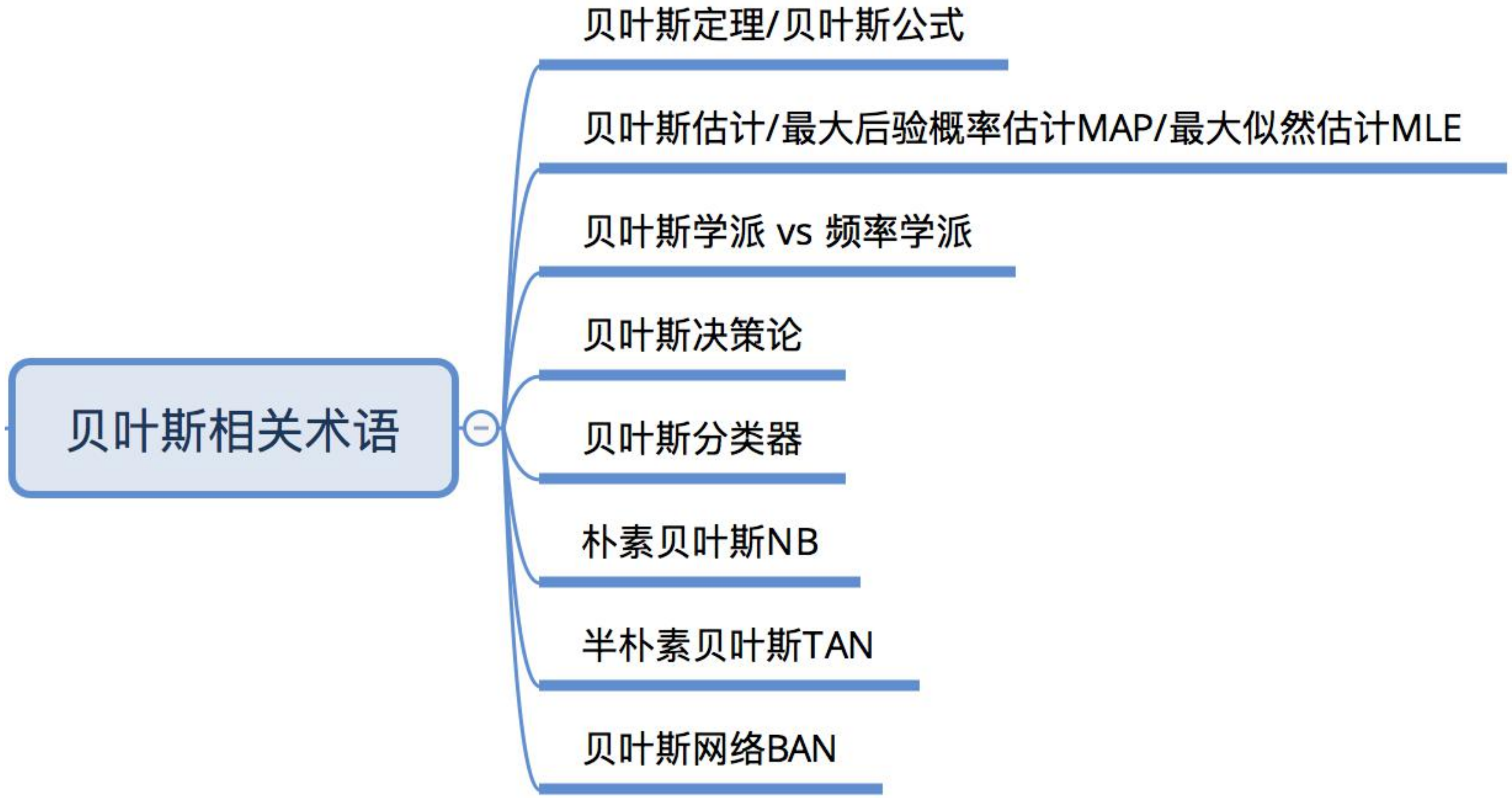
朴素贝叶斯

邱小丽 **2020-04-30**

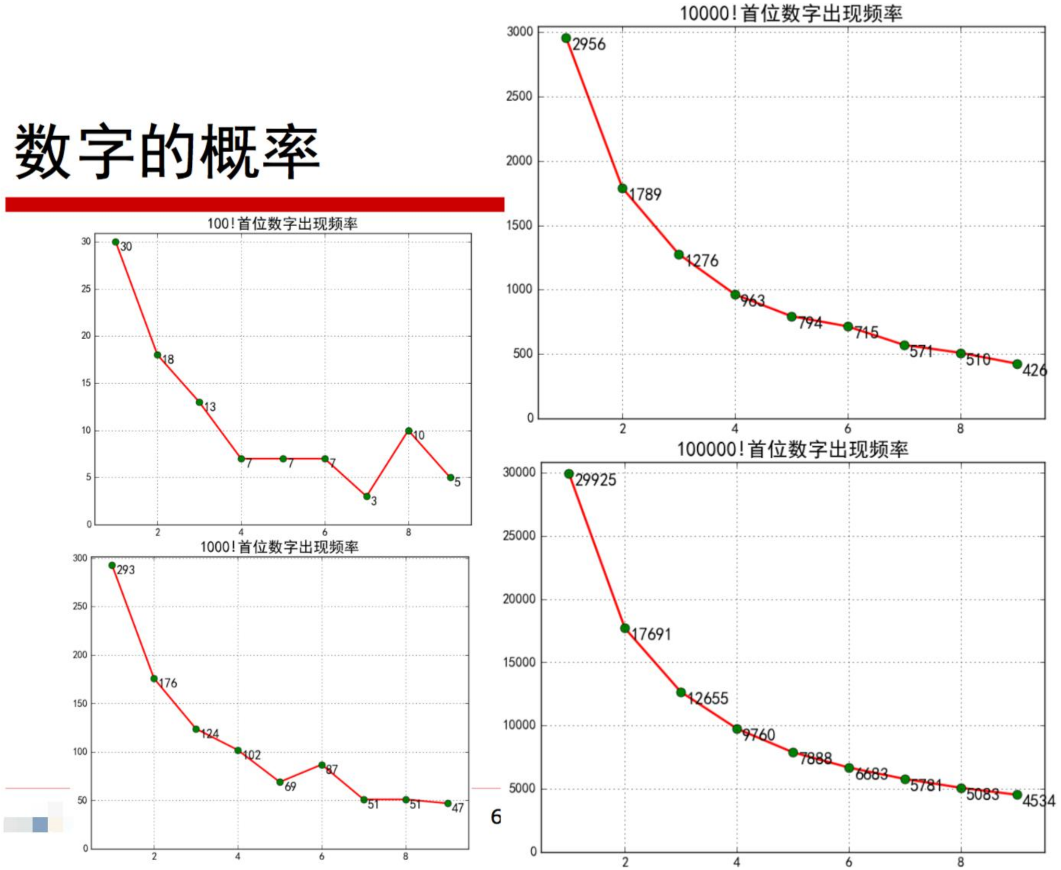

贝叶斯相关术语
贝叶斯定理/贝叶斯公式
贝叶斯估计/最大后验概率估计MAP/最大似然估计MLE
贝叶斯学派 vs 频率学派
贝叶斯决策论
贝叶斯分类器
朴素贝叶斯NB
半朴素贝叶斯TAN
贝叶斯网络BAN

朴素贝叶斯

- 相关统计学知识
  - 条件概率/先验概率/后验概率/全概率
- 贝叶斯定理
  - 贝叶斯公式定义
  - 应用案例
  - 贝叶斯定理解决问题思路
- 朴素贝叶斯算法
  - 模型、推理、参数估计
  - 算法过程
  - 几种常见的分类器
  - 小结
- 无处不在的贝叶斯
- 文档分类**DEMO**

概率与直观

- **本福特定律**，也称为本福特法则，说明一堆从实际生活得出的数据中，以**1**为首位数字的数的出现概率约为总数的三成，接近直觉得出之期望值**1/9**的**3**倍。推广来说，越大的数，以它为首几位的数出现的概率就越低。它可用于检查各种数据是否有造假 。



- 阶乘/素数数列/斐波那契数列
- 住宅地址号码
- 经济数据反欺诈
- 选举投票反欺诈

d	p
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

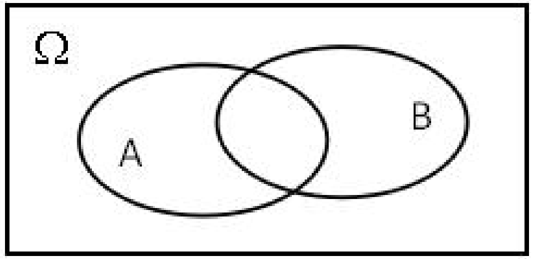
概率论只不过是把常识用数学公式表达了出来。——拉普拉斯

概率基础

- 条件概率

设**A,B**是两个事件，且**P(B)>0,**则在事件**B**发生的条件下，

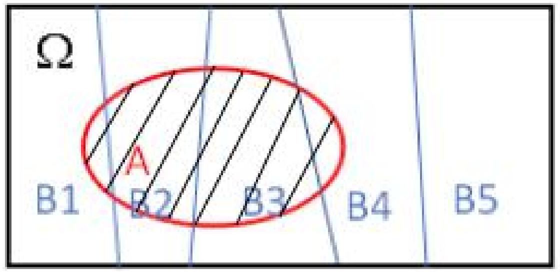
事件**A**发生的条件概率为：**P(A|B) = P(AB)/P(B)**



- 联合概率

联合概率指的是包含多个条件且所有条件同时成立的概率，记作**P(X=a,Y=b)**或**P(a,b)**，有的书上也习惯记作**P(ab)**

  - 与条件概率关系 **P(X|Y) =** $\frac{P(X,Y)}{P(Y)}$
  - 与边缘概率关系 $P(A=a)=\sum_b P(A=a,B=b)$ $P(A=b)=\sum_a P(A=a,B=b)$



- 联合概率分布：联合概率分布就是联合概率在样本空间中的分布情况。
- 乘法公式

P(AB) = P(B)P(A|B) = P(A)P(B|A)

乘法公式的推广：对于任何正整数**n≥2**，当**P(A1A2...An-1) > 0** 时，有：

P(A1A2...An-1An)=P(A1)P(A2|A1)P(A3|A1A2)...P(An|A1A2...An-1)

- 全概率公式 $P(A) = \sum_{i=1}^{\infty} P(B_i)P(A|B_i)$ P(A) = P(B1)P(A|B1) + P(B2)P(A|B2) *解决正向概率的问题*
- 先验概率

是指根据以往经验和分析得到的概率，如全概率公式，它往往作为"由因求果"问题中的"因"出现的概率。

- 后验概率

事情已经发生，要求这件事情发生的原因是由某个因素引起的可能性的大小。

后验概率是指在得到“结果”的信息后重新修正的概率，是“执果寻因”问题中的"果"。后验概率的计算要以先验概率为基础。

贝叶斯定理

- 由来

贝叶斯定理是**18**世纪英国数学家托马斯·贝叶斯（**Thomas Bayes**）提出得重要概率论理论。

> 所谓的贝叶斯定理源于他生前为解决一个“逆概”问题写的一篇文章，而这篇文章是在他死后才由他的一位朋友发表出来的。在贝叶斯写这篇文章之前，人们已经能够计算“正向概率”，如“假设袋子里面有 N 个白球，M 个黑球，你伸手进去摸一把，摸出黑球的概率是多大”。而一个自然而然的问题是反过来：“如果我们事先并不知道袋子里面黑白球的比例，而是闭着眼睛摸出一个（或好几个）球，观察这些取出来的球的颜色之后，那么我们可以就此对袋子里面的黑白球的比例作出什么样的推测”。这个问题，就是所谓的逆向概率问题。

正向概率

逆概率

- 贝叶斯公式

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

似然度

标准化常量

B要求解的问题 **A**已知信息

后验概率

先验概率

(先根据以往经验预估)

标准似然度/可能性函数

>1, 意味着先验概率被增强，事件**B**发生的可能性变大;

=1, 意味着事件**A**无助于判断事件**B**的可能性;

>1, 意味着先验概率被削弱，事件**B**发生的可能性变小

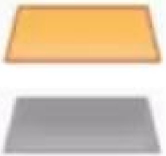
$$\mathbf{P}(B_i|A)=\frac{P(A|B_i)^*P(B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_j)}=\frac{likelihood\ ^*\ prior}{evidence}$$

给定某系统的若干样本**x**,计算该系统的参数θ（类别），即:

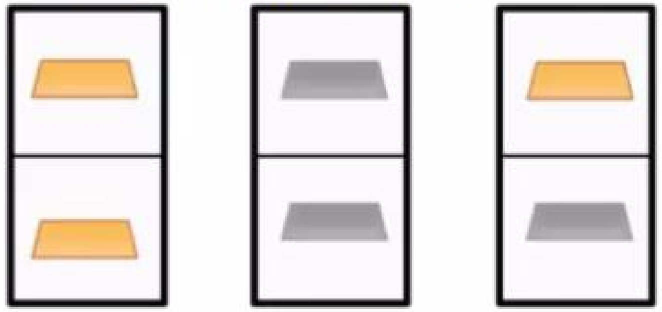
参数θ的概率分布：似然函数

$$\mathbf{P(\theta|x)}=\frac{P(x|\theta)\ P(\theta)}{P(x)}$$

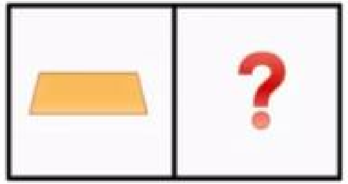
金条问题



- 现有三个箱子，每个箱子各有两块贵金属。
  - 三个箱子的金银条如下：



  - 现随机选择一个箱子的其中一块贵金属，发现是金条，请问，该箱子中另外一块仍然是金条的概率是多少？



- 举例:疾病检测

现在有种病的发病率是**0.001**，有一种试剂可以检测患者是否得病，准确率是**0.99**，他的误报率是**5**%，就是说被测者没有患病的情况下，他有**5%**的可能呈现阳性。现在有一个患者**检测结果是阳性，请问，他确实得病的可能性**有多大?

第一步：分解问题

- 要求解的问题

病人的检测结果是阳性（新的信息）为事件**B**；他得病记为事件**A**；那么求解的就是**P(A|B)**

- 已知信息

P(A)=0.001

P(B|A)＝0.99

P(B|A')＝5%

- 求解

P(B) = P(B|A)P(A) + P(B|A')P(A') = 0.99*0.001 + 0.05*0.999= 0.05

P(A|B) = 1.98%

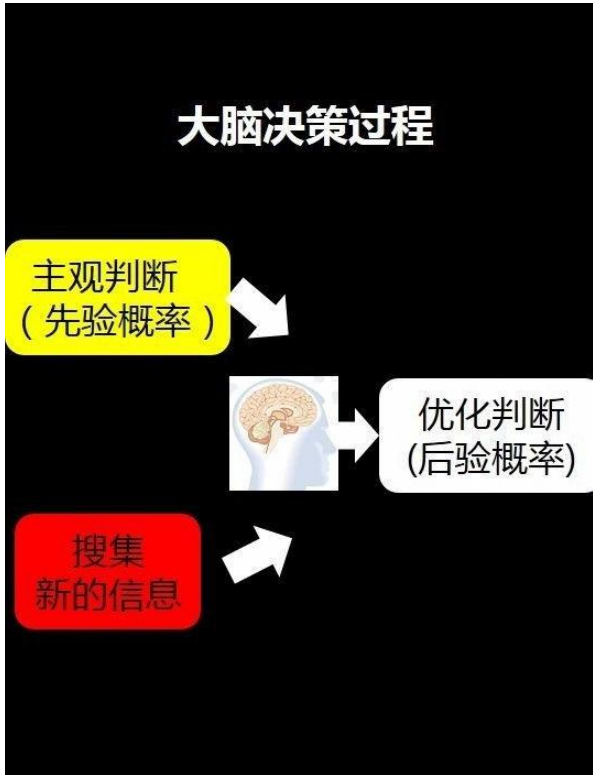
也就是说，虽然试剂的准确性为**99%**，但是通过检验判断有没有得病的概率只有**1.98%**

- 应该相信筛查结果吗?

提高先验概率，可以有效提高后验概率。这时就需要对检查出问题的人再重复检测一次，提高检测的准确率。

• 贝叶斯定理解决问题思路

如何在生活中优化你的决策



1、分解问题
先列出要解决的问题是什么?
已知的条件有哪些?
2、给出主观判断
不是瞎猜，是根据自己的经历和学识来给出主观判断，也就是给出先验概率
3、搜集新的信息，优化判断
持续关注你要解决问题的相关信息的最新动态，然后用获取到的新信息来不断调整第**2**步的主观判断。如果新信息符合这个主观判断，你就提高主观判断的可信度，如果不符合，就降低主观判断的可信度。

大胆假设，小心求证——胡适

朴素贝叶斯

- 假设
  - 一个特征出现的频率，与其他特征(条件)独立(特征独立性)
    - 其实是：对于给定分类的条件下，特征独立
  - 每个特征同等重要(特征均衡性)
- 朴素贝叶斯(**Naive Bayes,NB**)是基于“特征之间是独立的”这一朴素假设，应用贝叶斯定理的监督学习算法。

对于给定的特征向量 $x_1, x_2, \cdots, x_n$

类别y的概率可以根据贝叶斯公式得到：

$$P(y \mid x_1, x_2, \cdots, x_n) = \frac{P(y)P(x_1, x_2, \cdots, x_n \mid y)}{P(x_1, x_2, \cdots, x_n)}$$

- 典型的生成学习方法：由训练数据学习联合概率分布**P(X,Y)**然后求得后验概率分布**P(Y|X)**.计算联合概率**p(x,y)**，可以理解为对**p(x|y)**和**p(y)**同时进行建模;
- 概率估计方法可以是极大似然估计、最大后验概率估计。

推导

- 使用朴素的独立性假设：

$$P(x_i \mid y, x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) = P(x_i \mid y)$$

- 类别y的概率可简化为：

$$P(y \mid x_1, x_2, \cdots, x_n) = \frac{P(y)P(x_1, x_2, \cdots, x_n \mid y)}{P(x_1, x_2, \cdots, x_n)} = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, x_2, \cdots, x_n)}$$

- 在给定样本的前提下，$P(x_1, x_2, \cdots, x_n)$是常数：

$$P(y \mid x_1, x_2, \cdots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

- 从而：$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{n} P(x_i \mid y)$

朴素贝叶斯模型

假设分类模型样本为:

$(x_1^{(1)}, x_2^{(1)}, ... x_n^{(1)}, y_1), (x_1^{(2)}, x_2^{(2)}, ... x_n^{(2)}, y_2), ...(x_1^{(m)}, x_2^{(m)}, ... x_n^{(m)}, y_m)$

即**m**个样本，每个样本**n**个特征，特征输出为**K**个类别，定义为$C_1, C_2, ... C_k$。

从样本中可以学习到先验分布**P(Y=C_k)(k=1,2,...K)**

接着学习到条件概率分布 **P(X=x|Y=C_k) = P**$(X_1 = x_1, X_2 = x_2, ... X_n = x_n | Y = C_k)$

然后得到**X**和**Y**的联合分布**P(X,Y=C_k) = P(Y=C_k)P(X=x|Y=C_k)**

$$= P(Y=C_k)P(X_1 = x_1, X_2 = x_2, ... X_n = x_n | Y = C_k)$$

假设：**X**的**n**个维度之间互相独立

最大似然法

$$= P(Y=C_k)\ P(X_1 = x_1|Y=C_k)P(X_2 = x_2|Y=C_k)...P(X_n = x_n|Y=C_k)$$

给定测试集的一个新样本特征$(x_1^{(test)}, x_2^{(test)}, ... x_n^{(test)})$,如何判断属于哪个模型?

只需要计算出**K**个条件概率**P(Y=C_k|X=$X^{(test)}$)**，然后找出最大条件概率对应的类别。

朴素贝叶斯推导过程

我们预测的类别 C_{result} 是使P(Y=C_k|X=$X^{(test)}$)最大化的类别，即为：

$$C_{result} = \ argmax\,\mathrm{P}(\mathrm{Y}{=}C_k|\mathrm{X}{=}X^{(test)})$$

$$= argmax\,\frac{P(X{=}X^{(test)}\,|Y{=}C_k)\;P(Y{=}C_k)}{P(X{=}X^{(test)})}$$

对于所有的类别计算上式，分母都是一样的。因此可以简化为：

$$= argmax\,P(X = X^{(test)}\,|Y = C_k)\;P(Y = C_k)$$

接着利用朴素贝叶斯独立性假设，得到：

$$C_{result} = \ argmax\ P(Y = C_k)\prod_{j=1}^{n} P(X_j \ = x_j^{(test)}|\ Y = C_k)$$

朴素贝叶斯算法过程

假设训练集为**m**个样本**n**个维度，如下：

$(x_1^{(1)}, x_2^{(1)}, \ldots x_n^{(1)}, y_1),(x_1^{(2)}, x_2^{(2)}, \ldots x_n^{(2)}, y_2),\ldots(x_1^{(m)},x_2^{(m)},\ldots x_n^{(m)},y_m)$

特征输出为**K**个类别，定义为$C_1, C_2, \ldots C_k$。每个特征输出类别的样本个数为$m_1, m_2,\ldots,m_k$，

在第k个类别中，如果是离散特征，则特征X_j *各个类别取值为*m_{kjl}, 其中l 的取值为1, 2, $\ldots S_j$为特征j不同的取值数。

输出为实例$X^{(test)}$的分类。

算法流程如下：

1）计算Y的K个先验概率或者已有先验概率 P(**Y=**C_k) = $\frac{m_k+\lambda}{m_k+S_j\lambda}$ = $\frac{m_k+\lambda}{m+K\lambda}$

2) 分别计算第k个类别的第j维特征的第l个取值条件概率 **P(**X_j**=**x_{jl}**|Y=**C_k**)**

a) 如果是离散值

$$P(X_j = x_{jl}|Y = C_k) = \frac{m_{kjl} + \lambda}{m_k + S_j\lambda}$$

b) 如果是连续值，不需要计算各个l的取值概率，直接求正态分布的参数

$$P(X_j = x_j|Y = C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}exp\left(- \frac{(x_j - \mu_k)^2}{2\sigma_k^2}\right)$$

