南昌大学
NANCHANG UNIVERSITY

# 统计机器学习

主讲人：彭振华

数学与计算机学院

2026年

# 目录

CONTENTS

# 贝 叶 斯 分 类



- **贝叶斯决策论**（Bayesian decision theory）是在概率框架下实施决策的基本方法。
    - 在分类问题情况下，在所有相关概率都已知的理想情形下，贝叶斯决策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。
- 假设有 $N$ 种可能的类别标记，$y=\{c_1,c_2,\ldots,c_N\}$，$\lambda_{ij}$ 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 $P\{c_i \mid \mathbf{x}\}$可获得将样本$\mathbf{x}$分类为 $c_i$ 所产生的期望损失(expected loss)，即在样本上的“条件风险”(conditional risk)

$$R(c_i \mid \mathbf{x})=\sum_{j=1}^{N}\lambda_{ij}P(c_j \mid \mathbf{x}) \qquad (7.1)$$

- 我们的任务是寻找一个判定准则 $h: X \mapsto Y$ 以最小化总体风险

$$R(h)=\mathbf{E}_x\left[R(h(\mathbf{x}) \mid \mathbf{x})\right] \qquad (7.2)$$

# 贝 叶 斯 分 类



- 显然，对每个样本 $\mathbf{x}$，若 $h$ 能最小化条件风险 $R(h(\mathbf{x}) \mid \mathbf{x})$，则总体风险 $R(h)$也将被最小化。
- 这就产生了贝叶斯判定准则（Bayes decision rule）： 为最小化总体风险，只需在每个样本上选择那个能使条件风险$R(c \mid \mathbf{x})$ 最小的类别标记，即

$$h^*(x) = \underset{c \in y}{\operatorname{argmin}}\, R(c \mid x) \qquad (7.3)$$

  - 此时，被称为贝叶斯最优分类器(Bayes optimal classifier)，与之对应的总体风险 $R(h^*)$称为贝叶斯风险 (Bayes risk)
  - $1 - R(h^*)$ 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。

# 贝 叶 斯 分 类



- 具体来说，若目标是最小化分类错误率，则误判损失 $\lambda_{ij}$ 可写为

$$\lambda_{i,j}\begin{cases} 0, & \text{if} \quad i=j; \\ 1, & \text{otherwise,} \end{cases} \tag{7.4}$$

- 此时条件风险

$$R(c \mid \mathbf{x}) = 1 - P(c \mid \mathbf{x}) \tag{7.5}$$

- 于是，最小化分类错误率的贝叶斯最优分类器为

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}}\, P(c \mid x) \tag{7.6}$$

  - 即对每个样本 $\mathbf{x}$，选择能使后验概率 $P(c \mid \mathbf{x})$最大的类别标记。

# 贝 叶 斯 分 类



- 不难看出，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 $P(c \mid \mathbf{x})$。
- 然而，在现实中通常难以直接获得。机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率 $P(c \mid \mathbf{x})$。
- 主要有两种策略：
  - 判别式模型（discriminative models）
    - 给定 $\mathbf{x}$，通过直接建模 $P(c \mid \mathbf{x})$，来预测 $c$
    - 决策树，BP神经网络，支持向量机
  - 生成式模型（generative models）
    - 先对联合概率分布 $P(\mathbf{x}, c)$ 建模，再由此获得 $P(c \mid \mathbf{x})$
    - 生成式模型考虑

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \qquad (7.7)$$

# 贝 叶 斯 分 类



- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

□基于贝叶斯定理， $P(c \mid \mathbf{x})$ 可写成

类标记 $c$ 相对于样本 $\mathbf{x}$ 的“类条件概率” (class-conditional probability), 或称“似然”。

$$P(c \mid \mathbf{x}) = \frac{\boxed{P(c)}\boxed{P(\mathbf{x} \mid c)}}{\boxed{P(\mathbf{x})}} \tag{7.8}$$

