数据分析及实践
Analysis and Practice of the Data

第三章 数据统计

刘 淇
Email: qiliuql@ustc.edu.cn

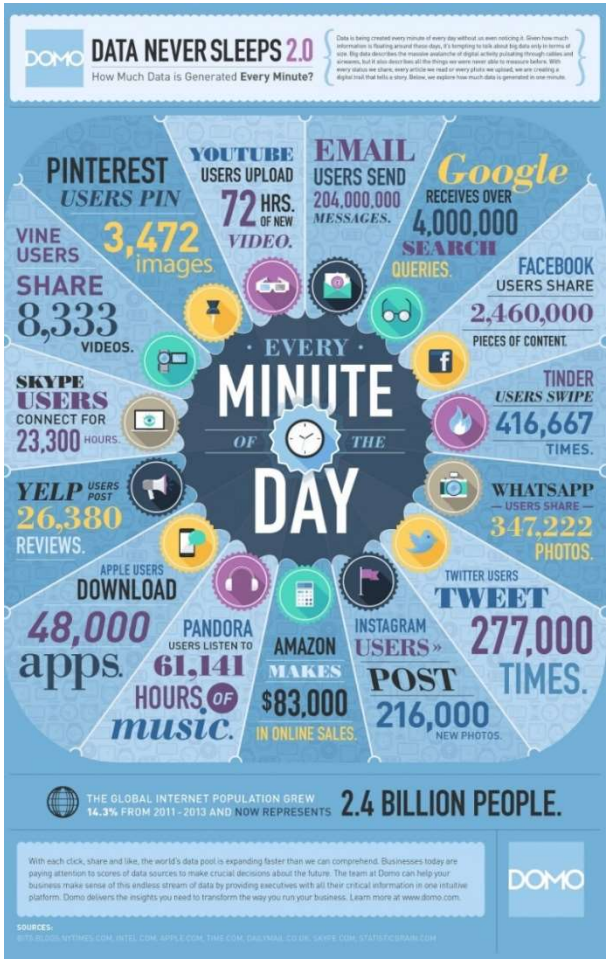
http://staff.ustc.edu.cn/~qiliuql/AD2021.html

4/12/2021

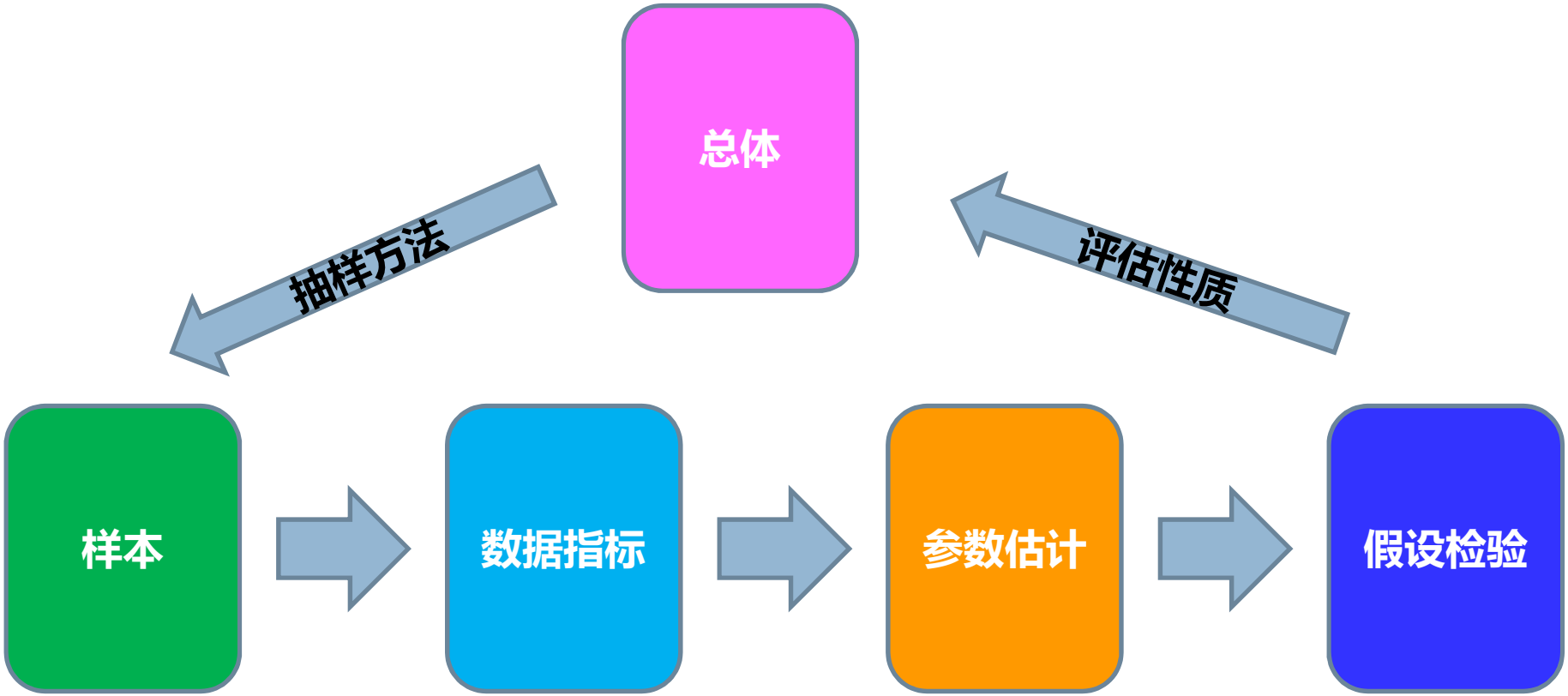
Data

- 大数据
  - 数据量大
  - 类型繁多
  - 时效性高
  - 价值密度低

- 大数据由于本身特性，通常处理代价巨大，可先利用统计手段了解数据基本信息
- 在实际处理大数据前，还可先在抽样得到的小型数据集上对总体进行推断



Data



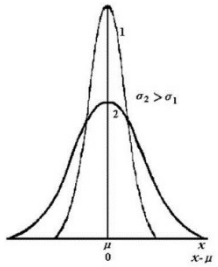
Data

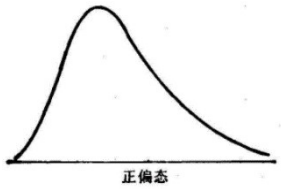
- 数据分布基本指标
- 参数估计
- 假设检验
- 抽样方法

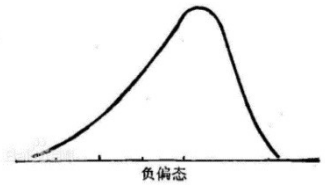
数据分布基本指标

- 在对大数据进行研究时，研究者往往希望知道所获得的数据的基本分布特征

- 数据分布的特征可以从三个方面进行测度和描述：
  - 描述数据分布的集中趋势：反映数据向其中心靠拢或聚集程度
  - 描述数据分布的离散程度：反映数据远离中心的趋势或程度
  - 描述数据分布的形状变化：反应数据分布的形状特征







数据分布基本指标

- 集中趋势
  - 集中趋势反映了一组数据的中心点位置所在及该组数据向中心靠拢或聚集的程度。

- 四种最常用的反映数据集中趋势的指标：
  - 平均数
  - 中位数
  - 分位数
  - 众数

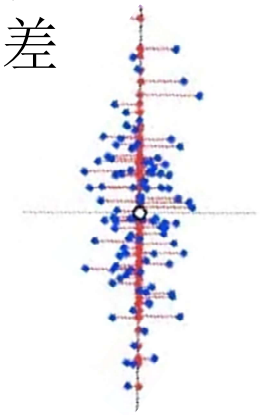
数据分布基本指标-离散程度

- 离散程度
  - 离散程度反映了各个数据属性值远离其中心值的程度，是数据分布的另一个重要特征。
  - 数据的离散程度越大，则集中趋势的测度值对该组数据的代表性就越差，反之亦然。

