# 贝叶斯分类器

## 贝叶斯定理

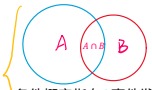
### 概率空间三要素

以经典的抛硬币为例

- 样本空间：样本空间就是事件条件下所得到的所有结果，因此掷一次硬币的样本空间为{ 正面，反面 }
- 事件：概率空间中的事件与我们平时生活中所说的事件没有任何分别。这里指的是此掷一次硬币
- 概率：概率是事件发生后，出现结果的个数与样本空间个数的比值

假如此次掷硬币的结果为正面，由于正面这个结果只发生了一次，而样本空间的个数是2，所以掷一次硬币出现正面的概率就为1/2

### 条件概率



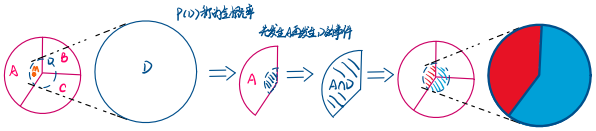

$$P(B|A)=\frac{P(A\cap B)}{P(A)}$$

条件概率指在A事件发生的情况下，B事件发生的概率。

"A事件发生的情况下" 代表A为样本空间，"B事件发生的概率",代表$A\cap B$为事件。因此$P(B|A)=\frac{P(A\cap B)}{P(A)}$。

做一下公式变形$P(A\cap B)=P(B|A)P(A)$

### 贝叶斯定理




假设某一样本空间S=A+B+C，且A，B，C都有一定概率发生D。M发生在D内，则M是由A发生的概率是多少？

可见 $P(D)=P(D\cap A)+P(D\cap B)+P(D\cap C)$ ，由条件概率的公式也可以写成$P(D)=P(D|A)P(A)+P(D|B)P(B)+P(D|C)P(C)$

算出来的结果就是事件D在样本空间S下发生的概率。那么M发生在A的概率为

$P(A|D)=\frac{P(A\cap D)}{P(D)}=\frac{P(D|A)P(A)}{P(D|A)P(A)+P(D|B)P(B)+P(D|C)P(C)}$ ，这就是贝叶斯公式

### 朴素贝叶斯算法

#### 问题引入

由上面的计算中我们知道贝叶斯的公式为$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$ ，不妨换个表达形式：$P(\text{类别}|\text{特征})=\frac{P(\text{特征}|\text{类别})\cdot P(\text{类别})}{P(\text{特征})}$

给定数据如下：

| 帅？ | 性格好？ | 身高？ | 上进？ | 嫁与否 |
|---|---|---|---|---|
| 帅 | 不好 | 矮 | 不上进 | 不嫁 |
| 不帅 | 好 | 矮 | 上进 | 不嫁 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 帅 | 好 | 高 | 不上进 | 嫁 |

现有一对情侣，男生特点分别是不帅，性格不好，身高矮，不上进，请判断女生是否嫁？

这是一个典型的分类问题，转为数学问题就是比较p(嫁|(不帅、性格不好、身高矮、不上进))与p(不嫁|(不帅、性格不好、身高矮、不上进))的概率哪个更大，就能给出嫁或者不嫁的答案

#### 朴素贝叶斯

联想到朴素贝叶斯公式：$P(\text{嫁}|\text{不帅,性格不好,身高矮,不上进})=\frac{P(\text{不帅,性格不好,身高矮,不上进}|\text{嫁})\cdot P(\text{嫁})}{P(\text{不帅,性格不好,身高矮,不上进})}$

我们需要求p(嫁|(不帅、性格不好、身高矮、不上进)，这是我们不知道的，但是通过朴素贝叶斯公式可以转化为好求的三个量，p(不帅、性格不好、身高矮、不上进|嫁)、p（不帅、性格不好、身高矮、不上进）、p(嫁)，就可将待求的量转化为其它可求的值，这就相当于解决了这个问题

所谓朴素贝叶斯，就假设各个特征之间相互独立，那么这个等式就成立了：p(不帅、性格不好、身高矮、不上进|嫁) = p(不帅|嫁)*p(性格不好|嫁)*p(身高矮|嫁)*p(不上进|嫁)