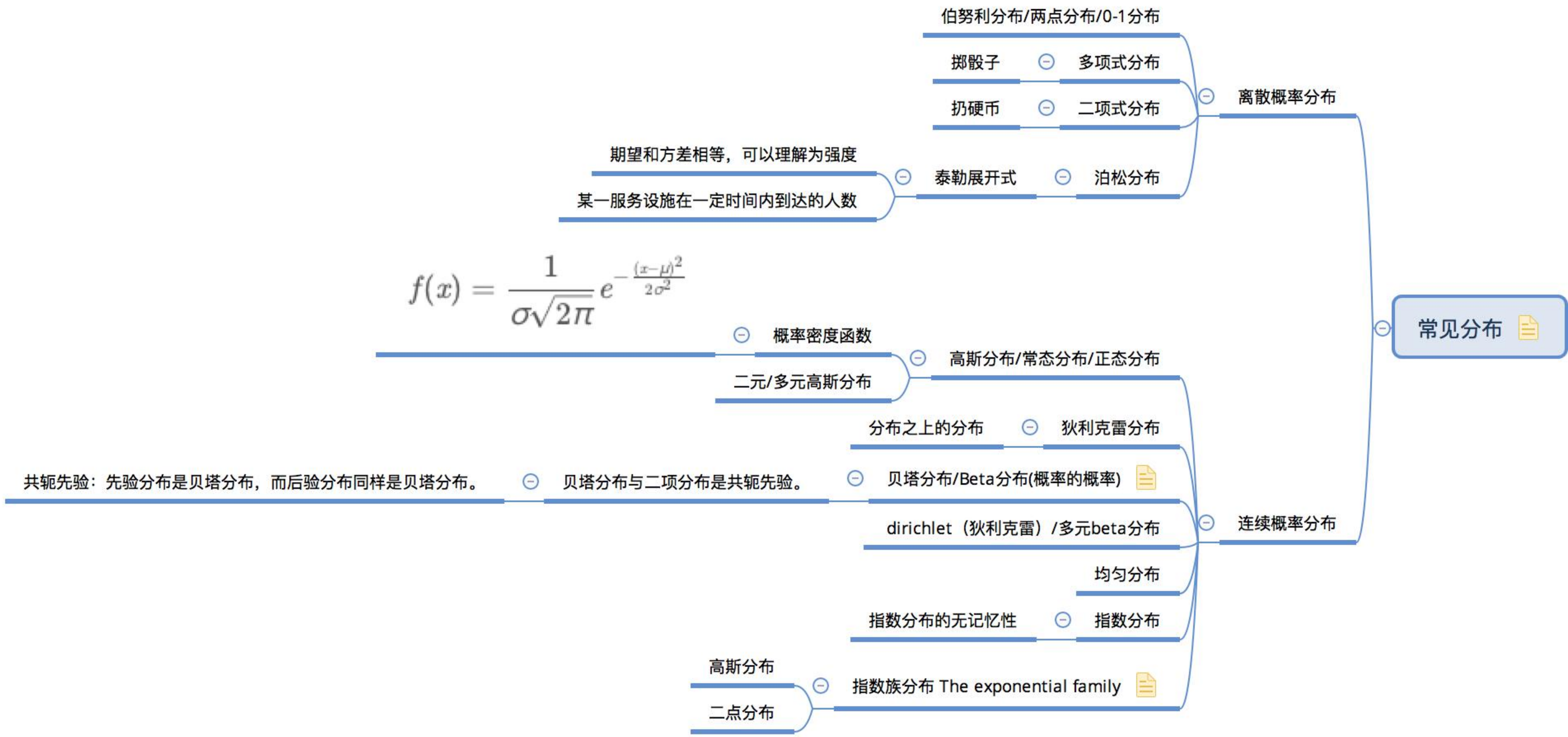
需要求出μ_k和σ_k^2。 μ_k为在样本类别C_k中，所有X_j的平均值。σ_k^2为在样本类别C_k中，所有X_j的方差。

3) 对于实例$X^{(test)}$，分别计算

$$P(Y = C_k)\prod_{j=1}^{n} P(X_j = x_j^{(test)}|Y = C_k)$$

4）确定实例$X^{(test)}$的分类 C_{result}

$$C_{result} = \underbrace{argmax}_{C_k} P(Y = C_k)\prod_{j=1}^{n} P(X_j = X_j^{(test)}|Y = C_k)$$


常见分布
离散概率分布
伯努利分布/两点分布/0-1分布
多项式分布
掷骰子
二项式分布
扔硬币
泊松分布
泰勒展开式
期望和方差相等，可以理解为强度
某一服务设施在一定时间内到达的人数
连续概率分布
高斯分布/常态分布/正态分布
概率密度函数
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
二元/多元高斯分布
狄利克雷分布
分布之上的分布
贝塔分布/Beta分布(概率的概率)
贝塔分布与二项分布是共轭先验。
共轭先验：先验分布是贝塔分布，而后验分布同样是贝塔分布。
dirichlet（狄利克雷）/多元beta分布
均匀分布
指数分布
指数分布的无记忆性
指数族分布 The exponential family
高斯分布
二点分布