- 四种最常用的反映数据离散程度的指标：
  - 方差和标准差
  - 极差和四分位差
  - 异众比率
  - 变异系数

数据分布基本指标-离散程度

- 方差和标准差
  - 在数值型数据中, 刻画数据围绕其中心位置附近分布的数字特征时，最重要且最常用的是方差(variance) 和标准差(standard deviation)。



  - 方差是各个变量与均值之差平方的平均数
    - 通过平方的方法消去差值中的正负号，再对其进行平均。
  - 方差的平方根即为标准差，两个指标均能较好地反映出数值型数据的离散程度。

数据分布基本指标-离散程度

- 方差
  - 对于使用简单平均数作为数据中心的未分组数据数据，$x_1, x_2, x_3, \ldots, x_n$，总体方差为:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

  - 对于使用加权平均数作为数据中心的分组数据，该组数据的总体方差为:

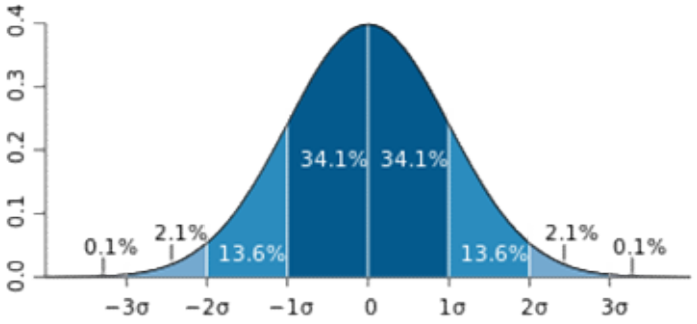
$$\sigma^2 = \frac{\sum_{i=1}^{k} (M_i - \mu)^2 f_i}{N}$$

数据分布基本指标-离散程度

- 标准差
  - 标准差为方差的算数平方根，是具有量纲(与原数据有相同单位)的。
  - 它与变量值的计量单位相同，实际意义比方差更清楚。
  - 对于未分组数据和加权的分组数据来说，其标准差的计算公式分别为：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(M_i - \mu)^2 f_i}{N}}$$



数据分布基本指标-离散程度

- 极差和四分位差
  - 在顺序数据中，当中位数作为数据中心位置的指标时，一般可用极差或四分位差反映数据的离散程度。
  - 极差：
    - 一组数据的最大值和最小值之差被称为极差(range)，也被称为全矩，用R表示，是描述数据离散程度的最简单的测度值。
    - 若一组数据中的最大值为$\max(x_i)$，最小值为$\min(x_i)$，则该组数据的极差R为：

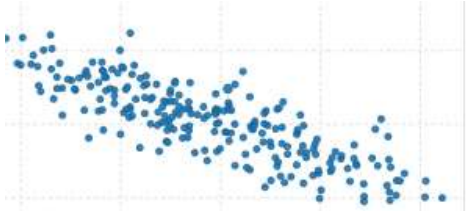
$$R = max(x_i) - min(x_i)$$

数据分布基本指标-离散程度

- 极差：
  - 一组数据的最大值和最小值之差被称为极差(range)，也被称为全矩，用R表示，是描述数据离散程度的最简单的测度值。
  - 若一组数据中的最大值为$\max(x_i)$，最小值为$\min(x_i)$，则该组数据的极差R为：

$$R = max(x_i) - min(x_i)$$

  - 极差即数据的振幅，振幅越大说明数据越分散，其直观意义非常明显。但由于极差只是利用了一组数据的两端信息，容易受极端值的影响，且不能反映出中间数据的分散状况、准确描述出数据的分散程度。

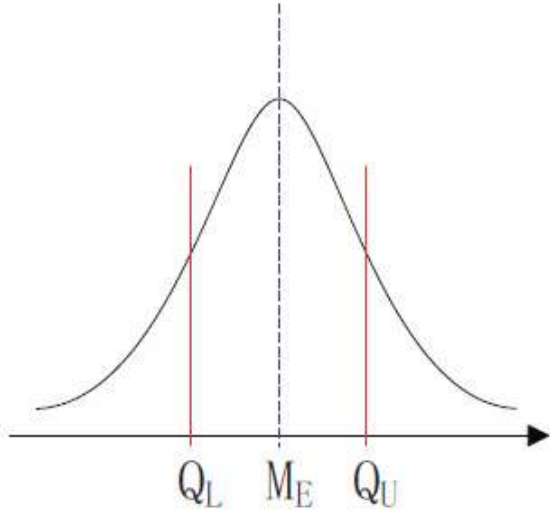


数据分布基本指标-离散程度

- 四分位差：
  - 一组数据的上四分位数和下四分位数的差值被称为四分位差(quartile deviation)，也被称为内矩，用H表示。
  - 若一组数据的上四分位数为Q_U，下四分位数为Q_L，则该数据的四分位差H为：

$$Q = Q_U - Q_L$$

  - 从定义可以看出，H是区间(Q_L, Q_U)的长度。
  - 且区间(Q_L, Q_U)正好含有50%的数据。
  - 不同于极差，四分位差不会受到数据中极端情况的影响。



数据分布基本指标-离散程度

- 异众比率
  - 在以众数作为数据中的的分类数据中，异众比率(variation ratio) 是指非众数组的频数占总频数的比率，用V_r 表示。
  - 主要用于衡量众数对一组数据的代表性程度。
  - 除了对于分类数据外，对于数值型数据和顺序数据也也可以计算其异众比率。计算公式为：

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i}$$

  - 其中，Σf_i 为变量值的总频数，f_m 为众数组的频数。异众比率越大，众数组的频数占总频数比率越小，数据离散程度越高，众数作为其中心的代表性越差。

数据分布基本指标-离散程度

- 变异系数
  - 当需要比较两组数据离散程度大小的时候，如果两组数据的测量尺度相差太大，或者数据量纲的不同，直接使用标准差来进行比较不合适，此时就应当消除测量尺度和量纲的影响。
  - 变异系数（Coefficient of Variation）是原始数据标准差与原始数据平均数的比。计算公式为：

$$c_v = \frac{\sigma}{\mu}$$

  - 在进行数据统计分析时，如果变异系数大于15%，则要考虑该数据可能不正常。

反映单位均值上的离散程度，常用在两个总体均值不等的离散程度的比较上

缺陷：当平均值接近于0的时候，微小的扰动也会对变异系数产生巨大影响，因此造成精确度不足。变异系数无法发展出类似于均值的置信区间的工具。

数据分布基本指标-形状变化

- 形状变化
  - 形状变化反映了一组数据分布的整体形状信息。

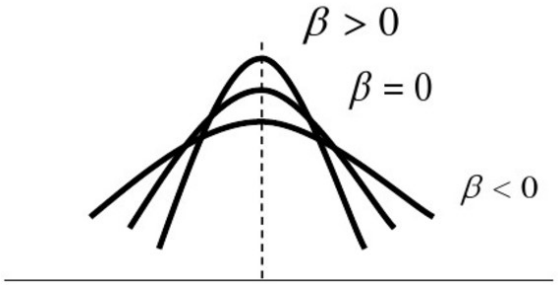
- 两种最常用的反映数据形状变化的指标：
  - 峰度
  - 偏度

数据分布基本指标-形状变化

- 峰度：度量数据在中心聚集程度
  - 峰度（Kurtosis）是描述总体中所有取值分布形态陡缓程度的统计量。
  - 峰度的具体计算公式为：

$$\text{Kurtosis} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4 / \text{SD}^4 - 3$$

  - 正态分布的峰度值为3
    - 但是SPSS等软件中将正态分布峰度值定为0，因为已经减去3，这样比较起来方便）
  - 需要与正态分布相比较
    - 峰度为0表示该总体数据分布与正态分布的陡缓程度相同
    - 峰度大于0表示该总体数据分布与正态分布相比较为陡峭，为尖顶峰；
    - 峰度小于0表示该总体数据分布与正态分布相比较为平坦，为平顶峰。

