# MLAPP 读书笔记 - 03 离散数据的生成模型(Generative models for discrete data)

> A Chinese Notes of MLAPP，MLAPP 中文笔记项目
> https://zhuanlan.zhihu.com/python-kivy

记笔记的人：cycleuser

2018年05月13日14:40:28

## 3.1 简介

在本书2.2.3.2中,提到了如何对一个特征向量 x 进行分类,使用了贝叶斯规则,构建了一个生成分类器(generative classifier),形式如下所示:

$$p(y=c|x,\theta) \propto p(x|y=c,\theta)p(y=c|\theta) \quad (3.1)$$

使用这样模型的管建就是要给类条件概率密度(class-conditional density)指定一个合适的形式$p(x|y=c,\theta)$,这定义了我们在每个类别中希望看到的数据种类.在本章中,观测数据是离散符号(discrete symbols).此外还要讨论如何推导这类模型的未知参数$\theta$.

## 3.2 贝叶斯概念学习(Bayesian concept learning)

想一下小朋友学习理解一个单词的意思,比如狗 dog 这个单词.可能家大人会指着一个这个名词指代概念的个体,然后告诉这东西是啥,比如说:"哎呀这是一只小狗啊"或者"小心有狗啊"之类的.不过他们往往都是给出正面例子,而不是给出反例,比如说:"看看这不是小狗啊".当然,在动态的学习过程中,反例是可能出现的,比如小朋友说:"哎呀这是小狗",然后家长更正说:"你个瓜娃这是小猫不是狗".不过心理学的研究表明,人类可以单纯从正面样例(positive examples)来学习概念(Xu and Tenenbaum 2007).

先把这种学习单词意思的学习过程就等价认为是概念学习(concept learning), 这样也就等价于二值化分类(binary classification).设如果 x 是概念 C 的一个具体实例,则$f(x)=1$,反之则$f(x)=0$.那么目标就是通过学习得到指示函数(indicator function) $f$,这个函数就是用来判断元素是否包含于集合 C 内的. 对$f$ 的定义可以允许存在不确定性,或者对 C 中的元素也可以有不确定性,这样就能用标准概率积分模拟模糊集合理论(fuzzy set theory).标准的二值化分类器方法需要正负两方面的样本.这次咱们要设计一个只需要正面样本的学习方法.

处于教学目的,这部分要参考 Josh Tenenbaum 1999年的博士论文中的内容. 这次样例是一个数字游戏,是概念学习的一个简单例子.这个游戏的内容如下,一个人甲选择一个简单算术概念组成的集合 C, 比如可以使素数或者是1到10之间的数字.然后给另一个人乙一个随机选择的正面样本系列$D = \{x_1, \ldots, x_N\}$,D 是从 C 中选取的,然后问乙新的测试样本$\hat{x}$是否属于 C, 也就是甲让乙去对$\hat{x}$进行分类.

此处参考原书图3.1

为了简单起见,就让所有数字都是1到100之间的整数.然后甲告诉乙16是这个概念的正面样本.那么有什么其他的数值你觉得也是正面样本呢? 17? 6? 32? 99? 只有一个样本分明很难对不对,所以预测起来也不靠谱,太难了.肯定得想和16更相似的数.比如17有点像,因为距离近,6也有所相似因为有一个同样的数字,32也相似,因为是偶数而且还是而被关系,不过99看着就不太像了.这样有的数字看上去比其他的看着更可能.这可以用一个概率分布$p(\hat{x} \mid D)$, 这个概率分布是数据集 D 当中任意$\hat{x} \in \{1, \ldots, 100\}$ 使得$\hat{x} \in C$的概率.这也叫做后验预测分布(posterior predictive distribution).图3.1展示的就是实验室中推出的一例预测分布.可以看到人们对于16相似的数的选择很相似,有某种程度的相似性.

然后甲告诉乙8,2,64也都是正面样本.这时候可能就会猜测隐藏概念是2的幂.这就是归纳法(induction) 的一个例子.基于这个假设,预测分布就很明确了,主要都集中在2的幂数上.

如果甲告诉乙数据集 $D = \{16, 23, 19, 20\}$就会得到不同的泛化梯度(generalization gradient),如图3.1底部所示.
在机器学习里面如何解释和模拟这种行为呢?
传统归纳方法是假设有一个概念假设空间(hypothesis space of concepts),$H$, 比如奇数/偶数/1-100之间的数/2的幂数/以 j 结尾的某个数等等.$H$的子集中与数据 D 一致的就被称为版本空间(version space).随着样本增多,版本空间缩小,对概念的确定性就随之增加(Mitchell 1997).
然而只有版本空间还不够.在看到了$D = \{16\}$之后,有很多都一直规则,怎么来结合起来去预测其他元素是否属于 C 即$\tilde{x} \in C$呢?另外在看到$D = \{16, 8, 2, 64\}$之后,为什么你就认为规则是2的幂数而不是所有偶数呢,或者也不是"除了32之外的所有2的幂数"呢?这几个都符合啊.接下来我们就用贝叶斯观点来对此进行解释.

## 3.2.1 似然率(likelihood)

在看到$D = \{16, 8, 2, 64\}$之后,为什么选择假设$h_{two} *$ = 2的幂数,而不选择$h_{even} *$ = 偶数呢?这两个假设都符合啊.关键就在于我们本能想要去避免可疑的巧合(suspicious coincidences).如果概念真是偶数,我们看到的怎么就都碰巧是2的幂数呢?

要用正规语言来说的话,假设样本是从概念的扩展集(extension)中随机抽取出来的.(概念的扩展及就是所有属于该概念的元素组成的集合,比如$h_{even}$的扩展及就是{2,4,6,...,98,100},以9结尾的数字的扩展及就是{9,19,...,99}.)Tenenbaum 称此为强抽样假设(strong sampling assumption).有了这

个假设之后,从h中可替换地独立抽取N个样本的概率就是:

$$p(D|h) = [\frac{1}{size(h)}]^N = [\frac{1}{|h|}]^N \text{(3.2)}$$

上面这个关键的等式体现了 Tenenbaum 所说的规模原则(size principle),也就意味着优先选择与数据样本一致且假设最少或者最简单的模型.这个原则也被通俗称为奥卡姆剃刀(Occam's razor).举个例子试试,设$D = \{16\}$,则$p(D|h_{two}) = 1/6$,这是因为在100以内有六个2的幂数;而$p(D|h_{even}) = 1/50$,这是因为1-100这个范围内有50个整数.所以$h = h_{two}$的可能性(likelihood)要比$h = h_{even}$高.有四个样本的时候,$h = h_{two}$的可能性是$(1/6)^4 = 7.7 \times 10^{-4}$,而$h = h_{even}$的可能性是$(1/50)^4 = 1.6 \times 10^{-7}$.这两者的概率比(likelihood ratio)高达5000:1了!自然好理解为啥人们都买$h = h_{two}$了.这也定量表明了之前的直觉(intuition),就是如果真的是$h = h_{even}$的话,那么$D = \{16, 8, 2, 64\}$就太巧合了.

## 3.2.2 先验(Prior)

若$D = \{16, 8, 2, 64\}$.那么$h_{32}$ = "除了32之外的2的幂数"会比$h_{two}$ = "2的幂数"有更高的概率,因为$h_{32}$不需要去解释为啥在样本集合中没有32.不过$h_{32}$ = "除了32之外的2的幂数"这个假设看上去太"不自然(conceptually unnatural)".我们之所以有这种直觉,是因为对不自然的概念赋予了低的先验概率(low prior probability).当然了,不同的人可能有不同的先验判断.这种主观色彩(subjective)也是贝叶斯估计有很多争议的一个原因,例如对于数值的判断来说,一个小朋友和一个数学教授就可能会有不同的答案.实际上这两者可能不仅先验判断不同,甚至连假设空间都不同.不过我们还是假设他们的假设空间一直,然后设置小朋友对于某些复杂概念上的先验权重为0.这样在先验和假设空间上就没有特别突兀的差别了.

虽然先验的主观性很惹争议,但还是很有用的.告诉你一串数字,其中有1200, 1500, 900, 1400,说取样自某个数学运算规则下,那你可能会认为400有可能符合这个规则,而1183就比较扯了.可要是告诉你这些数值来自健康胆固醇标准(healthy cholesterol levels),那就可能你觉得400概率不大而1183反而挺有可能出现了.因此不难发现,先验是利用背景知识来解决问题的机制.要是没有先验,快速学习就不可能实现了,比如从小规模样本进行学习等等.

那么针对例子中这种情况,咱们该用哪种先验呢?为了演示,咱们用最简单的先验,设定30中简单数学运算概念服从均匀分布,比如偶数/技术/质数/结尾有9的数等等.还可以让技术和偶数这两个概率更大.包括进去两个很不自然的概念,比如2的幂数加上37/除了32之外其他的2的幂数,但这两个给他们很低的先验权重.参考图3.2(a)是这个先验的图像.稍后会考虑更复杂情况.

此处参考原书图3.2

## 3.2.3 后验(Posterior)

后延就是可能性(似然率,likelihood)乘以先验,然后归一化.在本文就是:

$$p(h|D) = \frac{p(D|h)p(h)}{\sum_{\hat{h} \in H} p(D, \hat{h})} = \frac{p(h)I(D \in h)/|h|^N}{\sum_{\hat{h} \in H} p(\hat{h})I(D \in h)/|h|^N} \quad (3.3)$$

当且仅当所有数据都包含于假设h的扩展中的时候,其中的$I(D \in h) = 1$.图3.2所示为$D = \{16\}$时候的先验/似然率/后验的图像.很显然,后验是先验和似然率的结合.通常都假设先验是均匀分布,所以后验往往与似然率成正比.然而,一些"不自然"概念,比如底下那两个,2的幂数加37和除了32以外的2的幂数这些虽然似然率挺高的,但后验很低,因为对应的先验就很低.

图3.3所示为$D = \{16, 8, 2, 64\}$后的先验/似然率/后验.这时候在2的幂数上的概率就比之前高的多了,所以这就决定了后验概率的特征.

实际上人类学习者会有一个领悟时刻(aha moment),确定真正的概念.(这里就体现了对不自然概念的低先验的用处,否则如果选了"除了32以外的其他2的幂数"就明显会过拟合了.)

通常来说,只要有足够数据了,后验概率密度$p(h|D)$就会在一个单独概念位置有最大峰值,也就成了最大后验(MAP),即:

$$p(h|D) \rightarrow \delta_{\hat{h}^{MAP}}(h) \quad (3.4)$$

上式中$\hat{h}^{MAP} = \arg\max_h p(h|D)$是后验的模(posterior mode),其中的$\delta$是狄拉克测度(Dirac measure),定义如下:

$$\delta_x(A) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases} \quad (3.5)$$

最大后验(MAP)可以写作:

$$\hat{h}^{MAP} = \arg\max_h p(D|h)p(h) = \arg\max_h [\log p(D|h) + \log p(h)] \quad (3.6)$$

由于似然率项依赖于N的指数函数,而先验保持不变,所以随着数据越来越多,最大后验估计(MAP estimate)就收敛到最大似然估计(maximum likelihood estimate,MLE):

$$\hat{h}^{mle} * = \arg\max_h p(D|h) = \arg\max_h \log p(D|h) \quad (3.7)$$

也就是说,如果数据足够多了,就会发现数据特征盖过了先验.这种情况下最大后验估计(MAP)就朝着最大似然估计(MLE)收敛了.

如果假设空间里面包含了真实假设,那么最大后验估计/最大似然估计就都会收敛到这个假设.因此说贝叶斯推断(Bayesian inference)和最大似然估计是一致估计(consistent estimator),更多内容参考6.4.1.此事也说假设空间在限定范围内可识别(identifiable in the limit),意味着虽然受限于有限的数据也能够恢复出真实概念.如果假设类不足以全面表征真实情况(这也是常态),我们就只能收敛到尽可能接近真实概念的假设上.正规来说要用到亲密度(closeness)的概念,这超出了本章的范围了.

此处参考原书图3.3
此处参考原书图3.4

## 3.2.4 后验预测分布(Posterior predictive distribution)

后验其实是我们对外部世界的内在认知状态.怎么来测试我们所持信念呢?可以用他们来预测可观测的客观量(这也是科学方法的基础).比如上文所提到的后验预测分布可以写作:

$$p(\hat{x} \in C \mid D) = \sum_h p(y = 1 \mid \hat{x}, h) p(h \mid D) \quad (3.8)$$

这正好就是每个独立假设给出的预测的加权平均值,也叫做贝叶斯模型平均值(Bayes model averaging,缩写为 BMA,Hoeting et al. 1999).如图3.4所示,底部的实心点只是的是每个假设的预测,右边竖着的曲线是每个假设的权重.把每一列和权重相乘加到一起,就得到最顶部的分布了.

如果我们的数据集比较小或者存在模棱两可的不确定情况下,后延$p(h \mid D)$就很模糊了,就导致了宽泛预测分布(broad predictive distribution).不过,只要学习者弄明白了(figured things out),后验概率分布就成了以最大后验估计(MAP)为中心的$\delta$分布了.这时候,对应的预测分布就是:

$$p(\hat{x} \in C \mid D) = \sum_h p(\hat{x} \mid h) \delta_{\hat{h}}(h) = p(\hat{x} \mid \hat{h}) \quad (3.9)$$

这也叫做对预测密度的插值近似(plug-in approximation),特别简单所以用的很广泛.不过通常来说,这也代表了我们的不确定性,这种预测不会像使用贝叶斯模型均值(BMA)那样光滑.后文还会有更多相关例子.

虽然最大后验估计学习(MAP Learning)很简单,但不能解释从有不确定后验的相似度推理(similarity-based reasoning)到有确定后验的规则推理(rule-based reasoning)的这种渐进转变过程.比如,还说本章这个例子,先看到的是$D = \{16\}$,如果用上面的简单先验,最小一直假设就是"4的幂数",所以只有4和16会得到非零概率,被预测出来.这明显就是过拟合了.后面看到了$D = \{16, 8, 2, 64\}$,最大后验估计学习得到的假设就是"2的幂数".所以随着观测数据增多,差值预测分布会更宽或者维持原来的状态,最开始很窄,但随着观测数据增多逐渐变宽.与之相反,贝叶斯方法当中,最开始是很宽的估计,然后随着数据量增多逐渐收窄,这就很好理解了.比如最开始刚看到$D = \{16\}$的时候,有很多假设,都有不可忽略的后验支撑,所以预测分布也很宽.不过随着数据增多,$D = \{16, 8, 2, 64\}$,后验就集中在一个假设上了,预测分布也更窄了.因此差值近似和贝叶斯方法在小样本情况下是截然不同的,虽然二者随着数据规模扩大都会收敛到同样的答案.

## 3.2.5 更复杂的先验(A more complex prior)

为了对人类行为进行建模,Tenenbaum 在他的论文中用了一个更复杂的先验,这个先验是通过一个实验数据分析而推导得到的,这个实验测试了人们对数字相似性如何衡量,具体参考其博士论文的208页.结果就是得到了一个集合,跟前文的类似,也是数学概念组成的,相比之下多了一个就是在n到m中的所有间隔,$1 \leq n, m \leq 100$.(注意这些假设并不是互斥的.)所以这个先验实际上是两个先验

的混合,其中的一个是算数规则,另外一个是数字间隔:

$$p(h) = \pi_0 p_{rules}(h) + (1 - \pi_0) p_{interval}(h) \quad (3.10)$$

给定先验的两部分之后,这个模型中的唯一一个自由参数就是相对权重,$\pi_0$.只要$\pi_0 > 0.5$,结果就对这个值不是很敏感,反映的是人们更倾向于选择算数规则假设.这个模型的预测分布如图3.5所示,使用了更大的假设空间.这个分布和图3.1中人类的预测分布具有惊人的相似,虽然这个模型除了假设空间的选择之外并没有使用人类案例的数据进行拟合.

此处参考原书图3.5

# 3.3 $\beta$二项模型(beta-binomial model)

上面一节中讨论的数字游戏所涉及的是在得到一系列离散观察之后,从一个有限假设空间中推断一个离散变量的分布.这样计算其实都挺简单,只需要用到加法/乘法/除法.可是在实际应用中,有很多连续的未知参数,所以假设空间实际上是$R^K$或者其子集,其中的K就是参数个数.这样在数学上就复杂了,因为要用积分来替代之前离散情况下的加法.不过基本思想还是一样的.

举个例子,观测了一个硬币抛起落下的若干次实验之后,推测一个硬币人头朝上的概率.这看上去很小儿科,但这个模型是很多方法的基础,比如朴素贝叶斯分类器/马尔科夫模型等等.从历史角度来说这个实验也是很重要的,因为这正是1763年贝叶斯(Bayes)原版论文中用到的例子,贝叶斯的分析后来被皮埃尔-西蒙 拉普拉斯(Pierre-Simon Laplace)推广,建立成为了现在我们所知的贝叶斯规则,更多历史细节参考(Stigler 1986).

我们还是按照之前的套路,确定似然率(likelihood)和先验(prior),然后推导后验(posterior)和后验预测(posterior predictive).

## 3.3.1 似然率(Likelihood)

设 X 服从伯努利(Bernoulli)分布,即 $X_i \sim Ber(\theta)$,$X_i = 1$表示人头,$X_i = 0$表示背面,$\theta \in [0, 1]$是频率参数(人头出现的概率).如果实验事件是独立同分布的,那么似然率(likelihood)为:

$$p(D \mid \theta) = \theta^{N_1}(1 - \theta)^{N-0} \quad (3.11)$$

上式中的$N_1 = \sum_{i=1}^{N} I(x_i = 1)$对应人头,而$N_0 = \sum_{i=1}^{N} I(x_i = 0)$对应背面.这两个计数叫做数据的充分统计(sufficient statistics), 关于D我们只需要知道这两个量,就能推导$\theta$.充分统计集合也可以设置为$N_1$和$N = N_0 + N_1$.

正规表达下,若$p(\theta \mid D) = p(\theta \mid s(data))$,则就可以称$s(D)$是对数据D的一个充分统计.如果使用均匀分布作为先验,也就等价说$p(D \mid \theta) \propto p(s(D) \mid \theta)$.如果我们有两个集合,有同样的充分统计,就会推出同样的参数值$\theta$.

接下来设想在固定的总实验次数$N = N_0 + N_1$的情况下,数据中包含了人头朝上的次数为$N_1$.这时候就有$N_1$服从二项分布,即$N_1 \sim Bin(N, \theta)$,其中这个Bin的意思就是二项分布(binomial distribution),其概率质量函数(pmf)如下所示:

$$Bin(k|n,\theta)) * = \binom{n}{k}\theta^k(1-\theta)^{n-k} \quad (3.12)$$

因为$\binom{n}{k}$是独立于$\theta$的一个常数,所以二项取样模型的似然率和伯努利模型的似然率是一样的,所以我们对$\theta$的推断都是一样的,无论是观察一系列计数$D(N_1, N)$或者是有序的一系列测试
$D = \{x_1, \ldots, x_N\}$.

## 3.3.2 先验(prior)

需要一个定义在区间[0,1]上的先验.为了数学运算简便,可以让先验和似然率形式相同,也就是说对于参数为$\gamma_1, \gamma_2$的某个先验来说:
$p(\theta) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_2}$(3.13)$
这样的话,后验就很好估算了,只要合并指数就可以了:

$$p(\theta) \propto p(D|\theta)p(\theta) = \theta^{N_1}(1-\theta)^{N_0}\theta^{\gamma_1}(1-\theta)^{\gamma_2} = \theta^{N_1+\gamma_1}(1-\theta)^{N_0+\gamma_2} \quad (3.14)$$

这样先验和后验形式都一样了,就说这个先验是所对应似然率的共轭先验(conjugate prior).共轭先验用处很广泛,因为计算起来简单,也好理解.

使用伯努利分布的情况下,共轭先验就是$\beta$分布,在本书2.4.5就有提到:
$Beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$(3.15)

这个先验的参数叫超参数(hyper-parameters).可以根据我们事先持有的信念进行编码来对此进行设置.例如,如果我们认为$\theta$的均值应该是0.7,而标准差为0.2,就设定$a = 2.975, db = 1.275$(练习3.15).或者若认为$\theta$的均值应该是0.15而所属区间为开区间(0.05,0.30),就设定$a = 4.5, db = 25.5$(练习3.16).
如果关于参数$\theta$咱啥也不知道啊,只知道属于闭区间[0,1],就可以用均匀分布了,这也就是无信息先验(uninformative prior,更多细节参考本书5.4.2).均匀分布可以用一个$a = b = 1$的$\beta$分布来表示.

此处参考原书图3.6

## 3.3.3 后验(posterior)

把二项分布的似然率和$\beta$分布的先验乘到一起,就得到下面的后验了(参考公式3.14):

$$p(\theta|D) \propto Bin(N_1|\theta, N_0+N_1)Beta(\theta|a,b)Beta(\theta|N_1+a, N_0+b) \quad (3.16)$$

具体来说,这个后验是通过在经验计数(empirical counts)基础上加上了先验超参数(prior hyper-parameters)而得到的.因此将这些超参数称之为伪计数(pseudo counts).先验的强度,也是先验的有效取样规模(effective sample size)就是伪计数的和$a+b$;这个量起到的作用类似于数据集规模$N_1+N_0=N$.

图3.6(a)所示的例子中,弱先验Beta(2,2),似然率函数为单峰,对应一个大取样规模;从图中可见后验和似然率基本相符合:这是因为数据规模盖过了先验.图3.6(b)是使用了强先验Beta(5,2)来进行更新,也是一个单峰值似然率函数,可这时候很明显后验就是在先验和似然率函数之间的一个折中调和.

要注意的是按顺序对后验进行更新等价于单次批量更新.假设有两个数据集$D_a, D_b$,各自都有充分统计$N_1^a, N_0^a$和$N_1^b, N_0^b$.设$N_1=N_1^a+N_1^b, N_0=N_0^a+N_0^b$则是联合数据集的充分统计.在批量模式(batch mode)下则有:

$$p(\theta|D_a,D_b) \propto Bin(N_1|\theta,N_1+N_0)Beta(\theta|a,b) \propto Beta(\theta|N_1+a,N_0+b) \quad (3,17)$$

在序列模式(sequential mode)则有:

$$\begin{aligned} p(\theta|D_a,D_b) &\propto p(D_b|\theta)p(\theta|D_a) && (3.18)\\ &\propto Bin(N_1^b|\theta,N_1^b+N_0^b)Beta(\theta|N_1^a+a,N_0^a+b) && (3.19)\\ &\propto Beta(\theta|N^a a_1+N_1^b+a,N_0^a+N_0^b+b) && (3.20)\end{aligned}$$

这个性质使得贝叶斯推断很适合在线学习(online learning),后面会有更详细说明.

### 3.3.3.1 后验(posterior)的均值(mean)和模(mode)

参考等式2.62,最大后验估计(MAP)为:

$$\hat{\theta}_{MAP}=\frac{a+N_1-1}{a+b+N-2} \quad (3.21)$$

如果使用均匀分布先验,那么最大后验估计(MAP)就会降低成为最大似然估计(MLE),就正好是硬币人头朝上的经验分数(empirical):

$$\hat{\theta}_{MLE}=\frac{N_1}{N} \quad (3.22)$$

这个结论很符合直观印象,不过也可以通过应用基本微积分使等式3.11中的似然函数最大而推导出,参考练习3.1.

后验均值如下所示:

$$\bar{\theta}=\frac{a+N_1}{a+b+N} \quad (3.23)$$

这个区别后面有很大用处.后验均值是先验均值和最大似然估计的凸组合(convex combination),表示的就是在这两者之间进行折中,兼顾了先验的已有观点以及数据提供的信息.

设$\alpha_0 = a + b$是先验中的等效样本容量(equivalent sample size),控制的是先验强度,然后令先验均值(prior mean)为$m_1 = a/\alpha_0$.然后后验均值可以表示为:

$$E[] = \frac{\alpha_0 m_1 + N_1}{N + \alpha_0} = \frac{\alpha_0}{N + \alpha_0} m_1 + \frac{N}{N + \alpha_0} \frac{N_1}{N} = \lambda m_1 + (1 - \lambda)\hat{\theta}_{MLE} \quad (3.24)$$

上式中的$\lambda = \frac{\alpha_0}{N + \alpha_0}$为先验和后验的等效样本容量的比值.所以先验越弱,$\lambda$越小,而后验均值就更接近最大似然估计(MLE).

### 3.3.3.2 后验(posterior)的方差(variance)

均值和模都是点估计,还要知道可信程度.后验方差就是用来对此进行衡量的.\Beta后验的方差如下所示:

$$var[\theta | D] = \frac{(a + N_1)(b + N_0)}{(a + N_1 + b + N_0)^2 (a + N_1 + b + N_0 + 1)} \quad (3.25)$$

上面这个式子看着很麻烦,在$N >> a, b$的情况下可以对其进行近似以简化,得到的为:

$$var[\theta | D] \approx \frac{N_1 N_0}{NNN} = \frac{\bar{\theta}(1 - \bar{\theta})}{N} \quad (3.26)$$

其中的$\bar{\theta}$就是最大似然估计(MLE).然后能得到估计结果的"误差项(error bar)",也就是后验标准差:

$$\sigma = \sqrt{var[\theta | D]} \approx \sqrt{\frac{\bar{\theta}(1 - \bar{\theta})}{N}} \quad (3.27)$$

显然,不确定性以$1/\sqrt{N}$的速度降低.要注意这里的不确定性,也就是方差,在$\bar{\theta} = 0.5$的时候最大,在$\bar{\theta}$接近0或者1的时候最小.这意味着确定硬币是否有偏差要比确定硬币结果是否合理公平更容易(This means it is easier to be sure that a coin is biased than to be sure that it is fair).

## 3.3.4 后验预测分布(Posterior predictive distribution)

截至目前,我们关注的都是对未知参数的推导.这一节咱们回头来看对未来可观测数据的预测.

设预测一个硬币落地后人头朝上在未来单次实验中的概率服从后验分布$Beta(a, b)$.则有:

$$p(\bar{x} = 1 | D) = \int_0^1 p(x = 1 | \theta) p(\theta | D) d\theta \quad (3.28)$$

$$= \int_0^1 \theta Beta(\theta | a, b) d\theta = E[\theta | D] = \frac{a}{a + b} \quad (3.29)$$

这样就能发现,在这个案例中,后验预测分布(posterior predictive distribution)的均值(mean)和后验均值参数插值(plugging in the posterior mean parameters)是等价的:$p(\bar{x} | D) = Ber(\bar{x} | E[\theta | D])$.

### 3.3.4.1 过拟合与黑天鹅悖论

若不使用插值进行最大似然估计(MLE),也就是说使用$p(\bar{x}|D) \approx Ber(\bar{x}|\hat{\theta}_{MLE})$.很不幸,当样本规模小的时候这个近似表现很差.例如设一组实验$N=3$.那么最大似然估计(MLE)就是$\bar{\theta}=0/3=0$,这已经是最大程度利用了观测数据了.如果我们采信这个估计,就会认为硬币人头朝上是不可能事件了.这就叫做零计数问题(zero count problem)或者稀疏数据问题(sparse data problem),对小规模数据进行估计的时候会经常出现的.有人可能觉得在所谓大数据应用领域,就没必要太担心这种问题了,可是一定要注意,一旦我们对数据基于某些特定标准进行了人为划分,比如某个人从事某个活动的次数等等,样本规模就变小了很多了.这种问题就还会出现,比如在推荐个性化网页的时候就可能出现.所以即便是在所谓大数据时代,贝叶斯方法还是有用的 (Jordan 2011).

