

至道

本心、恒心、敬畏心

数据挖掘算法之深入朴素贝叶斯分类

2017年03月26日 22:04:44

阅读数：8200

写在前面的话：

我现在大四，毕业设计是做一个基于大数据的用户画像研究分析。所以开始学习数据挖掘的相关技术。这是我学习的一个新技术领域，学习难度比我以往学过的所有技术都难。虽然现在在一家公司实习，但是工作还是挺忙的，经常要加班，无论工作多忙，还是决定要写一个专栏，这个专栏就写一些数据挖掘算法、数据结构、算法设计和分析的相关文章。通过写博文来督促自己不断学习。以前对于数学没有多大的兴趣爱好，从小到大，学数学也是为了考试能考个好的成绩，学过的很多数学知识，并没有深刻的感受到它的用途，不用也就慢慢遗忘，但自从我看了数学之美这本书和开始学习数据挖掘后，使我对数学有了很大的兴趣。数学源于生活，用于生活。数据挖掘中涉及到很多统计学、线性代数、微积分等相关知识，而我的很多数学知识都已经还给我以前的老师了，所以现在只能慢慢一点一点捡起来。要感谢网上有很多作者写出的好的文章，让我受益匪浅，也算是站在他们的肩膀上学习。减少了我的学习困难，而我今天开始写的专栏里的一系列文章，很多例子都是借鉴于他们文章中的例子。想了想，这个专栏名称就叫<<算法大杂烩>>，以后我会把我工作中用到的、自己学习的新算法、以及回顾梳理的每一个算法的学习笔记和心得都更新到这个专栏里。写的博文难免会有写得不好的地方，欢迎大家指正，我也喜欢和有共同学习爱好的人一起学习交流。

不一定每天都更新，但是肯定会坚持写下去。

今天写的第一篇博文，是关于朴素贝叶斯分类的。几年前，我就听说过这个算法，只是稍微了解一点点，仅仅停留在只知道它是通过贝叶斯定理来分类的。写这篇文章之前，我看了很多的相关知识，包括书籍和网上的一些优秀的博文。哈哈，到现在也应该算对于这个算法入门了吧。后面的参考链接中会附上一些参考的文章地址。

朴素贝叶斯分类

引子

朴素贝叶斯分类是一种常用的分类算法，他根据研究对象的某些特征，来推断出该研究对象属于该研究领域的哪个类别。

数学是解决我们生活中产生的各种问题的。所以，数学源于生活，生活中也处处体现数学，我们编程，不过是把人能够理解的数学知识转换成计算机能够理解的形式来解决实际问题。拿朴素贝叶斯分类来说，其实生活中比比皆是，举个例子：

我们在大街上看到一个人，猜测他属于哪个职业。这就是一种分类，你是根据什么来判断的。可能是根据这个人的穿着打扮，言行举止。

穿着打扮：胡子拉碴、头发乱七八糟，背着大的电脑包

言行举止：双眼无神（估计在想哪个bug的解决办法），黑眼圈重，头发没洗。

所以，我大概能看出这个人职业是程序员（开个玩笑，这只是程序员自黑而已，我身边的程序员都不是这样的，当然也包括我）。

其实穿着打扮、言行举止就是人的特征属性

我们要对某个对象分类，必须根据他的特征属性来判断。

概述

要了解贝叶斯分类，必须了解贝叶斯定理，贝叶斯定理离不开条件概率

条件概率定义：

事件A在另外一个事件B已经发生条件下的发生概率。条件概率表示为 $P(A|B)$ ，读作“在B条件下A的概率”。根据文氏图，可以很清楚地看到在事件B发生的情况下，事件A发生的概率就是 $P(A \cap B)$ 除以 $P(B)$ 。