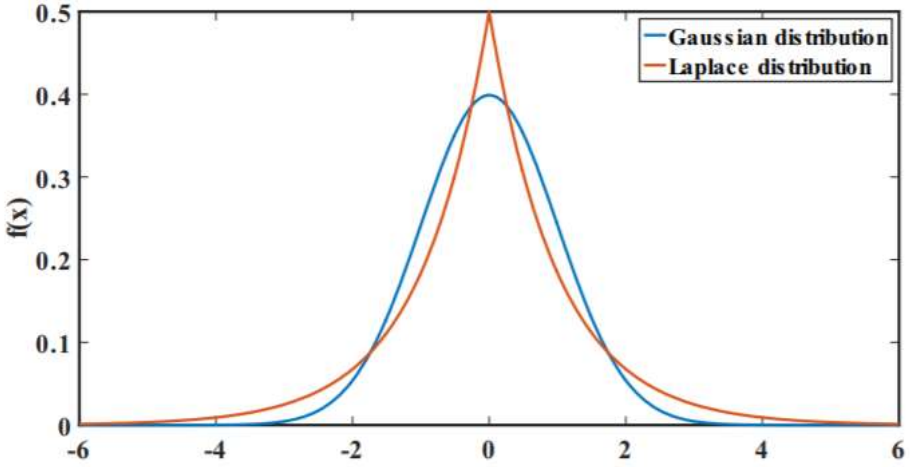


数据分布基本指标-形状变化

- 峰度： 度量数据在中心聚集程度
  - 峰度（Kurtosis）是描述总体中所有取值分布形态陡缓程度的统计量。
  - 峰度的具体计算公式为：

$$\text{Kurtosis} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4 / \text{SD}^4 - 3$$

  - 正态分布的峰度值为3
    - 但是SPSS等软件中将正态分布峰度值定为0，因为已经减去3，这样比较起来方便）



概率密度函数：

高斯分布 $p(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

Lapalace $p(x) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}}$

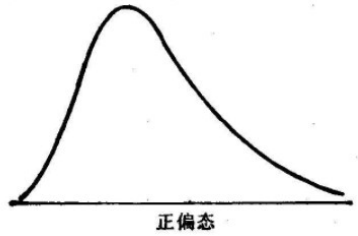
Sparse FM :Sparse Factorization Machines for Click-through Rate Prediction

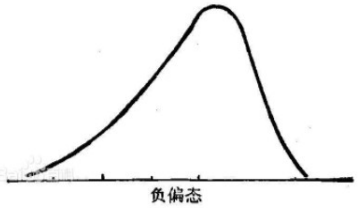
数据分布基本指标-形状变化

- 偏度
  - 偏度（Skewness）描述的是某总体取值分布的对称性
  - 偏度的具体计算公式为：

$$\text{Skewness} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3 / \text{SD}^3$$

  - 正态分布的偏度值为0
  - 某个总体
    - 偏度为0表示其数据分布形态与正态分布的偏斜程度相同；
    - 偏度大于0表示其数据分布形态与正态分布相比为正偏或右偏，即有一条长尾巴拖在右边，数据右端有较多的极端值
    - 偏度小于0表示其数据分布形态与正态分布相比为负偏或左偏，即有一条长尾拖在左边，数据左端有较多的极端值。

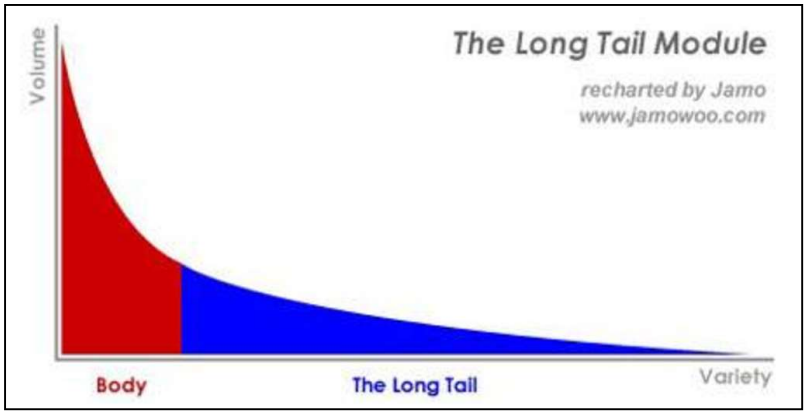


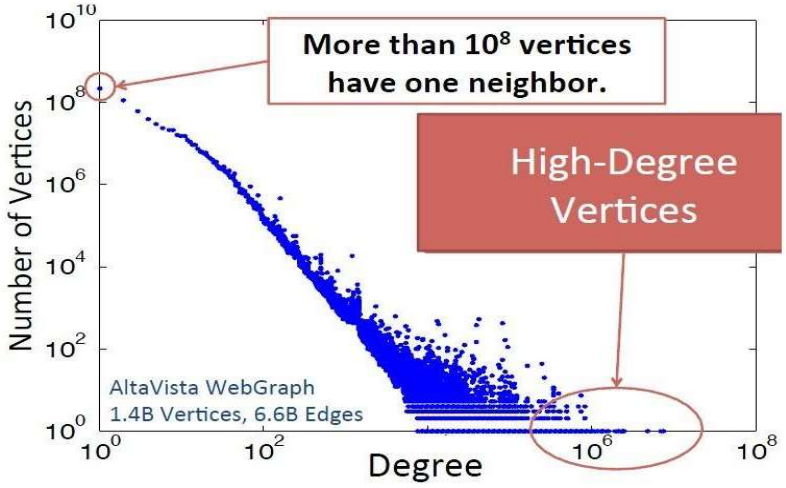


数据分布基本指标-形状变化

- 数据指标指导建模思路
  - 若均值与中位数接近，且偏度接近0，可知数据分布是近似对称的，建模时可考虑运用对称信息。
  - 若极差或四分位差较大，建模时需考虑数据是否有长尾现象。
    - 幂律分布：对数空间下呈现出线性关系(80-20法则)