零计数问题很类似一个叫做黑天鹅悖论的哲学问题.古代西方人的观念是所有天鹅都是白色的,就把黑天鹅当作不存在的事物的一个比喻.而在17世纪欧洲探索者在澳大利亚发现了黑天鹅.所以科学哲学家卡尔 波普(Karl Popper)就提出了黑天鹅悖论这个名词,另外也有个畅销书用黑天鹅做标题(Taleb 2007).这个悖论用于归纳问题,如何从过去的特定观察去得出对未来的一般性结论.

使用贝叶斯方法来推导一个对这个问题的解决方案.使用均匀先验,所以a=b=1.这样对后验均值插值就得到了拉普拉斯继承规则(Laplace's rule of succession):

$$p(\hat{x}=1|D)=\frac{N_1+1}{N_1+N_0+2} \quad (3.30)$$

上式中包含了一种实践中的常规做法,就是对经验计数(empirical counts)加1,归一化,然后插值,这也叫做加一光滑(add-one smoothing).要注意对最大后验估计(MAP)插值就不会有这种光滑效果,因为这时候模(mode)的形式$\hat{\theta}=\frac{N_1+a+1}{N+a+b+2}$,如果a=b=1就成了最大似然估计(MLE)了.

## 3.3.4.2 预测未来多次实验

设有 M 次未来实验,要去预测其中的人头朝上的次数 x.这个概率则为:

$$p(x|D,M)=\int_0^1 Bin(x|\theta,M)Beta(\theta|a,b)d\theta \quad (3.31)$$

$$=\binom{M}{x}\frac{1}{B(a,b)}\int_0^1 \theta^x(1-\theta)^{M-x}\theta^{a-1}(1-\theta)^{b-1}d\theta \quad (3.32)$$

这个积分正好就是$Beta(a+x,M-x+b)$这个分布的归一化常数.因此:
$(3.33)因此就能发现后验预测分布如下所示,是一个(复合)\beta-二项分布分布(beta-binomialdistribution):Bb(x|a,b,M)*= ({

$$M\backslash x$$

}) \frac{B(x+a,M-x+b)}{B(a,b)}(3.34)这个分布的均值和方差如下所示:
E[x]=M\frac{a}{a+b},var[x]=\frac{Mab}{(a+b)^2}\frac{(a+b+M)}{a+b+1}$(3.35)

如果$M = 1$,则$x \in \{0, 1\}$,均值就成了$E[x|D] = p(x = 1|D) = \frac{a}{a+b}$,和等式3.29一致.

这个过程如图3.7(a)所示.开始是用Beta(2,2)作为先验,投图的是$N_1 = 3, N_0 = 17$,即3次人头,17次背面的情况下的后验预测密度.图3.7(b)投图的是最大后验估计(MAP)插值近似.很明显贝叶斯预测(Bayesian prediction)有更长尾(longer tails)概率质量分布的更广泛,所以更不容易过拟合,也更不容易遇到黑天鹅悖论的情况.

此处查看原书图3.7

# 3.4 狄利克雷-多项式模型(Dirichlet-multinomial model)

上一节讲的抛硬币的概率问题只有两个状态,人头朝上或者背面朝上.本节要对上一节的结论推广到有K个面的骰子的k个状态上.这和另外一个玩具练习有点像,不过这一章要学到的方法还会广泛应用到文本/生物序列等数据的分析上.

## 3.4.1 似然率(Likelihood)

假设观测了N次掷骰子,得到的点数集合为$D = \{x_1, \ldots, x_N\}$,其中$x_i \in \{1, \ldots, K\}$.假设这个数据是独立同分布的(IID),那么似然率如下所示:

$p(D|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}$(3.36)

上式中的$N_k = \sum_{i=1}^{K} I(y_i = k)$是事件k出现的次数(这也是该模型的充分统计).多项式模型的似然率与上式形式相同,有不相关常数因子.(The likelihood for the multinomial model has the same form, up to an irrelevant constant factor.)

## 3.4.2 先验(Prior)

参数向量处在K维度概率单纯形(K-dimensional probability simplex)中,所以需要在这个单纯形上定义一个先验.理想情况下应该是共轭的(conjugate).很幸运的就是本书2.5.4中提到的狄利克雷分布就满足这两个条件.所以使用如下的先验:

$Dir(\theta|\alpha) = \frac{1}{B(\alpha)} \prod^{K} \theta_k^{\alpha_k - 1} I(x \in S_K)$(3.37)

## 3.4.3 后验(Posterior)

把先验和似然率相乘,就得到了后验了,也是一个狄利克雷分布:

$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{3.38}$$

$$\propto \prod_{k=1}^{K} \theta_k^{N_k} \theta_k^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1} \tag{3.39}$$

$$= Dir(\theta|\alpha_1 + N_1, \ldots, \alpha_K + N_K) \tag{3.40}$$

很明显这个后验是通过将先验的超参数（伪计数pseudo-counts）$\alpha_k$加到经验计数(empirical counts)$N_k$上而获得的。

可以通过积分来推导出这个后验的模,也就是最大后验估计(MAP estimate).不过还要必须强化约束条件$\sum_k \theta_k = 1$.可以通过拉格朗日乘数(Lagrange multiplier)来实现.受约束的目标函数,也叫拉格朗日函数(Lagrangian),可以通过对似然率取对数加上对先验取对数然后加上约束条件:
$l(\theta, \lambda) = \sum_k N_k \log\theta_k + \sum_k (\alpha_k - 1)\log\theta_k + \lambda(1 - \sum_k \theta_k)$(3.41)

为了简化表达,定义一个$\hat{N}_k * = N_k + \alpha_k - 1$.取关于$\lambda$的导数就得到了初始约束(original constraint):
$\frac{\partial l}{\partial \lambda} = (1 - \sum_k \theta_k) = 0$(3.42)

利用总和为1这个约束条件就可以解出来$\lambda$:
$\sum_k \hat{N}_k = \lambda \sum_k \theta_k$(3.45)

$N + \alpha_0 - K = \lambda$(3.46)
上式中的$\alpha_0 * = \sum_{k=1}^{K} \alpha_k$等于先验中的样本规模.这样最大后验估计(MAP estimate)为:
$\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$(3.47)

这与等式2.77相一致.如果使用均匀分布作为先验,即$\alpha_k = 1$,就能解出最大似然估计(MLE):
$\hat{\theta}_k = N_k / N$(3.48)

这正好是k面出现次数的经验分数(empirical fraction).

