

贝叶斯数据分析基础

A Tutorial on Bayesian Data Analytics

西安交通大学管理学院
信息管理与电子商务系
智能决策与机器学习研究中心
刘佳鹏

2021 年 3 月 22 日

概率论与数理统计回顾

- ▶ **随机试验：**如果一个试验具备以下特征：（1）可以在相同条件下重复进行；（2）每次试验的可能结果不止一个，但事先能明确全部可能的结果；（3）进行一次试验之前不能肯定哪一个结果会出现，则称这种试验为随机试验，简称试验
 - ▶ E_1 ：抛掷一枚硬币，观察正面、反面的出现情况
 - ▶ E_2 ：投掷一颗骰子，观察出现的点数
 - ▶ E_3 ：记录车站售票处一天内出售的车票数
 - ▶ E_4 ：从一大批元件中任意抽取一个，测试其使用寿命

概率论与数理统计回顾

- ▶ **样本空间**：对于一个试验 E ，虽然在一次试验之前不能肯定哪个结果会发生，但试验的一切可能结果是已知的，我们把 E 的所有可能的试验结果组成的集合称为 E 的**样本空间**，样本空间的元素（亦称 E 的每个可能结果）称为**样本点**。
- ▶ 用 Ω 表示样本空间。例如，上述 E_1 至 E_4 的样本空间分别是
 - ▶ $\Omega_1 = \{\omega_0, \omega_1\}$ ，其中 ω_0 表示“正面朝上”， ω_1 表示“反面朝上”
 - ▶ $\Omega_2 = \{1, 2, 3, 4, 5, 6\}$ ，其中数 i 表示“出现 i 点”， $i = 1, 2, 3, 4, 5, 6$
 - ▶ $\Omega_3 = \{0, 1, 2, \dots, n\}$ ，这里 n 是售票处一天内准备出售的车票数
 - ▶ $\Omega_4 = \{\omega | \omega \geq 0\}$ 或者 $\Omega_4 = [0, +\infty]$

概率论与数理统计回顾

- ▶ **随机事件**：在随机试验中，可能发生、也可能不发生的事情叫做随机事件，简称事件。
- ▶ 用大写字母 A, B, C, \dots 表示随机事件
- ▶ 在试验 E_2 中，存在如下随机事件
 - ▶ A 表示“掷出奇数点数”， $A = \{1, 3, 5\}$
 - ▶ B 表示“掷出偶数点数”， $A = \{2, 4, 6\}$
 - ▶ C 表示“掷出素数点数”， $A = \{2, 3, 5\}$

概率论与数理统计回顾

样本空间与随机事件的关系

- ▶ 对于一个试验 E ，它的样本空间 Ω 是由 E 的全部可能结果组成的集合
- ▶ 而 E 的一个随机事件 A 是由一部分可能结果组成的集合
- ▶ 因此事件 A 是样本空间 Ω 的子集，记为 $A \subset \Omega$
- ▶ 称事件 A 发生，当且仅当属于 A 的某一个样本点在试验中出现

概率论与数理统计回顾

样本空间与随机事件的关系

- ▶ 必然事件：每次试验中必然发生的事情，记为 Ω
- ▶ 不可能事件：每次试验中都不发生的事情，记为 \emptyset
- ▶ 例如在试验 E_2 中，事件 $A =$ “掷出的点数不超过6”是必然事件，这个事件包含 E_2 的所有可能结果， $A = \Omega = \{1, 2, 3, 4, 5, 6\}$ ，而事件 $B =$ “掷出的点数小于1”是不可能事件，这个事件不包含 E_2 的任何一个可能结果， $B = \emptyset$

