一文详解朴素贝叶斯(Naive Bayes)原理

一、历史背景解读

18世纪英国业余(一点都不业余好吗)数学家托马斯·贝叶斯(*Thomas Bayes，1702～1761*)提出过一种看似显而易见的观点：

“用客观的新信息更新我们最初关于某个事物的信念后，我们就会得到一个新的、改进了的信念。”

这个研究成果由于简单显得平淡无奇，直至他死后两年才于1763年由他的朋友理查德·普莱斯帮助发表。它的数学原理很容易理解，简单说就是，如果你看到一个人总是做一些好事，则会推断那个人多半会是一个好人。这就是说，**当你不能准确知悉一个事物的本质时，你可以依靠与事物特定本质相关的事件出现的多少去判断其本质属性的概率。**

用数学语言表达就是：**支持某项属性的事件发生得愈多，则该属性成立的可能性就愈大。**

与其他统计学方法不同，贝叶斯方法建立在主观判断的基础上，你可以先估计一个值，然后根据客观事实不断修正。

1774年，法国数学家皮埃尔-西蒙·拉普拉斯(*Pierre-Simon Laplace，1749-1827*)独立地再次发现了贝叶斯公式。拉普拉斯关心的问题是：当存在着大量数据，但数据又可能有各种各样的错误和遗漏的时候，我们如何才能从中找到真实的规律。拉普拉斯研究了男孩和女孩的生育比例。有人观察到，似乎男孩的出生数量比女孩更高。

这一假说到底成立不成立呢?

拉普拉斯不断地搜集新增的出生记录，并用之推断原有的概率是否准确。每一个新的记录都减少了不确定性的范围。拉普拉斯给出了我们现在所用的贝叶斯公式的表达：

$$P(A/B)=P(B/A)*P(A)/P(B)P(A/B)=P(B/A)*P(A)/P(B)$$

该公式表示在B事件发生的条件下A事件发生的条件概率，等于A事件发生条件下B事件发生的条件概率乘以A事件的概率，再除以B事件发生的概率。公式中，P(A)也叫做先验概率，P(A/B)叫做后验概率。严格地讲，贝叶斯公式至少应被称为“贝叶斯-拉普拉斯公式”。

二、原理推导

理论上，概率模型分类器是一个条件概率模型：

p(C|F1,...,Fn)p(C|F1,...,Fn)

独立变量C有若干类别，条件依赖于若干特征变量,但问题在于如果特征数量n的维度较大或者每个特征能取大量值时，基于概率模型列出概率表变得不现实。所以我们修改这个模型使之变得可行。 根据贝叶斯公式有以下式子：

p(C|F1,...,Fn)=p(C)∗p(F1,...Fn|C)p(F1,...,Fn)p(C|F1,...,Fn)=\frac{p(C)*p(F1,...Fn|C)}{p(F1,...,Fn)}

或者，这样表达比较简洁明了：

类别 特征 类别 类别 特征 特征 p(类别|特征)=p(类别)∗p(类别|特征)p(特征)p(类别|特征)=\frac{p(类别)*p(类别|特征)}{p(特征)}

其中， p(C)p(C) 为先验概率， p(C|F1,...,Fn)p(C|F1,...,Fn) 为后验概率；可以这么理解，再不知道需要预测的样本任何特征的时候，先判断该样本为某个类别的概率为 p(C)p(C) 为，再知道样本的特征之后，乘上 p(F1,...Fn|C)p(F1,...Fn)\frac{p(F1,...Fn|C)}{p(F1,...Fn)} 之后，得到该样本再知道 F1=f1,...,Fn=fnF1=f1,...,Fn=fn 之后，样本属于这个类别的条件概率。

这个乘上去的因子可能是起到促进的作用（当该因子大于1），也可能起到抑制的作用（当该因子小于1）。这个比较容易理解，比如没有任何信息的时候，可以判断一个官为贪官的概率为0.5，再知道该官员财产大于一千万后，则根据常理判断该官员为贪官的概率为0.8。

实际中，我们只关心分式中的分子部分 p(C)∗p(F1,...Fn|C)p(C)*p(F1,...Fn|C) ，因为分母不依赖于C,而且特征的值也是给定的，于是分母可以认为是一个常数。这样分子就等价于联合分布模型。

p(C,F1,...Fn)p(C,F1,...Fn)

$$
\begin{aligned}
&p(C,F1,...Fn)\\
&\propto p(C)*p(F1,...Fn|C)\\
&\propto p(C)*p(F1|C)*p(F2,...Fn|C,F1)\\
&\propto p(C)*p(F1|C)*p(F2|C,F1)*p(F3,...Fn|C,F1,F2)\\
&\propto p(C)*p(F1|C)*p(F2|C,F1)*p(F3|C,F1,F2)...p(Fn|C,F1,F2,...,Fn-1)
\end{aligned}
$$

现在，“朴素”的条件独立假设开始发挥作用了：假设每个特征,对于其他特征,是独立的，即特征之间相互独立，就有：

p(Fi|C,Fj)=p(Fi|C)p(Fi|C,Fj)=p(Fi|C)

这里还要再解释一下为什么要假设特征之间相互独立。

我们这么想，假如没有这个假设，在数据量很大的情况下，那么我们对右边这些概率的估计其实是不可做的，这么说，假设一个分类器有4个特征，每个特征有10个特征值，则这四个特征的联合概率分布是4维的，可能的情况就有 104=1000010^{4}=10000 种。

计算机扫描统计还可以，但是现实生活中，往往有非常多的特征，每一个特征的取值也是非常之多，那么通过统计来估计后面概率的值，变得几乎不可做，这也是为什么需要假设特征之间独立的原因，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。

有了特征相互独立的条件以后，对于,联合分布模型可表达为：

$$
\begin{aligned}
&p(C,F1,...Fn)\\
&\propto p(C)*p(F1,...Fn|C)\\
&\propto p(C)*p(F1|C)*p(F2,...Fn|C,F1)\\
&\propto p(C)*p(F1|C)*p(F2|C,F1)*p(F3,...Fn|C,F1,F2)\\
&\propto p(C)*p(F1|C)*p(F2|C,F1)*p(F3|C,F1,F2)...p(Fn|C,F1,F2,...,Fn-1)
\end{aligned}
$$

这就意味着，变量C的条件分布可以表达为：

$$p(C|F1,...,Fn)=\frac{1}{Z}p(C)*\prod_{i}^{n}p(Fi|C)$$

其中，Z只依赖 F1,...,FnF1,...,Fn ,当特征变量已知时Z是个常数。

至此，我们我们可以从概率模型中构造分类器，朴素贝叶斯分类器包括了这种模型和相应的决策规则。一个普通的规则就是选出最有可能的那个：这就是大家熟知的最大后验概率（MAP）决策准则。

相应的分类器便是如下定义的公式：

$$classify(f1,..,fn) = argmax\, p(C = c) \prod_i^n p(Fi = fi|C = c)$$

当特征值为离散型时：

类的先验概率可以通过训练集的各类样本的出现次数来估计，例如：

类的先验概率 $\frac{1}{\text{样本总数}}$ p(C=c1)=1样本总数p(C=c1)=\frac{1}{样本总数}

类条件概率= $\frac{\text{Fi=fi的样本总量}}{\text{样本总数}}$ p(Fi=fi|C=c1)=Fi=fi的样本总量样本总数 p(Fi=fi|C=c1)=\frac{Fi=fi的样本总量}{样本总数}

即可求得类的条件概率，最后比较各个类别的概率值的大小判断该测试样本应该属于哪个类别。

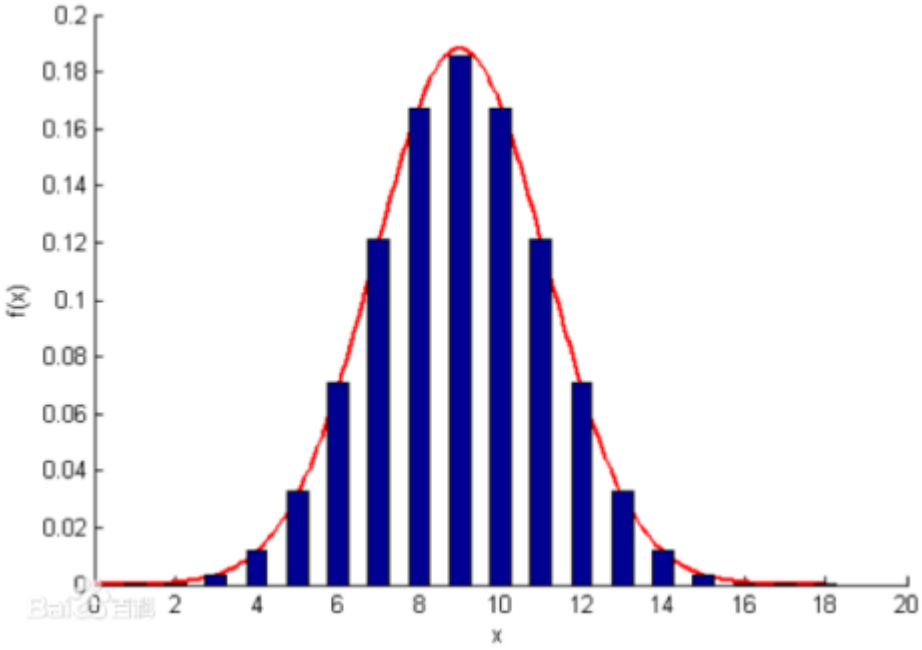
当特征值为连续型时：

通常的假设这些连续数值为高斯分布。例如，假设训练集中某个连续特征 xx 。首先我们对数据类别分类，然后计算每个类别中的 xx 均值和方差。令 μc\mu _{c} 表示为 xx 在**c**类上的均值，σc2\sigma _{c}^{2} 表示为 xx 在**c**类上的方差。在给定类中某个值的概率 （ ） p=（x=v|c）p=（x=v|c），可以通过将 vv 表示为均值为 μc\mu _{c} 方差为 σc2\sigma _{c}^{2} 的正态分布计算出来。如下，

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

处理连续数值问题的另一种常用的技术是通过离散化连续数值的方法，通常，当训练样本数量较少或者是精确的分布已知时，通过概率分布的方法是一种更好的选择。

在大量样本的情形下离散化的方法表现更优，因为大量的样本可以学习到数据的分布。由于朴素贝叶斯是一种典型的用到大量样本的方法（越大计算量的模型可以产生越高的分类精确度），所以朴素贝叶斯方法都用到离散化方法，而不是概率分布估计的方法。



三、实例（举个栗子）

这里举两个例子，一个离散型的，一个连续型的。

离散型实例，给定如下数据：

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	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
不帅	不好	矮	不上进	不嫁
帅	好	高	不上进	嫁
不帅	好	高	上进	嫁
帅	好	高	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

根据以上数据，现在有一堆男女朋友，男生向女生求婚，男生的四个特点分别是不帅，性格不好，身高矮，不上进，请判断女生是嫁还是不嫁?

该问题转换为数学问题就是比较

（嫁 不帅，性格不好，身高矮，不上进）p（嫁|不帅，性格不好，身高矮，不上进）p（嫁|不帅，性格不好，身高矮，不上进）

与

（不嫁 不帅，性格不好，身高矮，不上进）p（不嫁|不帅，性格不好，身高矮，不上进）p（不嫁|不帅，性格不好，身高矮，不上进） 的概率。

由贝叶斯公式得：

（嫁丨不帅，性格不好，身高矮，不上进）$\frac{\text{（不帅，性格不好，身高矮，不上进丨嫁）}\quad \text{嫁}}{\text{（嫁丨不帅，性格不好，身高矮，不上进}}$ p（嫁丨不帅，性格不好，身高矮，不上进）=p（不帅，性格不好，身高矮，不上进丨嫁）∗p(嫁)p（嫁丨不帅，性格不好，身高矮，不上进)p（嫁丨不帅，性格不好，身高矮，不上进）=\frac{p（不帅，性格不好，身高矮，不上进丨嫁）*p(嫁)}{p（嫁丨不帅，性格不好，身高矮，不上进)}

假设各个特征相互独立，即：

（嫁丨不帅，性格不好，身高矮，不上进）

$\frac{\text{不帅丨嫁}\quad \text{性格不好丨嫁}\quad \text{身高矮丨嫁}\quad \text{不上进丨嫁}\quad \text{嫁}}{\text{（不帅，性格不好，身高矮，不上进）}}$ p（嫁丨不帅，性格不好，身高矮，不上进）=p(不帅丨嫁)∗p(性格不好丨嫁)∗p(身高矮丨嫁)∗p(不上进丨嫁)∗p(嫁)p（不帅，性格不好，身高矮，不上进）p（嫁丨不帅，性格不好，身高矮，不上进）=\frac{p(不帅丨嫁)*p(性格不好丨嫁)*p(身高矮丨嫁)*p(不上进丨嫁)*p(嫁)}{p（不帅，性格不好，身高矮，不上进）}

首先我们整理训练数据中：

嫁的样本数总共有6个，则 $\overline{\text{嫁}}$ p(嫁)=612=12p(嫁)=\frac{6}{12}=\frac{1}{2} ；

不帅，也嫁了的样本数总共有6个，则 $\overline{\text{嫁}}$ p(嫁)=36=12p(嫁)=\frac{3}{6}=\frac{1}{2} ；

性格不好，也嫁了的样本数总共有1个，则 $\overline{\text{嫁}}$ p(嫁)=16p(嫁)=\frac{1}{6} ；

矮，也嫁了的样本数总共有1个，则 $\overline{\text{嫁}}$ p(嫁)=16p(嫁)=\frac{1}{6} ；

不上进，也嫁了的样本数总共有1个，则 $\overline{\text{嫁}}$ p(嫁)=16p(嫁)=\frac{1}{6} ；

$$p(\text{嫁}|\text{不帅，性格不好，身高矮，不上进})$$

$$=\frac{p(\text{不帅}|\text{嫁})*p(\text{性格不好}|\text{嫁})*p(\text{身高矮}|\text{嫁})*p(\text{不上进}|\text{嫁})*p(\text{嫁})}{p(\text{不帅，性格不好，身高矮，不上进})}$$

$$=\frac{\frac{1}{2}*\frac{1}{6}*\frac{1}{6}*\frac{1}{6}*\frac{1}{2}}{p(\text{不帅，性格不好，身高矮，不上进})}$$

$$=\frac{\frac{1}{864}}{p(\text{不帅，性格不好，身高矮，不上进})}$$

同理：

不嫁的样本数总共有6个，则 $\overline{\text{不嫁}}$ $p(\text{不嫁})=\frac{6}{12}=\frac{1}{2}$ ；

不帅，就不嫁的样本数总共有1个，则 $\overline{\text{不帅}|\text{不嫁}}$ $p(\text{不帅}|\text{不嫁})=\frac{1}{6}$ ；

性格不好，就不嫁的样本数总共有3个，则 $\overline{\text{不帅}|\text{不嫁}}$ $p(\text{不帅}|\text{不嫁})=\frac{3}{6}=\frac{1}{2}$ ；

矮，就不嫁的样本数总共有6个，则 $\overline{\text{矮}|\text{不嫁}}$ $p(\text{矮}|\text{不嫁})=\frac{6}{6}=1$ ；

不上进，就不嫁的样本数总共有3个，则 $\overline{\text{不上进}|\text{不嫁}}$ $p(\text{不上进}|\text{不嫁})=\frac{3}{6}=\frac{1}{2}$ ；

