贝叶斯定理与垃圾邮件过滤

文庆福
2011013239 thssvince@163.com
清华大学软件学院11班

2014年 6月 4日

目录

1 贝叶斯定理

1.1 历史

Thomas Bayes(1702-1763), 托马斯·贝叶斯是一位英国牧师数学家，1742年成为英国皇家学会会员，1761年4月7日逝世。他死后，理查德·普莱斯(Richard Price)于1763年将他的著作《机会问题的解法》发表。贝叶斯理论假设：如果事件的结果不确定，那么量化它的唯一方法就是事件的发生概率。如果过去试验中事件的出现率已知，那么根据数学方法可以计算出未来试验中事件出现的概率。贝叶斯定理可以用一个数学公式表达，即贝叶斯公式。

1.2 贝叶斯公式

那么，到底什么是贝叶斯公式呢？要理解贝叶斯公式必先了解什么是条件概率。
所谓“条件概率”（Conditional probability），就是指在事件B发生的情况下，事件A发生的概率，用$P(A|B)$来表示。
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$P(B)$，即$P(A|B) = \frac{P(A \cap B)}{P(B)}$， 故$P(A \cap B) = P(A|B)P(B)$。

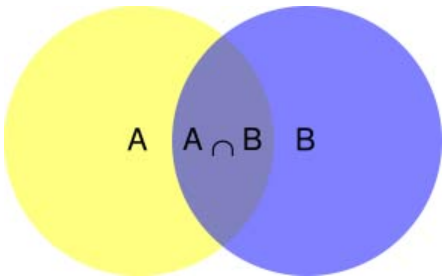

图 1: 条件概率

同理可知，$P(A \cap B) = P(B|A)P(A)$。 所以，

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

这便是贝叶斯公式。

2 垃圾邮件过滤

垃圾邮件过滤是具有相当难度的事情，垃圾邮件每天都在增加和变化。传统的垃圾邮件过滤方法，主要有“关键词法”和“校验码法”等。前者的过滤依据是特定的词

语；后者则是计算邮件文本的校验码，再与已知的垃圾邮件进行对比。它们的识别效果都不理想，而且很容易规避。垃圾邮件发送者只要简单的研究一下现在采用了哪些静态反垃圾邮件，然后相应的改变一下邮件的内容或发送方式，就可以逃避检查了。

因此，必须采用一种更强大的方法来克服静态反垃圾邮件的弱点，这种方法应该对垃圾邮件发送者的各种伎俩了如指掌，还要能适应不同用户对于反垃圾邮件的个性化需求。这种方法就是贝叶斯邮件过滤算法。

3 贝叶斯过滤算法

2002年，Paul Graham 提出使用“贝叶斯推断”过滤垃圾邮件[1]。实验表明，这种过滤方法的效果非常好，1000封垃圾邮件可以过滤掉995封，且没有一个误判。另外，这种过滤器还具有自我学习的功能，会根据新收到的邮件，不断调整。收到的垃圾邮件越多，它的准确率就越高。

3.1 建立邮件词频库

贝叶斯过滤器是一种统计学过滤器，建立在已有的统计结果之上。所以，我们必须预先提供两组已经识别好的邮件，一组是正常邮件，另一组是垃圾邮件。

我们用这两组邮件，对过滤器进行“训练”。这两组邮件的规模越大，训练效果就越好。Paul Graham使用的邮件规模，是正常邮件和垃圾邮件各4 000 封。

“训练”过程很简单。首先，解析所有邮件，提取每一个词。然后，计算每个词语在正常邮件和垃圾邮件中的出现频率。比如，我们假定“sex”这个词，在 4000 封垃圾邮件中，有 200 封包含这个词，那么它的出现频率就是 5%；而在4000封正常邮件中，只有2封包含这个词，那么出现频率就是0.05%。

3.2 贝叶斯公式的使用

现在，我们收到了一封新邮件。在未经统计分析之前，我们假定它是垃圾邮件的概率为 50%。（【注释】有研究表明，用户收到的电子邮件中，80% 是垃圾邮件。但是，这里仍然假定垃圾邮件的“先验概率”为 50%。）

我们用S表示垃圾邮件（spam），H表示正常邮件（healthy）。因此，$P(S)$和$P(H)$的先验概率，都是 50%。

$$P(H) = P(S) = 50\%$$

然后，对这封邮件进行解析，发现其中包含了sex这个词，请问这封邮件属于垃圾邮件的概率有多高？

我们用W表示“sex”这个词，那么问题就变成了如何计算$P(S|W)$的值，即在某个词语（W）已经存在的条件下，垃圾邮件（S）的概率有多大。

根据条件概率公式，马上可以写出

$$P(S|W) = \frac{P(W|S)P(S)}{P(W|S)P(S) + P(W|H)P(H)} \tag{2}$$

公式中，$P(W|S)$和$P(W|H)$的含义是，这个词语在垃圾邮件和正常邮件中，分别出现的概率。这两个值可以从历史资料库中得到，对sex这个词来说，上文假定它们分别等于5%和0.05%。另外，$P(S)$和$P(H)$的值，前面说过都等于50%。所以，马上可以计算$P(S|W)$的值：

$$P(S|W) = \frac{5\% * 50\%}{5\% * 50\% + 0.05\% * 50\%}$$

因此，这封新邮件是垃圾邮件的概率等于99%。这说明，sex这个词的推断能力很强，将50%的“先验概率”一下子提高到了99%的“后验概率”。

3.3 联合概率

做完上面一步，请问我们能否得出结论，这封新邮件就是垃圾邮件？

回答是不能。因为一封邮件包含很多词语，一些词语（比如sex）说这是垃圾邮件，另一些说这不是。你怎么知道以哪个词为准？

Paul Graham的做法是，选出这封信中$P(S|W)$最高的15个词，计算它们的联合概率。（如果有的词是第一次出现，无法计算$P(S|W)$，Paul Graham就假定这个值等于0.4。因为垃圾邮件用的往往都是某些固定的词语，所以如果你从来没见过某个词，它多半是一个正常的词。）