概率论与数理统计回顾

概率空间

- ▶ 概率空间是三位一体的研究对象 (Ω, F, P) ，其中 Ω 是样本空间， F 是事件域（随机事件全体，包含必然事件 Ω 和不可能事件 \emptyset ）， P 是定义在事件域 F 上的概率（测度），满足以下三条公理：
 - ▶ （1）非负性：对于任意事件 A ，其概率 $P(A) \geq 0$
 - ▶ （2）规范性：必然事件 Ω 的概率等于1，即 $P(\Omega) = 1$
 - ▶ （3）可列可加性： A_1, A_2, \dots, A_n 是一系列事件，满足 $A_i \cap A_j = \emptyset$ （称为两两不相容），则

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i)$$

条件概率公式

- ▶ 对于任意两个事件 A 和 B ，且 $P(A) > 0$ ，定义在 A 发生的条件下， B 发生的条件概率为

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

从而 $P(AB) = P(A)P(B|A)$ ，这就是乘法公式

- ▶ 推而广之，设 A_1, A_2, \dots, A_n 是任意 n 个随机事件，则有更一般的乘法公式

$$P(A_1 A_2 \cdots A_n) = P(A_1) P(A_2|A_1) P(A_3|A_1 A_2) \\ \cdots P(A_n|A_1 A_2 \cdots A_{n-1})$$

全概率公式

设 B_1, \dots, B_n 是样本空间 Ω 中的一个完备事件群（又称为 Ω 的一个划分）。换言之，它们满足下列条件：

(a) 两两不相交，即 $B_i \cap B_j = \emptyset$ ($i \neq j$)

(b) 它们的并(和)恰好是样本空间，即 $\sum_{i=1}^n B_i = \Omega$

设 A 为 Ω 中的一个事件，则全概率公式为

$$P(A) = P\left(\sum_{i=1}^n AB_i\right) = \sum_{i=1}^n P(A|B_i) P(B_i)$$

这个公式将整个事件 A 分解成一些两两不相交的事件之并(和)。直接计算 $P(A)$ 不容易，但分解后的那些事件的概率容易计算，从而使 $P(A)$ 的计算变得容易了

贝叶斯公式

- 在全概率公式的条件下，即存在样本空间 Ω 中的一个完备事件群 $\{B_1, \dots, B_n\}$ ，设 A 为 Ω 中的一个事件，且 $P(B_i) > 0$ ($i = 1, \dots, n$)， $P(A) > 0$ ，则按照条件概率的计算方法，有

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$$

贝叶斯公式

示例：一种诊断某癌症的试剂，经临床试验有如下记录：癌症病人试验结果是阳性的概率为95%，非癌症病人试验结果是阴性的概率为95%。现用这种试剂在某社区进行癌症筛查，该社区癌症发病率为0.5%，问某人反应为阳性时，该如何判断他是否患有癌症？

贝叶斯公式

解：设 A 表示“反应为阳性”的事件， B 表示“被诊断者患癌症”的事件，则 $B_1 = B$ 和 $B_2 = \bar{B}$ 构成一个完备事件群。由题意知

$$P(A|B_1) = 0.95, \quad P(A|B_2) = 1 - P(\bar{A}|B_2) = 1 - 0.95 = 0.05, \\ P(B_1) = 0.005, \quad P(B_2) = 0.995.$$

现在要算的是 $P(B_1|A)$ 和 $P(B_2|A)$ 。有贝叶斯公式易得

$$\begin{aligned} P(B_1|A) &= \frac{P(A|B_1) P(B_1)}{P(A|B_1) P(B_1) + P(A|B_2) P(B_2)} \\ &= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.05 \times 0.995} \approx 0.087 = 8.7\% \\ P(B_2|A) &= 1 - P(B_1|A) = 91.3\% \end{aligned}$$

贝叶斯公式

- ▶ 练习：试用贝叶斯公式解释“幸存者偏差”现象
- ▶ 幸存者偏差：二战期间，为了加强对战机的防护，英美军方调查了作战后幸存飞机上弹痕的分布，决定哪里弹痕多就加强哪里。然而统计学家亚伯拉罕·瓦尔德力排众议，指出更应该注意弹痕少的部位，因为这些部位受到重创的战机，很难有机会返航，而这部分数据被忽略了。事实证明，瓦尔德是正确的。

三种信息

总体信息

- ▶ 数理统计学的任务是要通过样本推断总体
- ▶ 样本有两重性，当把样本视为随机变量时，它有概率分布，称为总体分布。如果我们已经知道总体的分布形式，这就给了我们一种信息，称为总体信息
- ▶ 例如，若已知样本来自于正态总体，则它可以提供给我们很多信息，如它的密度函数是倒立的“钟”形曲线，它的所有阶矩存在，任何事件的概率都可以通过查表求出

三种信息

样本信息

- ▶ 另外一种信息是样本信息，就是从总体中抽取的样本所提供的信息。这是最“鲜活”的信息，样本越多，提供的信息越多，我们希望通过通过对样本的加工、整理，对总体的分布或某些数字特征作出统计推断
- ▶ 没有样本就没有统计推断

三种信息

- ▶ 总体信息和样本信息放在一起，称为抽样信息（sampling information）
- ▶ 基于总体信息和样本信息进行统计推断的理论和方法称为经典（古典）统计学（classical statistics）
- ▶ 它的基本观点是：把样本看成来自有一定概率分布的总体，所研究的对象是这个总体而不局限于数据本身
- ▶ 代表方法：极大似然估计、最小二乘法

三种信息

先验信息

- ▶ 另外一种信息称为先验信息(prior information), 就是在抽样之前, 有关统计推断问题中未知参数的一些信息
- ▶ 先验信息一般来自经验和历史资料, 例如“某人认为明天下雨的概率是0.6”、“某种疾病在中年男性中的发病率是5.2%”
- ▶ 下面两例说明先验信息是存在的且可被人们利用

三种信息

先验信息

- ▶ 英国统计学家Savage(1961)提出了一个令人信服的例子，可以说明先验信息有时是很重要的。看下面两个统计试验：
- ▶ （1）一个常饮牛奶和茶的女士说，她能辨别先倒进杯子里的是茶还是牛乳。对此做了10次试验，她都说对了
- ▶ （2）一位音乐家说，他能够从一页乐谱中辨别是海顿还是莫扎特的作品。在10次试验中，他都说对了

三种信息

- ▶ 基于上述三种信息进行统计推断的方法和理论称为贝叶斯统计学(Bayes statistics)
- ▶ 它与经典统计学的主要区别在于是否利用先验信息
- ▶ 在使用样本上也是存在差别的，贝叶斯方法重视已出现的样本，对尚未发生的样本值不予考虑
- ▶ 贝叶斯学派重视先验信息的收集、挖掘和加工，使之形成先验分布而参加到统计推断中来，以提高统计推断的效果
- ▶ 忽视先验分布的利用，有时是一种浪费！

古典学派与贝叶斯学派的争论

- ▶ 古典学派和贝叶斯学派是当今数理统计学的两大学派
- ▶ 凡是坚持概率的频率解释，对数理统计学中的概念、结果和方法性能的评价等都必须大量重复的意义上去理解的，都属于古典学派，亦称为频率学派
- ▶ 20世纪60年代以来贝叶斯学派迅速崛起，达到可以与频率学派分庭抗礼的程度
- ▶ 由于其发展较新，因此贝叶斯学派常常把频率学派称为古典学派
- ▶ 贝叶斯统计与机器学习

古典学派与贝叶斯学派的争论

两大学派的主要分歧

- ▶ (1) 对于概率含义的解释
 - ▶ 古典学派：一个事件的概率可以用大量重复试验下的频率来解释
 - ▶ 贝叶斯学派：将主观概率认为是认识主体对事件发生机会的相信程度，因为有些事件不可重复
- ▶ (2) 对于参数的理解
 - ▶ 古典学派：参数是一个固定值，虽然可能未知，但可以推断
 - ▶ 贝叶斯学派：参数是随机变量，具有特定分布

古典学派与贝叶斯学派的争论

- ▶ 虽然两个学派有很多不同观点，但也存在不少共同点，如都承认样本有概率分布，概率的计算遵循共同的原则
- ▶ 对上述争论有一个至高无上的“裁判者”，即应用的结果如何，统计方法无论在理论上如何精细高明，总要用实践来检验
- ▶ 迄今为止，实践显示这两派的得分都不低。也正是因为它们在实际应用上的表现不错，才能各自聚合了一批追随者而形成学派
- ▶ 作为一名研究人员，可以不执著于任何一派的观点，而是各取其长，为我所用

极大似然估计

- ▶ 示例：掷硬币试验，掷出 n 次，设随机变量 X 表示正面向上的次数，因此随机变量 X 服从二项分布 $B(n, \theta)$ ， θ 是硬币正面向上的概率，概率分布如下

$$p(X = x|\theta) = C_n^x \theta^x (1 - \theta)^{n-x}, \quad x = 0, 1, \dots, n,$$

其中 x 表示观测到正面向上的次数

- ▶ 使用极大似然估计法估计参数 θ ：
似然函数

$$L(x|\theta) = C_n^x \theta^x (1 - \theta)^{n-x}$$

对数似然函数

$$LL(x|\theta) = \log(C_n^x) + x \log \theta + (n - x) \log(1 - \theta)$$

极大似然估计

- ▶ 对数似然函数关于参数 θ 求导并置零

$$\frac{\partial LL(x|\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$$

得到

$$\hat{\theta}_{MLE} = \frac{x}{n}$$

贝叶斯参数估计

- ▶ 示例：掷硬币试验，掷出 n 次，设随机变量 X 表示正面向上的次数，因此随机变量 X 服从二项分布 $B(n, \theta)$ ， θ 是硬币正面向上的概率，概率分布如下

$$p(X = x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad x = 0, 1, \dots, n,$$

其中 x 表示观测到正面向上的次数

- ▶ 贝叶斯参数估计是基于贝叶斯公式的参数估计方法

The diagram illustrates the Bayesian formula for parameter estimation. The formula is
$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)} = \frac{p(x | \theta)p(\theta)}{\int p(x | \theta)p(\theta)d\theta}$$
 Each term is labeled with a yellow callout box:

- $p(\theta | x)$ is labeled "参数 θ 的后验分布" (Posterior distribution of parameter θ).
- $p(x | \theta)$ is labeled "x关于 θ 的似然函数" (Likelihood function of x given θ).
- $p(\theta)$ is labeled "参数 θ 的先验分布" (Prior distribution of parameter θ).
- $p(x)$ is labeled "x的边缘分布, 亦称归一化因子" (Marginal distribution of x , also known as the normalization factor).

贝叶斯参数估计

- ▶ x 关于参数 θ 的似然函数

$$p(x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

- ▶ 参数 θ 的先验分布：选取 $[0,1]$ 区间上的均匀分布

$$p(\theta) = 1, \quad 0 \leq \theta \leq 1$$

- ▶ x 的边缘分布（归一化因子）

$$p(x) = \int_0^1 p(x|\theta)p(\theta)d\theta = \int_0^1 \binom{n}{x} \theta^x (1 - \theta)^{n-x} d\theta = \frac{1}{1+n}$$

- ▶ 将上述三项代入贝叶斯公式，得到参数 θ 的后验分布

$$p(\theta|x) = (1+n) \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

贝叶斯参数估计

Beta分布

- ▶ Beta分布是一组定义在 $[0,1]$ 区间上的连续概率分布
- ▶ Beta分布的概率密度函数

$$\text{Beta}(\theta|a, b) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$$

