机器学习

第五章 贝叶斯分类

鲍鹏
北京交通大学

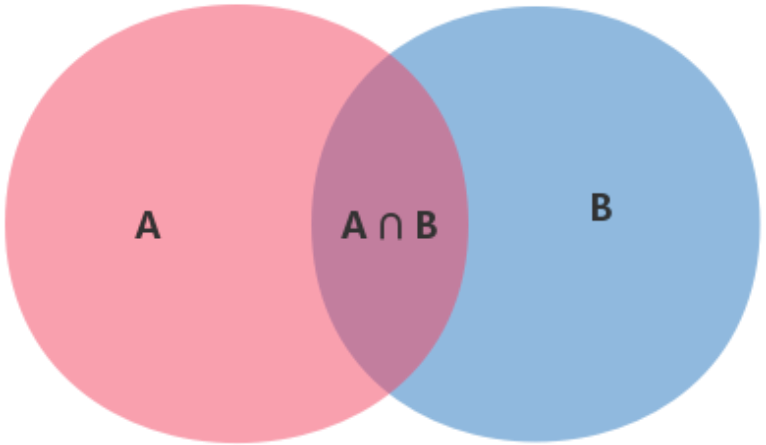
目录

目录

01 概率知识回顾

条件概率

条件概率（Conditional Probability）是指在事件B发生的情况下，事件A发生的概率，用$P(A|B)$表示。



上方的文氏图中，描述了两个事件A和B与它们的交集$A \cap B$，根据条件概率公式，可推出事件A与事件B同时发生的概率为

$$P(A \cap B) = P(A|B)P(B)$$

01 概率知识回顾

条件概率

$$P(A \cap B) = P(A|B)P(B)$$

对以上公式稍稍变换可得

$$P(A \cap B) = P(B|A)P(A)$$

由上式可推知

$$P(A|B)P(B) = P(B|A)P(A)$$

01 概率知识回顾

- 举例

假设两个人在扔两个六面的骰子D1与D2，预测D1与D2的向上面的结果的概率。

Table 1

+		D2					
		1	2	3	4	5	6
D1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

在Table1中描述了一个含有36个结果的样本空间，标红处为D1的向上面为2的6个结果，其概率为

$$P(D1 = 2) = \frac{6}{36} = \frac{1}{6}$$

01 概率知识回顾

Table 1

+		D2					
		1	2	3	4	5	6
D1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Table 2

+		D2					
		1	2	3	4	5	6
D1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Table2描述了D1 + D2 <= 5的概率，一共10个结果，用条件概率公式表示为

$$P(D1 + D2 \leq 5) = \frac{10}{36}$$

01 概率知识回顾

Table 1

+		D2					
		1	2	3	4	5	6
D1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Table 2

+		D2					
		1	2	3	4	5	6
D1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Table 3

+		D2					
		1	2	3	4	5	6
D1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Table3描述了满足Table2的条件同时也满足D1 = 2的结果，它选中了Table2中的3个结果，用条件概率公式表示为

$$P(D1 = 2|D1 + D2 \leq 5) = \frac{3}{10}$$

01 概率知识回顾

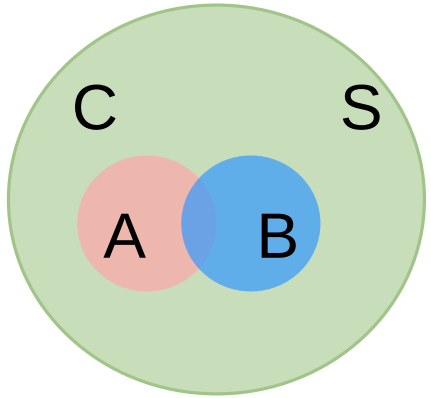
全概率公式：是将边缘概率与条件概率关联起来的基本规则，它表示了一个结果的总概率，可以通过几个不同的事件来实现。

全概率公式将对一复杂事件的概率求解问题转化为了在不同情况下发生的简单事件的概率的求和问题，公式为

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

01 概率知识回顾

假定一个样本空间S，它是两个事件A与C之和，同时事件B与它们两个都有交集，如下图所示



事件B的概率可以表示为

$$P(B) = P(B \cap A) + P(B \cap C)$$

通过条件概率，可以推断出$P(B \cap A) = P(B|A)P(A)$，所以

$$P(B) = P(B|A)P(A) + P(B|C)P(C)$$

这就是全概率公式，即事件B的概率等于事件A与事件C的概率分别乘以B对这两个事件的条件概率之和

01 概率知识回顾

举例：

假设有两家工厂生产并对外提供电灯泡，工厂X生产的电灯泡在99%的情况下能够工作超过5000小时，工厂Y生产的电灯泡在95%的情况下能够工作超过5000小时。工厂X在市场的占有率为60%，工厂Y为40%，如何推测出购买的灯泡的工作时间超过5000小时的概率是多少呢？

01 概率知识回顾

购买到工厂X制造的电灯泡的概率

$$\mathrm{P}(B_x) = \frac{6}{10}$$

购买到工厂Y制造的电灯泡的概率

$$\mathrm{P}(B_y) = \frac{4}{10}$$

工厂x制造的电灯泡工作时间超过5000小时的概率

$$\mathrm{P}(A|B_x) = \frac{99}{100}$$

工厂Y制造的电灯泡工作时间超过5000小时的概率

$$\mathrm{P}(A|B_y) = \frac{95}{100}$$

01 概率知识回顾

$\Pr(B_x)=\frac{6}{10}$，$\Pr(B_y)=\frac{4}{10}$，$\Pr(A|B_x)=\frac{99}{100}$，$\Pr(A|B_y)=\frac{95}{100}$

运用全概率公式可得：

$$\Pr(A)=\Pr(A|B_x)\cdot\Pr(B_x)+\Pr(A|B_y)\cdot\Pr(B_y)=\frac{99}{100}\cdot\frac{6}{10}+\frac{95}{100}\cdot\frac{4}{10}$$

$$=\frac{594+380}{1000}=\frac{974}{1000}$$

01 概率知识回顾

贝叶斯(Thomas Bayes,1701-1761)英国牧师、业余数学家。在《论机会学说中一个问题的求解》中给出了贝叶斯定理。

贝叶斯公式

$$P(A|B) = \frac{P(B,A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

01 概率知识回顾

它解决了两个事件条件概率的转换问题

- 先验概率：由以往的数据分析得到的概率
- 后验概率：得到“结果”的信息后重新修正的概率

简单的说，贝叶斯定理是基于假设的先验概率、给定假设下观察到不同数据的概率，提供了一种计算后验概率的方法。

01 概率知识回顾

贝叶斯定理

贝叶斯定理是关于随机事件A和B的条件概率的一则定理。通常，事件A在事件B（发生）的条件下的概率，与事件B在事件A（发生）的条件下的概率是不一样的，但它们两者之间是有确定的关系的，贝叶斯定理陈述了这个关系。

条件概率公式

$$P(A|B)P(B) = P(B|A)P(A)$$

对上式稍稍变换可得

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

01 概率知识回顾

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$：在事件B发生条件下的事件A发生的概率，在贝叶斯定理中，条件概率也被称为**后验概率**，即在事件B发生之后，对事件A发生概率的重新评估。

$P(B|A)$：在事件A发生条件下的事件B发生的概率，与上一条同理。

$P(A)$**与**$P(B)$**被称为先验概率（也被称为边缘概率）**，$P(A)$是指事件B发生之前，对事件A概率的一个推断（不考虑任何事件B方面的因素），$P(B)$同理。

$P(B|A)/P(B)$**被称为标准相似度**，它是一个调整因子，主要为保证预测概率更接近真实概率。

根据以上术语，**贝叶斯定理可表述为：后验概率 = 标准相似度 * 先验概率**

01 概率知识回顾

贝叶斯定理

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

根据全概率公式可得

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

因此

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

01 概率知识回顾

举例：

假设某种疾病的发病率为0.001（1000个人中会有1个人得病），现有一种试剂在患者确实得病的情况下，有99%的几率呈现为阳性，而在患者没有得病的情况下，它有5%的几率呈现为阳性（假阳性），如有一位病人的检验成果为阳性，那么他的得病概率是多少呢？

