隐变量 z_{jk} 表示，其定义如下：

$$z_{jk} = \begin{cases} 1, & \text{第} j \text{个观测数据来自第} k \text{个高斯分布模型} \\ 0, & \text{otherwise.} \end{cases}$$

- ✓ 有了观测数据 y_j 以及隐变量 z_{jk} ，那么完全数据是 $(y_1, z_{j1}, z_{j2}, \dots, z_{jK};), j = 1, 2, \dots, N$ 。

四、EM算法：在高斯混合模型学习中的应用

- 于是，完全数据的似然函数为

$$\begin{aligned} P(y, z | \theta) &= \prod_{j=1}^N P(y_j, z_{j1}, z_{j2}, \dots, z_{jK} | \theta) \\ &= \prod_{k=1}^K \prod_{j=1}^N [\alpha_k \phi(y_j | \varphi_k)]^{z_{jk}} \\ &= \prod_{k=1}^K \alpha_k^{n_k} \prod_{j=1}^N [\phi(y_j | \varphi_k)]^{z_{jk}}, \quad (\text{其中 } n_k = \sum_{j=1}^N z_{jk}, \sum_{k=1}^K n_k = N) \\ &= \prod_{k=1}^K \alpha_k^{n_k} \prod_{j=1}^N \left[\frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y_j - \mu_k)^2}{2\sigma_k^2}\right) \right]^{z_{jk}} \end{aligned}$$

- $\log P(y, z | \theta) = \sum_{k=1}^K n_k \log \alpha_k + \sum_{j=1}^N z_{jk} \left[\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2} \right]$

四、EM算法：在高斯混合模型学习中的应用

• 2. EM算法的E步：确定Q函数

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= \mathbb{E}_z[\log P(y, z | \theta) | y, \theta^{(i)}] \\ &= \mathbb{E}_z \left\{ \sum_{k=1}^K \left\{ n_k \log \alpha_k + \sum_{j=1}^N z_{jk} \left[\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2} \right] \right\} \right\} \\ &= \sum_{k=1}^K \left\{ \log \alpha_k \mathbb{E}_z[n_k] + \sum_{j=1}^N \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2} \right) \mathbb{E}_z[z_{jk}] \right\} \\ &= \sum_{k=1}^K \left\{ \log \alpha_k \mathbb{E}_z \left[\sum_{j=1}^N z_{jk} \right] + \sum_{j=1}^N \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2} \right) \mathbb{E}_z[z_{jk}] \right\} \\ &= \sum_{k=1}^K \left\{ \log \alpha_k \left(\sum_{j=1}^N \mathbb{E}_z[z_{jk}] \right) + \sum_{j=1}^N \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2} \right) \mathbb{E}_z[z_{jk}] \right\} \end{aligned}$$

四、EM算法：在高斯混合模型学习中的应用

- 这里需要计算 $\mathbb{E}_z[z_{jk}]$ ，记为 \hat{z}_{jk} ，

$$\hat{z}_{jk} = \mathbb{E}_z[z_{jk}] = P(z_{jk} = 1 \mid y_j, \theta^{(i)})$$

$$\begin{aligned} &= \frac{P(z_{jk} = 1, y_j \mid \theta^{(i)})}{\sum_{k=1}^K P(z_{jk} = 1, y_j \mid \theta^{(i)})} \\ &= \frac{P(y_j \mid z_{jk} = 1, \theta^{(i)}) P(z_{jk} = 1 \mid \theta^{(i)})}{\sum_{k=1}^K P(y_j \mid z_{jk} = 1, \theta^{(i)}) P(z_{jk} = 1 \mid \theta^{(i)})} \\ &= \frac{\alpha_k \phi(y_j \mid \varphi_k^{(i)})}{\sum_{k=1}^K \alpha_k \phi(y_j \mid \varphi_k^{(i)})} \end{aligned}$$

- \hat{z}_{jk} 是在当前模型参数下，第 j 个观测数据来自第 k 个高斯分布模型的概率。

$$\begin{aligned} &P(B_i \mid A) \\ &= \frac{P(A, B_i)}{P(A)} \\ &= \frac{P(A, B_i)}{\sum_j P(A, B_j)} \\ &= \frac{P(A \mid B_i) P(B_i)}{\sum_j P(A \mid B_j) P(B_j)} \end{aligned}$$

四、EM算法：在高斯混合模型学习中的应用

令 $\hat{n}_k = \sum_{j=1}^N \mathbb{E}_z[z_{jk}]$ ，并将 $\hat{z}_{jk} = \mathbb{E}_z[z_{jk}]$ 代入得：

$$Q(\theta, \theta^{(i)}) = \sum_{k=1}^K \hat{n}_k \log \alpha_k + \sum_{j=1}^N \hat{z}_{jk} \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2} \right).$$

四、EM算法：在高斯混合模型学习中的应用

• 3. 确定EM算法的M步

迭代的M步是求函数 $Q(\theta, \theta^{(i)})$ 对 θ 的极大值，即求新一轮迭代的模型参数：

$$\theta^{(i+1)} = \arg \max_{\theta} Q(\theta, \theta^{(i)})$$

$$= \arg \max_{\theta} \sum_{k=1}^K \hat{n}_k \log \alpha_k + \sum_{j=1}^N \hat{z}_{jk} \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2} \right)$$

$$\frac{\partial Q(\theta, \theta^{(i)})}{\partial \mu_k} = 0 \Rightarrow \mu_k^{(i+1)} = \frac{\sum_{j=1}^N \hat{z}_{jk} y_j}{\sum_{j=1}^N \hat{z}_{jk}}$$

$$\frac{\partial Q(\theta, \theta^{(i)})}{\partial \sigma_k} = 0 \Rightarrow \sigma_k^{(i+1)} = \frac{\sum_{j=1}^N \hat{z}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^N \hat{z}_{jk}}$$

本章的主要内容

- 贝叶斯决策论
- 朴素贝叶斯分类器
- EM算法
- 小结

小结

- **贝叶斯决策论**为概率框架下的决策提供了理论基础，其基于先验和数据观察的假定，赋予每个假设一个后验概率。
- 贝叶斯方法确定的**极大后验概率假设**是最可能成为最优假设的假设。
- 朴素贝叶斯分类器**不考虑属性间的依赖性**。
- EM算法是最常见的**隐变量估计**方法，在机器学习中有极为广泛的用途，例如常被用来学习高斯混合模型 (GMM) 的参数。

本章作业

• 本章作业

(必做) 试证明EM算法的收敛性。

(选做) 基于EM的高斯混合模型参数估计：假设我们有一组从高斯混合模型中独立采样得到的样本 $X = \{x_1, x_2, \dots, x_n\}$ ，使用EM算法估计高斯混合模型的参数。

$$p(x | \theta) = \sum_c \omega(c) N(x | \mu_c, \sigma_c), \sum_{c=1}^C \omega(c) = 1$$

谢谢！

李爽

E-mail: shuangli@bit.edu.cn



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY