

朴素贝叶斯分类器的应用

作者： 阮一峰

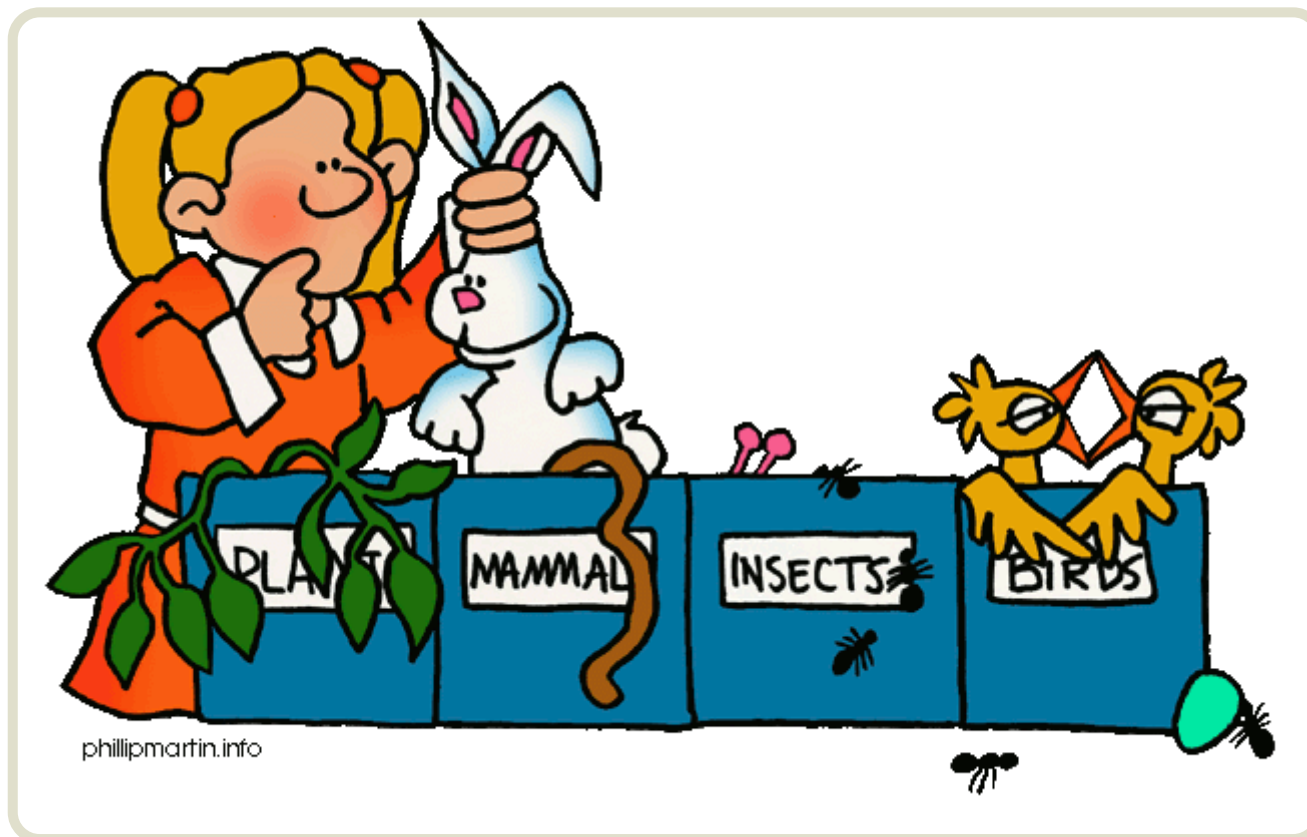
日期： 2013年12月16日



本站由 珠峰培训（专业前端培训）独家赞助

生活中很多场合需要用到分类，比如新闻分类、病人分类等等。

本文介绍[朴素贝叶斯分类器](#)（Naive Bayes classifier），它是一种简单有效的常用分类算法。



一、病人分类的例子

让我从一个例子开始讲起，你会看到贝叶斯分类器很好懂，一点都不难。

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

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

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

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

$$/ P(\text{打喷嚏}) \times P(\text{建筑工人})$$

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。

这就是贝叶斯分类器的基本方法：在统计资料的基础上，依据某些特征，计算各个类别的概率，从而实现分类。

二、朴素贝叶斯分类器的公式

假设某个体有 n 项特征（Feature），分别为 F_1 、 F_2 、...、 F_n 。现有 m 个类别（Category），分别为 C_1 、 C_2 、...、 C_m 。贝叶斯分类器就是计算出概率最大的那个分类，也就是求下面这个算式的最大值：

$$\begin{aligned} & P(C | F_1 F_2 \dots F_n) \\ &= P(F_1 F_2 \dots F_n | C) P(C) / P(F_1 F_2 \dots F_n) \end{aligned}$$