#### 特征之间相互独立

1）若没有这个假设，我们对右边这些概率的估计其实是不可做的。例子中共有4个特征，这四个特征的联合概率分布总共是4维空间。然而现实中往往会有更多的特征，每一个特征的取值也是非常之多，通过统计来估计后面概率的值几乎不可做，这也是为什么需要假设特征之间独立的原因。

2）若没有假设特征之间相互独立，那么我们统计的时候就需要在整个特征空间中去找：在嫁的条件下，去找四种特征全满足分别是不帅，性格不好，身高矮，不上进的人的个数。由于数据的稀疏性，很容易统计到0的情况，而这样是不合适的。

综上，朴素贝叶斯法对条件概率分布做了条件独立性的假设，由于这是一个较强的假设，朴素贝叶斯也由此得名。这一假设使得朴素贝叶斯法变得简单，但有时会牺牲一定的分类准确率。

#### 公式整理

此时可以将上面公式整理为：$P=\frac{P(\text{不帅}|\text{嫁})\cdot P(\text{性格不好}|\text{嫁})\cdot P(\text{身高矮}|\text{嫁})\cdot P(\text{不上进}|\text{嫁})\cdot P(\text{嫁})}{P(\text{不帅})\cdot P(\text{性格不好})\cdot P(\text{身高矮})\cdot P(\text{不上进})}$

由于在数据量很大的时，根据中心极限定理，频率是等于概率的，所以在数据集中，分别统计出p(嫁)=? p(不帅|嫁)=? p(性格不好|嫁)= ？,……再求出分母p(不帅)，p（性格不好），p（矮），p（不上进）。这样，要求p(不帅、性格不好、身高矮、不上进|嫁)的所需项全部求出来了，再带入进p即可算出p(嫁|(不帅、性格不好、身高矮、不上进)，最后用同理计算得到p(不嫁|(不帅、性格不好、身高矮、不上进)，比较两者的大小，就可以知道到底是嫁还是不嫁

## 贝叶斯决策论

### 描述

对分类任务来说，在所有相关概率都已知的理想情况下，贝叶斯决策论考虑如何基于这些概率和误判来选择最优的类别标记

### 基本原理

在上述问题中，$y=\{\text{嫁},\text{不嫁}\}$

假设有N个类别标记，即$\gamma=\{c_1,c_2,\cdots c_N\}$，$\lambda_{ij}$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$P(c_i|x)$可获得将样本x分类为ci所产生的期望损失，即在样本x上的条件风险$R(c_i|x)=\sum_{j=1}^{N}\lambda_{ij}P(c_j|x)$

我们的任务是寻找一个判定准则$h:\mathcal{X}\mapsto\mathcal{Y}$以最小化总体风险$R(h)=E_x[R(h(x)|x)]$。显然，对每个样本x，若h能最小化条件风险，则总体风险也将被最小化。通过输入特征值，判断为嫁或不嫁 {if $P(\text{嫁}|x)>P(\text{不嫁}|x)$: return 嫁 else: return 不嫁}

这就产生了贝叶斯判定准则：为最小化总体风险，只需在每个样本上选择那个使条件风险$R(c|x)$最小的类别标记，即$h^*(x)=\underset{c\in\mathcal{Y}}{\operatorname{argmin}}R(c|x)$，此时h*称为贝叶斯最优分类器，与之对应的总体风险R(h*)称为贝叶斯风险。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。

（若$P(\text{不嫁}|x)=0.6>P(\text{嫁}|x)$，机器判断为不嫁。但仍有0.4的概率为嫁。）

### 公式转换

具体地，若目标是最小化分类器错误率，则误判损失λij可写为$\lambda_{ij}=\begin{cases}0, & \text{if } i=j\\ 1, & \text{else}\end{cases}$，此时条件风险$R(c|x)=1-P(c|x)$

于是，最小化分类错误率的贝叶斯最优分类器为$h^*(x)=\underset{c\in\mathcal{Y}}{\operatorname{argmax}}P(c|x)$，即对每个x，选择能使后验概率P(c|x)最大的类别标记。基于贝叶斯定理，P(c|x)可写为$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$。其中，P(c)是类先验概率，P(x|c)是样本x相对于类标记c的类条件概率，或称“似然”；P(x)是用于归一化的“证据”因子。

基于先验信息，对事件发生概率进行估计，不考虑任何新的信息，$\Rightarrow P(A)$

后验概率：若观测到新现象可证据后，对事件发生概率进行修正或更新的概率 $\Rightarrow P(A|B)$

# 极大似然估计

## 公式转换

于是，最小化分类错误率的贝叶斯最优分类器为 $h^*(x)=\underset{c\in\mathcal{Y}}{\arg\max}\, P(c|x)$，即对每个x，选择能使后验概率P(c|x)最大的类别标记。基于贝叶斯定理，P(c|x)可写为 $P(c|x)=\frac{P(x|c)P(c)}{P(x)}$。其中，P(c)是类先验概率，P(x|c)是样本x相对于类标记c的类条件概率，或称“似然”；P(x)是用于归一化的“证据”因子。

对给定样本x，证据因子P(x)与类标记无关，因此估计P(c|x)的问题就转化为如何基于训练数据D来估计先验P(c)和似然P(x|c)。

根据大数定律，当训练集包含充足的独立分布样本使，P(c)可通过各类样本出现的频率来进行估计。

对类条件概率P(x|c)来说，由于它涉及关于x所有属性的联合概率，直接根据样本出现的频率来估计将会遇到严重的困难，直接使用频率来估计P(x|c)显然不行，因为“未被观测到”和“出现概率为零”通常是不同的。

## 解决问题

### 似然

假设现有一枚特殊的硬币，每次抛出后得到是花或字的概率可能并不为0.5。在这种情况下，通过抛硬币的方式，如果抛100次硬币的结果都是花，则可认为这枚硬币的两面都是花的可能性最大。这种通过事实反过来猜测硬币的情况就是似然。通过事实推断出最有可能的硬币情况，就是极大似然估计。

### 似然和概率

已知硬币具有花和字两面，推测抛硬币的各种情况的可能性，称为概率。

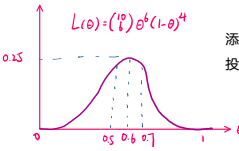
如果对硬币的两面是什么并不清楚，需要通过抛硬币的情况来确定，称为似然。

假设在一次实验中，10次抛硬币，有6次是花，则硬币抛出后是花的概率(以下简称为硬币的参数)有多大?

所谓极大似然估计，就是假设硬币的参数，然后计算实验结果的概率是多少。通过比对设置不同参数下得到的概率，概率越大的越可能是正确的。

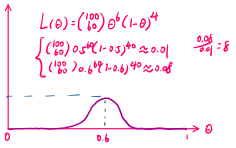
用0.5作为参数，概率为 $\binom{10}{6}0.5^6(1-0.5)^4\approx0.21$，再用0.6作参数: $\binom{10}{6}0.6^6(1-0.6)^4\approx0.25$，可见0.25>0.21，$\frac{0.25}{0.21}\approx1.2$ 可以认为0.6作为参数的可能性是0.5作为参数的可能性的1.2倍。

设置硬币的参数为θ，可以得到似然函数:




添加更多的实验结果:

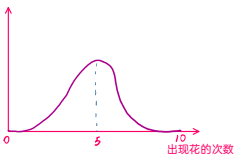
投掷100次，出现了60次花



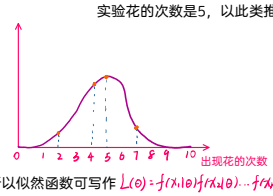

此时，0.6作为参数的可能性是0.5作为参数的可能性的8倍，新的实验结果更加支持0.6这个参数。

图像有明显的缩窄，可理解为可选参数的分布更集中了。通过投入更多的是实验结果，参数会越来越明确。

此时如果打开上帝视角，可以观察到抛出后是花的概率即硬币的参数其实是θ=0.5，所以这里绘制出抛10次硬币后出现花的次数的概率分布图




但没有上帝视角加持的我们事先并不知道硬币的参数，只能通过实验的方式来推断。这里进行了6次实验后（一次实验抛10次硬币，统计花的次数），实验结果为[4,5,5,2,7,4]，表示第一次实验花的次数是4，第二次实验花的次数是5，以此类推，可在图中画出对应的点




