贝叶斯分类器

贝叶斯定理

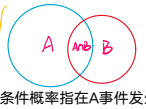
概率空间三要素

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



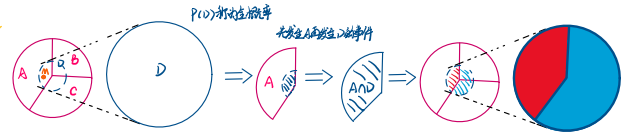
$P(B|A)=\frac{P(A\cap B)}{P(A)}$

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

由上面的计算中我们知道贝叶斯的公式为 $P(B|A)=\frac{P(A|B)P(B)}{P(A)}$ ，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 $Y=\{c_1,c_2,\cdots,c_N\}$，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

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 $h:X\mapsto Y$ 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，判断的嫁或不嫁 $\begin{cases}\text{if } P(\text{嫁}|x)>P(\text{不嫁}|x): \text{return 嫁}\\ \text{else: return 不嫁}\end{cases}$

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 $h^*(x)=\underset{c\in Y}{\operatorname{argmin}}R(c|x)$，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的概率是嫁）

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 $h^*(x)=\underset{c\in Y}{\operatorname{argmax}}P(c|x)$，即对每个x，选择能使后验概率P(c|x)最大的类别标

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。

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 $h^*(x)=\underset{c\in\mathcal{Y}}{\arg\max}\, P(c|x)$，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。

极大似然估计

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1/2。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，就是最大似然估计。

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。

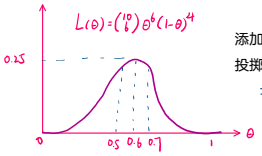
解决问题

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。

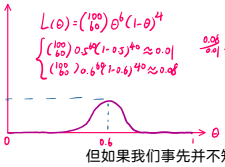
用0.5作为参数，概率为 $\binom{10}{6}0.5^6(1-0.5)^4\approx0.21$，再用0.6作参数：$\binom{10}{6}0.6^6(1-0.6)^4\approx0.25$，可见0.25>0.21，$\frac{0.25}{0.21}\approx1.2$ 可以认为0.6作为参数的可能性是0.5作为参数的可能性的1.2倍。

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

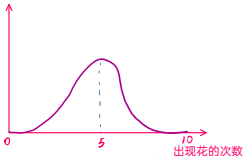


$\frac{0.08}{0.01}=8$

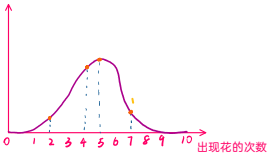
此时，0.6作为参数的可能性是0.5作为参数的可能性的8倍，新的实验结果更加支持0.6这个参数。

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。

实际上，如果这是一枚普通的硬币，我们知道，抛出后是花的概率即硬币的参数θ=0.5，这里绘制出抛10次硬币后出现花的次数的概率分布图



但如果我们事先并不知道它是否是普通的硬币，只能通过实验的方式来推断。我们进行了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 $\{4,5,5,3,2,7,4\}$，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 μ_c 和 σ_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

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

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

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

$$h_{nb}(x)=\underset{c\in\mathcal{Y}}{\arg\max}\, P(c)\prod_{i=1}^{d}P(x_i|c)$$

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 $P(x_i|c)=\frac{|D_{c,x_i}|}{|D_c|}$ [对于离散属性]

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	…	…	…	…	…	…	…	…
⋮									

朴素贝叶斯分类器

例：西瓜数据集

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	?

注意：共属性值为连续值

P(c)
- $P(好瓜=是)=\frac{8}{17}\approx 0.471$
- $P(好瓜=否)=\frac{9}{17}\approx 0.529$

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)=$\frac{3}{8}=0.375$
- P青绿|否=P(色泽=青绿|好瓜=否)=$\frac{3}{9}\approx 0.333$
- P蜷缩|是=P(根蒂=蜷缩|好瓜=是)=$\frac{5}{8}=0.625$
- P蜷缩|否=P(根蒂=蜷缩|好瓜=否)=$\frac{3}{9}\approx 0.333$
- ……
- P硬滑|否=P(触感=硬滑|好瓜=否)=$\frac{6}{9}\approx 0.667$

连续属性：均值，正态分布，标准差/方差

- p密度:0.697|是=p(密度=0.697|好瓜=是)
  $=\frac{1}{\sqrt{2\pi}\times 0.129}\exp\left(-\frac{(0.697-0.574)^2}{2\times 0.129^2}\right)\approx 1.959$
- p密度:0.697|否=p(密度=0.697|好瓜=否)
  $=\frac{1}{\sqrt{2\pi}\times 0.195}\exp\left(-\frac{(0.697-0.496)^2}{2\times 0.195^2}\right)\approx 1.203$
- ……
- p含糖:0.460|否=p(含糖率=0.460|好瓜=否)
  $=\frac{1}{\sqrt{2\pi}\times 0.101}\exp\left(-\frac{(0.697-0.154)^2}{2\times 0.101^2}\right)\approx 0.066$

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”