根据文氏图，可以很清楚地看到在事件B发生的情况下，事件A发生的概率就是 $P(A \cap B)$ 除以 $P(B)$ 。

$$P(A|B) = P(A \cap B) / P(B)$$

因此，

$$P(A \cap B) = P(A|B)P(B)$$

所以，

$$P(A|B)P(B) = P(B|A)P(A)$$

即

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

上面的公式就是贝叶斯定理

生活中，我们经常知道这种情况， $P(A|B)$ ，但是不知道 $P(B|A)$ 。比如：

A：表示用户收入高

B：表示订购2G流量套餐

$P(A|B)$ 表示订购2G流量套餐的用户收入高的概率，这个可以通过统计的样本算出得到。

但是现在有一个用户收入高（A），他购买2G流量套餐(B)的概率是多少，即 $P(B|A)$ ，这才是我们关注

所以通过贝叶斯定理，我们可以把这两者挂上钩，求出我们想知道的 $P(B|A)$

病人疾病预测分类例子

看一个简单的形式化的例子，来说明贝叶斯分类的作用

这个例子来自：阮一峰老师的介绍贝叶斯应用博文中的一个病人分类的例子

如下：其中特征属性是症状和职业，类别是疾病（包括感冒、过敏、脑震荡）
某个医院早上收了六个门诊病人，如下表。

症状	职业	疾病
打喷嚏	护士	感冒
打喷嚏	农夫	过敏
头痛	建筑工人	脑震荡
头痛	建筑工人	感冒
打喷嚏	教师	感冒
头痛	教师	脑震荡

现在又来了第七个病人，是一个打喷嚏的建筑工人。请问他患上感冒的概率有多大？

根据贝叶斯定理：

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

可得

$$\begin{aligned} 1 & P(\text{感冒} | \text{打喷嚏} \times \text{建筑工人}) \\ 2 & = P(\text{打喷嚏} \times \text{建筑工人} | \text{感冒}) \times P(\text{感冒}) \\ 3 & / P(\text{打喷嚏} \times \text{建筑工人}) \end{aligned}$$

假定"打喷嚏"和"建筑工人"这两个特征是独立的，因此，上面的等式就变成了

$$\begin{aligned} 1 & P(\text{感冒} | \text{打喷嚏} \times \text{建筑工人}) \\ 2 & = P(\text{打喷嚏} | \text{感冒}) \times P(\text{建筑工人} | \text{感冒}) \times P(\text{感冒}) \\ 3 & / P(\text{打喷嚏}) \times P(\text{建筑工人}) \end{aligned}$$

这是可以计算的。

$$\begin{aligned} P(\text{感冒} | \text{打喷嚏} \times \text{建筑工人}) \\ &= 0.66 \times 0.33 \times 0.5 / 0.5 \times 0.33 \\ &= 0.66 \end{aligned}$$

因此，这个打喷嚏的建筑工人，有66%的概率是得了感冒。同理，可以计算这个病人患上过敏或脑震荡的概率。比较这几个概率，就可以知道他最可能得什么病。

依据已有的统计数据，根据属性特征计算分类的概率，概率最大的类为分类结果

这就是贝叶斯分类器的基本方法：**在统计资料的基础上，依据找到的一些特征属性，来计算各个类别的概率，找到概率最大的类，从而实现分类。**

贝叶斯分类的定义

1、设 $x = \{a_1, a_2, \dots, a_m\}$ 为一个待分类项，而每个 a 为 x 的一个特征属性。

2、有类别集合 $C = \{y_1, y_2, \dots, y_n\}$ 。

3、计算 $P(y_1|x), P(y_2|x), \dots, P(y_n|x)$ 。

4、求出在 X 个属性条件下，所有类别的概率，选取概率最大的。则 X 属于概率最大的类别

$$P(y_k|x) = \max\{P(y_1|x), P(y_2|x), \dots, P(y_n|x)\}, \text{ 则 } x \in y_k。$$

根据贝叶斯定理，要求 $P(A|B)$ ，只要求出 $P(B|A)$ 即可。这里 Y 指 A , X 指 B 。把 B 分解为各个特征属性，求出每个类别的每个特征属性即可，如下

1、找到一个已知分类的待分类项集合，这个集合叫做训练样本集。

2、统计得到在各类别下各个特征属性的条件概率估计。即 $P(a_1|y_1), P(a_2|y_1), \dots, P(a_m|y_1); P(a_1|y_2), P(a_2|y_2), \dots, P(a_m|y_2); \dots; P(a_1|y_n), P(a_2|y_n), \dots, P(a_m|y_n)$ 。

3、如果各个特征属性是条件独立的，则根据贝叶斯定理有如下推导：

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

因为分母对于所有类别为常数，因为我们只要将分子最大化皆可。又因为各特征属性是条件独立的，所以有：

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i)\dots P(a_m|y_i)P(y_i) = P(y_i) \prod_{j=1}^m P(a_j|y_i)$$