我们用x1,x2,⋯,xn表示每次实验结果，因为每次实验都是独立的，所以似然函数可写作 $L(\theta)=f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$，其中 $f(x_n|\theta)$ 表示在同一个参数下的结果。

随着试验次数的增加，θ的值会逐渐逼近真实值。但由于实验本身具有二项随机性，可能会导致它与真实值存在细微的误差。

### 极大似然估计 Maximum Likelihood Estimation

记关于类别c的类条件概率为P(x|c)，假设P(x|c)具有确定的形式并且被参数向量θc唯一确定，则我们的任务就是利用训练集D估计参数θc。这个概率模型训练的过程就是参数估计过程，通过极大似然估计法解决问题。

令Dc表示训练集D中第c类样本组成的集合，假设这些样本是独立同分布的，则参数θc对于数据集Dc的似然是: $P(D_c|\theta_c)=\prod_{x\in D_c}P(x|\theta_c)$

对θc进行极大似然估计，就是去寻找能最大化似然P(Dc|θc)的参数 $\hat{\theta}_c$。直观上看，极大似然估计是试图从θc所有可能的取值中，找到一个能使数据出现的“可能性”最大的值。为了防止上式连乘操作造成下溢，通常使用对数似然: $LL(\theta_c)=\log P(D_c|\theta_c)=\sum_{x\in D_c}\log P(x|\theta_c)$

此时参数θc的极大似然估计 $\hat{\theta}_c$ 为 $\hat{\theta}_c=\underset{\theta_c}{\arg\max}\,LL(\theta_c)$。

例如，在连续属性情形下，假设概率密度函数 $p(x|c)\sim\mathcal{N}(\mu_c,\sigma_c^2)$，则参数 $\mu_c$ 和 $\sigma_c^2$ 的极大似然估计为 $\hat{\mu}_c=\frac{1}{|D_c|}\sum_{x\in D_c}x$，$\hat{\sigma}_c^2=\frac{1}{|D_c|}\sum_{x\in D_c}(x-\hat{\mu}_c)(x-\hat{\mu}_c)^T$

也就是说，通过极大似然估计法得到的正态分布均值就是样本均值，方差就是 $(x-\hat{\mu}_c)(x-\hat{\mu}_c)^T$ 的均值，这显然是一个符合直觉的结果。在离散属性情形下，也可通过类似的方式估计类条件概率。

应注意，由于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，所以在现实应用中，还需要利用关于应用任务本身的经验知识，否则若仅凭“猜测”来假设概率分布形式，很可能产生误导性的结果。

# 朴素贝叶斯

## 公式

### 属性条件假设

基于属性条件独立性假设P(c|x)可重写为

$$P(c|\boldsymbol{x})=\frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})}=\frac{P(c)}{P(\boldsymbol{x})}\prod_{i=1}^{d}P(x_i|c)$$

### 朴素贝叶斯分类器表达式

其中d为属性数目，xi为x在第i个属性上的取值

由于对所有类别来说P(x)相同，因此可将贝叶斯判定准则改写为

$$h_{nb}(\boldsymbol{x})=arg_{c\in\mathcal{Y}}\,maxP(c)\prod_{i=1}^{d}P(x_i|c)$$

### 求解关键

显然，朴素贝叶斯分类器的训练过程就是基于训练集D来估计类先验概率P(c)，并为每个属性估计条件概率P(xi|c)。

令Dc表示训练集D中第c类样本组成的集合，若有充足的独立同分布样本，则可容易地估计先验概率 $P(c)=\frac{|D_c|}{|D|}$

则条件概率P(xi|c)可估计为 $P(x_i|c)=\frac{|D_{c,x_i}|}{|D_c|}$ 〔对离散属性〕

对连续属性可考虑概率密度函数，假定 $p(x_i|c)\sim\mathcal{N}(\mu_{c,i},\sigma_{c,i}^2)$，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 $p(x_i|c)=\frac{1}{\sqrt{2\pi}\sigma_{c,i}}\exp\left(-\frac{(x_i-\mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$