## 3.4.4 后验预测分布(Posterior predictive)

对一个单次多重伯努利实验(single multinoulli trial),后验预测分布如下所示:

$$p(X = j|D) = \int p(X = j|\theta)p(\theta|D)d\theta \tag{3.49}$$

$$= \int p(X = j|\theta_j)[\int p(\theta_{-j}, \theta_j|D)d\theta_{-j}]d\theta_j \tag{3.50}$$

$$= \int \theta_j p(\theta_j|D)d\theta = E[\theta_j|D] = \frac{\alpha_j + N_j}{\sum_k (\alpha_k + N_k)} = \frac{\alpha_j + N_j}{\alpha_0 + N} \tag{3.51}$$

上式中的$\theta_{-j}$是除了$\theta_j$之外的其他所有$\theta$的成员,参考练习3.13.

上面的表达式避免了零计数问题(zero-count problem),参考3.3.4.1.实际上,贝叶斯光滑(Bayesian smoothing)在多项分布情况比二值分布中更重要,因为一旦将数据分成许多类别了,数据稀疏的可能性就增加了.

### 3.4.4.1 实例:使用单词袋的语言模型

使用狄利克雷-多项模型进行贝叶斯光滑的一个应用就是语言建模(language modeling),就是预测一个序列中下一个位置出现什么词.
设第i个词为$X_i \in \{1, \ldots K\}$,使用多重伯努利分布$Cat(\theta)$从所有其他词汇中独立取样.这就叫单词袋模型(bag of words model).知道了已经出现的单词序列之后,如何预测下一个单词可能是什么呢?

假设我们观察到的是下面这个序列,来自一段儿歌:

Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow

然后设我们的词汇表包含下面的单词:

| mary | lamb | little | big | fleece | white | black | snow | rain | unk |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

上面表格中的 unk表示的是未知词汇,也就是所有没在列表中出现的其他词汇.要对儿歌进行编码,先去掉标点符号,然后去掉停止词(stop words)比如a/as/the等等.这就要进行词干化(stemming),意思就是把所有词汇恢复原形,去掉复数,去掉ing恢复动词本身等等.不过这个儿歌里面到没有这么麻烦的.把每个单词用词汇表中的索引号进行编码就得到了:
1 10 3 2 3 2 3 2
1 10 3 2 10 5 10 6 8
接下来忽略单词排序,只数一下每个单词出现的次数,得到一个频率分布表:

| 单词 | mary | lamb | little | big | fleece | white | black | snow | rain | unk |
|---|---|---|---|---|---|---|---|---|---|---|
| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 次数 | 2 | 4 | 4 | 0 | 1 | 1 | 0 | 1 | 0 | 4 |

上面的计数记作$N_j$.如果用一个狄利克雷函数$Dir(\alpha)$作为$\theta$的先验,后验预测分布为:

$$p(\tilde{X} = j \mid D) = E[\theta_j \mid D] = \frac{\alpha_j + N_j}{\sum_{j'} \alpha_{j'} + N_{j'}} = \frac{1 + N_j}{10 + 17} \quad (3.52)$$

如果设$\alpha_j = 1$,则有:

$$p(\tilde{X} = j \mid D) = (3/27, 5/27, 5/27, 1/27, 2/27, 2/27, 1/27, 2/27, 1/27, 5/27) \quad (3.53)$$

上面这个预测分布的模(mode)是$X = 2(\text{“lamb”}), X = 10(\text{“unk”})$.这里要注意,有的单词虽然没出现在当前看过的词汇序列中,但依然被预测了非零概率,也就是以后有可能出现,比如big/black/rain这几个词.后面还有更复杂的语言模型.

# 3.5 朴素贝叶斯分类器(Naive Bayes classifiers)

本节讨论的是对离散值特征向量进行分类,其中特征$x \in \{1, \dots, K\}^D$,K是每个特征的数值个数,D是特征数.这次要用一种通用方法.这就需要确定类条件分布(class conditional distribution)$p(x \mid y = c)$.最简单的方法,在给定类标签的情况下,假设所有特征有条件独立.这样就可以将类条件密度(class conditional density)写成一维密度的乘积:

$$p(x \mid y = c, \theta) = \prod_{j=1}^{D} p(x_j \mid y = c, \theta_{jc}) \quad (3.54)$$

这样得到的模型就叫做朴素贝叶斯分类器(naive Bayes classifier,缩写为 NBC).

称之为"朴素(naive)"是因为我们并不指望各个特征独立,甚至即便在类标签上也未必有条件独立.不过即便朴素贝叶斯假设不成立,得到的分类结果也还都不错(Domingos and Pazzani 1997).一个原因就是这个模型特别简单,对于C个类D个特征的情况只有$O(CD)$个参数,所以相对来说不容易过拟合.

类条件密度的形式取决于每个特征的类型,下面给出一些可能的情况:
* 如果特征向量是实数值的,可以用高斯分布,也就是正态分布:$p(x \mid y = c, \theta) = \prod^{D} N(x_j \mid \mu_{jc}, \sigma_{jc}^2)$,其中的$\mu_{jc}$是类c对象中特征j的均值,$\sigma_{jc}^2$是方差.
* 如果特征是二值化的,即$x_j \in \{0, 1\}$,可以用伯努利分布:$p(x \mid y = c, \theta) = \prod_{j=1}^{D} Ber(x_j \mid \mu_{jc})$,其中的$\mu_{jc}$是特征j出现在类别c的概率.这有时候也叫做多元伯努利朴素贝叶斯模型(multivariate Bernoulli naive Bayes model).
* 如果是分类特征,$x_j \in \{1, \dots, K\}$,可以用多重伯努利(multinoulli)分布:
$p(x \mid y = c, \theta) = \prod_{j=1}^{D} Cat(x_j \mid \mu_{jc})$,其中的$\mu_{jc}$是类中的$x_j$的K个可能值的频数(histogram).

当然还可以处理其他类型的特征,或者使用不同的分布假设.另外也可以对不同类型特征进行混合和匹配.

## 3.5.1 模型拟合

接下来就要"训练"一个朴素贝叶斯分类器了.一般都是对参数进行最大似然估计(MLE)或者最大后验估计(MAP estimate).不过本节还要讲如何对整个后验$p(\theta \mid D)$进行计算.

### 3.5.1.1 朴素贝叶斯分类器的最大似然估计

单数据情况(single data case)下的概率为:

$p(x_i, y_i|\theta) = p(y_i|\pi)\prod_j p(x_{ij}|\theta_j) = \prod_c \pi_c^{I(y_i=c)} \prod_j \prod_c p(x_{ij}|\theta_{jc})^{I(y_i=c)}$(3.55)

这样则有对数似然率(log-likelihood):
$\log p(D|\theta) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{D} \sum_{c=1}^{C} \sum_{i:y_i=c} \log p(x_{ij}|\theta_{jc})$(3.56)

很明显上面这个表达式可以拆解成一系列子项,其中有一个包含了$\pi$,另外的DC项目包含了$\theta_{jc}$.所以可以对所有这些参数分开优化.

从等式3.48得知,分类先验(class porior)的最大似然估计(MLE)为:
$\hat{\pi}_c = \frac{N_c}{N}$(3.57)

上式中的$N_c * = \sum_i I(y_i = c)$,是类c中的样本个数.