01 概率知识回顾

假定事件A表示为得病的概率，P(A) = 0.001，它是在病人在实际注射试剂（缺乏实验的结果）之前预计的发病率

假定事件B为试剂结果为阳性的概率，需要计算条件概率$P(A|B)$，即病人在注射试剂之后（得到实验结果）得出的发病率

假设事件C为未得病的先验概率$P(C)=1-P(A)=0.999$，$P(B|C)$表示未得病条件下的试剂结果为阳性的概率，应用全概率公式就可得出最终结果

$P(A|B)=(P(B|A)P(A))/(P(B))=(P(B|A)P(A))/(P(B|A)P(A)+P(B|C)P(C))=(0.99\times0.001)/(0.99\times0.001+0.05\times0.999)\approx0.019$

目录

02 贝叶斯决策论

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。

02 贝叶斯决策论

以一个多分类任务为例：假设有N种可能的类别标记，即$y = \{c_1, c_2, \dots, c_N\}$

- **【损失】**：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。
- **【条件风险】**：单个样本$\boldsymbol{x}$的期望损失（expected loss），即在样本上的“条件风险”（conditional risk）

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

其中，$P(c_j|\boldsymbol{x})$为后验概率。

- **【总体风险】**：全部样本构成的总体风险为

$$R(h) = \mathbb{E}_{\boldsymbol{x}}[R(h(\boldsymbol{x})|\boldsymbol{x})]$$

其中，h为分类器（模型）。显然，分类效果越准确的h，条件风险和总体风险越小。

02 贝叶斯决策论

贝叶斯判定准则（Bayes decision rule）：为最小化总体风险$R(h)$，只需要在每个样本上选择哪个能使条件风险$R(c|\boldsymbol{x})$最小的类别标记，即

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

此时，h^*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。

02 贝叶斯决策论

具体来说，若目标是最小化分类错误率，则λ_{ij}误判损失可写为

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

此时单个样本$\boldsymbol{x}_i$的条件风险

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

$$R(c_i|\boldsymbol{x}) = 1 * P(c_1|\boldsymbol{x}) + \cdots + 1 * P(c_{i-1}|\boldsymbol{x}) + 0 * P(c_i|\boldsymbol{x}) + \cdots + 1 * P(c_N|\boldsymbol{x})$$

又$\sum_{j=1}^{N} P(c_j|\boldsymbol{x}) = 1$，则

$$R(c_i|\boldsymbol{x}) = 1 - P(c_i|\boldsymbol{x})$$

02 贝叶斯决策论

于是，按照贝叶斯判定准则，最小化分类错误率的贝叶斯最优分类器为

$$h^*(\boldsymbol{x}) = \underset{c \in y}{\operatorname{argmin}} R(c|\boldsymbol{x}) = \underset{i \in \{1,2,\ldots,N\}}{\operatorname{argmin}} R(c_i|\boldsymbol{x})$$

$$h^*(\boldsymbol{x}) = \underset{i \in \{1,2,\ldots,N\}}{\operatorname{argmin}} 1 - P(c_i|\boldsymbol{x})$$

$$h^*(\boldsymbol{x}) = \underset{i \in \{1,2,\ldots,N\}}{\operatorname{argmax}} P(c_i|\boldsymbol{x})$$

$$h^*(\boldsymbol{x}) = \underset{c \in y}{\operatorname{argmax}} P(c|\boldsymbol{x})$$

后验概率最大化

又每个样本$\boldsymbol{x}$，选择后验概率$P(c_i|\boldsymbol{x})$最大的类别c_i作为标记

02 贝叶斯决策论

主要有两种策略：

➢**判别式模型（discriminative models）**：给定$\boldsymbol{x}$，通过直接建模$P(c|\boldsymbol{x})$, 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$P(\boldsymbol{x}, c)$建模，然后再由此获得$P(c|\boldsymbol{x})$。生成式模型考虑

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

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）

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

“证据”（evidence）因子，与类标记无关

目录

03 极大似然估计

参数估计基本概念

1. 统计量——样本中包含着总体的信息，我们希望通过样本集把有关信息抽取出来，就是说针对不同要求构造出样本的某种函数，这种函数在统计学中称为统计量。

2. 参数空间——如上所述，在参数估计中，我们总是假设总体概率密度函数的形式已知，而未知的仅是分布中的几个参数，将未知参数记为θ，在统计学中，我们将总体分布未知参数θ的全部可容许值组成的集合称为参数空间。

03 极大似然估计

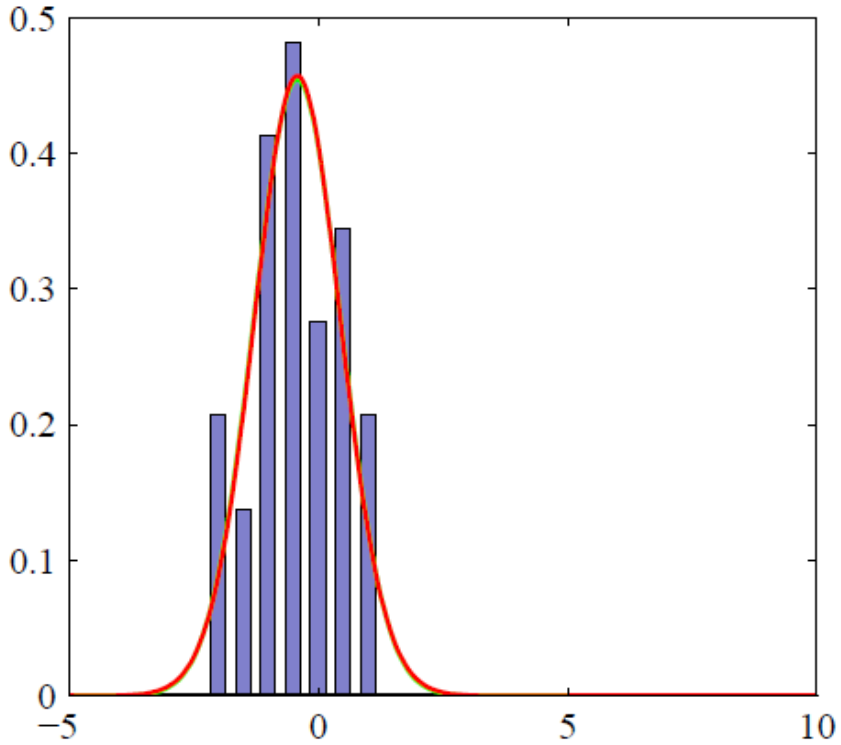
参数估计基本概念

3. 点估计、估计量和估计值——点估计问题就是要构造一个统计量$d(x_1, \ldots, x_N)$作为参数θ的估计$\hat{\theta}$，在统计学中称$\hat{\theta}$为θ的估计量。如果, …, $x_N{}^{(i)}$是属于类别w_i的几个样本观察值，代入统计量d就得到对于第i类的$\hat{\theta}$的具体数值，这个数值在统计学中称为θ的估计值。

4. 区间估计——除点估计外，还有另一类估计，它要求用区间 (d_1, d_2)作为θ可能取值范围的一种估计。这个区间称为置信区间，这类估计问题称为区间估计。

03 极大似然估计

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$P(\boldsymbol{x}|c)$，假设$P(\boldsymbol{x}|c)$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$\boldsymbol{\theta}_c$。

03 极大似然估计

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的方案：

- 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。

03 极大似然估计

令D_c表示训练集D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

$$P(D_c|\boldsymbol{\theta}_c) = \prod_{\boldsymbol{x}\in D_c} P(\boldsymbol{x}|\boldsymbol{\theta}_c)$$