其中 $B(a, b)$ 是Beta函数，定义为

$$B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}$$

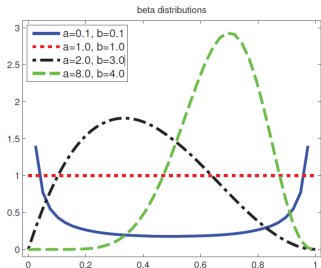
其中 $\Gamma(\cdot)$ 是Gamma函数，定义为

$$\Gamma(s) = \int_0^{\infty} x^{s-1} e^{-x} dx, \quad s > 0$$

贝叶斯参数估计

Beta分布

- ▶ 参数a和b控制着Beta分布的形式



- ▶ 特别地，当 $a = b = 1$ 时，Beta分布就是 $[0,1]$ 区间上的均匀分布
- ▶ Beta分布通常作为二项分布(Binomial distribution)的参数的先验分布使用
- ▶ Beta分布的期望、众数、方差

$$\text{mean} = \frac{a}{a+b} \quad \text{mode} = \frac{a-1}{a+b-2} \quad \text{var} = \frac{ab}{(a+b)^2(a+b+1)}$$

贝叶斯参数估计

回到掷硬币试验

- ▶ 将参数 θ 的先验分布设定为Beta分布

$$\text{Beta}(\theta|a, b) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$$

- ▶ 当 $a = b = 1$ 时, Beta分布就是 $[0,1]$ 区间上的均匀分布

贝叶斯参数估计

- x 的边缘分布 (归一化因子) 可以写为

$$\begin{aligned} p(x) &= \int_0^1 p(x|\theta) p(\theta) d\theta \\ &= \int_0^1 \underbrace{\binom{n}{x} \theta^x (1-\theta)^{n-x}}_{p(x|\theta)} \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}}_{p(\theta)} d\theta \\ &= \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{a+x-1} (1-\theta)^{b+n-x-1} d\theta \\ &= \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} \\ &\quad \underbrace{\int_0^1 \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)} \theta^{a+x-1} (1-\theta)^{b+n-x-1} d\theta}_{\text{Beta}(x|a+x, b+n-x)} \\ &= \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} = \binom{n}{x} \frac{B(a+x, b+n-x)}{B(a, b)} \end{aligned}$$

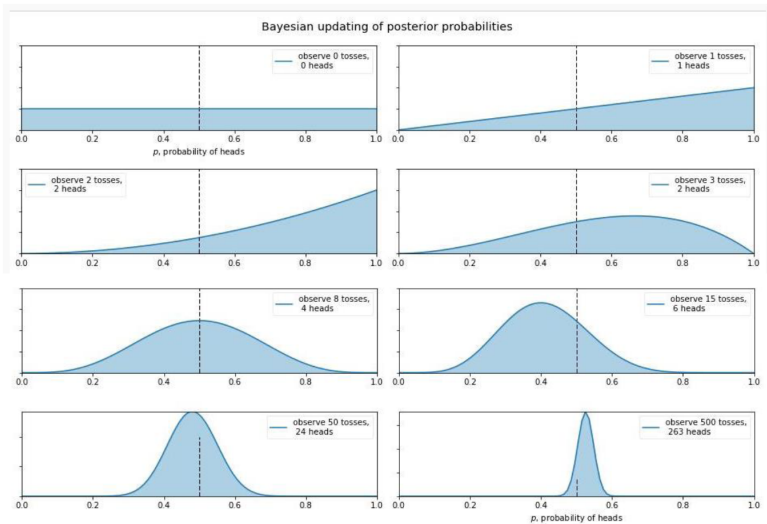
贝叶斯参数估计

- ▶ 将 x 的边缘分布 $p(x)$ 代入贝叶斯公式

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\ &= \frac{\underbrace{\binom{n}{x} \theta^x (1-\theta)^{n-x}}_{p(x|\theta)} \underbrace{\frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}}_{p(\theta)}}{\underbrace{\binom{n}{x} \frac{B(a+x, b+n-x)}{B(a,b)}}_{p(x)}} \\ &= \frac{\theta^{a+x-1} (1-\theta)^{b+n-x-1}}{B(a+x, b+n-x)} \\ &= \text{Beta}(\theta|a+x, b+n-x) \end{aligned}$$