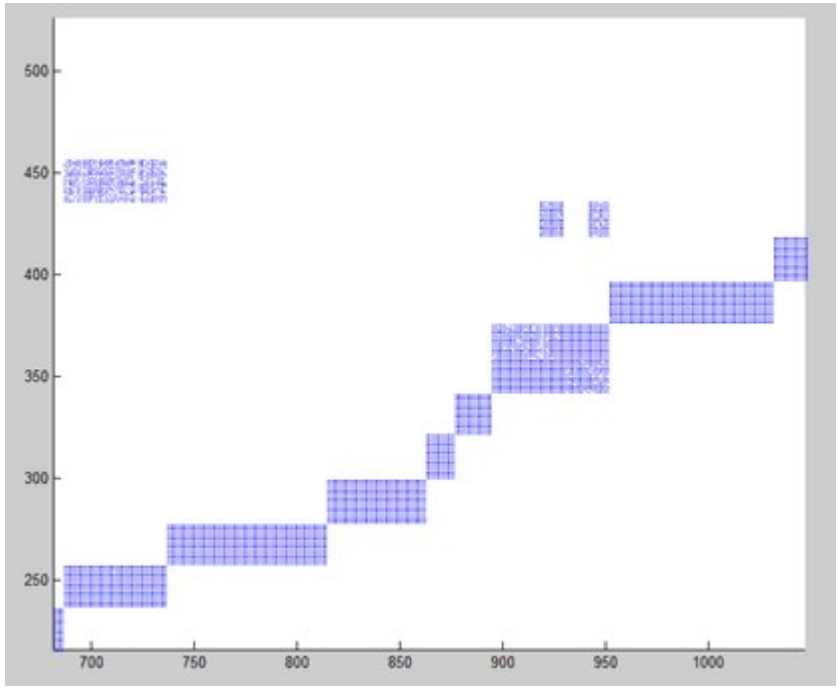
$$y = cx^{-r}$$

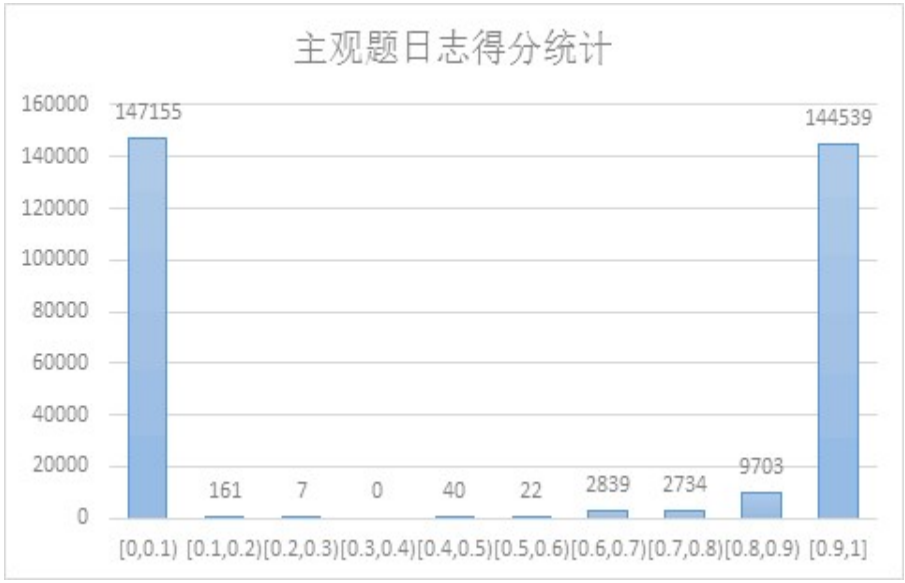




数据分布基本指标—举例

- 其它指标和现象观察
  - 教育数据





数据分布基本指标—举例

- 以旅游套餐数据为例

Niagara Falls Discovery

(Tour style-Culture & History, Wildlife & Nature), *8* days, from *$1260.00*

This ***eastern*** travel experiences the biggest, boldest and brightest of American destinations. From New York City, Niagara to Cambridge and Washington DC. Experience ***American life*** in full and gain perspective among giant ***monuments,*** stunning ***skyscrapers,*** fascinating ***history*** and spectacular ***natural*** wonders. Day 1 ***New York***: Enter a neon jungle at ***Times square,*** find a quiet corner in ***Central Park*** or watch the sunset from atop the ***Empire State Building.*** Days 2-3 ***Washington DC***: See all the big names - the ***White House,*** the ***Lincoln Memorial, Washington Monument*** and ***Capitol Hill.*** Day 4 ***Finger Lakes***: ***Finger Lakes,*** go swimming or hiking. Day 5 ***Niagara Falls***: ***Niagara Falls*** is a favorite for lovers and lovers of nature alike. Days 6-7 ***Boston***: Retrace the nation's revolutionary past by walking the ***Freedom Trail,*** or visit bustling ***North End*** for Italian feasts. Day 8 ***New York***: Continue to buzzing New York and travel to ***Coney Island,*** the ***Met*** or see a ***Broadway*** show.

Accommodation: Multishare hostels/cabins. **Size:** 13 travelers per group. **What to budget:** Allow USD $160 for meals not included.....

Figure 1. An example of the travel package document, where the landscapes are represented by the words in red.

数据分布基本指标—举例

- 以旅游套餐数据为例

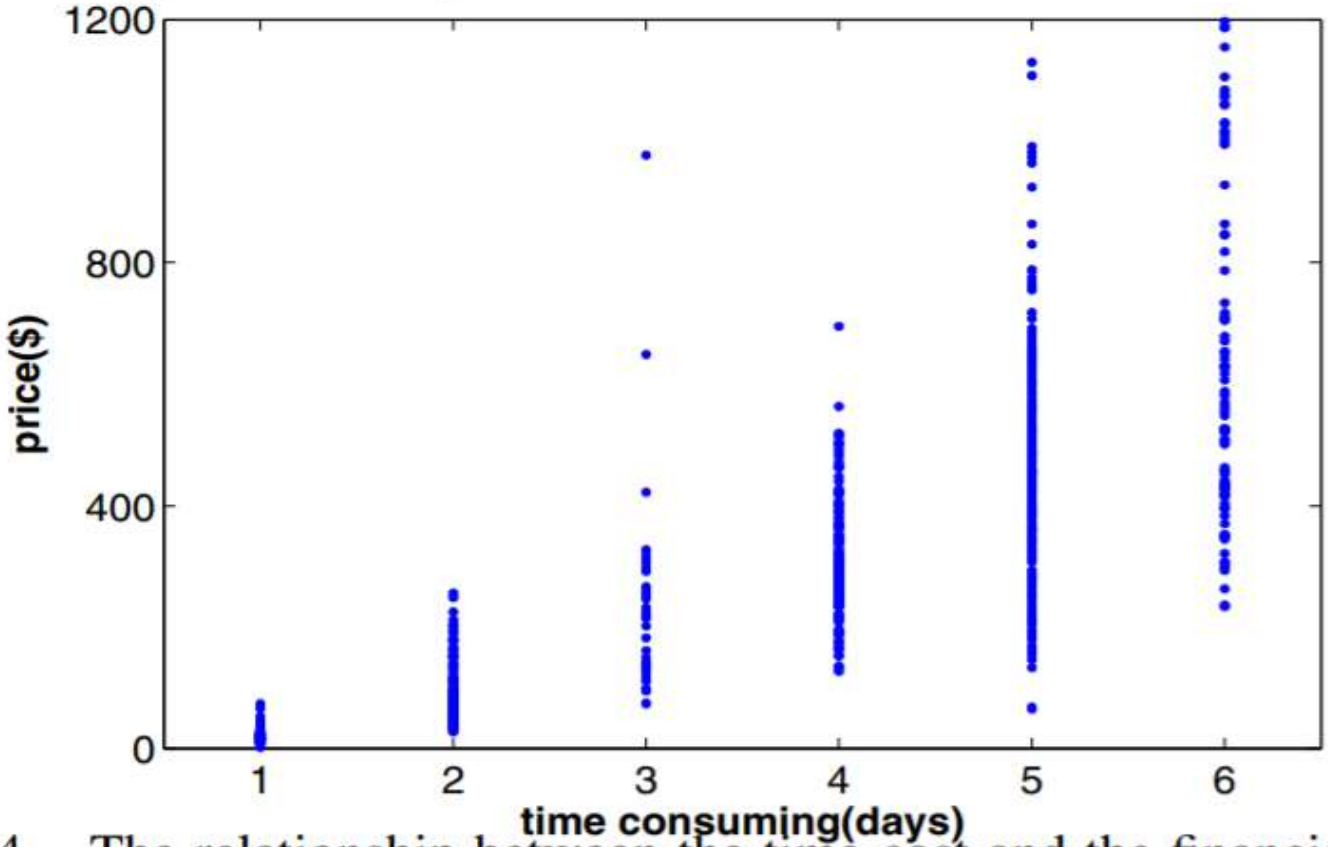


Figure 4. The relationship between the time cost and the financial cost in travel packages.

参数估计

- 参数
  - 参数 (parameter) 是用来描述总体特征的概括性数字度量。
- 统计量
  - 统计量 (statistic) 是用来表述样本特征的概括性数字度量，它完全由所抽取的样本计算得出，不依赖于任何其他未知的量（特别是不能依赖于总体分布中所包含的未知参数）。
- 参数估计
  - 参数估计 (parameter estimation) 是统计推断的基本问题之一，就是用样本统计量估计总体的参数。

参数估计

- 点估计
  - 点估计 (point estimate) 就是用样本统计量 $\hat{\theta}$ 的某个取值直接作为总体参数 θ 的估计值。

- 三个常用的点估计方法
  - 矩估计
  - 极大似然估计
  - 贝叶斯估计

还有一个常用的方法是最大后验概率估计（**MAP**），它介于极大似然估计和贝叶斯估计之间

参数估计

- 矩估计
  - 每一个随机变量X的矩都告诉你一些关于X分布的信息。
  - 随机变量X的矩：
    - K阶原点矩：$E(X^k)$
    - K阶中心矩：$E([X-E(X)]^k)$
    - K是正整数，且假设上述期望均存在。
  - 随机变量的一阶原点矩就是均值，二阶中心矩就是方差。