似然函数

对$\boldsymbol{\theta}_c$进行极大似然估计，寻找能**最大化似然$\boldsymbol{P(D_c|\theta_c)}$的参数值$\widehat{\boldsymbol{\theta}_c}$**。直观上看，极大似然估计是试图在$\boldsymbol{\theta}_c$所有可能的取值中，找到一个使数据出现的“可能性”最大值。连乘操作易造成下溢，通常使用对数似然(log-likelihood)

$$LL(\boldsymbol{\theta}_c) = logP(D_c|\boldsymbol{\theta}_c) = \sum_{\boldsymbol{x}\in D_c} \log P(\boldsymbol{x}|\boldsymbol{\theta}_c)$$

此时参数$\boldsymbol{\theta}_c$的极大似然估计$\widehat{\boldsymbol{\theta}_c}$为：$\widehat{\boldsymbol{\theta}_c} = \underset{\boldsymbol{\theta}_c}{\operatorname{argmax}} LL(\boldsymbol{\theta}_c)$

03 极大似然估计

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

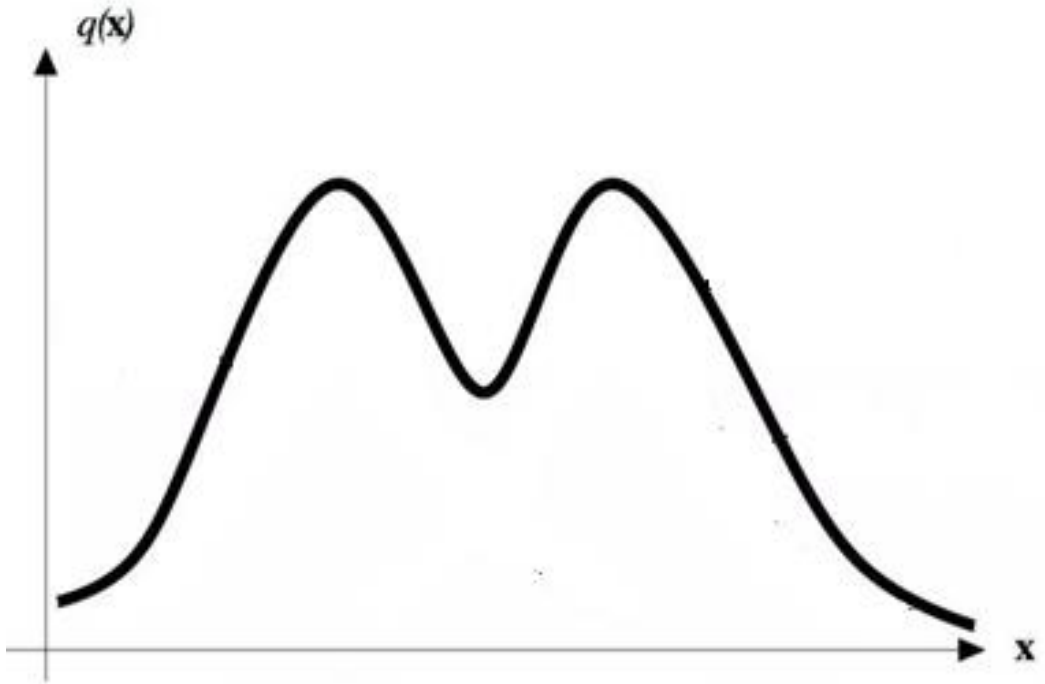
$$\widehat{\mu_c}=\frac{1}{|D_c|}\sum_{\boldsymbol{x}\in D_c}\boldsymbol{x}$$

$$\widehat{\sigma_c^2}=\frac{1}{|D_c|}\sum_{\boldsymbol{x}\in D_c}(\boldsymbol{x}-\widehat{\mu_c})\,(\boldsymbol{x}-\widehat{\mu_c})^{\mathrm{T}}$$

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$(\boldsymbol{x}-\widehat{\mu_c})(\boldsymbol{x}-\widehat{\mu_c})^{\mathrm{T}}$的均值，这显然是一个符合直觉的结果。

03 极大似然估计

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。**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。



目录

04 朴素贝叶斯分类器

设输入空间$\mathcal{X} \subseteq \boldsymbol{R^n}$为$n$维向量的集合，输出空间为类标记集合$\mathcal{Y} = \{c_1, c_2, \dots, c_K\}$，输入为特征向量$x \in \mathcal{X}$，输出为类标记$y \in \mathcal{Y}$。

- X是定义在输入空间$\mathcal{X}$上的随机向量，Y是定义在输入空间$\mathcal{Y}$上的随机变量。
- $P(X, Y)$是X和Y的联合概率分布。

训练数据集$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$由$P(X, Y)$独立同分布产生。

$$h^*(\boldsymbol{x}) = \underset{c \in y}{\operatorname{argmax}} P(c|\boldsymbol{x})$$

后验概率最大化

$$h^*(x) = \underset{c_k \in \mathcal{Y}}{\operatorname{argmax}} P(Y = c_k | X = x)$$

04 朴素贝叶斯分类器

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。

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)P(X=x|Y=c_k)}{P(X=x)}$$

条件独立性假设

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\middle|Y=c_k\right)$$

$$=\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$$

分类的特征在类确定的条件下都是条件独立的。

04 朴素贝叶斯分类器

计算后验概率

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}{\sum_{k=1}^{K}P(Y=c_k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}$$

朴素贝叶斯分类器表示

$$y=h_{nb}(x)=\underset{c_k}{\operatorname{argmax}}\frac{P(Y=c_k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}{\sum_{k=1}^{K}P(Y=c_k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}$$

分母对所有c_k都是相同的，所以

$$y=\underset{c_k}{\operatorname{argmax}}P(Y=c_k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$$

04 朴素贝叶斯分类器

在朴素贝叶斯法中，**学习意味着估计**$P(Y=c_k)$**和**$P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$**，可应用极大似然估计法估计。**

先验概率$P(Y=c_k)$的极大似然估计：

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)}{N}\ ,\ k=1,2,\cdots K$$

04 朴素贝叶斯分类器

条件概率$P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$的极大似然估计

设第j个特征$x^{(j)}$可能取值的集合为$\{a_{j1},a_{j2},\cdots,a_{jS_j}\}$，条件概率$P\left(X^{(j)}=a_{jl}\middle|Y=c_k\right)$的极大似然估计为

$$P\left(X^{(j)}=a_{jl}\middle|Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(x_i^{(j)}=a_{jl},y_i=c_k\right)}{\sum_{i=1}^{N}I(y_i=c_k)}$$

式中，$x_i^{(j)}$是第i个样本的第j个特征；a_{jl}是第j个特征可能取的第l个值；

$j=1,2,\ldots,n$　　$l=1,2,\ldots,S_j$　　$k=1,2,\ldots,K$

I为指示函数$I_A(x)=\begin{cases}1, & x\in A\\ 0, & x\notin A\end{cases}$

04 朴素贝叶斯分类器

例子：

大小	颜色	形状	大小
大	红色	圆形	?

c_1=是（好果）　c_2=否（坏果）

$$P(Y=c_1)=\frac{4}{10}\ ,\ P(Y=c_2)=\frac{6}{10}$$

编号	大小	颜色	形状	好果
1	小	青	非规则	否
2	大	红	非规则	是
3	大	红	圆形	是
4	大	青	圆形	否
5	大	青	非规则	否
6	小	红	圆形	是
7	大	青	非规则	否
8	小	红	非规则	否
9	小	青	圆形	否
10	大	红	圆形	是