在计算联合概率过程中，我们需要作出了个非常简单粗暴的假设：邮件中每个单词的出现是独立事件（显然，这种假设是不完全正确的，但是这并不会影响邮件过滤）。基于这一假设（邮件中每个单词的出现是独立事件），根据贝叶斯定理可以推导出最终的计算公式：

$$P = \frac{P_1 P2 \cdots P_n}{P_1 P2 \cdots P_n + (1-P_1)(1-P_2)\cdots(1-P_n)} \tag{3}$$

其中，P_n表示$P(S|W_1)$，知道包含第一个单词W_1情况下，这封邮件是垃圾邮件的概率。其推导流程如下：

$$\begin{aligned} P &= P(S|W_1, W_2, \ldots, W_n) \\ &= \frac{P(W_1, W_2, \ldots, W_n|S)P(S)}{P(W_1, W_2, \ldots, W_n)} \\ &= \frac{P(W_1, W_2, \ldots, W_n|S)P(S)}{P(W_1, W_2, \ldots, W_n|S)P(S) + P(W_1, W_2, \ldots, W_n|H)P(H)} \\ &= \frac{1}{1 + \frac{P(W_1, W_2, \ldots, W_n|H)P(H)}{P(W_1, W_2, \ldots, W_n|S)P(S)}} \end{aligned}$$

由于每个单词出现事件W_i是相互独立的，所以上式可以改为：

$$\begin{aligned} P &= \frac{1}{1 + \frac{P(W_1, W_2, \ldots, W_n|H)P(H)}{P(W_1, W_2, \ldots, W_n|S)P(S)}} \\ &= \frac{1}{1 + \frac{P(W_1|H)P(W_2|H)\cdots P(W_n|H)P(H)}{P(W_1|S)P(W_2|S)\cdots P(W_n|S)P(S)}} \\ &= \frac{1}{1 + \frac{P(W_1|H)P(W_2|H)\cdots P(W_n|H)}{P(W_1|S)P(W_2|S)\cdots P(W_n|S)}} \end{aligned}$$

注：P(H)=P(S) = 0.5

又因为

$$P_i = P(S|W_i) = \frac{P(W_i|S)}{P(W_i|S) + P(W_i|H)}$$

所有

$$\frac{P(W_i|H)}{P(W_i|S)} == \frac{1}{P_i} - 1$$

将上式代入公式有：

$$P = \frac{1}{1 + (\frac{1}{P_1} - 1)(\frac{1}{P_2} - 1)\cdots(\frac{1}{P_n} - 1)} = \frac{P_1 P2 \cdots P_n}{P_1 P2 \cdots P_n + (1 - P_1)(1 - P_2)\cdots(1 - P_n)} \tag{4}$$

这样，计算出来的概率P值通常与给定的阈值进行比较来决定是否为垃圾邮件。如果P小于阈值，邮件被认为是正常邮件，否则被认为是垃圾邮件。

4 算法结果

下面以我今天所收到一封邮件为例：

2013～2014学年春季学期第二次SRT项目报名工作开始，学生报名时间为2014年6月3日至2014年6月8日，立项人集中审核学生报名时间为2014年6月9日至2014年6月11日。项目详细情况可在网上直接查询。报

名学生和立项人请用自己的学号或工作证号登录综合信息服务系统进行网上报名及报名审核。

报名及审核注意事项如下： 1．报名学生填写报名申请表后，要进行确认，只有在确认后立项人才能看到报名申请表； 2．在立项人同意接收之前或者不同意接收之后，学生可以取消自己的申请并申请其他项目，但在立项人同意接收之后，就不能再申请其他项目； 3．没有进行网上报名的学生，即使完成了SRT项目，也不能获得学分。 4．由立项人网上审批学生的报名申请表； 5. 学生立项人不用再登录系统报名，系统默认学生立项人为项目参与人之一。只有立项人参与的项目，立项人不用进行报名审核。 6.本次报名开放的项目为2012－2013学年春季学期———2013－2014学年春季学期立项的未结题项目，若项目接纳人数已满，不再接收学生报名，立项人可登录系统，点击“报名审核”，针对项目进行“结束报名”操作。

说明：自本学期开始，SRT管理启用新版系统，采用新的流程。每学期有两批项目立项、两次学生报名。本次开放报名为本学期的第二次项目报名。

下面是该邮件经过过滤算法测试之后的结果： $P(S|W)$最高的15个词以及其对应

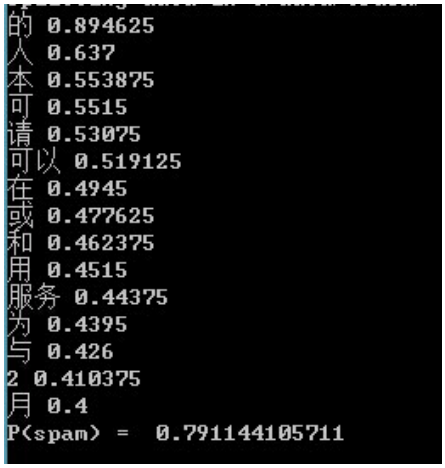

图 2: 运行结果

的$P(S|W)$，最后的概率是该邮件时垃圾邮件的概率。

5 参考资料

[1]. Paul Graham, a plan for spam, http://www.paulgraham.com/index.html

[2]. http://en.wikipedia.org/wiki/Bayesian_spam_filtering