先验概率

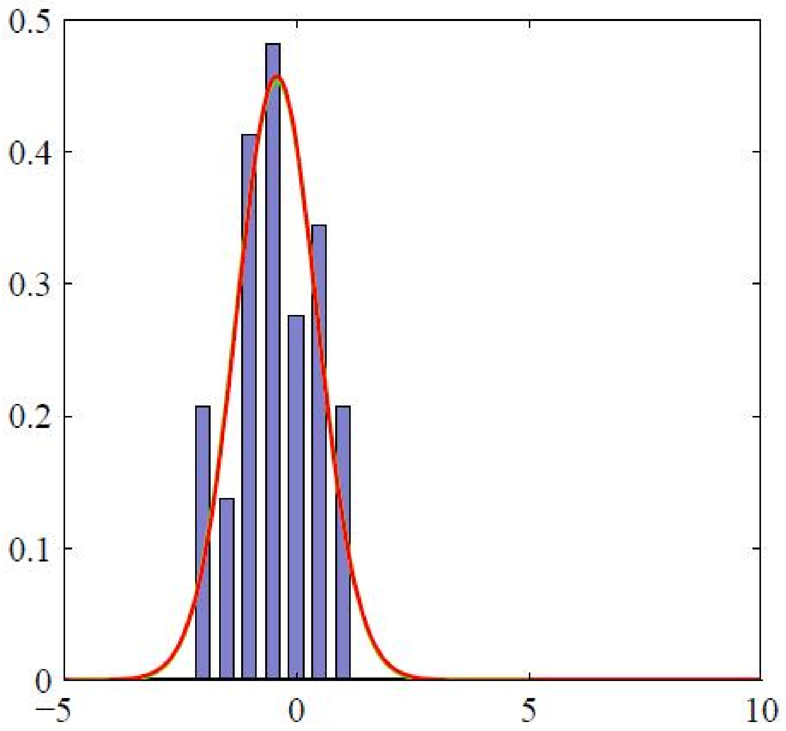
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

“证据” （evidence）因子，与类标记无关

# 贝 叶 斯 分 类



- 估计类条件概率的常用策略：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参数估计。




- 记关于类别 $c$ 的类条件概率为 $P(\mathbf{x} \mid c)$，
  - 假设 $P(\mathbf{x} \mid c)$ 具有确定的形式被参数 $\boldsymbol{\theta}_c$ 唯一确定，我们的任务就是利用训练集 $D$ 估计参数 $\boldsymbol{\theta}_c$

# 贝 叶 斯 分 类



- 估计类条件概率的常用策略：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参数估计。
- 记关于类别$c$的类条件概率为 $P(\mathbf{x} \mid c)$，
  - 假设 $P(\mathbf{x} \mid c)$ 具有确定的形式被参数 $\boldsymbol{\theta}_c$ 唯一确定，我们的任务就是利用训练集 $D$ 估计参数 $\boldsymbol{\theta}_c$
- 概率模型的训练过程就是参数估计过程，统计学界的两个学派提供了不同方案：
  - 频率主义学派 (Frequentist)认为参数虽然未知，但却存在客观值，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 (Bayesian)认为参数是未观察到的随机变量、其本身也可由分布，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计算参数的后验分布。

- 令 $D_c$ 表示训练集中第 $c$ 类样本组合的集合，假设这些样本是独立同分布的，则参数$\boldsymbol{\theta}_c$ 对于数据集 $D_c$ 的似然是

$$P(D_c \mid \boldsymbol{\theta_c}) = \prod_{\mathbf{x} \in D_c} P(\mathbf{x} \mid \boldsymbol{\theta_c}) \qquad (7.9)$$

  - 对 $\boldsymbol{\theta}_c$ 进行极大似然估计，寻找能最大化似然 $P(D_c \mid \theta_c)$ 的参数值 $\hat{\theta}_c$ 。直观上看，极大似然估计是试图在 $\theta_c$ 所有可能的取值中，找到一个使数据出现的“可能性”最大值。

- 式 (7.9)的连乘操作易造成下溢，通常使用对数似然(log-likelihood)

$$\begin{aligned} LL(\boldsymbol{\theta_c}) &= \log P(D_c \mid \boldsymbol{\theta_c}) \\ &= \sum_{\mathbf{x} \in D_c} \log P(\mathbf{x} \mid \boldsymbol{\theta}_c) \end{aligned} \qquad (7.10)$$