04 朴素贝叶斯分类器

条件独立假设：

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\middle|Y=c_k\right)$$

$$=\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)$$

$$P\left(X^{(1)}=\text{大},X^{(2)}=\text{红},X^{(3)}=\text{圆形}|Y=c_1\right)=P\left(X^{(1)}=\text{大}\,|Y=c_1\right)*P\left(X^{(2)}=\text{红}|Y=c_1\right)*P\left(X^{(3)}=\text{圆形}|Y=c_1\right)$$

$$=\frac{2}{4}*\frac{3}{4}*\frac{2}{4}=\frac{3}{16}$$

$$P\left(X^{(1)}=\text{大},X^{(2)}=\text{红},X^{(3)}=\text{圆形}|Y=c_2\right)=P\left(X^{(1)}=\text{大}\,|Y=c_2\right)*P\left(X^{(2)}=\text{红}|Y=c_2\right)*P\left(X^{(3)}=\text{圆形}|Y=c_2\right)$$

$$=\frac{4}{6}*\frac{2}{6}*\frac{3}{6}=\frac{1}{9}$$

04 朴素贝叶斯分类器

朴素贝叶斯算法

输入：训练数据$D=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$，其中$x_i=(x_i^{(1)},x_i^{(2)},\dots,x_i^{(n)})^{\mathrm{T}}$，$x_i^{(j)}$是第$i$个样本的第$j$个特征，$x_i^{(j)}\in\{a_{j1},a_{j2},\dots,a_{jS_j}\}$，$a_{jl}$是第$j$个特征可能取得第$l$个值，$j=1,2,\dots,n$，$l=1,2,\dots,S_j$，$y_i\in\{c_1,c_2,\dots,c_K\}$；实例$x$。

输出：实例x的分类。

04 朴素贝叶斯分类器

（1）计算先验概率及条件概率

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N}, \quad k=1,2,\dots,K$$

$$P\left(X^{(j)}=a_{jl}|Y=c_k\right)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)},$$

$$j=1,2,\dots,n, \qquad l=1,2,\dots,S_j, \qquad k=1,2,\dots,K$$

04 朴素贝叶斯分类器

（2）对于给定的实例$x = (x^{(1)}, x^{(2)}, \dots, x^{(n)})^{\mathrm{T}}$，计算

$$P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k), \qquad k = 1,2, \dots, K$$

（3）确实实例x**的类**

$$y = \underset{c_k}{\operatorname{argmax}} P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)$$

04 朴素贝叶斯分类器

例子：试由下表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，$Y\in C=\{1,-1\}$。

	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

04 朴素贝叶斯分类器

解：根据朴素贝叶斯算法，首先计算下列概率

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

$$P\left(X^{(1)}=1|Y=1\right)=\frac{2}{9}\qquad P\left(X^{(1)}=2|Y=1\right)=\frac{3}{9}\qquad P\left(X^{(1)}=3|Y=1\right)=\frac{4}{9}$$

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

$$P\left(X^{(1)}=1|Y=-1\right)=\frac{3}{6}\qquad P\left(X^{(1)}=2|Y=-1\right)=\frac{2}{6}\qquad P\left(X^{(1)}=3|Y=-1\right)=\frac{1}{6}$$

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

04 朴素贝叶斯分类器

对于给定的$x=(2,S)^{\mathrm{T}}$，计算

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

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

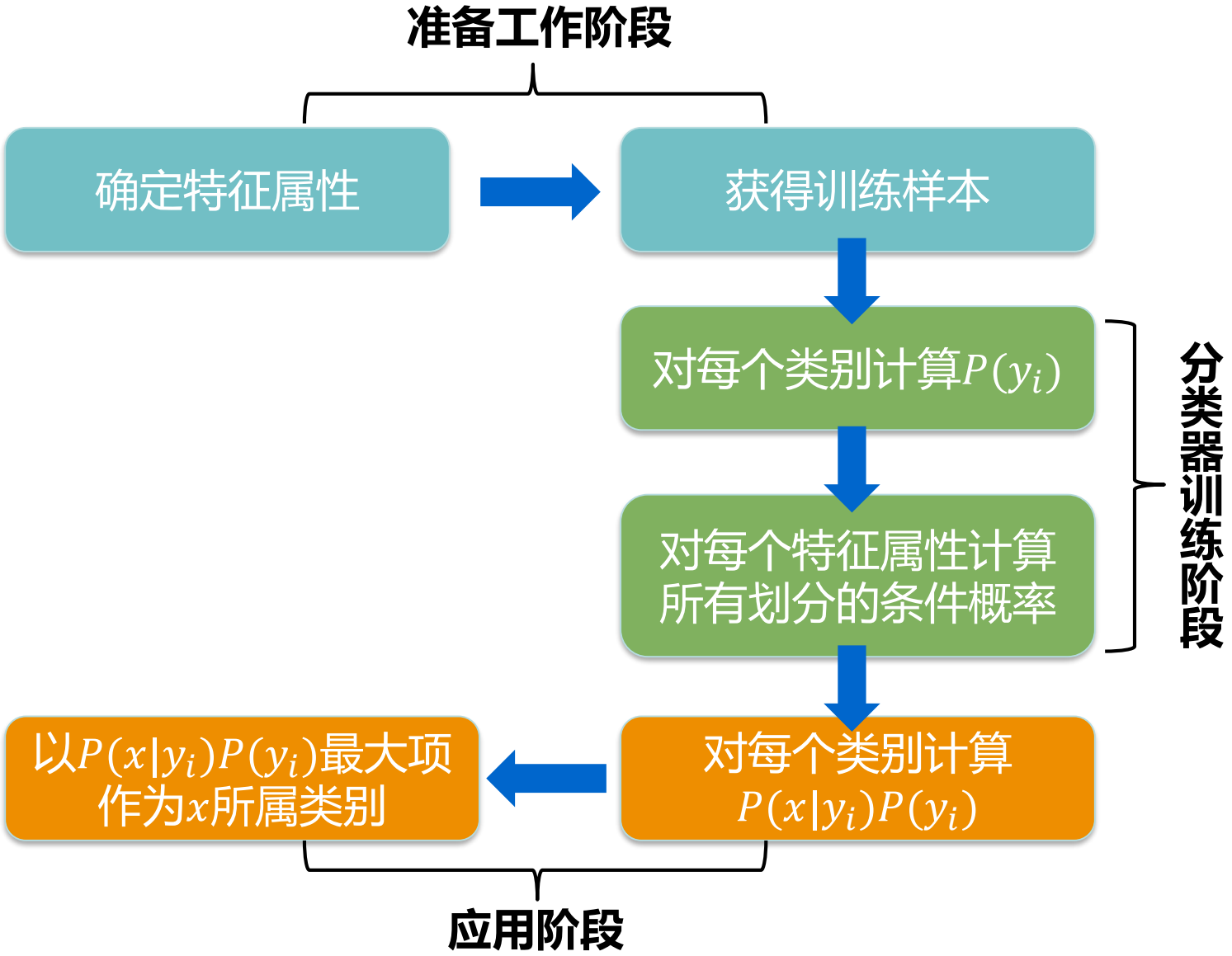
由于$P(Y=-1)P\left(X^{(1)}=2\middle|Y=-1\right)P\left(X^{(2)}=S\middle|Y=-1\right)$最大，所以$y=-1$。

04 朴素贝叶斯分类器

- 朴素贝叶斯法是基于贝叶斯定理和特征条件独立假设分类方法。对于给定训练集，首先基于特征条件独立性的假设，学习输入/输出联合概率（计算出先验概率和条件概率，然后求出联合概率）。然后基于此模型，给定输入x，利用贝叶斯概率定理求出最大的后验概率作为输出y。朴素贝叶斯法实现简单，学习和预测效率都很高，是一种常用的分类方法。
- **基本思想**：对于给定的待分类项x，求解在此样本出现的条件下各个类别出现的概率，计算出每一个类别的$P(y_i|x), i = 1,2,\dots,k$，根据**一定的决策规则**，决定此样本归属于哪个类别。

04 朴素贝叶斯分类器

- 朴素贝叶斯分类的流程



04 朴素贝叶斯分类器

整个朴素贝叶斯分类分为三个阶段：

- 第一阶段——准备工作阶段，这个阶段的任务是为朴素贝叶斯分类做必要的准备，主要工作是根据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。这一阶段的输入是所有待分类数据，输出是特征属性和训练样本。这一阶段是整个朴素贝叶斯分类中唯一需要人工完成的阶段，其质量对整个过程将有重要影响，分类器的质量很大程度上由特征属性、特征属性划分及训练样本质量决定。
- 第二阶段——分类器训练阶段，这个阶段的任务就是生成分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。其输入是特征属性和训练样本，输出是分类器。这一阶段是机械性阶段，根据前面讨论的公式可以由程序自动计算完成。
- 第三阶段——应用阶段。这个阶段的任务是使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。这一阶段也是机械性阶段，由程序完成。