西瓜数据集（1-8号为好瓜，9-17号为坏瓜）

| 编号 | 色泽 | 根蒂 | 敲声 | 纹理 | 脐部 | 触感 | 密度 | 含糖率 | 好瓜 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 青绿 | 蜷缩 | 浊响 | 清晰 | 凹陷 | 硬滑 | 0.697 | 0.460 | 是 |
| 2 | 乌黑 | … | … | … | … | … | … | … | … |
| 17 | 浅白 | … | … | … | … | … | … | … | 否 |

目标：根据属性判断是否是好瓜

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

# 斯分类器

例：西瓜数据集

目标：根据属性判断是否是好瓜

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

注意：共属性值为连续值

P(c)
- P(好瓜=是) = 8/17 ≈ 0.471
- P(好瓜=否) = 9/17 ≈ 0.529

P(xi|c)

离散属性：为每个属性估计条件概率 $P(x_i|c)$
- P青绿|是=P(色泽=青绿|好瓜=是)= 3/8 = 0.375
- P青绿|否=P(色泽=青绿|好瓜=否)= 3/9 ≈ 0.333
- P蜷缩|是=P(根蒂=蜷缩|好瓜=是)= 5/8 = 0.625
- P蜷缩|否=P(根蒂=蜷缩|好瓜=否)= 3/9 ≈ 0.333
- ……
- P硬滑|否=P(触感=硬滑|好瓜=否)= 6/9 ≈ 0.667

连续属性：
- p密度:0.697|是=p(密度=0.697|好瓜=是) = 1/(√(2π)×0.129) exp(−(0.697−0.574)²/(2×0.129²)) ≈ 1.959（均值、正态分布、方差、概率密度）
- p密度:0.697|否=p(密度=0.697|好瓜=否) = 1/(√(2π)×0.195) exp(−(0.697−0.496)²/(2×0.195²)) ≈ 1.203
- ……
- p含糖:0.460|否=p(含糖率=0.460|好瓜=否) = 1/(√(2π)×0.101) exp(−(0.697−0.154)²/(2×0.101²)) ≈ 0.066

P(c)*P(xi|c)
- 于是就有
- P(好瓜=是)*P青绿|是*P蜷缩|是*……*p含糖：0.460|是≈0.038
- P(好瓜=否)*P青绿|否*P蜷缩|否*……*p含糖：0.460|否0≈6.80*10^-5

得出结论
- 由于0.038>6.80*10^-5，因此朴素贝叶斯分类器
- 将测试样本Test判别为“好瓜”

# 拉普拉斯修正

解决问题：需注意，若某个属性值在训练集中没有与某个类同时出现过，则直接基于上式进行概率估计，再进行判别将出现问题。例如，在使用西瓜数据集训练朴素贝叶斯分类器时，对一个“敲声=清脆”的测试例，有P清脆|是=P(敲声=清脆|好瓜=是)=0/8=0，由于练乘式计算出的概率值为0，因此无论该样本的其他属性是什么，哪怕在其他属性上明显像好瓜，分类的结果都将是“好瓜=否”，这显然不太合理。

为了避免其他属性携带的星系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“平滑”，常用“拉普拉斯修正”。具体来说，令N表示训练集D中可能的类别数，Ni表示第i个属性可能的取值数，则P(c)，P(x|c)分别修正为 $\hat{P}(c)=\frac{|D_c|+1}{|D|+N}$，$\hat{P}(x_i|c)=\frac{|D_{c,x_i}|+1}{|D_c|+N_i}$

公式转换：此时类先验概率可估计为 $\hat{P}(\text{好瓜}=\text{是})=\frac{8+1}{17+2}\approx 0.474$，$\hat{P}(\text{好瓜}=\text{否})=\frac{9+1}{17+2}\approx 0.526$
类似地，P青绿|是和P青绿|否可估计为 $\hat{P}(\text{青绿}|\text{好瓜}=\text{是})=\frac{3+1}{8+3}\approx 0.364$，$\hat{P}(\text{青绿}|\text{好瓜}=\text{否})=\frac{3+1}{9+3}\approx 0.333$
同时上文提到的概率P清脆|是可估计为 $\hat{P}(\text{敲声}=\text{清脆}|\text{好瓜}=\text{是})=\frac{0+1}{8+3}\approx 0.091$
显然，拉普拉斯修正避免了因训练集样本不充分而导致概率估计为零的问题，并且在训练集变大时，修正过程所引入的先验的影响也会逐渐变得可忽略，是的估计值趋向于实际概率值