- 此时参数 $\boldsymbol{\theta}_c$ 的极大似然估计 $\hat{\boldsymbol{\theta}}_c$ 为

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{c}} = \underset{\boldsymbol{\theta}_c}{\operatorname{argmax}}\, LL(\boldsymbol{\theta}_c) \quad (7.11)$$

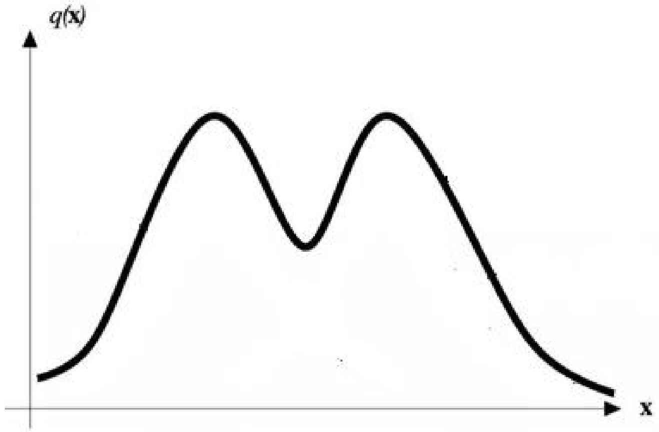
# 贝 叶 斯 分 类



- 例如，在连续属性情形下，假设概率密度函数 $p(\mathbf{x} \mid c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$，则参数 $\boldsymbol{\mu}_c$ 和 $\boldsymbol{\sigma}_c^2$ 的极大似然估计为

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{|D_c|} \sum_{\mathbf{x} \in D_c} \mathbf{x} \tag{7.12}$$

$$\hat{\boldsymbol{\sigma}}_c^2 = \frac{1}{|D_c|} \sum_{\mathbf{x} \in D_c} (\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}} \tag{7.13}$$




- 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 $(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}$ 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。例如下图真实数据分布为双峰分布，若采取单峰的高斯分布进行估计，则准确性较低。

# 贝 叶 斯 分 类



- 估计后验概率 $P(c \mid \mathbf{x})$ 主要困难：类条件概率 $P(\mathbf{x} \mid c)$是所有属性上的联合概率难以从有限的训练样本估计获得。
  - 例如：d个二值属性 → $2^d$可能取值， $2^d >> $ 样本数m
- **朴素贝叶斯分类器**(Naïve Bayes Classifier)采用了“属性条件独立性假设”(attribute conditional independence assumption)：**对已知类别**，所有属性互相独立。
- 基于属性条件独立性假设，(7.8)可重写为

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.14)$$

  - 其中$d$为属性数目，$x_i$ 为$\mathbf{x}$在第$i$ 个属性上的取值。

# 贝 叶 斯 分 类



$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.14)$$

由于对所有类别来说 $P(x)$ 相同，因此基于式 (7.6)的贝叶斯判定准则有

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\arg\max}\, P(c) \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.15)$$

- 这就是**朴素贝叶斯分类器**的表达式

- 朴素贝叶斯分类器的训练器的训练过程就是基于训练集 $D$ 估计类先验概率 $P(c)$ 并为每个属性估计条件概率 $P(x_i \mid c)$。
  - 令 $D_c$ 表示训练集 $D$ 中第 $c$ 类样本组合的集合，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

$$P(c) = \frac{|D_c|}{D} \qquad (7.16)$$

  - 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 $P(x_i \mid c)$ 可估计为

$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{D} \qquad (7.17)$$

  - 对连续属性而言可考虑概率密度函数，假定 $p(x_i \mid c) \sim N(\mu_{c,i}, \sigma_{c,i}^2)$，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

$$P(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp(-\frac{(x_i - \mu_{c,i})^2}{2{\sigma_{c,i}}^2}) \qquad (7.18)$$

# 贝 叶 斯 分 类



**例 4.1** 试由表 4.1 的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^{\mathrm{T}}$ 的类标记 $y$. 表中 $X^{(1)}$, $X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$, $A_2=\{S,M,L\}$, $Y$ 为类标记，$Y\in C=\{1,-1\}$.