- ▶ θ 的后验分布是参数为 $a+x$ 和 $b+n-x$ 的Beta分布

贝叶斯参数估计



贝叶斯参数估计

- ▶ θ 的后验分布是参数为 $a + x$ 和 $b + n - x$ 的Beta分布
- ▶ 后验概率密度最大的点(众数mode)是

$$\hat{\theta}_{MAP} = \frac{a + x - 1}{a + b + n - 2}$$

称之为极大后验估计(maximum a posterior probability estimation, MAP)

- ▶ 回忆: 极大似然估计(maximum likelihood estimation, MLE)的结果为

$$\hat{\theta}_{MLE} = \frac{x}{n}$$

- ▶ 后验众数可以看成极大似然估计结果和先验众数的加权组合

$$\frac{a + x - 1}{a + b + n - 2} = w \times \frac{x}{n} + (1 - w) \times \frac{a - 1}{a + b - 2}$$

其中 $w = \frac{n}{a + b + n - 2}$

贝叶斯参数估计

$$\frac{a + x - 1}{a + b + n - 2} = w \times \frac{x}{n} + (1 - w) \times \frac{a - 1}{a + b - 2}$$

其中 $w = \frac{n}{a+b+n-2}$

- ▶ 当 n 变大, w 趋向于 1, 后验众数趋向于极大似然估计结果
- ▶ 当 $a = b = 1$ 时, $w = 1$, 后验众数等于极大似然估计结果, 贝叶斯参数估计结果与极大似然估计结果相同

贝叶斯参数估计

- ▶ 若取后验均值作为贝叶斯参数估计的结果

$$\hat{\theta}_{Mean} = \frac{a + x}{a + b + n}$$

- ▶ 若先验取为[0,1]区间上的均匀分布($a=b=1$ 的Beta分布), 有

$$\hat{\theta}_{Mean} = \frac{1 + x}{2 + n}$$

- ▶ 对比极大似然估计的结果

$$\hat{\theta}_{MLE} = \frac{x}{n}$$

- ▶ $\hat{\theta}_{Mean}$ 在小样本情形下比 $\hat{\theta}_{MLE}$ 更合理
- ▶ 当试验次数 n 增加时, $\hat{\theta}_{Mean}$ 趋向于 $\hat{\theta}_{MLE}$
- ▶ 为什么要用先验? 因为有些试验不能大量重复进行!

贝叶斯参数估计

共轭先验

- ▶ 在硬币试验中，参数 θ 的先验分布 $p(\theta)$ 和后验分布 $p(\theta|x)$ 都是Beta分布
- ▶ 称Beta分布是二项分布(Binomial distribution)的共轭先验分布
- ▶ 当先验分布和后验分布是同一种分布，称先验分布是似然函数的共轭先验分布(conjugate prior)
- ▶ When the prior and the posterior have the same form, we say that the prior is a conjugate prior for the corresponding likelihood.
- ▶ 共轭先验可以简化计算，且易于解释

贝叶斯参数估计

共轭先验

- ▶ 只有给定似然函数，才能确定其共轭先验分布
- ▶ 也就是说，必须根据问题的性质选取其共轭先验分布
- ▶ 常见的共轭先验分布如下

似然函数	参数	共轭先验分布
二项分布	成功概率	Beta分布
多项分布	成功概率	Dirichlet分布
泊松分布	期望	Gamma分布
指数分布	期望的倒数	Gamma分布
正态分布(方差已知)	均值	正态分布
正态分布(方差未知)	方差	逆Gamma分布