目录

05 EM算法

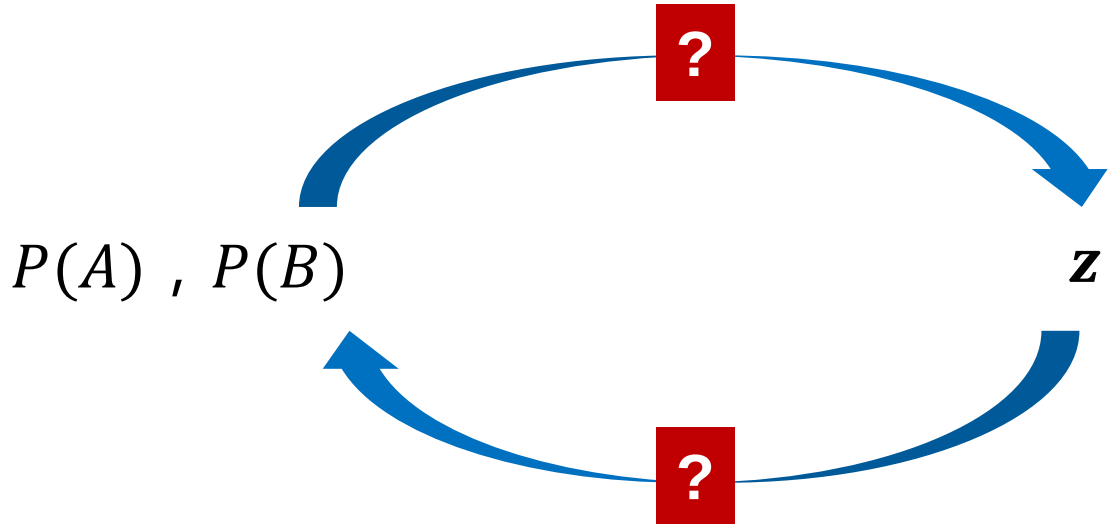
三硬币模型：假设有3枚硬币，分别记作A，B，C。这些硬币正面出现的概率分别是π，p和q。进行如下掷硬币试验：先掷硬币A，根据其结果选出硬币B或硬币C，正面选硬币B，反面选硬币C；然后投掷选出的硬币，根据投掷硬币的结果，出现整门记作1，出现反面记作0；独立地重复n次试验（这里，$n=10$），观测结果如下：

1，1，0，1，0，0，1，0，1，1

假设只能观测到投掷硬币的结果，不能观测投掷硬币的过程。问如何估计三硬币整门出现的概率，即三硬币模型的参数。

05 EM算法

假设我们把每一次硬币的种类设为z，则这10次实验生成了一个10维的向量$z = \{z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8, z_9, z_{10}\}$，现在问题来了，如果我们要根据观测结果去求出$P(A)$，$P(B)$，那么首先需要知道$z$，但是如果用最大似然估计去估计$z$，又要先求出$P(A)$，$P(B)$。这就产生了一个循环。



解：三硬币模型可以写作

$$P(y|\theta)=\sum_z P(y,z|\theta)=\sum_z P(z|\theta)\,P(y|z,\theta)$$

$$=\pi p^y(1-p)^{1-y}+(1-\pi)\,q^y(1-q)^{1-y}$$

这里，随机变量y是观测变量，表示一次试验观测的结果是1或者0；随机变量z是隐变量，表示未观测到的投掷硬币A的结果；$\theta=(\pi,p,q)$是模型参数。

注意，随机变量y的数据可以观测，随机变量z的数据不可观测。

05 EM算法

将观测数据表示为$Y = (Y_1, Y_2, \dots, Y_n)^{\mathrm{T}}$，未观测数据表示为$Z = (Z_1, Z_2, \dots, Z_n)^{\mathrm{T}}$，则观测数据的似然函数为

$$P(Y|\theta) = \sum_Z P(Z|\theta)P(Y|Z,\theta)$$

即

$$P(Y|\theta) = \prod_{j=1}^{n}[\pi p^{y_j}(1-p)^{1-y_j} + (1-\pi)q^{y_j}(1-q)^{1-y_j}$$

考虑求模型参数$\theta = (\pi, p, q)$的极大似然估计，即

$$\hat{\theta} = \arg\max_{\theta} \log P(Y|\theta)$$

这个问题没有解析解，只有通过迭代的方法求解。

05 EM算法

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θ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θ做极大似然估计(M步)

05 EM算法

EM算法

输入：观测变量数据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）选择参数的初值$\theta^{(0)}$,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

$$Q\left(\theta,\theta^{(i)}\right)=E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}]$$

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

这里，$P(Y,Z|\theta)$是在给定观测数据Y和当前的参数估计$\theta^{(i)}$下隐变量数据Z的条件概率分布。

EM算法

（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$\theta^{(i+1)}$

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

（4）重复第2步和第3步，直到收敛。

【Q函数】：完全数据的对数似然函数$\log P(Y, Z|\theta)$关于在给定观测数据Y和当前参数$\theta^{(i)}$下对未观测数据Z的条件概率分布$P(Z|Y, \theta^{(i)})$的期望称为Q函数，即

$$Q(\theta, \theta^{(i)}) = E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}]$$

05 EM算法

针对前面的三硬币模型例子：

EM算法首先选取参数的初值，记作$\theta^{(0)} = (\pi^{(0)}, p^{(0)}, q^{(0)})$

E步：计算在模型参数$\pi^{(i)}, p^{(i)}, q^{(i)}$下观测数据$y_i$来自投掷硬币B的概率

$$\mu_j^{(i+1)} = \frac{\pi^{(i)}\left(p^{(i)}\right)^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}\left(p^{(i)}\right)^{y_j}(1-p^{(i)})^{1-y_j}+(1-\pi^{(i)})\left(q^{(i)}\right)^{y_j}(1-q^{(i)})^{1-y_j}}$$

M步：计算模型参数的新估计值

$$\pi^{(i+1)} = \frac{1}{n}\sum_{j=1}^{n}\mu_j^{(i+1)} \qquad p^{(i+1)} = \frac{\sum_{j=1}^{n}\mu_j^{(i+1)}y_j}{\sum_{j=1}^{n}\mu_j^{(i+1)}} \qquad q^{(i+1)} = \frac{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})y_j}{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})}$$

05 EM算法

针对前面的三硬币模型例子：

假设模型参数的初值取为，$\pi^{(0)}=0.5, p^{(0)}=0.5, q^{(0)}=0.5$

对$y_j=1$与$y_j=0$，均值$\mu_j^{(1)}=0.5$

利用迭代公式得到：$\pi^{(1)}=0.5, p^{(1)}=0.6, q^{(1)}=0.6$

根据上述结果计算：均值$\mu_j^{(2)}=0.5$

继续迭代，得：$\pi^{(1)}=0.5, p^{(1)}=0.6, q^{(1)}=0.6$

无变化

于是得到模型参数θ的极大似然估计：$\hat{\pi}=0.5$，$\hat{p}=0.6$，$\hat{q}=0.6$

05 EM算法

思考：若$\pi^{(0)}, p^{(0)}, q^{(0)}$取初值变化，估计结果会有变化么？

假设模型参数的初值取为：

$$\pi^{(0)} = 0.4, p^{(0)} = 0.6, q^{(0)} = 0.7$$

于是得到模型参数θ的极大似然估计：

$$\hat{\pi} = 0.406\ ,\ \hat{p} = 0.5368\ ,\ \hat{q} = 0.6432$$

EM算法与初值的选择有关，选择不同的初值可能得到不同的参数估计值。

05 EM算法

EM算法几点说明

（1）步骤（1）中参数的初值可以任意选择，但需要注意EM算法对初值是敏感的。

（2）步骤（2）中E步求$Q(\theta,\theta^{(i)})$。Q函数式子中Z是未观测数据，Y是观测数据。注意，$Q(\theta,\theta^{(i)})$的第1个变元表示要极大化的参数，第2个变元表示参数的当前估计值。每次迭代实际在求Q函数及其极大。