参数估计

- 矩估计

数学上，“矩”是一组点组成的模型的特定的数量测度。在力学和统计学中都有用到“矩”。

- 如果这些点代表“质量”，那么：

零阶矩表示所有点的 质量；
一阶矩表示 质心；
二阶矩表示 转动惯量。

- 如果这些点代表“概率密度”，那么：

零阶矩表示这些点的 总概率（也就是1）；
一阶矩表示 期望；
二阶（中心）矩表示 方差；
三阶（中心）矩表示 偏斜度；
四阶（中心）矩表示 峰度；

- 这个数学上的概念和物理上的“矩”的概念关系密切

参数估计

- 矩估计
  - 矩估计法的基本思想是替换原理，即用样本矩替换同阶总体矩。

设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本，$X(f,\theta), \theta \in \Theta$，其中 $\theta = \theta_1, \theta_2, \ldots, \theta_k$ 为未知分布参数，Θ 为 k 维欧式空间的一个子集。记 $\mu_i = E(X^i)$ 为总体第 i 阶原点矩，$m_i = \frac{1}{n}\sum_{j=1}^{n} x_j^i$ 为样本第 i 阶原点矩 $(i = 1, 2, \ldots, k)$。替换原理即为，若参数 θ_i 能表示为 $\theta_i = g_i(\mu_1, \mu_2, \ldots, \mu_k)(i = 1, 2, \ldots, k)$，其中 $g_1, g_2, \ldots, g_k$ 为 k 个多源的已知函数，则可用 m_i 替换 $\mu_i(i = 1, 2, \ldots, k)$，得到 $\hat{\theta}_i = g_i(m_1, m_2, \ldots, m_k)$，即为 θ_i 的估计 $(i = 1, 2, \ldots, k)$。

直接把样本当成总体然后求解参数

参数估计

- 例子：黑白球（矩估计）
  - 假如有一个罐子，里面有黑白两种颜色的球，数目多少不知，两种颜色的比例也不知。
  - 每次任意从已经摇匀的罐中拿一个球出来，记录球的颜色，然后把拿出来的球再放回罐中。
  - 假如在前面的一百次重复记录中，有七十次是白球。请问罐中白球所占的比例是多少？

解：用样本中白球比例的均值作为估计代替总体均值。即估计结果为罐中白球所占的比例70%。符合直观。

理论根源是辛钦大数定律，样本之间是独立同分布，当数据样本量很大的时候，样本观测值的平均值和总体的数学期望是在一个极小的误差范围内。

参数估计

- 极大似然估计、最大后验概率估计(MAP)，贝叶斯估计
  - 贝叶斯公式：

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}$$

$$\mathrm{P}(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$

  - 因果概率

$$P(Cause \mid Effect) = \frac{P(Effect \mid Cause)P(Effect)}{P(Effect)}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

思路：在给定样本的前提下求最有可能的参数情况

参数估计

- 极大似然估计、最大后验概率估计(MAP)，贝叶斯估计
  - 给定：
    - A doctor knows that meningitis(脑膜炎) causes stiff neck(颈部僵硬) 50% of the time
    - Prior probability of any patient having meningitis is 1/50,000
    - Prior probability of any patient having stiff neck is 1/20
  - If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

E.g., let *M* be meningitis（脑膜炎）, *S* be stiff neck（脖子僵硬）:

Note: posterior probability of meningitis still very small!

参数估计

- 极大似然估计、最大后验概率估计(MAP)，贝叶斯估计
  - 贝叶斯公式：

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

  - 它将后验概率转化为基于似然函数和先验概率的计算表达式：

$$posterior = \frac{likelihood \cdot prior}{evidence}$$

需要估计的值

Reference: https://blog.csdn.net/yangliuy/article/details/8296481

参数估计

- 极大似然估计、最大后验概率估计(MAP)，贝叶斯估计
  - 贝叶斯公式：

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

  - 极大似然估计(MLE)：用似然函数取到最大值时的参数值作为估计值

$$L(\theta|X) = p(X|\theta) = \prod_{x \in X} p(X = x|\theta)$$

利用已知的样本结果，反推最有可能（最大概率）导致这样结果的参数值(样本分布已知，参数未知）

参数估计

- 极大似然估计、最大后验概率估计(MAP)，贝叶斯估计
  - 极大似然估计(MLE)：用似然函数取到最大值时的参数值作为估计值

$$L(\theta|X) = p(X|\theta) = \prod_{x \in X} p(X = x|\theta)$$

  - 由于有连乘运算，通常对似然函数取对数(log或者ln)计算简便(连乘运算变为连加运算)，即对数似然函数。所以，极大似然估计问题可以写成

$$\hat{\theta}_{ML} = argmax_{\theta} L(\theta|X) = argmax_{\theta} \sum_{x \in X} \log p(x|\theta)$$

这是一个关于θ的函数，求解这个优化问题通常对θ求导，得到导数为0的极值点。该函数取得最大值是对应的θ的取值就是我们估计的模型参数。

参数估计

- 极大似然估计:以扔硬币的伯努利实验为例子，N次实验的结果服从二项分布，参数为p，即每次实验事件发生的概率，设为得到正面的概率。为了估计p，采用极大似然估计，对数似然函数：

$$
\begin{aligned}
L &= \log \prod_{i=1}^{N} p(C = c_i|p) = \sum_{i=1}^{N} \log p(C = c_i|p) \\
&= n^{(1)} \log p(C = 1|p) + n^{(0)} \log p(C = 0|p) \\
&= n^{(1)} \log p + n^{(0)} \log(1 - p)
\end{aligned}
$$

其中n^i表示实验结果为i的次数。下面求似然函数的极值点，有 $\frac{\partial L}{\partial p} = \frac{n^{(1)}}{p} - \frac{n^{(0)}}{1-p} = 0$

得到参数p的最大似然估计值为 $\hat{p}_{ML} = \frac{n^{(1)}}{n^{(1)} + n^{(0)}} = \frac{n^{(1)}}{N}$

可以看出二项分布中每次事件发的概率p就等于做N次独立重复随机试验中事件发生的概率。如果我们做20次实验，出现正面12次，反面8次:

那么根据极大似然估计得到参数值p为12/20 = 0.6

参数估计

- 极大似然估计、最大后验概率估计(MAP)，贝叶斯估计
  - 贝叶斯公式：

$$p(\theta|X) = \frac{p(X|\theta)\cdot p(\theta)}{p(X)}$$

- 最大后验概率估计(MAP)：与极大似然估计不同点在于估计θ的函数中允许加入一个先验 **p(θ)**，也就是说此时不是要求似然函数最大，而是要求由贝叶斯公式计算出的整个后验概率最大，即

$$\begin{aligned}\hat{\theta}_{MAP} &= argmax_\theta \frac{p(X|\theta)p(\theta)}{p(X)} \\ &= argmax_\theta p(X|\theta)p(\theta) \\ &= argmax_\theta \{L(\theta\ X) + \log p(\theta)\} \\ &= argmax_\theta \{\sum_{x\in X} \log p(x|\theta) + \log p(\theta)\}\end{aligned}$$

注意这里p(X)与参数θ无关，因此等价于要使分子最大

与极大似然估计相比，多加上了一个先验分布概率的对数

参数估计

- 最大后验概率估计(MAP)
  - 在实际中，这个先验**p(θ)**可以用来描述人们已知或者接受的普遍规律。例如在扔硬币的试验中，每次抛出正面发生的概率应该服从一个概率分布，这个概率在0.5处取得最大值，这个分布就是先验分布。先验分布的参数(一个或多个)我们称为超参(hyperparameter)即

$$p(\theta) = p(\theta|\alpha)$$

  - 当上述后验概率取得最大值时，我们就得到根据MAP估计出的参数值。给定观测到的样本数据，一个新的值 $\tilde{x}$ 发生的概率可以用以下公式来估计：

$$p(\tilde{x}|X) = \int_{\theta \in \Theta} p(\tilde{x}|\hat{\theta}_{MAP}) p(\theta|X) d\theta = p(\tilde{x}|\hat{\theta}_{MAP})$$

参数估计

- 最大后验概率估计(MAP)
  - 扔硬币的例子：我们期望先验概率（待估计的参数θ即为p）分布在0.5处取得最大值，可以选用Beta分布（θ服从Beta分布）即：

$$p(p|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \triangleq Beta(p|\alpha,\beta)$$

  - 其中Beta函数展开是 $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ ，Gamma函数 $\Gamma(n) = (n-1)!$
  - Beta分布的随机变量范围是[0,1]，下图给出了不同参数情况下的Beta分布的概率密度函数

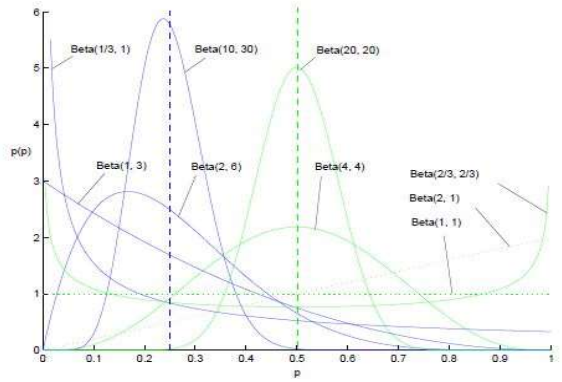


Fig. 1. Density functions of the beta distribution with different symmetric and asymmetric parametrisations.