由于分母都相同，且分子 $\overline{\text{嫁}}$ $\overline{\text{不嫁}}$ $\frac{1}{864}<\frac{1}{48}$ ，所以走后得出的结论是该女生不嫁给这个男生。

连续型实例，给定训练数据如下：

性别	身高(英尺)	体重(磅)	脚的尺寸(英寸)
男	6	180	12
男	5.92 (5'11")	190	11
男	5.58 (5'7")	170	12
男	5.92 (5'11")	165	10
女	5	100	6
女	5.5 (5'6")	150	8
女	5.42 (5'5")	130	7
女	5.75 (5'9")	150	9

通过以上人体测量特征，包括身高、体重、脚的尺寸，判断一个人是男性还是女性。

测试样本：

性别	身高(英尺)	体重(磅)	脚的尺寸(英尺)
sample	6	130	8

首先，假设人的身高，体重，脚的尺寸都满足高斯分布，分别计算各个特征的均值和方差，得到下表：

性别	均值(身高)	方差(身高)	均值(体重)	方差(体重)	均值(脚的尺寸)	方差(脚的尺寸)
男性	5.855	3.5033e−02	176.25	1.2292e+02	11.25	9.1667e−01
女性	5.4175	9.7225e−02	132.5	5.5833e+02	7.5	1.6667e+00

其次，我们认为先验概率是男性或者是女性是等概率的，

即 p(male)=p(female)=0.5p(male)=p(female)=0.5 ，或者通过统计样本中男女比例来作为先验概率也可以，本例得到的结果是一样的。

判断该条测试样本属于男性还是女性，就等价于比较是男性的后验概率和女性的后验概率哪个大。

$$posterior(male) = \frac{p(male) * p(height|male) * p(weight|male) * p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{p(female) * p(height|female) * p(weight|female) * p(footsize|female)}{evidence}$$

分母是个常数，只需要比较分子就行，这里给出分母的值：

$$evidence = p(male) * p(height|male) * p(weight|male) * p(footsize|male) +$$

$$p(female) * p(height|female) * p(weight|female) * p(footsize|female)$$

计算男性的后验概率： p(male)=0.5p(male)=0.5 ，其中

$$p(height|male) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(6-\mu)^2}{2\sigma^2}} \approx 1.5789, \quad 其中\mu = 5.855, \quad \sigma^2 = 3.5033e^{-2},$$