**表 4.1 训练数据**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X^{(1)}$ | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| $X^{(2)}$ | $S$ | $M$ | $M$ | $S$ | $S$ | $S$ | $M$ | $M$ | $L$ | $L$ | $L$ | $M$ | $M$ | $L$ | $L$ |
| $Y$ | −1 | −1 | 1 | 1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | −1 |

# 贝 叶 斯 分 类



$$P(Y=1)=\frac{9}{15}, \quad P(Y=-1)=\frac{6}{15}$$

$$P(X^{(1)}=1|Y=1)=\frac{2}{9}, \quad P(X^{(1)}=2|Y=1)=\frac{3}{9}, \quad P(X^{(1)}=3|Y=1)=\frac{4}{9}$$

$$P(X^{(2)}=S|Y=1)=\frac{1}{9}, \quad P(X^{(2)}=M|Y=1)=\frac{4}{9}, \quad P(X^{(2)}=L|Y=1)=\frac{4}{9}$$

$$P(X^{(1)}=1|Y=-1)=\frac{3}{6}, \quad P(X^{(1)}=2|Y=-1)=\frac{2}{6}, \quad P(X^{(1)}=3|Y=-1)=\frac{1}{6}$$

$$P(X^{(2)}=S|Y=-1)=\frac{3}{6}, \quad P(X^{(2)}=M|Y=-1)=\frac{2}{6}, \quad P(X^{(2)}=L|Y=-1)=\frac{1}{6}$$

对于给定的 $x=(2,S)^{\mathrm{T}}$ 计算：

$$P(Y=1)P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1)=\frac{9}{15}\cdot\frac{3}{9}\cdot\frac{1}{9}=\frac{1}{45}$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)=\frac{6}{15}\cdot\frac{2}{6}\cdot\frac{3}{6}=\frac{1}{15}$$

因为 $P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$ 最大，所以 $y=-1$． ■

# 贝 叶 斯 分 类



- 练习：用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类 (p151, 西瓜数据集 p84 表4.3)

| 编号 | 色泽 | 根蒂 | 敲声 | 纹理 | 脐部 | 触感 | 密度 | 含糖率 | 好瓜 |
|---|---|---|---|---|---|---|---|---|---|
| 测 1 | 青绿 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 0.697 | 0.460 | ? |

# 贝 叶 斯 分 类



- 若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题. 比如“敲声=清脆”测试例，训练集中没有该样例，因此连乘式计算的概率值为0，无论其他属性上明显像好瓜，分类结果都是“好瓜=否”，这显然不合理。
- 为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“拉普拉斯修正”（Laplacian correction）
  - 令 $N$ 表示训练集 $D$ 中可能的类别数，$N_i$ 表示第 $i$ 个属性可能的取值数，则式(7.16)和 (7.17)分别修正为

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \quad (7.19)$$

$$\hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \quad (7.20)$$

- 现实任务中，朴素贝叶斯分类器的使用：速度要求高，“查表”；任务数据更替频繁，“懒惰学习”（lazy learning）；数据不断增加，增量学习等等。

# 贝 叶 斯 分 类



**例 4.1** 试由表 4.1 的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^{\mathrm{T}}$ 的类标记 $y$. 表中 $X^{(1)}$, $X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$, $A_2=\{S,M,L\}$, $Y$ 为类标记，$Y\in C=\{1,-1\}$.

**表 4.1 训练数据**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X^{(1)}$ | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| $X^{(2)}$ | $S$ | $M$ | $M$ | $S$ | $S$ | $S$ | $M$ | $M$ | $L$ | $L$ | $L$ | $M$ | $M$ | $L$ | $L$ |
| $Y$ | −1 | −1 | 1 | 1 | −1 | −1 | −1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | −1 |