参数估计

- 最大后验概率估计(MAP)

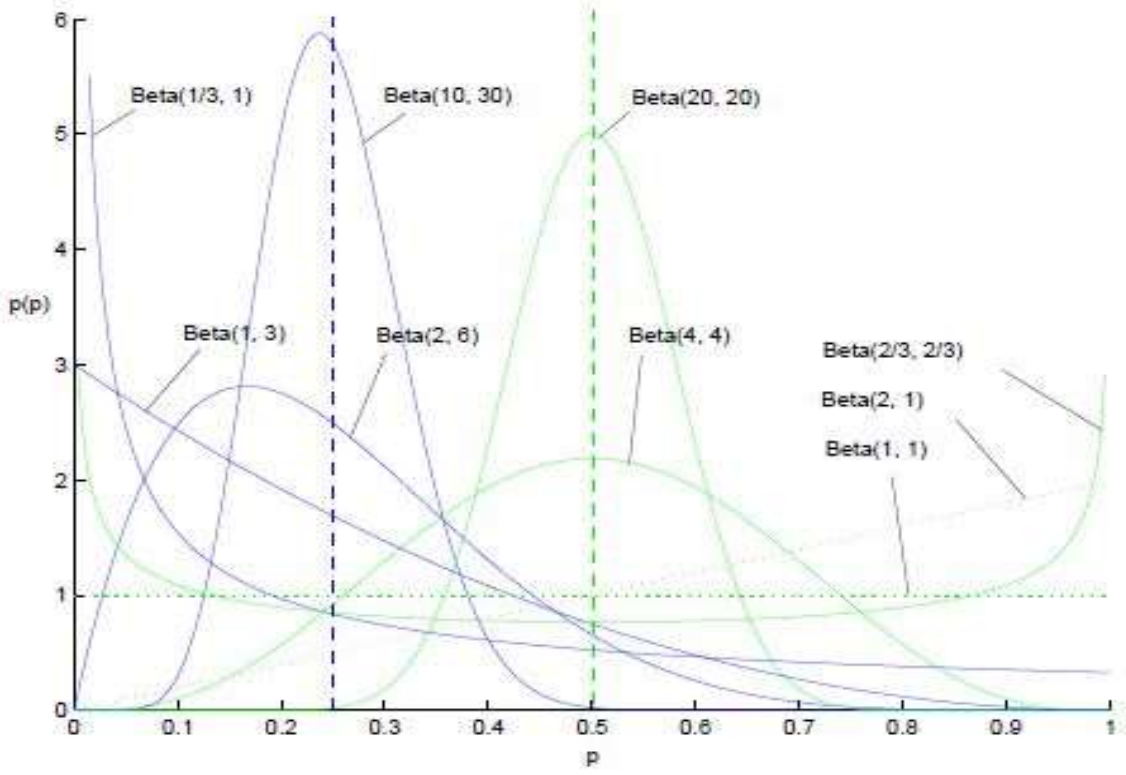


Fig. 1. Density functions of the beta distribution with different symmetric and asymmetric parametrisations.

参数估计

- 最大后验概率估计(MAP)
  - 仍以扔硬币的例子来说明，我们期望先验概率（待估计的参数θ即为这里的p）分布在0.5处取得最大值，可以选用Beta分布即：

$$p(p|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \triangleq Beta(p|\alpha,\beta)$$

  - 我们取 $\alpha = \beta = 5$ ，这样先验分布在0.5处取得最大值，现在我们来求解MAP估计函数的极值点，同样对p求导数我们有（练习）？

$$\hat{\theta}_{MAP} = argmax_{\theta}\{\sum_{x \in X} \log p(x|\theta) + \log p(\theta)\}$$

$\prod_{x \in X} p(X = x|\theta)$

$p(p|\alpha,\beta)$

参数估计

- 最大后验概率估计(MAP)
  - 仍以扔硬币的例子来说明，我们期望先验概率（待估计的参数θ即为这里的p）分布在0.5处取得最大值，可以选用Beta分布即：

$$p(p|\alpha,\beta)=\frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1}\triangleq Beta(p|\alpha,\beta)$$

  - 我们取 $\alpha=\beta=5$ ，这样先验分布在0.5处取得最大值，现在我们来求解MAP估计函数的极值点，同样对p求导数我们有

$$\hat{\theta}_{MAP}=argmax_\theta\{\sum_{x\in X}\log p(x|\theta)+\log p(\theta)\}$$

$$\prod_{x\in X}p(X=x|\theta)$$

$$p(p|\alpha,\beta)$$

$$\frac{\partial\hat{\theta}_{MAP}}{\partial p}=\frac{n^{(1)}}{p}-\frac{n^{(0)}}{1-p}+\frac{\alpha-1}{p}-\frac{\beta-1}{1-p}=0$$

参数估计

- 最大后验概率估计(MAP)
  - 仍以扔硬币的例子来说明，我们期望先验概率（待估计的参数θ即为这里的**p**）分布在0.5处取得最大值，可以选用Beta分布即：

$$p(p|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \triangleq Beta(p|\alpha,\beta)$$

设超参：
$\alpha = \beta = 5$

  - 得到参数**p**的的最大后验估计值为

$$\hat{p}_{MAP} = \frac{n^{(1)} + \alpha - 1}{n^{(1)} + n^{(0)} + \alpha + \beta - 2} = \frac{n^{(1)} + 4}{n^{(1)} + n^{(0)} + 8}$$

  - 和极大似然估计的结果对比可以发现结果中多了 $\alpha-1,\alpha+\beta-2$ 这样的pseudo-counts,这就是先验在起作用。并且超参数越大，为了改变先验分布传递的belief所需要的观察值就越多，此时对应的Beta函数越聚集，紧缩在其最大值两侧。

参数估计

- 最大后验概率估计(MAP)
  - 得到参数p的的最大后验估计值为

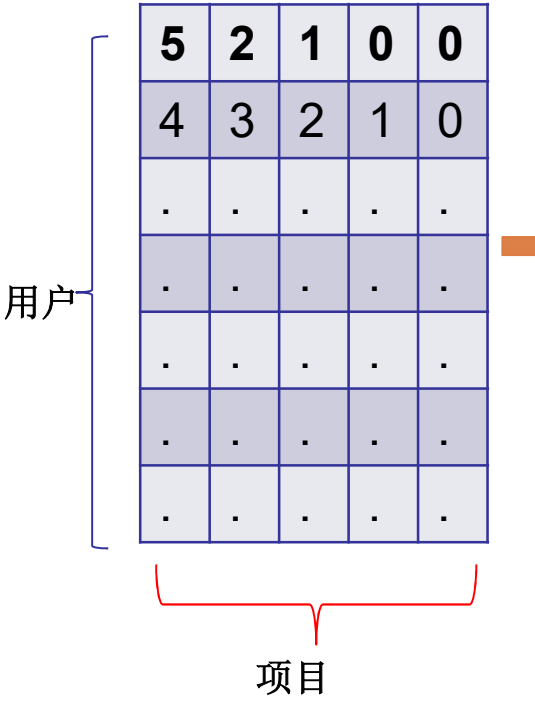
$$\hat{p}_{MAP} = \frac{n^{(1)} + \alpha - 1}{n^{(1)} + n^{(0)} + \alpha + \beta - 2} = \frac{n^{(1)} + 4}{n^{(1)} + n^{(0)} + 8}$$