（3）步骤（3）中M步求$Q(\theta,\theta^{(i)})$的极大化，得到$\theta^{(i+1)}$，完成一次迭代$\theta^{(i)} \to \theta^{(i+1)}$。

（4）步骤（4）给出停止迭代的条件，一般是对较小的正数ε_1，ε_2，若满足

$$||\theta^{(i+1)} - \theta^{(i)}|| < \varepsilon_1 \quad 或 \quad ||Q(\theta^{(i+1)},\theta^{(i)}) - Q(\theta^{(i)},\theta^{(i)})|| < \varepsilon_2$$

则停止迭代。

EM算法的导出

面对一个含有隐变量的概率模型，目标是极大化观测数据（不完全数据）Y关于参数θ的对数似然函数，即极大化

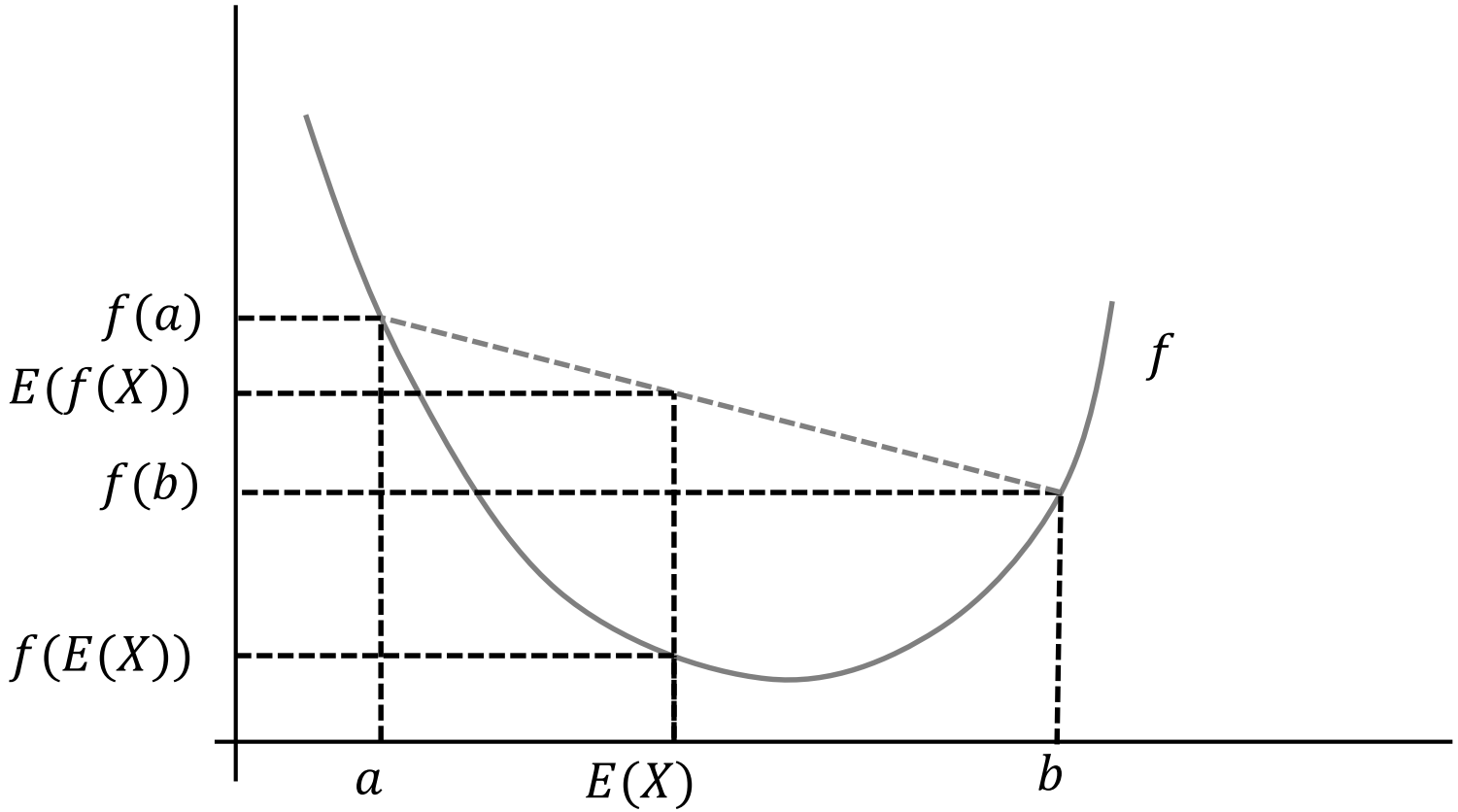
$$L(\theta) = \log P(Y|\theta) = \log \sum_Z P(Y,Z|\theta) = \log \left(\sum_Z P(Y|Z,\theta) P(Z|\theta) \right)$$

$$L(\theta) - L\left(\theta^{(i)}\right) = \log \left(\sum_Z P(Y|Z,\theta) P(Z|\theta) \right) - \log P(Y|\theta^{(i)})$$

05 EM算法

- **Jensen不等式**

如果$f(x)$是凸函数，x是随机变量，$\frac{\sum_{i=1}^{n} f(x_i)}{n} \geq f(\frac{\sum_{i=1}^{n} x_i}{n})$，即$E(f(X)) \geq f(E(X))$



05 EM算法

- 为什么EM算法能近似实现对观测数据的极大似然估计？
- 极大化观测数据（不完全数据）Y关于参数θ的对数似然函数：

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

$$= \log \left(\sum_{Z} P(Y|Z, \theta) P(Z|\theta) \right)$$

- 难点：有未观测数据，包含和（或积分）的对数
- EM算法通过迭代逐步近似极大化$L(\theta)$，希望

$$L(\theta) > L\left(\theta^{(i)}\right)$$

05 EM算法

考虑二者的差：

$$L(\theta)-L\left(\theta^{(i)}\right)=\log\left(\sum_{Z}P(Y|Z,\theta)P(Z|\theta)\right)-\log P(Y|\theta^{(i)})$$

利用Jensen不等式（Jensen inequality）得到其下界

$$L(\theta)-L\left(\theta^{(i)}\right)=\log\left(\sum_{Z}P\left(Y\middle|Z,\theta^{(i)}\right)\frac{P(Y|Z,\theta)P(Z|\theta)}{P\left(Y\middle|Z,\theta^{(i)}\right)}\right)-\log P(Y|\theta^{(i)})$$

$$\geq\sum_{Z}P(Z|Y,\theta^{(i)})\log\left(\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}\right)$$

$$=\sum_{Z}P\left(Z\middle|Y,\theta^{(i)}\right)\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P\left(Z\middle|Y,\theta^{(i)}\right)P(Y|\theta^{(i)})}$$

05 EM算法

令

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

则

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

$L(\theta)$的一个下界

$$L(\theta^{(i)}) = B(\theta^{(i)}, \theta^{(i)})$$

因此，任何可以使$B(\theta,\theta^{(i)})$增大的θ也可以使$L(\theta)$增大。为了使$L(\theta)$有尽可能大的增长，选择$\theta^{(i+1)}$使$B(\theta,\theta^{(i)})$达到极大，即