这里大于1是因为是概率密度函数，而不是概率分布函数，所以大于1也是合理的。

$$(weight|male) = 5.9881e^{-6}$$

$$(footsize|male) = 1.3112e^{-3}$$

$$posterior(male) = \frac{6.1984e^{-9}}{evidence}$$

$$p(female) = 0.5$$

$$(height|female) = 2.2346e^{-1}$$

$$(weight|female) = 1.6789e^{-2}$$

$$(footsize|female) = 2.8669e^{-1}$$

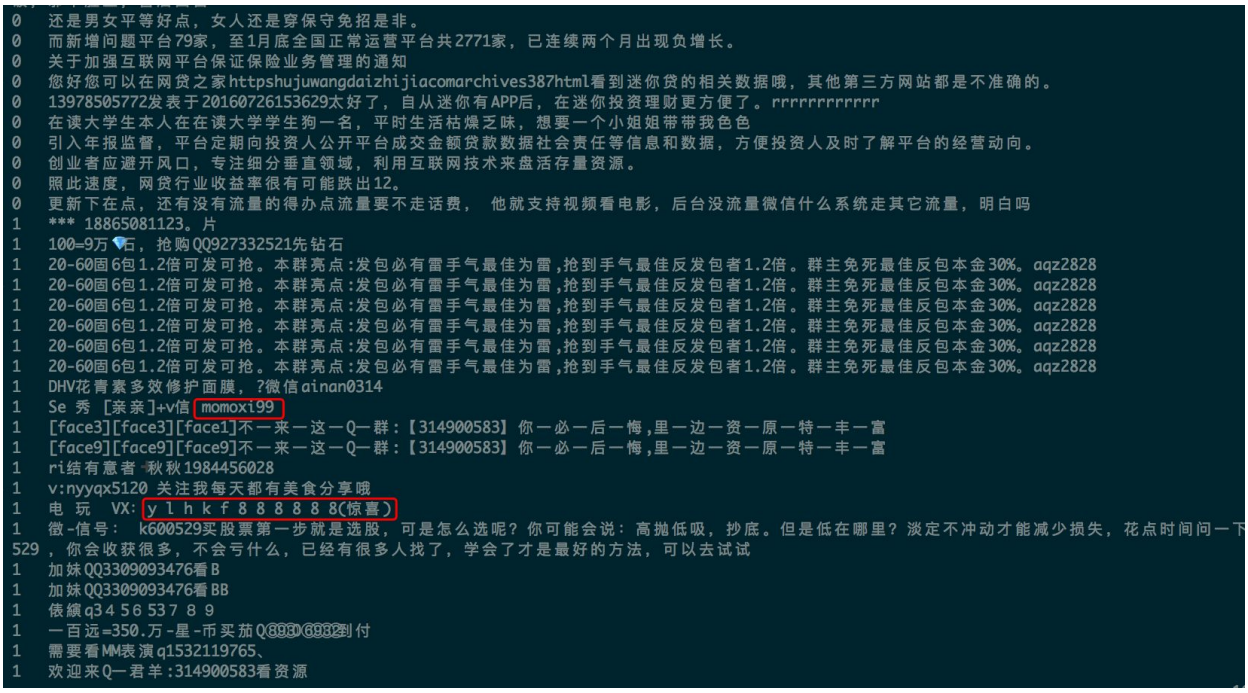
$$posterior(female) = \frac{5.3778e^{-4}}{evidence}$$

由于女性的后验概率分子较大，所以我们预测这个样本的为女性。

四、实战运用

达观数据作为在文本智能处理领域的领先企业，在文本挖掘领域有着深厚的技术底蕴。在实际的工作中有大量的文本分类的实际运用场景。下面将举一个常见的例子：广告识别。

这里有一批已经标注好的用户评论数据如下图，我们定义有联系方式的评论为广告，即广告识别等价于识别是否包含联系方式。



```
0   还是男女平等好点，女人还是穿保守免招是非。
0   而新增问题平台79家，至1月底全国正常运营平台共2771家，已连续两个月出现负增长。
0   关于加强互联网平台保证保险业务管理的通知
0   您好您可以在网贷之家httpshujuwangdaizhijiacomarchives387html看到迷你贷的相关数据哦，其他第三方网站都是不准确的。
0   13978505772发表于20160726153629太好了，自从迷你有APP后，在迷你投资理财更方便了。rrrrrrrrrrr
0   在读大学生本人在在读大学学生狗一名，平时生活枯燥乏味，想要一个小姐姐带带我色色
0   引入年报监督，平台定期向投资人公开平台成交金额贷款数据社会责任等信息和数据，方便投资人及时了解平台的经营动向。
0   创业者应避开风口，专注细分垂直领域，利用互联网技术来盘活存量资源。
0   照此速度，网贷行业收益率很有可能跌出12。
0   更新下在点，还有没有流量的得办点流量要不走话费， 他就支持视频看电影，后台没流量微信什么系统走其它流量，明白吗
1   *** 18865081123。片
1   100=9万💎石，抢购QQ927332521先钻石
1   20-60固6包1.2倍可发可抢。本群亮点:发包必有雷手气最佳为雷,抢到手气最佳反发包者1.2倍。群主免死最佳反包本金30%。aqz2828
1   20-60固6包1.2倍可发可抢。本群亮点:发包必有雷手气最佳为雷,抢到手气最佳反发包者1.2倍。群主免死最佳反包本金30%。aqz2828
1   20-60固6包1.2倍可发可抢。本群亮点:发包必有雷手气最佳为雷,抢到手气最佳反发包者1.2倍。群主免死最佳反包本金30%。aqz2828
1   20-60固6包1.2倍可发可抢。本群亮点:发包必有雷手气最佳为雷,抢到手气最佳反发包者1.2倍。群主免死最佳反包本金30%。aqz2828
1   20-60固6包1.2倍可发可抢。本群亮点:发包必有雷手气最佳为雷,抢到手气最佳反发包者1.2倍。群主免死最佳反包本金30%。aqz2828
1   20-60固6包1.2倍可发可抢。本群亮点:发包必有雷手气最佳为雷,抢到手气最佳反发包者1.2倍。群主免死最佳反包本金30%。aqz2828
1   DHV花青素多效修护面膜，?微信ainan0314
1   Se 秀 [亲亲]+v信 momoxi99
1   [face3][face3][face1]不一来一这一Q一群：【314900583】你一必一后一悔,里一边一资一原一特一丰一富
1   [face9][face9][face9]不一来一这一Q一群：【314900583】你一必一后一悔,里一边一资一原一特一丰一富
1   ri结有意者 秋秋1984456028
1   v:nyyqx5120 关注我每天都有美食分享哦
1   电 玩  VX: y l h k f 8 8 8 8 8 8(惊喜)
1   微-信号：  k600529买股票第一步就是选股，可是怎么选呢？你可能会说：高抛低吸，抄底。但是低在哪里？淡定不冲动才能减少损失，花点时间问一下
529 ，你会收获很多，不会亏什么，已经有很多人找了，学会了才是最好的方法，可以去试试
1   加妹QQ3309093476看B
1   加妹QQ3309093476看BB
1   俵縯q34 56 537 8 9
1   一百远=350.万-星-币买茄Q⑧⑨③⑥⑨③②到付
1   需要看MM表演q1532119765、
1   欢迎来Q一君羊:314900583看资源
```