说明：拉普拉斯修正实质上假设了属性值与类别均匀分布，这是在朴素贝叶斯学习过程中额外引入的关于数据的先验

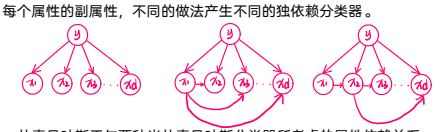
# 半朴素贝叶斯分类器

解决问题：虽然朴素贝叶斯公式采用了属性条件独立性假设来降低贝叶斯公式中估计后验概率的困难，但在现实任务中这个假设往往很难成立。于是，人们尝试对属性条件独立性假设进行一定程度的放松，由此产生了一类称为“半朴素贝叶斯分类器”的学习方法。

*基本思想*：半朴素贝叶斯分类器的基本思想是适当考虑一部分属性间的相互依赖信息，从而既不需及进行完全联合概率计算，又不至于彻底忽略了比较强的属性依赖关系。

独依赖估计：独依赖估计是半朴素贝叶斯分类器最常用的一种策略，就是假设每个属性在类别之外最多仅依赖于一个其它属性，即 $P(c|\boldsymbol{x}) \propto P(c)\prod_{i=1}^{d} P(x_i|c,pa_i)$

其中pai为属性xi所依赖的属性，称为xi的父属性。此时，对每个属性xi若其父属性已知，则可采用拉普拉斯修正的方法来估计概率值P(xi|c,pai)。于是，问题的关键就转化为如何确定每个属性的副属性，不同的做法产生不同的独依赖分类器。



朴素贝叶斯于与两种半朴素贝叶斯分类器所考虑的属性依赖关系

问题引入：继续以上例中的硬币为例，假设现在我们拥有两枚这样特殊的硬币，每次抛掷后得到是花或字的概率可能并不为0.5。记这两枚硬币（编号1和2）抛掷后是花的概率分别为P1,P2。为了估计这两个概率，现做实验，每次取其中一枚硬币，连掷5下并记录下结果。

| 硬币 | 结果 | 统计 |
|---|---|---|
| 1 | 花花字花字 | 3花-2字 |
| 2 | 字字花花字 | 2花-3字 |
| 1 | 花字字字字 | 1花-4字 |
| 2 | 花字字花花 | 3花-2字 |
| 1 | 字花花字字 | 2花-3字 |

可以很容易地估计出P1和P2：

P1=(3+1+2)/15=0.4（硬币1出现花的次数/硬币1总抛掷次数）
P2=(2+3)/10=0.5

加入隐变量z：在上面的问题中，抹去每轮投掷时使用的硬币标记，如下：

| 硬币 | 结果 | 统计 |
|---|---|---|
| Nan | 花花字花字 | 3花-2字 |
| Nan | 字字花花字 | 2花-3字 |
| Nan | 花字字字字 | 1花-4字 |
| Nan | 花字字花花 | 3花-2字 |
| Nan | 字花花字字 | 2花-3字 |

现在目标仍未变，还是估计P1和P2，该如何实现？

显然此时我们多了一个隐变量z，可以把它认为是一个5维的向量(z1,z2,z3,z4,z5)，代表每次投掷时所使用的硬币，比如z1代表第一轮使用的硬币是1还是2。但是这个变量z并不知道，就无法去估计P1,P2，因此我们必须先估计出z，才能进一步估计P1和P2

# EM算法

## 加入隐变量z

| | | | |
|---|---|---|---|
| Nan | 字字花花字 | 2花-3字 | 显然此时我们多了一个隐变量z，可以把它认为是一个5维的向量 |
| Nan | 花字字字字 | 1花-4字 | (z1,z2,z3,z4,z5)，代表每次投掷时所使用的硬币，比如z1代表第一轮使用的硬 |
| Nan | 花字字花花 | 3花-2字 | 币是1还是2。但是这个变量z并不知道，就无法去估计P1,P2，因此我们必须先 |
| Nan | 字花花字字 | 2花-3字 | 估计出z，才能进一步估计P1和P2 |