# 贝 叶 斯 分 类



$$P(Y=1)=\frac{10}{17},\quad P(Y=-1)=\frac{7}{17}$$

$$P(X^{(1)}=1\mid Y=1)=\frac{3}{12},\quad P(X^{(1)}=2\mid Y=1)=\frac{4}{12},\quad P(X^{(1)}=3\mid Y=1)=\frac{5}{12}$$

$$P(X^{(2)}=S\mid Y=1)=\frac{2}{12},\quad P(X^{(2)}=M\mid Y=1)=\frac{5}{12},\quad P(X^{(2)}=L\mid Y=1)=\frac{5}{12}$$

$$P(X^{(1)}=1\mid Y=-1)=\frac{4}{9},\quad P(X^{(1)}=2\mid Y=-1)=\frac{3}{9},\quad P(X^{(1)}=3\mid Y=-1)=\frac{2}{9}$$

$$P(X^{(2)}=S\mid Y=-1)=\frac{4}{9},\quad P(X^{(2)}=M\mid Y=-1)=\frac{3}{9},\quad P(X^{(2)}=L\mid Y=-1)=\frac{2}{9}$$

对于给定的 $x=(2,S)^{\mathrm{T}}$ 计算：

$$P(Y=1)P(X^{(1)}=2\mid Y=1)P(X^{(2)}=S\mid Y=1)=\frac{10}{17}\cdot\frac{4}{12}\cdot\frac{2}{12}=\frac{5}{153}=0.0327$$

$$P(Y=-1)P(X^{(1)}=2\mid Y=-1)P(X^{(2)}=S\mid Y=-1)=\frac{7}{17}\cdot\frac{3}{9}\cdot\frac{4}{9}=\frac{28}{459}=0.0610$$

由于 $P(Y=-1)P(X^{(1)}=2\mid Y=-1)P(X^{(2)}=S\mid Y=-1)$ 最大，所以 $y=-1$. ■

# 贝 叶 斯 分 类



- 为了降低贝叶斯公式中估计后验概率的困难，朴素贝叶斯分类器采用的属性条件独立性假设；对属性条件独立假设记性一定程度的放松，由此产生了一类称为“半朴素贝叶斯分类器” (semi-naïve Bayes classifiers)

- 半朴素贝叶斯分类器最常用的一种策略： “独依赖估计”(One-Dependent Estimator, ODE)，假设每个属性在类别之外最多仅依赖一个其他属性，即

$$P(c \mid x) \propto P(c) \prod_{i=1}^{d} P(x_i \mid c, pa_i)$$

  - 其中 $pa_i$ 为属性 $x_i$ 所依赖的属性，称为 $x_i$ 的父属性

- 对每个属性 $x_i$，若其父属性 $pa_i$ 已知，则可估计概率值 $P(x_i \mid c, pa_i)$，于是问题的关键转化为如何确定每个属性的父属性

## □ EM算法

□ “不完整”的样本：西瓜已经脱落的根蒂，无法看出是“蜷缩”还是“坚挺”，则训练样本的“根蒂”属性变量值未知，如何计算？

□未观测的变量称为“隐变量”(latent variable)。令$\mathbf{X}$表示已观测变量集，$\mathbf{Z}$表示隐变量集，若预对模型参数$\Theta$做极大似然估计，则应最大化对数似然函数

$$LL(\mathbf{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta}) \tag{7.34}$$

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

$$LL(\mathbf{\Theta} \mid \mathbf{X}) = \ln P(\mathbf{X} \mid \mathbf{\Theta}) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta}) \tag{7.35}$$

# 贝 叶 斯 分 类



EM (Expectation-Maximization)算法 [Dempster et al., 1977] 是常用的估计参数隐变量的利器。

- 当参数 $\Theta$ 已知 - > 根据训练数据推断出最优隐变量 $\mathbf{Z}$ 的值(E步)
- 当 $\mathbf{Z}$ 已知 - > 对 $\Theta$ 做极大似然估计(M步)

于是，以初始值$\Theta^0$为起点，对式子(7.35),可迭代执行以下步骤直至收敛：

- 基于$\Theta^t$ 推断隐变量 $\mathbf{Z}$ 的期望,记为$\mathbf{Z}^t$；
- 基于已观测到变量$\mathbf{X}$和 $\mathbf{Z}^t$ 对参数$\Theta$做极大似然估计，记为 $\Theta^{t+1}$;
- 这就是EM算法的原型。

进一步，若我们不是取Z的期望，而是基于$\Theta^t$计算隐变量$\mathbf{Z}$的概率分布$P(\mathbf{Z} \mid \mathbf{X}, \Theta^t)$，则EM算法的两个步骤是：