上式等号右边的每一项，都可以从统计资料中得到，由此就可以计算出每个类别对应的概率，从而找出最大概率的那个类。

注意：各个特征属性是条件独立的，这是朴素贝叶斯所要求的，如果各个特征属性不独立，就不属于朴素贝叶斯，属于贝叶斯网络，后面的文章会讲解。

账号真假检测例子

再看一个例子，该例子来自网上张阳的算法杂货铺博文

根据某社区网站的抽样统计，该站10000个账号中有89%为真实账号（设为C0），11%为虚假账号（设为C1）。

C0 = 0.89

C1 = 0.11

接下来，就要用统计资料判断一个账号的真实性。假定某一个账号有以下三个特征：

F1: 日志数量/注册天数

F2: 好友数量/注册天数

F3: 是否使用真实头像（真实头像为1，非真实头像为0）

F1 = 0.1

F2 = 0.2

F3 = 0

请问该账号是真实账号还是虚假账号？

方法是使用朴素贝叶斯分类器，计算下面这个计算式的值。

$P(F1|C)P(F2|C)P(F3|C)P(C)$

虽然上面这些值可以从统计资料得到，但是这里有一个问题：F1和F2是连续变量，不适宜按照某个特定值计算概率。

一个技巧是将连续值变为离散值，计算区间的概率。比如将F1分解成[0, 0.05]、(0.05, 0.2)、[0.2, +∞]三个区间，然后计算每个区间的概率。在我们这个例子中，F1等于0.1，落在第二个区间，所以计算的时候，就使用第二个区间的发生概率。

根据统计资料，可得：

$P(F1|C0) = 0.5, P(F1|C1) = 0.1$

$P(F2|C0) = 0.7, P(F2|C1) = 0.2$

$P(F3|C0) = 0.2, P(F3|C1) = 0.9$

?

因此，

$$\begin{aligned} P(F1|C0) P(F2|C0) P(F3|C0) P(C0) \\ = 0.5 \times 0.7 \times 0.2 \times 0.89 \\ = 0.0623 \end{aligned}$$

$$\begin{aligned} P(F1|C1) P(F2|C1) P(F3|C1) P(C1) \\ = 0.1 \times 0.2 \times 0.9 \times 0.11 \\ = 0.00198 \end{aligned}$$

可以看到，虽然这个用户没有使用真实头像，但是他是真实账号的概率，比虚假账号高出30多倍，因此判断这个账号为真。

贝叶斯分类的含义

长久以来，人们对一件事情发生或不发生的概率 θ ，只有固定的0和1，即要么发生，要么不发生，从来不会去考虑某件事情发生的概率有多大，不发生的概率又是多大。比如如果问那时的人们一个问题：“有一个袋子，里面装着若干个白球和黑球，请问从袋子中取得白球的概率 θ 是多少？”他们会想都不用想，会立马告诉你，取出白球的概率 θ 就是1/2，要么取到白球，要么取不到白球，即 θ 只能有一个值，不是1/2，就是0，而且不论你取了多少次，取得白球的概率 θ 始终都是1/2，即不随观察结果 X 的变化而变化。

直到贝叶斯定理的出现，贝叶斯定理不把概率 θ 看做一个固定的值（比如上面取白球的概率一直都是1/2），而看做一个随机变量，他会随着观察结果变化

贝叶斯及贝叶斯派提出了一个思考问题的固定模式：

- 先验分布 $\pi(\theta)$ + 样本信息 $\chi \Rightarrow$ 后验分布 $\pi(\theta|x)$

上述思考模式意味着，新观察到的样本信息将修正人们以前对事物的认知。换言之，在得到新的样本信息之前，人们对 θ 的认知是先验分布 $\pi(\theta)$ ，在得到新的样本信息 χ 后，人们对 θ 的认知为 $\pi(\theta|x)$ 。

条件概率公式进行变形，可以得到如下形式：

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

我们把 $P(A)$ 称为“先验概率”，即在事件 B 发生之前，事件 A 发生的概率，在事件 B 发生之前，它是一个无条件分布，因为 A 还没有与事件 B 关联上，他是先验分布。

$P(A|B)$ 称为“后验概率”（Posterior probability），即在 B 事件发生之后，我们对 A 事件概率的重新评估。 $P(B|A)/P(B)$ 称为“可能性函数”（Likelihood），这是一个调整因子，使得预估概率更接近真实概率。它的分布就是后验分布