由于 $P(F_1F_2...F_n)$ 对于所有的类别都是相同的，可以省略，问题就变成了求

$$P(F_1F_2...F_n|C)P(C)$$

的最大值。

朴素贝叶斯分类器则是更进一步，假设所有特征都彼此独立，因此

$$\begin{aligned} &P(F_1F_2...F_n|C)P(C) \\ &= P(F_1|C)P(F_2|C) \dots P(F_n|C)P(C) \end{aligned}$$

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。

虽然"所有特征彼此独立"这个假设，在现实中不太可能成立，但是它可以大大简化计算，而且有研究表明对分类结果的准确性影响不大。

下面再通过两个例子，来看如何使用朴素贝叶斯分类器。

三、账号分类的例子

本例摘自张洋的[《算法杂货铺----分类算法之朴素贝叶斯分类》](#)。

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 C_0 ），11%为虚假账号（设为 C_1 ）。

$$C_0 = 0.89$$

$$C_1 = 0.11$$

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）

$$F_1 = 0.1$$

$$F_2 = 0.2$$

$$F_3 = 0$$

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。

$$P(F_1|C)P(F_2|C)P(F_3|C)P(C)$$

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 $[0, 0.05]$ 、 $(0.05, 0.2)$ 、 $[0.2, +\infty]$ 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：

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

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

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

因此，

$$P(F1|C0) \quad P(F2|C0) \quad P(F3|C0) \quad P(C0)$$

$$= 0.5 \times 0.7 \times 0.2 \times 0.89$$

$$= 0.0623$$

$$P(F1|C1) \quad P(F2|C1) \quad P(F3|C1) \quad P(C1)$$

$$= 0.1 \times 0.2 \times 0.9 \times 0.11$$

$$= 0.00198$$

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。

四、性别分类的例子

本例摘自[维基百科](#)，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。

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

$$P(\text{身高}|\text{性别}) \times P(\text{体重}|\text{性别}) \times P(\text{脚掌}|\text{性别}) \times P(\text{性别})$$

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）。

$$p(\text{height}|\text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6 - \mu)^2}{2\sigma^2}\right) \approx 1.5789$$

有了这些数据以后，就可以计算性别的分类了。

$$P(\text{身高}=6|\text{男}) \times P(\text{体重}=130|\text{男}) \times P(\text{脚掌}=8|\text{男}) \times P(\text{男})$$

$$= 6.1984 \times e^{-9}$$



$$P(\text{身高}=6|\text{女}) \times P(\text{体重}=130|\text{女}) \times P(\text{脚掌}=8|\text{女}) \times P(\text{女})$$

$$= 5.3778 \times e^{-4}$$

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。

（完）

文档信息

- 版权声明：自由转载-非商用-非衍生-保持署名（创意共享3.0许可证）
- 发表日期：2013年12月16日
- 更多内容：档案 » 算法与数学
- 文集：《前方的路》，《未来世界的幸存者》
- 社交媒体： twitter,  weibo



育精英前端 冲月薪3万



Tencent 腾讯



硅谷名企前端开发认证课程

免费试听 >

硅谷实战项目 | 人工逐行代码审阅 | 导师一对一辅导



相关文章

- **2017.12.13:** [图像与滤波](#)

我对图像处理一直很感兴趣，曾经写过好几篇博客（1，2，3，4）。

- **2017.08.02:** [正态分布为什么常见？](#)

统计学里面，正态分布（normal distribution）最常见。男女身高、寿命、血压、考试成绩、测量误差等等，都属于正态分布。

- **2017.07.13:** [神经网络入门](#)

眼下最热门的技术，绝对是人工智能。

- **2016.07.22:** [如何识别图像边缘？](#)

图像识别（image recognition）是现在的热门技术。

广告（购买广告位）