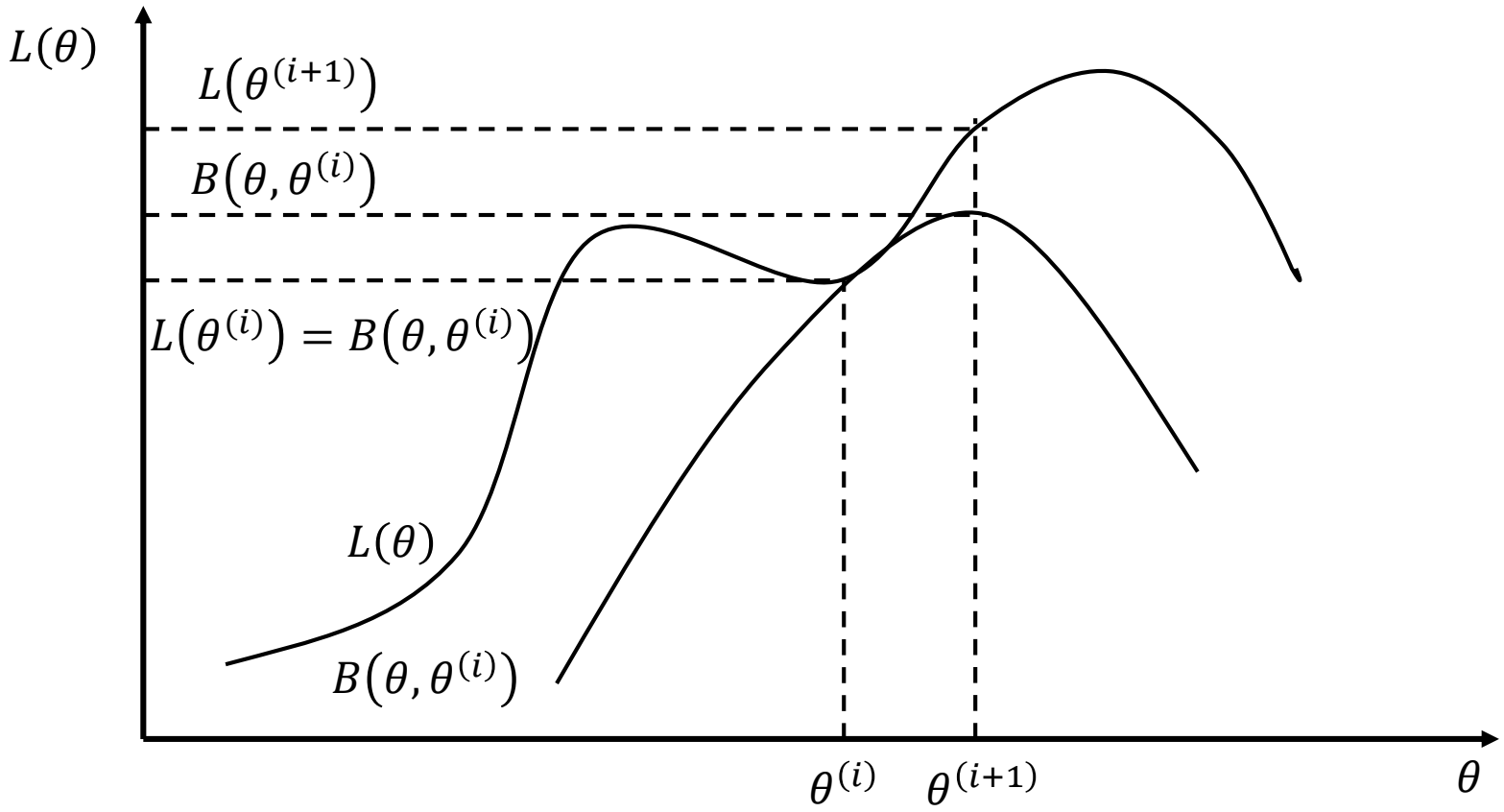
$$\theta^{(i+1)}=\arg\max_{\theta} B(\theta,\theta^{(i)})$$

现求$\theta^{(i+1)}$的表达式。省去对θ的极大化而言是常数的项：

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

05 EM算法

EM不能保证找到全局最优解



05 EM算法

- EM算法提供一种近似计算含有隐变量概率模型的极大似然估计的方法
- EM算法最大优点：简单性和普适性
- 疑问：
- 1. EM算法得到的估计序列是否收敛？
- 2. 如果收敛，是否是全局极大值或局部极大值？

05 EM算法

EM算法的收敛性

【定理1】设$P(Y|\theta)$为观测数据的似然函数，$\theta^{(i)}(i=1,2,\dots)$为EM算法得到的估计序列，$P(Y|\theta^{(i)})(i=1,2,\dots)$为对应的似然函数序列，则$P(Y|\theta^{(i)})$是单调递增的，即

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

- 证明：由于

$$P(Y|\theta)=\frac{P(Y,Z|\theta)}{P(Z|Y,\theta)}$$

取对数有

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

由：

$$Q(\theta,\theta^{(i)})=\sum_{Z}logP(Y|Z,\theta)P(Z|Y,\theta^{(i)})$$

于是对数似然函数可以写成：

$$logP(Y|\theta)=Q(\theta,\theta^{(i)})-H(\theta,\theta^{(i)})$$

得：

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

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

05 EM算法

- 只需证右端非负
- 右端第一项， 由于$\theta^{(i+1)}$使$Q\left(\theta,\theta^{(i)}\right)$达到极大，所以

$$Q\left(\theta^{(i+1)},\theta^{(i)}\right)-Q\left(\theta^{(i)},\theta^{(i)}\right)\geq 0$$

- 第二项:

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

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

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

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

EM算法的收敛性

【定理2】设$L(\theta) = \log P(Y|\theta)$为观测数据的对数似然函数，$\theta^{(i)}(i = 1,2,\dots)$为EM算法得到的估计序列，$L\left(\theta^{(i)}\right)(i = 1,2,\dots)$为对应的对数似然函数序列。

（1）如果$P(Y|\theta)$有上界，则$L\left(\theta^{(i)}\right) = \log P(Y|\theta^{(i)})$收敛到某一值$L^*$；

（2）在函数$Q(\theta,\theta')$与$L(\theta)$满足一定条件的情况下，由EM算法得到的参数估计序列$\theta^{(i)}$的收敛值θ^*是$L(\theta)$的稳定点。

EM算法在高斯混合模型学习中的应用

高斯混合模型

- **定义**：高斯混合模型是指具有如下形式的概率分布模型：

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

系数： $\alpha_k \geq 0$， $\sum_{k=1}^{K} \alpha_k = 1$

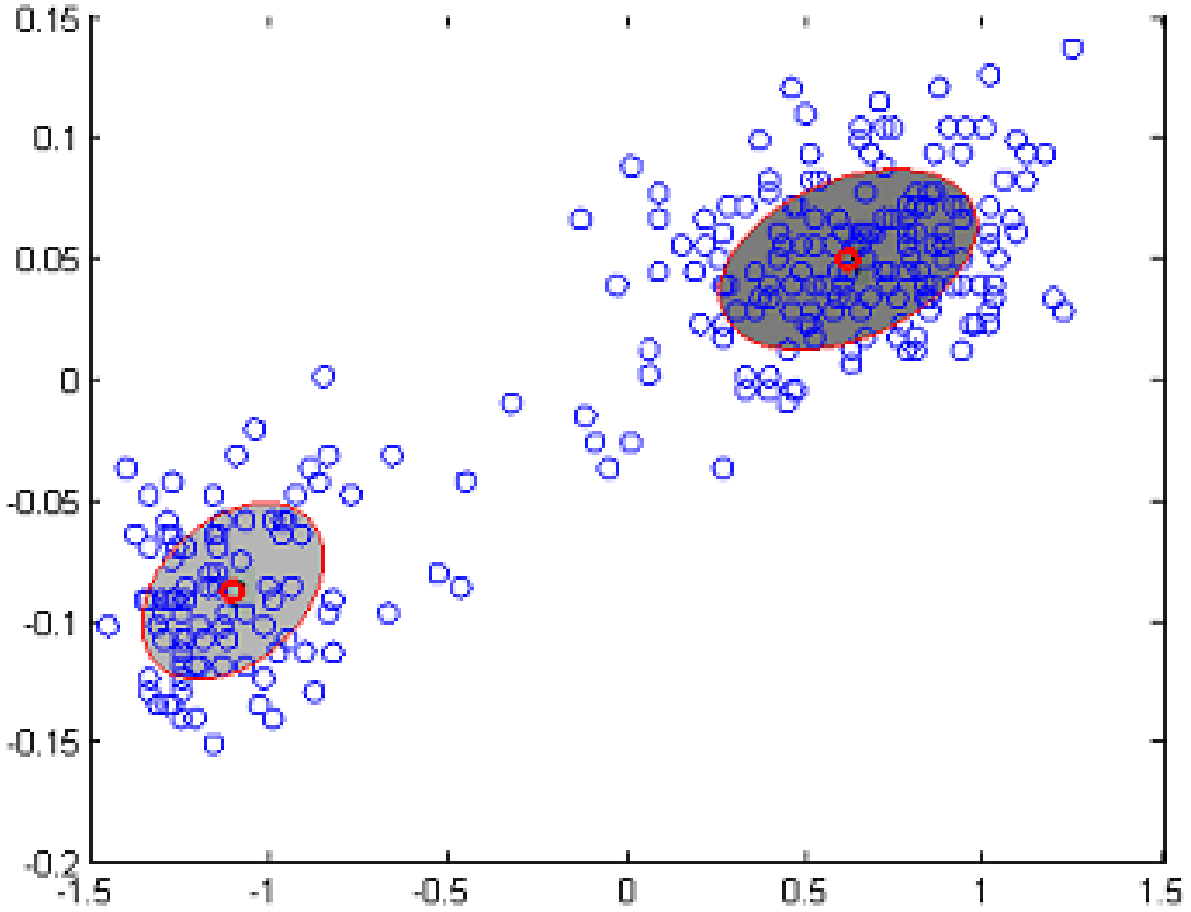
高斯分布密度： $\phi(y|\theta_k)$， $\theta_k = (\mu_k, \sigma_k^2)$