所以，条件概率可以理解成下面的式子：

$$\text{后验概率} = \text{先验概率} \times \text{调整因子}$$

这就是贝叶斯推断的含义。我们先预估一个“先验概率”，然后加入实验结果，看这个实验到底是增强还是削弱了“先验概率”，由此得到更接近事实的“后验概率”。

在这里，如果“可能性函数” $P(B|A)/P(B) > 1$ ，意味着“先验概率”被增强，事件 A 的发生的可能性变大；如果“可能性函数” $= 1$ ，意味着 B 事件无助于判断事件 A 的可能性；如果“可能性函数” < 1 ，意味着“先验概率”被削弱，事件 A 的可能性变小。

水果糖问题例子

这个例子同样来自阮一峰老师的博文，加深对贝叶斯推断的理解

第一个例子。两个一模一样的碗，一号碗有30颗水果糖和10颗巧克力糖，二号碗有水果糖和巧克力糖各20颗。现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大？

我们假定，H1表示一号碗，H2表示二号碗。由于这两个碗是一样的，所以 $P(H1)=P(H2)$ ，也就是说，在取出水果糖之前，这两个碗被选中的概率相同。因此， $P(H1)=0.5$ ，我们把这个概率就叫做"先验概率"，即没有做实验之前，来自一号碗的概率是0.5。

再假定，E表示水果糖，所以问题就变成了在已知E的情况下，来自一号碗的概率有多大，即求 $P(H1|E)$ 。我们把这个概率叫做"后验概率"，即在E事件发生之后，对 $P(H1)$ 的修正。

根据贝叶斯定理，得到

$$P(H1|E)=P(H1) * P(E|H1)/P(E)$$

已知， $P(H1)$ 等于0.5， $P(E|H1)$ 为一号碗中取出水果糖的概率，等于0.75，那么求出 $P(E)$ 就可以得到答案。根据全概率公式（不懂全概率公式的可以查找相关资料理解），

$$P(E)=P(E|H1) * P(H1) + P(E|H2)*P(H2)$$

所以，将数字代入原方程，得到

$$P(H1|E)=0.5* 0.75/0.625=0.6$$

这表明，来自一号碗的概率是0.6。也就是说，取出水果糖之后，H1事件的可能性得到了增强。

连续型特征属性和零概率事件处理

上面讲的特征属性值，都是离散的，账号真假检测例子中把连续的转换成区间，每个区间也可以看成离散的，但是如果在不能这样转换的情况下怎么解决？如果特征属性值是不是离散的，而是连续的怎么办？

我们是站在巨人的肩膀上，前人早已经为我们找到了应对之策

当特征属性为连续值时，通常假定其值服从高斯分布（也称正态分布）。即：

$$g(x, \eta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\eta)^2}{2\sigma^2}}$$

$$\text{而 } P(a_k|y_i) = g(a_k, \eta_{y_i}, \sigma_{y_i})$$

因此只要计算出训练样本中各个类别中此特征项划分的各均值和标准差，代入上述公式即可得到需要的估计值。（ a_k 为观察到的属性值）

另一个需要讨论的问题就是当 $P(a|y)=0$ 怎么办，当某个类别下某个特征项划分没有出现时，就是产生这种现象，这会令分类器质量大大降低。为了解决这个问题，引入了**拉普拉斯校准**，它的思想非常简单，就是对没类别下所有划分（概率为零的）的计数加1，这样如果训练样本集数量充分大时，并不会对结果产生影响，并且解决了上述频率为零的尴尬局面。

买电脑是否和收入相关的例子

验证买电脑，是否和收入有关的场景下：

类 buys_computer=yes包含1000个元组，有0个元组income=low ,990个元组 income=medium,10个元组income=high,这些事件的概率分别是0, 0.990, 0.010.