首先，将每条评论转换为词向量，这里我们使用python的分词包jieba来进行分词，例如上图中的某条评论：”您好您可以在网贷之家httpshujuwangdaizhijiacomarchives387html看到迷你贷的相关数据哦，其他第三方网站都是不准确的。”

分词结果为：“您好/您/可以/在/网贷/之家/httpshujuwangdaizhijiacomarchives387html/看到/迷你/贷/的/相关/数据/哦/，/其他/第三方/网站/都/是/不/准确/的/。”

```
[siterecoffline - src]$ python jieba_fenci.py
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 0.303 seconds.
Prefix dict has been built succesfully.
您好/您/可以/在/网贷/之家/httpshujuwangdaizhijiacomarchives387html/看到/迷你/贷/的/相关/数据/哦/，/其他/第三方/网站/都/是/不/准确/的/。
```

然后，将语料库(corpus)中所有的评论分词后的词语作为一个集合，称为词袋(Bag of words)。计算每条评论中每个词语的TF-IDF值，TF-IDF公式为：

$$TF = \frac{\text{该词语出现在本评论中的次数}}{\text{总评论数}},\quad \text{这里除以总评论数是将词频归一化了.}$$

$$IDF = log^{\frac{\text{总评论数}}{\text{包含该词语的评论数+1}}},\quad \text{这里+1 是为了避免分母为 0.}$$

$$TF-IDF = TF * IDF$$

TF-IDF值的意义是：一个词语在评论中出现次数越多, 同时在所有评论中出现次数越少, 越能够代表该评论。每条评论根据词袋构建一个数字向量，向量长度为词袋的词语总数，每个词对应一维特征，如果该词在这条评论中，则这维特征是这个词语的TF-IDF值；如果该词不在评论中，则这维特征为0。这样就将每条评论转换为一个向量。

假设特征之间是相互独立的，该例子就转换为连续型的贝叶斯分类器。当然，这里也可以用这个词语的词频来作为特征，这时该例子为离散型的贝叶斯分类器，这里我们用TF-IDF值。