贝叶斯参数估计

共轭先验

- ▶ 对于一般形式的似然函数，共轭先验分布可能不存在
- ▶ 若选取某种分布作为参数 θ 的先验分布， x 的边缘分布（归一化因子）很有可能没有解析表达式

$$p(x) = \int_0^1 p(x|\theta) p(\theta) d\theta$$

- ▶ 这将导致参数 θ 的后验分布没有解析表达式

$$p(\theta|x) = \frac{p(x|\theta) p(\theta)}{p(x)}$$

- ▶ 解决方法：（1）Markov Chain Monte Carlo (MCMC) （2）Variational Inference (VI)

贝叶斯方法的应用

潜在狄利克雷分配模型 (LDA)

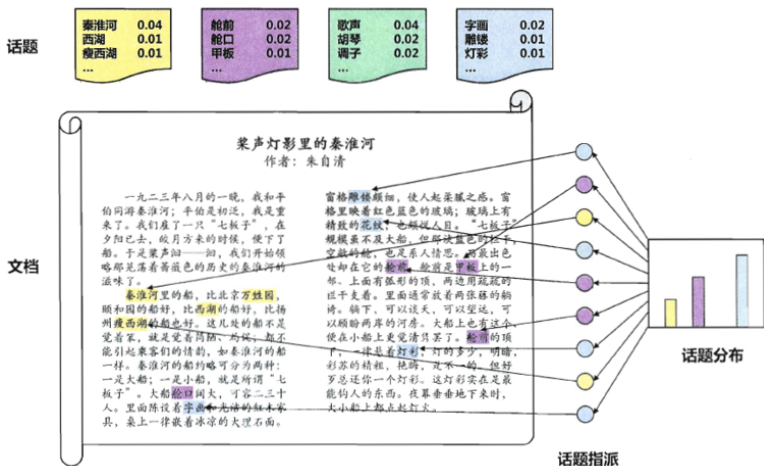
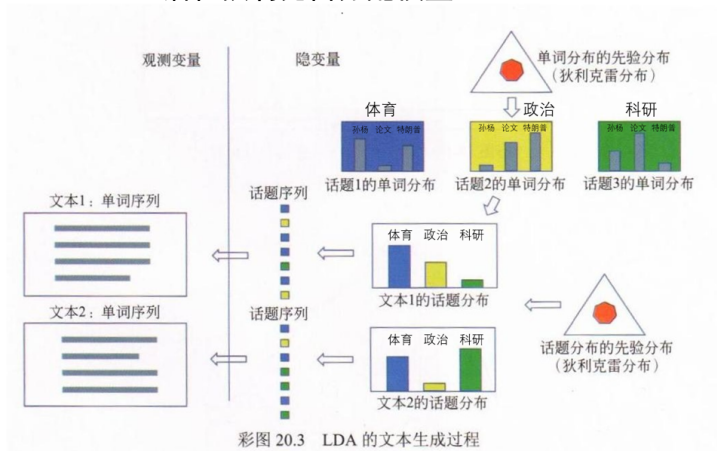


图 14.11 LDA 的文档生成过程示意图

贝叶斯方法的应用

潜在狄利克雷分配模型 (LDA)



贝叶斯方法的应用

潜在狄利克雷分配模型（LDA）

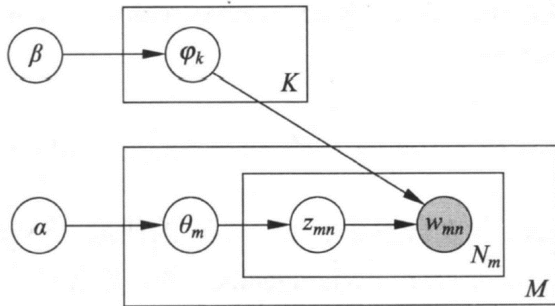


图 20.4 LDA 的板块表示

同学们，明年见！

See you next year !