## 解决思路

要估计z，我们又得知道P1和P2，这样我们才能用最大似然概率法则去估计z，但是我们的目标就是估计P1和P2，现在又需要让它作为条件，陷入了循环中。

突破问题的关键，就是先随机初始化一个P1和P2，用它来估计z，然后基于z，还是按照最大似然反过来去估计新的P1和P2。

当新的P1和P2与我们的初始值一样时，说明了我们初始的P1和P2很可能是真实的值；当新估计出来的P1与P2与初始化差别很大时，继续用新的P1,P2迭代，直至收敛。

## EM初级版

不妨给P1，P2赋初始值：P1=0.2,P2=0.7，看看第一轮最可能是哪个硬币：

1）如果是硬币1，得出“花花字花字”的概率为0.2*0.2*0.2*0.8*0.8=0.00512

2）如果是硬币2，得出“花花字花字”的概率为0.7*0.7*0.7*0.3*0.3=0.03087

一次求出其它4轮中的相应概率，做成表格如下：按照极大似然法则：

| 轮数 | 若是硬币1 | 若是硬币2 | |
|---|---|---|---|
| 1 | 0.00512 | 0.03087 | 第1轮中最有可能的是硬币2 |
| 2 | 0.02048 | 0.01323 | 第2轮中最有可能的是硬币1 |
| 3 | 0.08192 | 0.00567 | 第3轮中最有可能的是硬币1 |
| 4 | 0.00512 | 0.03087 | 第4轮中最有可能的是硬币2 |
| 5 | 0.02048 | 0.01323 | 第5轮中最有可能的是硬币1 |

把上面的值作为z的估计值，然后按照极大似然概率法则来估计新的P1，P2:

P1=(2+1+2)/15=0.33,P2=(3+3)/10=0.6

现在打开上帝视角，可以知道P1，P2的极大似然估计其实就是上面例子中的0.4和0.5（下文中将这两个值称为P1,P2的真实值）。对比我们初始化的P1,P2和新估计出的P1,P2:

| 初始化的P1 | 估计出的P1 | 真实的P1 | 初始化的P2 | 估计出的P2 | 真实的P2 |
|---|---|---|---|---|---|
| 0.2 | 0.33 | 0.4 | 0.7 | 0.6 | 0.5 |

不难看出，我们估计的P1,P2相比于它们的初始值，更接近它们的真实值了。可以期待，我们继续按照上面的思路，用估计出的P1和P2再来估计z，再用z来估计新的P1和P2，反复迭代下去，就可以最终得到P1=0.4，P2=0.5。此时无论怎样迭代，P1和P2的值都会保持0.4和0.5不变，这样我们就找到了P1和P2的极大似然估计。

## 注意事项

1）数学证明，新估计出的P1和P2一定会更接近真实的P1和P2

2）迭代不一定会收敛于真实的P1,P2，这取决于P1,P2的初始值

## EM进阶版

我们用极大似然估计法则估计出z的值，然后再用z值按照极大似然故概率估计法则估计出新的P1和P2。也就是说，我们使用了一个最可能的z值，而不是所有可能的z值。

如果考虑所有可能的z值，对每一个z值都估计出一个新的P1和P2，将每一个z值概率大小作为权重，将所有新的P1和P2分别加权相加，这样的P1和P2应该会更好。

在本例中，z值共有2^5=32种，而我们实际上并不需要算这么多次，只需要用期望来简化运算。

| 轮数 | 若是硬币1 | 若是硬币2 |
|---|---|---|
| 1 | 0.00512 | 0.03087 |
| 2 | 0.02048 | 0.01323 |
| 3 | 0.08192 | 0.00567 |
| 4 | 0.00512 | 0.03087 |
| 5 | 0.02048 | 0.01323 |

利用这张表，我们可以计算出每轮投掷种使用硬币1或硬币2的概率。比如第1轮，使用硬币1的概率是：0.00512/(0.00512+0.03087)=0.14,使用硬币2的概率是1-0.14=0.86。可以依次计算出其它4轮的概率：

| 轮数 | zi=硬币1 | zi=硬币2 |
|---|---|---|
| 1 | 0.14 | 0.86 |
| 2 | 0.61 | 0.39 |
| 3 | 0.94 | 0.06 |
| 4 | 0.14 | 0.86 |
| 5 | 0.61 | 0.39 |