高斯朴素贝叶斯Gaussian Naive Bayes

- 根据样本使用MAP(Maximum A Posteriori)估计$P(y)$，建立合理的模型估计$P(x_i \mid y)$，从而得到样本的类别。

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

- 假设特征服从高斯分布，即：

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

  - 参数使用MLE估计即可。

多项分布朴素贝叶斯Multinomial Naive Bayes

- 假设特征服从多项分布，从而，对于每个类别y，参数为$\theta_y=(\theta_{y1},\theta_{y2},\cdots,\theta_{yn})$，其中n为特征的数目，$P(x_i\mid y)$的概率为$\theta_{yi}$。
- 参数θ_y使用MLE估计的结果为：$\hat{\theta}_{yi}=\dfrac{N_{yi}+\alpha}{N_y+\alpha\cdot n}$，$\alpha\ge 0$
- 假定训练集为T，有：

$$\begin{cases} N_{yi}=\sum\limits_{x\in T} x_i & \text{训练集}T\text{上特征}i\text{在类别}y\text{中出现的次数} \\ N_y=\sum\limits_{i=1}^{|T|} N_{yi} & \text{类别}y\text{的所有特征的总数} \end{cases}$$

- 其中，
  - $\alpha=1$称为Laplace平滑，
  - $\alpha<1$称为Lidstone平滑。

$$\boxed{\hat{y}=\arg\max_y P(y)\prod_{i=1}^{n}P(x_i\mid y)}$$

MLE

以抛硬币为例，假设我们有一枚硬币，现在要估计其正面朝上的概率 θ。为了对 θ 进行估计，我们进行了10次实验（独立同分布，i.i.d.），这组实验记为 $X = x_1$，x_2，$\ldots$，x_{10}，其中正面朝上的次数为6次，反面朝上的次数为4次，结果为 $(1,0,1,1,0,0,0,1,1,1)$。

MLE的思想是使得观测数据（样本）发生概率最大的参数就是最好的参数。

$$L(X;\theta) = \prod_{i=0}^{n} P(x_i|\theta) = \theta^6(1-\theta)^4 \qquad \hat{\theta} = 0.6$$

$$lnL(X;\theta) = ln\prod_{i=0}^{n} P(x_i|\theta) = \sum_{i=0}^{n} ln(P(x_i|\theta)) = 6ln(\theta) + 4ln(1-\theta)$$

MLE的求解步骤：

- 确定似然函数
- 将似然函数转换为对数似然函数
- 求对数似然函数的最大值（求导，解似然方程）

MAP

最大后验概率估计的求解步骤：

- 确定参数的先验分布以及似然函数
- 确定参数的后验分布函数
- 将后验分布函数转换为对数函数
- 求对数函数的最大值（求导，解方程）

$$\underset{\theta}{argmax}\, P(\theta|X) = \underset{\theta}{argmax}\, \frac{P(X|\theta)P(\theta)}{P(X)} \propto \underset{\theta}{argmax}\, \underbrace{P(X|\theta)P(\theta)}$$

假设 θ~N(0.5,0.1)

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\theta-\mu)^2}{2\sigma^2}} = \frac{1}{10\sqrt{2\pi}}e^{-50(\theta-0.5)^2}$$

$$P(X|\theta)P(\theta) = \theta^6 \times (1-\theta)^4 \times \frac{1}{10\sqrt{2\pi}} \times e^{-50(\theta-0.5)^2}$$

$$ln(P(X|\theta)P(\theta)) = ln(\theta^6 \times (1-\theta)^4 \times \frac{1}{10\sqrt{2\pi}} \times e^{-50(\theta-0.5)^2}) = 6ln(\theta) + 4ln(1-\theta) + ln(\frac{1}{10\sqrt{2\pi}}) - 50(\theta-0.5)^2$$

$$\hat{\theta} \approx 0.529$$

假设 θ~Beta(3,3)

$$P(X|\theta)P(\theta) = \theta^6 \times (1-\theta)^4 \times \frac{1}{B(\alpha,\beta)} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\hat{\theta} = \frac{\alpha+5}{\alpha+\beta+8} = \frac{8}{3+3+8} \approx 0.57$$

BE

共轭先验：

在贝叶斯统计中，如果后验分布与先验分布属于同类，则先验分布与后验分布被称为共轭分布，而先验分布被称为似然函数的共轭先验

目的：

估计θ的分布P(θ|X)；如果后验分布的范围较窄，则估计值的准确度相对较高，反之，如果后验分布的范围较广，则估计值的准确度就较低

似然函数服从二项分布(共轭先验为**beta**分布) $P(\theta) \sim Beta(\alpha, \beta)$

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\int_\Theta P(X|\theta)P(\theta)d\theta} = \frac{\theta^6(1-\theta)^4 \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\int_\Theta \theta^6(1-\theta)^4 \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} d\theta} = Beta(\theta|\alpha+6, \beta+4)$$