本例子使用scikit-learn中的Naive Bayes模块，这个模块中有三个训练模块：GaussianNB、MultinomialNB、BernoulliNB，分别是高斯朴素贝叶斯、多项式分布朴素贝叶斯和伯努利朴素贝叶斯。多项式分布是将重复词语是为其重复多次，伯努利朴素贝叶斯是将重复的词语视为其只出现1次，本例子是连续型的这里我们用高斯朴素贝叶斯。将转换后的数据90%作为训练集，10%作为测试集，部分代码如图：

```
def train(self, file_name):
    corpus, wordseg, target = self.load_corpus(file_name)

    kf = KFold(corpus.shape[0], n_folds=3, random_state=1)
    bad_case_after = set([])
    bad_case_before = set([])
    score = 0.0
    post_score = 0.0
    for train_index, test_index in kf:
        y_post_process = []
        X_train, X_test = corpus[train_index], corpus[test_index]
        y_train, y_test = target[train_index], target[test_index]
        seg_test = wordseg[test_index]

        clf = GaussianNB().fit(X_train, y_train.ravel())
        predict_proba = clf.predict(X_test)
        y_predict = np.argmax(predict_proba, axis=1)
        bad_case_before = bad_case_before.union(set(self.bad_case(seg_test, y_test, y_predict)))

        print 'Test without post process:'
        print metrics.classification_report(y_test, y_predict, target_names=self.mlb.classes_)
        score += metrics.accuracy_score(y_test, y_predict)

        self.ppcount = 0
        for (scores_list, seg_result) in zip(predict_proba, seg_test):
            words = seg_result.split(' ')
            score_list = self.post_process(scores_list, words)
            y_post_process.append(np.argmax(score_list))

        y_post_process = np.array(y_post_process)
        bad_case_after = bad_case_after.union(set(self.bad_case(seg_test, y_test, y_post_process)))

        print 'Test with post process:'
        print metrics.classification_report(y_test, y_post_process, target_names=self.mlb.classes_)
        post_score += metrics.accuracy_score(y_test, y_post_process)
```

进行交叉验证后得到的结果如图：

```
Test with post process:
              precision    recall  f1-score   support

           0       0.86      0.86      0.86      2165
           1       0.79      0.79      0.79      1414

avg / total       0.83      0.83      0.83      3579

Test without post process:
              precision    recall  f1-score   support

           0       0.82      0.87      0.85      2051
           1       0.82      0.75      0.78      1528

avg / total       0.82      0.82      0.82      3579

Test with post process:
              precision    recall  f1-score   support

           0       0.82      0.87      0.85      2051
           1       0.82      0.75      0.78      1528

avg / total       0.82      0.82      0.82      3579

k-fold cross validation time cost: 1312.03389168ms
model saved
```

五、综述

贝叶斯分类器是一种生成式模型，通过计算概率来进行分类，可以用来处理多分类问题，对于小规模的数据预测，同样表现良好。

贝叶斯分类器适合多分类任务，适合增量式训练，对于大规模数据，计算复杂度较低，同时算法原理比较简单易懂。但缺点是，对输入数据比较敏感，而且贝叶斯分类器是假设特征之间相互独立，而往往实际例子中特征之间都有相互联系，所以对于特征之间相关性较强的运用场景，准确率上会有一定损失；并且连续型的特征是假设该特征满足高斯分布，同样会带来一定准确率上的损失。

所以在实际运用当中，充分考虑特征之间的相关性和特征的分布情况是至关重要的。

References

1. 李贤平.概率论基础（第三版）[M]. 高等教育出版社，2010.
2. 李航．统计学习方法[M]. 北京：清华大学出版社 2012.
3. 钟波 刘琼荪．数理统计[M]. 高等教育出版社，2012.

4. Domingos, Pedro; Pazzani, Michael. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. 1997, 29: 103–137.
5. Webb, G. I.; Boughton, J.; Wang, Z. Not So Naive Bayes: Aggregating One-Dependence Estimators. Machine Learning (Springer). 2005, 58 (1): 5–24. doi:10.1007/s10994-005-4258-6.