该表中的左右两列表示期望值。第一行当中，0.86表示从期望的角度看，这轮抛掷使用硬币2的概率是0.86。相比于前面的方法，这种说法更加严谨。这样我们在估计P1或P2时，就可以用上全部的数据，而不是部分的数据。

在这一步种，我们实际上是估计出了z的概率分布，这步被称作E步。

结合下表，我们按照期望最大似然估计概率的法则来估计新的P1和P2:

| 硬币 | 结果 | 统计 |
|---|---|---|
| Nan | 花花字花字 | 3花-2字 |
| Nan | 字字花花字 | 2花-3字 |
| Nan | 花字字字字 | 1花-4字 |
| Nan | 花字字花花 | 3花-2字 |
| Nan | 字花花字字 | 2花-3字 |

以P1估计为例，这里我们要做的是在之前推算出来的使用硬币1的概率下，去计算出现在这种抛掷情况下的可能性，就是计算一下期望看看有多少次等价的硬币1抛出来为为正，多少次抛出来为负。以第一轮为例，硬币1在第一轮中被使用的概率是0.14，现在出现了3花2字的情况，相当于：0.14*3=0.42花，0.14*2=0.28字。

依次算出其它四轮，列表如下：(由掷硬币，抛硬币果次说的是期次数)

| 轮数 | 花 | 字 |
|---|---|---|
| 1 | 0.42 | 0.28 |
| 2 | 1.22 | 1.83 |
| 3 | 0.94 | 3.76 |
| 4 | 0.42 | 0.28 |
| 5 | 1.22 | 1.83 |
| 总计 | 4.22 | 7.98 |

计算P1=4.22/(4.22+7.98)=0.35

可以看到，改变了z值的估计方法后，新估计出的P1更加接近真实值0.4。原因就是我们使用了所有抛掷的数据，而不再是之前只使用了部分的数据。

这步中，我们根据E步中求出的z的概率分布，根据极大似然概率法则估计P1和P2，被称作M步。

## EM算法

$Expectation-Maximization$

在上述例子中，我们抹去了硬币的标记，这种“未观测变量”叫做“隐变量”。令X表示已观测变量集，Z表示隐变量集，Θ表示模型参数。若对Θ做极大似然估计，则应最大化对数似然：$\mathrm{LL}(\Theta|X,Z)=\ln \mathrm{P}(X,Z|\Theta)$

然而由于Z是隐变量，上式无法直接求解，此时我们可通过对Z计算期望，来最大化已观测数据的对数“边际似然”：$\mathrm{LL}(\Theta|X)=\ln \mathrm{P}(X|\Theta)=\ln \sum_{Z} P(X,Z|\Theta)$

EM算法是常用的估计参数隐变量的利器，其基本思想是：若参数Θ已知，则可根据训练数据推出最优隐变量Z的值(E步)；反之，若Z的值已知，则可方便地对参数Θ做极大似然估计(M步)。

在“EM初级版”中，其实就是以初始值$\Theta^0$为起点，对LL(Θ|X)，可迭代执行以下步骤直至收敛：

1）基于Θt判断隐变量Z的期望，记为Zt

2）基于已观测变量X和Zt对参数Θ做极大似然估计，记为Θt+1

进一步，在“EM进阶版”当中，我们不是取Z的期望，而是基于Θt计算隐变量Z的概率分布P(Z|X,Θt)，则EM算法的两个步骤是：

1）以当前参数Θt推断隐变量分布P(Z|X,Θt)，并计算对数似然LL(Θ|X,Z)关于Z的期望：$\mathrm{Q}(\theta|\theta^t)=\mathrm{E}_{Z|X,\theta^t}\mathrm{LL}(\Theta|X,Z)$

2）寻找参数最大化期望似然，即 $\Theta^{t+1}=\arg\max_{\theta} Q(\Theta|\theta^t)$

EM算法使用两个步骤交替计算，新得到的参数值重新被用于E步，直至收敛到局部最优解。