对似然率的最大似然估计(MLE)依赖于我们对特征所选的分布类型.简单起见,假设所有特征都是二值化的,这样使用伯努利分布,即$x_j|y = c \sim Ber(\theta_{jc})$.这时候最大似然估计(MLE)则为:
$\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$(3.58)

这个模型拟合过程的实现特别简单,可以参考本书算法8作为伪代码,或者MATLAB代码中的naiveBayesFit.这个算法的复杂度是O(ND).此方法也很容易泛化拓展到对混合类型特征的应用上.由于简单,应用广泛.

图3.8给出的例子中,有2个类,600个二值特征,这些特征表示的是一个词是否出现在一个词汇袋模型中.图中对两类中的$\theta_c$向量进行了可视化.107位置上的特别高端峰值对应的是单词"subject",在两类中出现的概率都是1.在3.5.4会讲到如何"滤除(filter out)"这些非信息特征.

此处查看原书图3.8

## 3.5.1.2 使用贝叶斯方法的朴素贝叶斯(Bayesian naive Bayes)

最大似然估计有个麻烦就是可能会过拟合.比如图3.8中的例子,"subject"这个词,假设作为特征j,在两类中都出现,所以对应这个特征j的$\hat{\theta}_{jc} = 1$.那如果收到一个新邮件其中不包含这个词会怎么办?算法就会崩溃了,因为发现对于两个类来说此事都有$p(y = c|x, \hat{\theta}) = 0$.这也是3.3.4.1当中提到的黑天鹅悖论的另一种体现.

避免过拟合的简单解决方案就是使用贝叶斯方法.简单起见,使用一个因式化先验:
$p(\theta) = p(\pi)\prod_{j=1}^{D}\prod_{c=1}^{C} p(\theta_{jc})$(3.59)

对于$\pi$使用狄利克雷先验$Dir(\alpha)$,对每个参数$\theta_{jc}$采用$\beta$分布$Beta(\beta_0, \beta_1)$.通常就设$\alpha = 1, \beta = 1$对应的是加一光滑或者拉普拉斯光滑.

结合等式3.56当中的因式化似然率与因式化先验,就得到了下面的因式化后验:

$$p(\theta|D) = p(\pi|D)\prod_{j=1}^{D}\prod_{c=1}^{C}p(\theta_{jc}|D) \quad (3.60)$$

$$p(\pi|D) = Dir(N_1 + \alpha_1, \ldots, N_C + \alpha_C) \quad (3.61)$$

$$p(\theta_{jc}|D) = Beta((N_c - N_{jc}) + \beta_0, N_{jc} + \beta_1) \quad (3.62)$$

换句话说,要计算这个后验,只需要用似然率中的经验计数(empirical counts)更新先验计数(prior counts).修改一下算法8就可以进行这个版本的模型拟合了.

## 3.5.2 使用模型做预测

再测试的时候,目标是计算:

$$p(y = c|x, D) \propto p(y = c|D)\prod_{j=1}^{D}p(x_j|y = c|D) \quad (3.63)$$

正确的贝叶斯步骤就是使用积分排除掉未知参数:

$$p(y = c|x, D) \propto [\int Cat(y = c|\pi)p(\pi|D)d\pi] \quad (3.64)$$

$$\prod_{j=1}^{D}[\int Ber(x_j|y = c, \theta_{jc})p(\theta_{jc}|D)] \quad (3.65)$$

还好这个比较好实现,至少在后验为狄利克雷分布的时候挺简单的.参考等式3.51,我们知道后验预测密度可以通过插入后验均值参数$\theta$来获得.因此有:

$$p(y = c|x, D) \propto \bar{\pi}_C\prod_{j=1}^{D}(\bar{\theta}_{jc})^{I(x_j=1)}(1 - \bar{\theta}_{jc})^{I(x_j=0)} \quad (3.66)$$

$$\bar{\theta}_{jk} = \frac{N_{jc} + \beta_1}{N_c + \beta_0 + \beta_1} \quad (3.67)$$

$$\bar{\pi}_c = \frac{N_c + \alpha_c}{N + \alpha_0} \quad (3.68)$$

上式中$\alpha_0 sum_c \alpha_c$.
如果我们通过单个点估计了后验,$p(\theta|D) \approx \delta_{\hat{\theta}}(\theta)$,其中的$\hat{\theta}$可以使最大似然估计(MLE)或者最大后验估计(MAP),然后就可以通过对参数插值来得到后验预测密度了,生成的是一个虚拟一致规则(virtually identical rule):

$$p(y = c|x, D) \propto \hat{\pi}_c\prod_{j=1}^{D}(\hat{\theta}_{jc})^{I(x_j=1}(1 - \hat{\theta}_{jc})^{I(x_j=0)} \quad (3.69)$$

唯一具备就是把后验均值的$\bar{\theta}$换成了后验模或者最大似然估计$\hat{\theta}$.不过这差别虽然小,实践中的影响可能很大,因为后验均值更不容易过拟合,参考本书3.4.4.1.

## 3.5.3 求对数-相加-幂运算组合技巧(log-sum-exp trick)

接下来要讨论的是一个在使用各种通用分类器的时候都很有用的重要应用细节.对类标签的后验计算可以使用等式2.13,使用合适的类条件密度(class-conditional density)(以及插值近似).然而很不幸,直接使用等式2.13进行计算可能会因为数值向下溢出(numerical underflow)而失败.这是因为概率$p(x|y=c)$通常都是非常非常小的数值,尤其是如果x是高维度向量的时候更是如此.而概率总和必然是1,即$\sum_x p(x|y)=1$,所以任何特定的高维度向量被观测到的概率都是很小的.要解决这个问题,就需要在应用贝叶斯规则的时候先取对数,如下所示:

$$\log p(y=c|X) = b_c - \log[\sum \hat{} C_{c'=1} e\hat{} b_{c'}] \qquad (3.70)$$

$$b_c * = \log p(x|y=c) + \log p(y=c) \qquad (3.71)$$

然而这需要我们计算下面这个表达式:

$\log[\sum_{c'} e^{b_{c'}}] = \log[\sum_{c'} p(y=c', x)] = \log p(x)$(3.72)

可是这算起来挺麻烦的,不过好在可以找到最大因子项,然后提取出来,用这一项来表示其他的,如下所示:

$\log(e+e^{-121}) = \log e - 120(e^0+e^{-1}) = \log(e^0+e^{-1}) - 120$(3.73)

通常来说就得到下面这种:

$\log\sum_c e^{b_c} = \log[(\sum_c e^{b_c-B})e^B] = [\log(\sum_c e^{b_c-B})] + B$(3.74)

其中的最大公因式项$B = \max_c b_c$.这种手法就是求对数-相加-幂运算组合技巧(log-sum-exp trick),用的很广泛,PMTK3中的logsumexp就是一个实例.

这个方法用在了算法1中,算法以的伪代码是使用朴素贝叶斯分类器来计算$p(y_i|x_i,\hat{\theta})$.PMTK3中的naiveBayesPredict是MATLAB代码.如果只要计算$\hat{y}_i$其实并不需要这样做,因为直接将未归一化的量$\log p(y_i=c) + \log p(x_i|y=c)$. 最大化就可以了.