有概率为0,肯定不行.使用拉普拉斯校准，对每个收入-值对应加1个元组，分别得到如下概率

$$\begin{aligned} 1/1003 &= 0.001 \\ 991/1003 &= 0.998 \\ 11/1003 &= 0.011 \end{aligned}$$

这些校准的概率估计与对应的未校准的估计很接近，但是避免了零概率值

性别分类的例子

再看一个阮一峰老师的朴素贝叶斯应用一文中摘自维基百科的例子，关于处理连续变量的另一种方法。

下面是一组人类身体特征的统计资料。

1	性别	身高（英尺）	体重（磅）	脚掌（英寸）
2	男	6	180	12
3	男	5.92	190	11
4	男	5.58	170	12
5	男	5.92	165	10
6	女	5	100	6
7	女	5.5	150	8
8	女	5.42	130	7
9	女	5.75	150	9

已知某人身高6英尺、体重130磅，脚掌8英寸，请问该人是男是女？

根据朴素贝叶斯分类器，计算下面这个式子的值。

P(身高|性别) x P(体重|性别) x P(脚掌|性别) x P(性别)

这里的困难在于，由于身高、体重、脚掌都是连续变量，不能采用离散变量的方法计算概率。而且由于样本太少，所以也无法分成区间计算。怎么办？

这时，可以假设男性和女性的身高、体重、脚掌都是正态分布，通过样本计算出均值和方差，也就是得到正态分布的密度函数。有了密度函数，就可以把值代入，算出某一点的密度函数的值。

比如，男性的身高是均值5.855、方差0.035的正态分布。所以，男性的身高为6英尺的概率的相对值等于1.5789（大于1并没有关系，因为这里是密度函数的值，只用来反映各个值的相对可能性）。

有了这些数据以后，就可以计算性别的分类了。
对于连续型特征属性，对每个特征属性计算正态分布的密度函数，得到该特征属性发生的概率

$$\begin{aligned} &P(\text{身高}=6|\text{男}) \times P(\text{体重}=130|\text{男}) \times P(\text{脚掌}=8|\text{男}) \times P(\text{男}) \\ &= 6.1984 \times e^{-9} \end{aligned}$$

$$P(\text{身高}=6|\text{女}) \times P(\text{体重}=130|\text{女}) \times P(\text{脚掌}=8|\text{女}) \times P(\text{女}) \\ = 5.3778 \times 10^{-4}$$

可以看到，女性的概率比男性要高出将近10000倍，所以判断该人为女性。

下一篇会写贝叶斯网络。

参考文章：

http://www.ruanyifeng.com/blog/2011/08/bayesian_inference_part_one.html

http://www.ruanyifeng.com/blog/2013/12/naive_bayes_classifier.html

<http://www.cnblogs.com/leoo2sk/archive/2010/09/17/naive-bayesian-classifier.html>

http://blog.csdn.net/zdy0_2004/article/details/41096141

参考书籍：数据挖掘概念与技术。数据挖掘十大算法，统计学概率论方面的数学知识

码字不易，转载请指明出自<http://blog.csdn.net/tanggao1314/article/details/66478782>

版权声明：本文为博主原创文章，转载请指明 <http://blog.csdn.net/tanggao1314/> <https://blog.csdn.net/tanggao1314/article/details/66478782>

个人分类：[算法大杂烩](#)

所属专栏：[大数据生态系统技术](#)

相关热词：[数据挖掘算法](#) [八大数据挖掘算法](#) [数据挖掘算法入门](#) [数据挖掘算法对错](#) [数据挖掘算法题](#)


[上一篇](#) [低版本中mysql不支持在limit语句中有子查询](#)

[下一篇](#) [数据挖掘算法之贝叶斯网络](#)


想对作者说点什么？

[我来说一句](#)

 TTcccCarrie 2017-06-06 21:58:46 #5楼
贝叶斯网络 没听过。长见识了。

 穷理何须格物 2017-03-31 13:42:01 #4楼
买电脑的例子中：“ $1/1003=0.001$ $999/1003=0.998$ $11/1003=0.011$ ” 这里增加一个元组后 $999/1003$ 应该是 $991/1003$ 吧？

[查看回复\(1\)](#)

 野木香 2017-03-29 21:50:51 #3楼
写的很好，贝叶斯网络 呢？

[查看回复\(1\)](#)

 野木香 2017-03-29 20:30:46 #2楼
 $P(A|B)P(B)=P(B|A)P(A)$ 如果更细致一些，更好。同理得出 $P(B|A)=P(B|A)P(A)$ 因为 $P(A \cap B)=P(B|A)P(A)$ 且 $P(B|A)=P(A \cap B)$ 所以 $P(A|B)P(B)=P(B|A)P(A)$

 江流天地外 2017-03-27 23:18:52 #1楼
谢谢博主，写得很精彩。就是有好多图片都看不到.....希望博主能再上传分享一下，谢谢

[查看回复\(1\)](#)