- E步(Expectation):以当前参数$\Theta^t$推断隐变量分布$P(\mathbf{Z} \mid \mathbf{X}, \Theta^t)$，并计算对数似然$LL(\Theta \mid \mathbf{X}, \mathbf{Z})$关于$\mathbf{Z}$的期望:

$$Q(\mathbf{\Theta} \mid \mathbf{\Theta}^t) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\mathbf{\Theta}^t} LL(\mathbf{\Theta} \mid \mathbf{X}, \mathbf{Z}) \qquad (7.36)$$

- M步(Maximization):寻找参数最大化期望似然，即

$$\mathbf{\Theta}^{t+1} = \underset{\mathbf{\Theta}}{\arg\max}\, Q(\mathbf{\Theta} \mid \mathbf{\Theta}^t) \qquad (7.37)$$

- EM算法是在最大化对数似然下界，可以保证收敛，但无法保证找到全局最优。

- 两个收敛定理：
- 定理9.1：设P(Y|Θ)为观测数据的似然函数Θ⁽ⁱ⁾(i=1,2..）为EM参数估计序列 $P(Y|\theta^{(i)})\ (i=1,2,\cdots)$ 为对应的似然函数序列，则P(Y|Θ⁽ⁱ⁾)是单调递增的，即：

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

- 证明：由 $P(Y|\theta)=\dfrac{P(Y,Z|\theta)}{P(Z|Y,\theta)}$

$$\log P(Y|\theta)=\log P(Y,Z|\theta)-\log P(Z|Y,\theta)$$

- 由： $Q(\theta,\theta^{(i)})=\sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(i)})$
- 令： $H(\theta,\theta^{(i)})=\sum_Z \log P(Z|Y,\theta)P(Z|Y,\theta^{(i)})$
- 则： $\log P(Y|\theta)=Q(\theta,\theta^{(i)})-H(\theta,\theta^{(i)})$

# 贝 叶 斯 分 类



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

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

- 只需证右端非负

前半部分，Θ$^{(i+1)}$为极大值，所以 $Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)}) \geqslant 0$

后半部分：

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

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

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

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

- 定理9.2:
- 设$L(\Theta)=\log P(Y|\Theta)$,为观测数据的对数似然函数，$\Theta^{(i)}(i=1,2..)$ 为EM算法得到的参数估计序列，$L(\Theta^{(i)})$为对应的对数似然函数序列，
- 1、如果$P(Y|\Theta)$有上界，则$L(\Theta^{(i)})=\log P(Y|\Theta^{(i)})$收敛到某一值$L^*$；
- 2、在函数$Q(\Theta,\Theta')$与$L(\Theta)$满足一定条件下，由EM算法得到的参数估计序列$\Theta^{(i)}$的收敛值$\Theta^*$是$L(\Theta)$的稳定点。

# 贝 叶 斯 分 类



- 高斯混合模型：概率分布模型

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

- 系数：

$$\alpha_k \geqslant 0,\quad \sum_{k=1}^{K}\alpha_k=1$$

- 高斯分布密度：

$$\phi(y|\theta_k)\qquad \theta_k=(\mu_k,\sigma_k^2)$$

- 第K个分模型：

$$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

# 贝 叶 斯 分 类



- 假设观测数据$y_1,y_2,\ldots,y_N$由高斯混合模型生成：

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

$$\theta=(\alpha_1,\alpha_2,\cdots,\alpha_K;\theta_1,\theta_2,\cdots,\theta_K)$$

- 用EM算法估计参数；
- 1、明确隐变量，写出完全数据的对数似然函数：
  - 设想观测数据yi是依概率$a_k$选择第k个高斯分模型 $\phi(y|\theta_k)$ 生成，隐变量

$$\gamma_{jk}=\begin{cases}1, & \text{第 } j \text{ 个观测来自第 } k \text{ 个分模型}\\ 0, & \text{否则}\end{cases}$$