假设 $\alpha = 3, \beta = 3$，在这种情况下，先验分布会在 0.5 处取得最大值

可以用后验分布的期望作为θ的估计值 $\hat{\theta} = \int_\Theta \theta P(\theta|X)d\theta = E(\theta) = \frac{\alpha}{\alpha+\beta} = \frac{9}{9+7} = 0.5625$

注：二项分布参数的共轭先验是**Beta**分布，多项式分布参数的共轭先验是**Dirichlet**分布，指数分布参数的共轭先验是**Gamma**分布，高斯分布均值的共轭先验是另一个高斯分布，泊松分布的共轭先验是**Gamma**分布。

MLE & MAP & BE

Type	MLE	MAP	BE
$\hat{\theta}$	0.6	0.57	0.5625
f	$P(X\|\theta)$	$P(X\|\theta)P(\theta)$	$\frac{P(X\|\theta)P(\theta)}{P(X)}$

- 从MLE、MAP到BE，从上表可以看出的θ估计值是逐渐接近0.5的。

从公式的变化可以看出，使用的信息是逐渐增多的。

- MLE、MAP中都是假设θ未知，但是确定的值，都将使函数取得最大值的作为估计值，区别在于最大化的函数不同，最大后验概率估计使用了θ的先验概率。
- BE中，假设参数θ是未知的随机变量，不是确定值，求解的是参数θ在样本X上的后验分布。

一句话总结:

- MLE: 寻找使模型产出观测数据的概率最大的一组模型参数
- MAP:寻找对于已知先验概率以及观测数据最适合的一组模型参数
- BE: 估计参数的分布

以文本

- 样本：
  邮件
- 分类
  圾邮
- 方法：

- 类别c：垃圾邮件c_1，非垃圾邮件c_2
- 词汇表，两种建立方法：
  - 使用现成的单词词典；
  - 将所有邮件中出现的单词都统计出来，得到词典。
  - 记单词数目为N
- 将每个邮件m映射成维度为N的向量**x**
  - 若单词w_i在邮件m中出现过，则$x_i=1$，否则，$x_i=0$。即邮件的向量化：$m \rightarrow (x_1, x_2 \ldots\ldots x_N)$
- 贝叶斯公式：$P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$
  - $P(c_1|\mathbf{x})=P(\mathbf{x}|c_1)*P(c_1) / P(\mathbf{x})$
  - $P(c_2|\mathbf{x})=P(\mathbf{x}|c_2)*P(c_2) / P(\mathbf{x})$
    - 注意这里**x**是向量

- $P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$
- $P(\mathbf{x}|c)=P(x_1,x_2 \ldots x_N|c)=P(x_1|c)*P(x_2|c) \ldots P(x_N|c)$
  - 特征条件独立假设
- $P(\mathbf{x})=P(x_1,x_2 \ldots x_N)=P(x_1)*P(x_2) \ldots P(x_N)$
  - 特征独立假设
- 带入公式：$P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$

无处不在的贝叶斯

- 机器学习：垃圾邮件分类/文本分类
  - 如果样本特征的分布大部分是连续值，则用**GaussianNB**比较好；
  - 特征分布大部分是多元离散，用**MultinomialNB**比较合适
  - 样本特征是二元离散或者很稀疏的多元离散，用**BernoulliNB;**
  - 如果**x**即有连续又有离散，一般选择高斯朴素贝叶斯
- 自然语言处理：中文分词
- 统计机器翻译
- 图像识别：贝叶斯图像识别
- **EM**算法与基于模型的聚类
- 推荐系统
- 博弈论

朴素贝叶斯算法小结

- 优点：

  1）朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。

  2）对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。

  3）对缺失数据不太敏感，算法也比较简单，常用于文本分类。

- 朴素贝叶斯的主要缺点有：

  1） 理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为朴素贝叶斯模型给定输出类别的情况下,假设属性之间相互独立，这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。

  2）需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。

  3）由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。

  4）对输入数据的表达形式很敏感。

文本分类demo

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

实验数据：新闻组中的20个类别，原始文本数目约两万个，根据新闻组中文本的时间前后，划分成训练集(60%)和测试集(40%)。

数据获取：

- 可使用sklearn.datasets.fetch_20newsgroups获取原始文本
- 或者使用sklearn.datasets.fetch_20newsgroups_vectorized返回文本向量

http://qwone.com/~jason/20Newsgroups/

Thanks & Questions