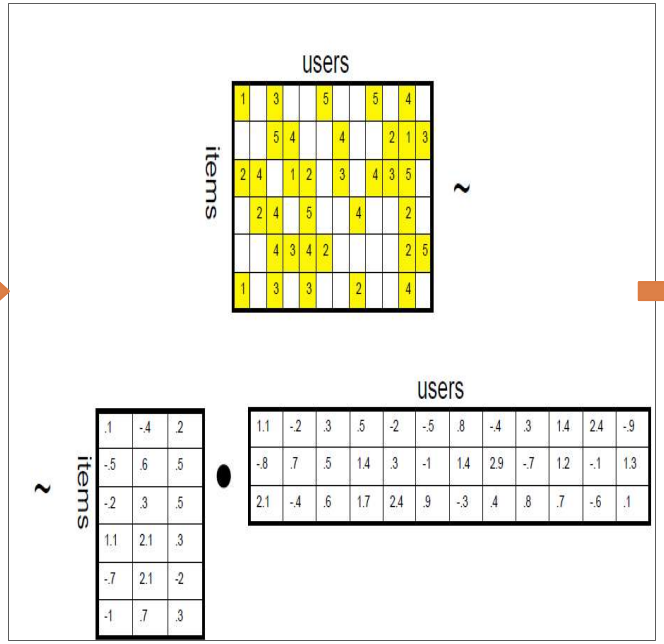
  - 和极大似然估计的结果对比可以发现结果中多了 $\alpha-1,\alpha+\beta-2$ 这样的pseudo-counts,这就是先验在起作用。
- 如果我们做20次实验，出现正面12次，反面8次，那么根据MAP估计出来的参数p为16/28 = 0.571,小于最大似然估计得到的值0.6
  - 这也显示了“硬币一般是两面均匀的”这一先验对参数估计的影响。

小结：MAP的后验概率的最大化是和先验分布紧密相关的。故而MAP可以看作是极大似然估计的正则化。

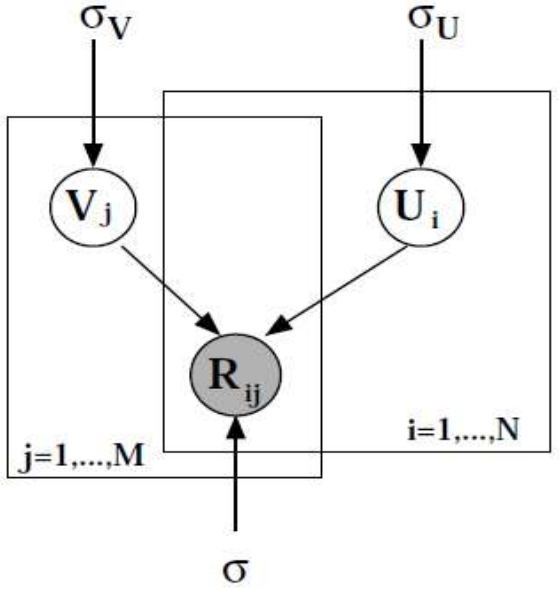
参数估计

- 基于矩阵分解的协同过滤算法
  - 面向评分预测的模型





基本分解思想



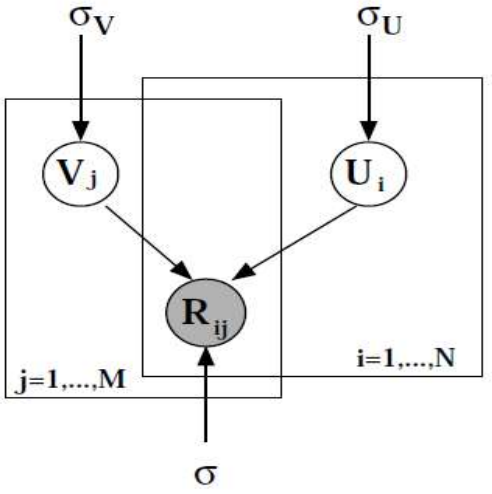
PMF（概率矩阵分解）

Probabilistic Matrix Factorization

参数估计：评分预测算法设计

- 基于矩阵分解的协同过滤算法
  - PMF Solution
    - 似然（预测）函数

$$p(R|U,V,\sigma^2)=\prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}(R_{ij}|U_i^TV_j,\sigma^2)\right]^{I_{ij}}$$



  - How to get U and V (θ)? --- MAP
    - The log-posterior of user and item features over fixed parameters

$$p(U,V\mid R,\sigma^2,\sigma_U^2,\sigma_V^2)$$

$$\propto p(R\mid U,V,\sigma^2)*p(U\mid\sigma_U^2)*p(V\mid\sigma_V^2)$$

$$p(V|\sigma_V^2)=\prod_{j=1}^{M}\mathcal{N}(V_j|0,\sigma_V^2\mathbf{I})$$

$$p(U|\sigma_U^2)=\prod_{i=1}^{N}\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})$$

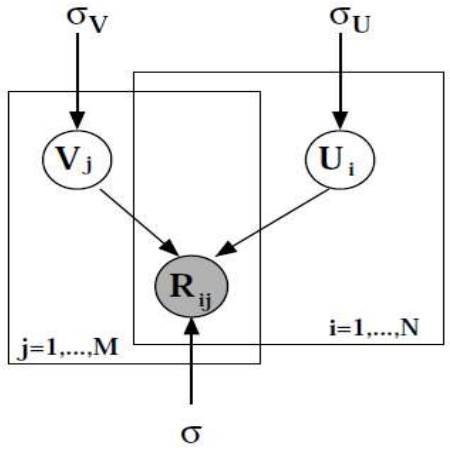
Likelihood!

Prior

超参

参数估计：评分预测算法设计

- 概率矩阵分解
  - MAP learning
    - **Maxmize the *ln*-posterior ?**

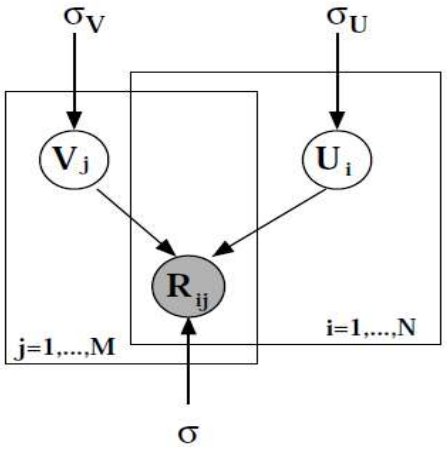


$$\ln p(U,V|R,\sigma^2,\sigma_V^2,\sigma_U^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij}-U_i^T V_j)^2 - \frac{1}{2\sigma_U^2}\sum_{i=1}^{N} U_i^T U_i - \frac{1}{2\sigma_V^2}\sum_{j=1}^{M} V_j^T V_j$$
$$-\frac{1}{2}\left(\left(\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\right)\ln\sigma^2 + ND\ln\sigma_U^2 + MD\ln\sigma_V^2\right) + C, \qquad (3)$$

参数估计：评分预测算法设计

- 概率矩阵分解
  - MAP learning

$$\ln p(U,V|R,\sigma^2,\sigma_V^2,\sigma_U^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij}-U_i^T V_j)^2 - \frac{1}{2\sigma_U^2}\sum_{i=1}^{N} U_i^T U_i - \frac{1}{2\sigma_V^2}\sum_{j=1}^{M} V_j^T V_j - \frac{1}{2}\left(\left(\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\right)\ln\sigma^2 + ND\ln\sigma_U^2 + MD\ln\sigma_V^2\right) + C, \quad (3)$$



  - Equivalent to minimize sum-of-squared-errors with quadratic regularization terms.