# 贝 叶 斯 分 类



- 1、明确隐变量，写出完全数据的对数似然函数：
- 完全数据： $(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK}),\quad j=1,2,\cdots,N$
- 似然函数：

$$P(y,\gamma\mid\theta)=\prod_{j=1}^{N}P(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK}\mid\theta)$$

$$=\prod_{k=1}^{K}\prod_{j=1}^{N}[\alpha_k\phi(y_j\mid\theta_k)]^{\gamma_{jk}}$$

$$n_k=\sum_{j=1}^{N}\gamma_{jk}\qquad =\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}[\phi(y_j\mid\theta_k)]^{\gamma_{jk}}$$

$$\sum_{k=1}^{K}n_k=N\qquad =\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y_j-\mu_k)^2}{2\sigma_k^{\,2}}\right)\right]^{\gamma_{jk}}$$

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

- 2、EM算法的E步，确定Q函数

$$Q(\theta,\theta^{(i)}) = E[\log P(y,\gamma|\theta)|y,\theta^{(i)}]$$

$$= E\left\{\sum_{k=1}^{K} n_k \log \alpha_k + \sum_{j=1}^{N} \gamma_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log \sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}$$

$$= \sum_{k=1}^{K}\left\{\sum_{j=1}^{N}(E\gamma_{jk})\log \alpha_k + \sum_{j=1}^{N}(E\gamma_{jk})\left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log \sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}$$

需要计算 $E(\gamma_{jk}|y,\theta)$，记为 $\hat{\gamma}_{jk}$

- 第j个观测数据来自第k个分模型的概率，称为分模型k对观测数据$y_j$的响应度。

- 2、EM算法的E步，确定Q函数

$$\hat{\gamma}_{jk} = E(\gamma_{jk} \mid y, \theta) = P(\gamma_{jk} = 1 \mid y, \theta)$$

$$= \frac{P(\gamma_{jk} = 1, y_j \mid \theta)}{\sum_{k=1}^{K} P(\gamma_{jk} = 1, y_j \mid \theta)}$$

$$= \frac{P(y_j \mid \gamma_{jk} = 1, \theta) P(\gamma_{jk} = 1 \mid \theta)}{\sum_{k=1}^{K} P(y_j \mid \gamma_{jk} = 1, \theta) P(\gamma_{jk} = 1 \mid \theta)}$$

$$= \frac{\alpha_k \phi(y_j \mid \theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j \mid \theta_k)}, \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K$$

将 $\hat{\gamma}_{jk} = E\gamma_{jk}$ 及 $n_k = \sum_{j=1}^{N} E\gamma_{jk}$ 代入 $Q(\theta, \theta^{(i)}) = \sum_{k=1}^{K} n_k \log \alpha_k + \sum_{k=1}^{N} \hat{\gamma}_{jk} \left[ \log\left( \frac{1}{\sqrt{2\pi}} \right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right]$

- 3、确定EM算法的M步：
- 求：

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

用 $\hat{\mu}_k$，$\hat{\sigma}_k^2$ 及 $\hat{\alpha}_k$，$k=1,2,\cdots,K$，表示 $\theta^{(i+1)}$

- 采用求导的方法：

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}} \qquad \hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^{N} \hat{\gamma}_{jk}} \qquad \hat{\alpha}_k = \frac{n_k}{N} = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}$$

# 贝 叶 斯 分 类



- 输入：观测数据$y_1, y_2, \ldots y_N$，高斯混合模型
- 输出：高斯混合模型参数
- 1、设定初始值开始迭代
- 2、E步，响应度计算

$$\hat{\gamma}_{jk} = \frac{\alpha_k \phi(y_j \mid \theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j \mid \theta_k)}$$

# 贝 叶 斯 分 类



- 输入：观测数据$y_1,y_2,\ldots y_N$，高斯混合模型
- 输出：高斯混合模型参数
- 3、M步，计算新一轮迭代的模型参数：

$$\hat{\mu}_k=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}y_j}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}\qquad \hat{\sigma}_k^2=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}(y_j-\mu_k)^2}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}\qquad \hat{\alpha}_k=\frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}}{N}$$

- 4、重复2，3步直到收敛

# 感 谢 观 看

## 统 计 机 器 学 习

主讲人：彭振华

数学与计算机学院

2026年