## 3.5.4 使用互信息量进行特征选择

朴素贝叶斯分类器是对一个在多个潜在特征上的联合分布进行拟合,所以可能会过拟合.另外其运算上的开销是O(D),对于某些情况下可能太高开销了.一种解决这些问题的常见方法就是进行特征选择(feature selection),移除一些对于分类问题本身没有太大帮助的"不相关(irrelevant)"信息.最简单的信息选择方法就是单独评价每个特征的相关性(relevance),然后选择最大的K个,K是根据精确度和复杂度之间的权衡来选择的.这种方法也叫做变量排序/过滤/筛选.

衡量相关性的一个手段就是利用互信息量(mutual information),参考本书2.8.3.要计算特征$X_j$和分类标签Y之间的互信息量:

$$I(X, Y) = \sum_{x_j}\sum_{y} p(x_j, y)\log\frac{p(x_j, y)}{p(x_j)p(y)} \quad (3.75)$$

互信息量可以被理解为在观测了特诊j的值的时候标签分布上的信息熵降低.如果特征是二值化的,就很明显可以用下面的公式来计算(参见练习3.2.1):

$$I_j = \sum_c [\theta_{jc}\pi_c \log\frac{\theta_{jc}}{\theta_j} + (1 - \theta_{jc})\pi)c\log\frac{1-\theta_{jc}}{1-\theta_j}] \quad (3.76)$$

上式中的$\pi_c = p(y = c), \theta_{jc} = p(x_j = 1 \mid y = c), \theta_j = p(x_j = 1) = \sum_c \pi_c \theta_{jc}$,所有这些量都可以在拟合朴素贝叶斯分类器的时候作为副产品被计算出来.

表3.1展示的是将这个方法用于图3.8所示的二值化词汇袋得到的结果.从表中可以看到有最高互信息量的单词会比常见单词有更大区别作用(discriminative).例如两类中最常见的单词就是"subject"主体,者经常出现因为这两份都是新闻组数据,总会包含一个主题行.不过很明显这个词没有什么区别作用.带有分类标签的最高互信息量的词汇按照降序排列为"windows", “microsoft”, “DOS”,“motif”这就合理了,因为这两个分类对应的是微软的Windows和X Windows.

## 3.5.5 使用词汇袋进行文档分类

文档分类(Document classification)问题是要把文本文档分成不同类别.一个简单方法就是把每个文档都看做二值化向量,每个单词是否出现是值,当且仅当单词j出现在文档i当中 $x_{ij} = 1$,否则 $x_{ij} = 0$ .就可以用下面的分类条件密度了:

$$p(x_i \mid y_i = c, \theta) = \prod_{j=1}^{D} Ber(x_{ij} \mid \theta_{jc} = \prod_{j=1}^{D} \theta_{jc}^{I(x_{ij})}(1 - \theta_{jc})^{I(1-x_{ij})} \quad (3.77)$$

这也叫做伯努利乘积模型(Bernoulli product model,),或者叫二值独立模型(binary independence model).

可是上面这个模型只有是否包含单词这个信息,缺失了单词出现次数,丢失了很多信息(McCallum and Nigam 1998).更精确的表示需要记录每个单词出现的次数.具体来说设$x_i$是文档i的技术向量,所以有$x_{ij} \in \{0, 1, \ldots, N_i\}$,其中$N_i$是文档i中的词汇总数,所以有$\sum_{j=1}^{D} x_{ij} = N_i$.对于类条件密度,可以使用多项式分布(multinomial distribution):

$$p(x_i \mid y_i = c, \theta) = Mu(x_i \mid N_i, \theta_c) = \frac{N_i!}{\prod_{j=1}^{D} x_{ij}!}\prod_{j=1}^{D} \theta_{jc}^{x_{ij}} \quad (3.78)$$

上式中我们隐含着假设了文档长度$N_i$与类别不相关.其中的$\theta_{jc}$是c类文档中生成单词j的的概率;因此对于每个类别c,这些参数都要满足归一化约束$\sum_{j=1}^{D} \theta_{jc} = 1$

虽然多项式分类器(multinomial classifier)训练起来简单,测试的时候用着也简单,但对于文档分类来说并不太适合.一个原因就是这个模型没有考虑单词使用的时候的突发性(burstiness ).这是指单词突发出现的现象,有的单词可能之前从来没在给定的文档中出现过,但是一旦出现一次,之后就可能出现不止一次了.

多项式模型不能捕获这种单词突发现象.具体原因可以参考等式3.78当中的一项为$\theta_{jcv}^{N_{ij}}$,对于罕见词汇来说,$\theta_{jc} << 1$,越来越不可能出现很多了.对于更多常见词,这个衰减速率(decay rate)就没那么显著了.直观理解的话,要注意大多数常见词都是功能词,比如助词介词之类的,并不能够对应特定文档类别.比如 and 这个词出现的概率基本是固定的,不受之前出现多少次(文档长度模数 modulo document length)的影响,所以与文档类别不想关掉假设对于这些常见词就更适合一些.不过罕见词对文档分类来说更重要,我们建模的时候要对这些罕见次仔细对待.

为了改善多项式文档分类器的效果,已经有很多特征启发式(ad hoc heuristics)的方法被开发了出来(Rennie et al. 2003).有另外一种类条件密度(class conditional density)方法,性能表现和特征启发方法一样好,而可以从概率角度解释 (Madsen et al. 2005).

若把多项式类条件密度替换成狄利克雷符合多项式密度(Dirichlet Compound Multinomial,缩写为DCM),定义如下所示:

$$p(x_i|y_i = c, \alpha) =\in Mu(x_i|N_i, \theta_c)Dir(\theta_c|\alpha_c)d\theta_c = \frac{N_i!}{prod_{j=1}^{D} x_{ij}!}\frac{B(x_i+\alpha_c)}{B(\alpha_c)} \quad (3.79)$$

这个等式是在等式5.24中推导出来的.令人惊讶的是只要稍作上述改动就可以捕获到罕见词的突发现象.直观来说理由如下:看到一个出现过的词之后,比如词汇j,后验技术$\theta_j$就更新了,让单词j的另一次出现的概率提高.相反,如果$\theta_j$是固定的,那么每个词出现就是独立的.这个多项模型对应的就像是从一个瓮(坛子)里有K种颜色的球当中抽取一个球,记录下颜色,然后替换掉.与之对比的狄利克雷复合多项模型(DCM)对应的是抽取一个球,记录下颜色,然后用一个额外的副本替换掉它;这也叫做波利亚瓮(Polya urn)模型.

如 (Madsen et al. 2005)所述,使用DCM比单纯用多项式的效果更好,并且和现代方法性能表现相当.唯一的劣势就是狄利克雷复合多项式模型(DCM)拟合起来更复杂了些,更多细节参考(Minka 2000e; Elkan 2006) .

# 练习 3

尼玛太长了,回头再说吧,先去弄下一章了.