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\left(R_{ij}-U_i^T V_j\right)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|_{Fro}^2 + \frac{\lambda_V}{2}\sum_{j=1}^{M}\|V_j\|_{Fro}^2$$

$$\lambda_U = \frac{\sigma^2}{\sigma_U^2}, \lambda_V = \frac{\sigma^2}{\sigma_V^2}$$

regularization term to avoid over fitting

参数估计：评分预测算法设计

- 概率矩阵分解

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\left(R_{ij} - U_i^T V_j\right)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{N} \| U_i \|_{Fro}^2 + \frac{\lambda_V}{2}\sum_{j=1}^{M} \| V_j \|_{Fro}^2$$

1)Intialize U ,V with small, random values

2)repeat

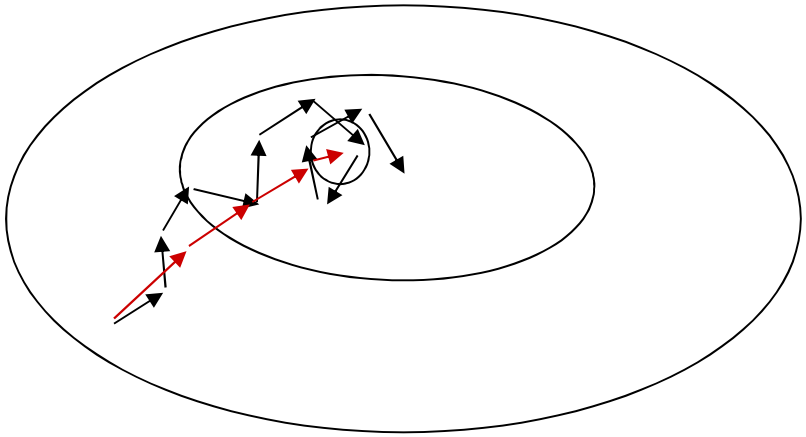
for each record in the training data

2.a) $U_i = U_i - a\frac{\partial E}{\partial U_i} = U_i - a\left(\sum_j I_{ij}(R_{ui} - U_i^T V_j)(-V_j) + \lambda_U U_i\right)$

2.b) $V_j = V_j - a\frac{\partial E}{\partial V_j} = V_j - a\left(\sum_i I_{ij}(R_{ui} - U_i^T V_j)(-U_i) + \lambda_V V_j\right)$

until convergence

Optimize: stochastic gradient descent



stochastic updates

full updates (averaged over all data-items)

参数估计

- MAP的优点
  - 引入了先验知识
  - 在数据量较小时更稳定
- MAP的缺点
  - 和MLE一样，只返回参数的单值估计
    - 导致后验在单值附近有明显尖峰
    - 预测结果不是不确定性上的平均（而是基于单值参数的推断）
  - 当我们用不同的参数去表示同一分布时，MAP会对这种表示的参数（超参数）的改变较为敏感
- 当先验分布均匀之时，MAP 估计与 MLE 相等
  - 最大似然方法可被看作一种特殊的 MAP，“让观察数据自己说话”

参数估计

- 极大似然估计、最大后验概率估计（MAP），贝叶斯估计
  - 贝叶斯公式：

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

  - 贝叶斯估计是贝叶斯统计框架下的一种参数估计方式，它与MAP一样**将参数视为随机变量**。不同的是，算法不是直接估计参数的值，而是估计**参数的概率分布**。
  - 每个样本的独立性假设 -> 条件独立性假设 $P(X|\theta)$
  - 和MAP/MLE不同，贝叶斯方法并不试图去寻找一个“最好的参数”$\hat{\theta}$，而是使用概率去描述我们对参数的初始不确定性（**先验**），并在见到证据后利用贝叶斯规则将观测纳入估计的过程（**后验**）。

参数估计

- 贝叶斯估计
  - 为了执行贝叶斯估计，首先需要在参数θ和数据X上描述一个联合分布（记住参数此时也是随机变量）$\mathrm{P}(X,\theta)$，易得:

$$\mathrm{P}(X,\theta) = P(X|\theta)P(\theta)$$

第一项刚好是我们之前描述的似然，后一项为先验

  - 由似然和先验，容易由贝叶斯法则导出后验:

$$\mathrm{P}(\theta|X) = \frac{P(X|\theta)P(\theta)}{p(X)}$$

其中$P(X) = \int_{\Theta} P(X|\theta)P(\theta)d\theta$为似然在所有可能参数赋值上的积分

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{\Theta} P(X|\theta)P(\theta)d\theta}$$

可看出，贝叶斯估计的求解非常复杂，因此选择合适的先验分布就非常重要

一般来说，计算积分是不可能的

https://www.jianshu.com/p/9c153d82ba2d

参数估计

- 贝叶斯估计
  - 下面仍然以抛硬币为例，此时我们选择Beta分布作为先验，类似之前的MAP，我们有：

$$p(p|\alpha,\beta)=\frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1}\triangleq Beta(p|\alpha,\beta)$$

$$P(\theta)=\gamma\theta^{\alpha-1}(1-\theta)^{\beta-1}$$ 其中γ为归一化常数

  - Beta分布在这里作为先验来做参数估计尤为有用，假设我们现在只有先验，没有数据，此时来考虑一次单独的硬币投掷$\mathbf{X}_1$,那么贝叶斯方法预测该硬币朝上的概率为：

$$P(x_1=1)=\int_0^1 P(x_1=1|\theta)P(\theta)d\theta$$

$$=\int_0^1 \theta P(\theta)d\theta=\int_0^1 \theta\gamma\theta^{\alpha-1}(1\text{-}\theta)^{\beta-1}d\theta$$

积分后(积分过程较复杂，此处省略)可得： $P(x_1=1)=\frac{\alpha}{\alpha+\beta}$

该结论可看出，Beta分布作为先验表明（假设）我们已经看到α次正面朝上和β次反面朝上。

参数估计

- 贝叶斯估计
  - 现在，让我们在先验的基础上加入更多观测，假设我们的数据集X中有M_1次正面($\mathbf{n^{(1)}}$)，M_2次反面($\mathbf{n^{(0)}}$)，那么后验即为：

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$
$$\propto \theta^{M_1}(1-\theta)^{M_2}\theta^{\alpha}(1-\theta)^{\beta}$$
$$= \theta^{\alpha+M_1}(1-\theta)^{\beta+M_2}$$

在把后验概率推导为和先验概率一样的分布形式的时候，贝叶斯公式分母p(X)可以看做一个常数，往往充当了一个normalize，归一化的作用。

更正：公式里指数应当是α-1， β-1

如果“先验概率”和“后验概率”都服从同样的分布类型（参数当然是不同的），那么那么计算先验概率和似然概率的乘积就很方便了，只需要将指数相加即可

显然可观察到，在抛硬币的实验中（likelihood $P(X|\theta)$ 为伯努利分布），**当先验(prior)** $P(\theta)$**为Beta分布时，后验（posterior）** $P(\theta|X)$ **也为Beta分布**，即更新后的参数服从一个新的Beta($\alpha+M_1$, $\beta+M_2$)分布，这种情况我们称之为Beta分布是伯努利似然函数$P(X|\theta)$的**共轭**。

参数估计

- 贝叶斯估计
  - 因此，在应用中，我们常常使用likelihood的共轭分布作为参数的先验分布，以取得计算和形式上的便利。一个常见的例子是描述文本主题分布中，我们对多项式(Multinomial)分布的似然选取参数服从迪利克雷（Dirichlet）分布作为先验
    - 先验分布叫做似然函数P(X|θ)的共轭先验分布
    - 共轭分布总是针对分布中的某个参数θ而言。之所以采用共轭先验的原因是可以使得先验分布和后验分布的形式相同，但是参数不同。

似然　　常见共轭先验分布

总体分布	参数	共轭先验分布
二项分布	成功概率p	β分布β(α,β)
泊松分布	均值λ	Γ分布Γ(αβ)
指数分布	均值的倒数λ	Γ分布Γ(αβ)
正态分布（方差已知）	均值μ	正态分布N(μ,σ²)
正态分布（均值已知）	方差σ²	倒Γ分布

Beta(α, β)

https://blog.csdn.net/sthp888/article/details/90636368