第K个分模型：

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

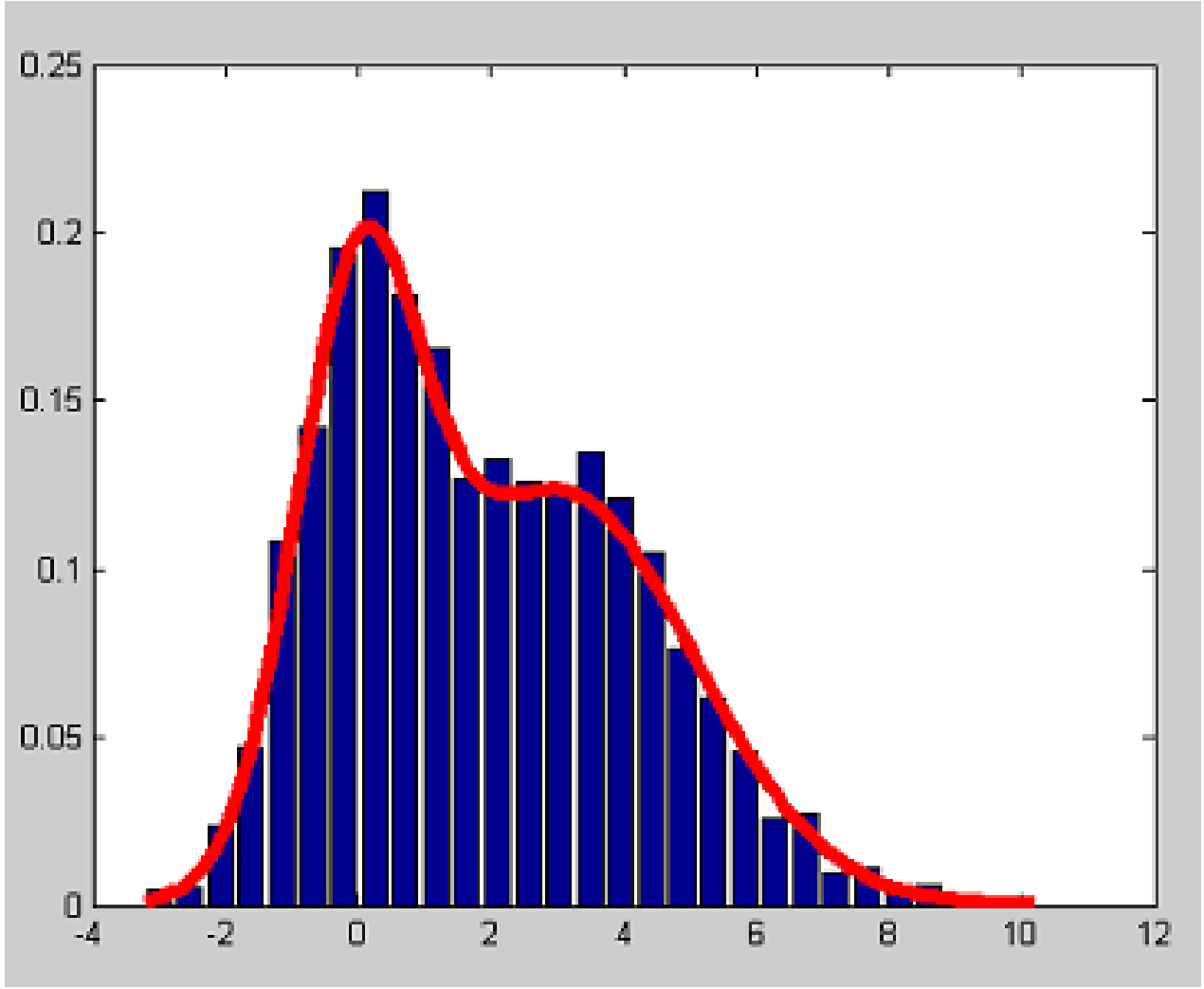
05 EM算法

◆高斯混合模型用于聚类



05 EM算法

◆高斯混合模型用于概率密度估计



05 EM算法

◆高斯混合模型参数估计的EM算法

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

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

$$\theta = (\alpha_1, \alpha_2, \dots, \alpha_k; \theta_1, \theta_2, \dots, \theta_k)$$

- 用EM算法估计参数θ

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$y_j, j = 1,2, \dots, N$，是依概率α_k选择第K个高斯分模型$\phi(y|\theta_k)$生成，隐变量

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

05 EM算法

◆高斯混合模型参数估计的EM算法

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

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

$$\theta = (\alpha_1, \alpha_2, \dots, \alpha_k; \theta_1, \theta_2, \dots, \theta_k)$$

- 用EM算法估计参数θ

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$y_j, j = 1,2, \dots, N$，是依概率α_k选择第K个高斯分模型$\phi(y|\theta_k)$生成，隐变量

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

05 EM算法

- 完全数据：$(y_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jK}), j = 1, 2, \cdots, N$
- 似然函数：

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

$$= \prod_{K=1}^{K} \prod_{j=1}^{N} [\alpha_k \phi(y_j|\theta_K)]^{\gamma_{jK}}$$

$$= \prod_{k=1}^{K} \alpha_k^{n_k} \prod_{j=1}^{N} [\phi(y_j|\theta_K)]^{\gamma_{jK}}$$

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

$$n_k = \sum_{j=1}^{N} \gamma_{jK}$$

$$\sum_{k=1}^{K} n_k = N$$

05 EM算法

- 完全数据的对数似然函数为：

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

2. EM算法的E步：确定Q函数

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

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

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

需要计算$\mathrm{E}\left(\gamma_{jk}|y,\theta\right)$，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。

05 EM算法

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

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

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

$$= \frac{\alpha_k \phi(y_j|\theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j|\theta_k)}\ ,\ j = 1,2,\dots,N; k = 1,2,\dots,K$$

05 EM算法

将$\hat{\gamma}_{jk} = E\gamma_{jk}$及$n_k = \sum_{j=1}^{N} E\gamma_{jK}$代入

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

3. 确定EM算法的M步

求新一轮迭代的模型参数：

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

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

采用求偏导数的方法：

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}$$

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

05 EM算法

◆高斯混合模型参数估计的EM算法

- **算法**
- 输入：观测数据$y_1, y_2, \dots, y_N$，高斯混合模型；
- 输出：高斯混合模型参数

（1）取参数的初始值开始迭代

（2）E步：依据当前模型参数，计算分模型k对观测数据y_j的相应度

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

$$j = 1,2,\dots,N; k = 1,2,\dots,K$$

05 EM算法

（3）M步：计算新一轮迭代的模型参数

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}$$

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

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}$$

（4）重复第2、3步直到收敛

总结

- 贝叶斯决策论
- 朴素贝叶斯分类算法流程
- EM算法流程

谢 谢！