




贝叶斯算法(bayesian)介绍

00748009 李怡文

- 
- 1、 贝叶斯算法概况简介
 - 2、 贝叶斯过滤算法的主要步骤
 - 3、 贝叶斯过滤算法举例
 - 4、 贝叶斯算法的应用
 - 5、 总结

贝叶斯算法概况简介

- 贝叶斯是基于概率的一种算法，是Thomas Bayes：一位伟大的数学大师所创建的。
- 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1763年4月7日逝世，1763年，他发表了贝叶斯统计理论，即根据已经发生的事件来预测事件发生的可能性，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。

贝叶斯算法概况简介

- 目前此种算法用于过滤垃圾邮件得到了广泛地好评。贝叶斯过滤器是基于“自我学习”的智能技术，能够使自己适应垃圾邮件制造者的新把戏，同时为合法电子邮件提供保护。在智能邮件过滤技术中，贝叶斯（Bayesian）过滤技术取得了较大的成功，被越来越多地应用在反垃圾邮件的产品中。

贝叶斯过滤算法的主要步骤

- 算法的基本思想：
- 根据已有的垃圾邮件和非垃圾邮件建立贝叶斯概率库，利用贝叶斯概率库分析预测新邮件为垃圾邮件的概率
- 优点：
- 1、纯粹根据统计学规律运作
- 2、可计算性强

贝叶斯过滤算法的主要步骤

- 1. 收集大量的垃圾邮件和非垃圾邮件，建立垃圾邮件集和非垃圾邮件集。
- 2. 提取邮件主题和邮件体中的独立字符串，例如 ABC32， ¥ 234等作为TOKEN串并统计提取出的TOKEN串出现的次数即字频。按照上述的方法分别处理垃圾邮件集和非垃圾邮件集中的所有邮件。

贝叶斯过滤算法的主要步骤

- 3. 每一个邮件集对应一个哈希表，
hashtable_good对应非垃圾邮件集而
hashtable_bad对应垃圾邮件集。表中存
储TOKEN串到字频的映射关系。

贝叶斯过滤算法的主要步骤

- 4. 计算每个哈希表中TOKEN串出现的概率
- $P = (\text{某TOKEN串的字频}) / (\text{对应哈希表的长度})$

贝叶斯过滤算法的主要步骤

- 5. 综合考虑hashtable_good和hashtable_bad，推断出当新来的邮件中出现某个TOKEN串时，该新邮件为垃圾邮件的概率。数学表达式为：
- A 事件 ---- 邮件为垃圾邮件；
- t_1, t_2, \dots, t_n 代表TOKEN 串

则 $P(A|t_i)$ 表示在邮件中出现TOKEN 串 t_i 时，该邮件为垃圾邮件的概率。

设 $P1(t_i) = (\text{t}_i \text{ 在 hashtable_good 中的值})$

$P2(t_i) = (\text{t}_i \text{ 在 hashtable_bad 中的值})$

则 $P(A|t_i) = P2(t_i) / [P1(t_i) + P2(t_i)]$;

贝叶斯过滤算法的主要步骤

- 6. 建立新的哈希表hashtable_probability存储TOKEN串 t_i 到 $P(A|t_i)$ 的映射
- 7. 至此，垃圾邮件集和非垃圾邮件集的学习过程结束。根据建立的哈希表hashtable_probability可以估计一封新到的邮件为垃圾邮件的可能性。

贝叶斯过滤算法的主要步骤

- 当新到一封邮件时，按照步骤2，生成TOKEN串。查询hashtable_probability得到该TOKEN 串的键值。
- 假设由该邮件共得到N个TOKEN 串，
 t_1, t_2, \dots, t_n , hashtable_probability中对应的值为 P_1 ,
 P_2 , P_N , $P(A|t_1, t_2, t_3, \dots, t_n)$ 表示在邮件中
同时出现多个TOKEN串 t_1, t_2, \dots, t_n 时，该邮件为垃圾邮件的概率。
- 由复合概率公式可得
$$P(A|t_1, t_2, t_3, \dots, t_n) = \frac{P_1 * P_2 * \dots * P_N}{P_1 * P_2 * \dots * P_N + (1 - P_1) * (1 - P_2) * \dots * (1 - P_N)}$$

设B: t1 ,t2, t3.....tn全发生

$$P(A|t1 ,t2, t3.....tn)=P(A|B) = P(AB)/P(B)$$

$$P(AB)=P(A t1)*P(A t2)...P(A tn)=P1*P2*...PN$$

$$P(A_cB)=P(A_c t1)*...*P(A_c tn) =[1-P(A t1)]*...*[1-P(A tn)] = (1-P1) * (1-P2) *..... (1-PN)$$

$$P(B)=P(AB)+P(A_c B) = P1*P2*.....PN+ (1-P1) * (1-P2) *..... (1-PN)$$

$$\Rightarrow P(A|t1 ,t2, t3.....tn)=P(AB)/P(B)= (P1*P2*.....PN) / [P1*P2*.....PN+ (1-P1) * (1-P2) *..... (1-PN)]$$

贝叶斯过滤算法的主要步骤

- 当 $P(A|t_1, t_2, t_3, \dots, t_n)$ 超过预定阈值时，就可以判断邮件为垃圾邮件。
- 预定阈值可以根据对垃圾邮件的限制级别设计。

贝叶斯过滤算法举例

- 例：
- 你的邮箱里现有两份邮件：
- 一封是含有“法轮功”字样的垃圾邮件 A
- 另一封是含有“法律”字样的非垃圾邮件 B。

贝叶斯过滤算法举例

- 下面根据这两份已有邮件来建立贝叶斯概率库
- 根据邮件 A 生成 hashtable_bad , 该哈希表中的记录为
- 法： 1 次
- 轮： 1 次
- 功： 1 次

贝叶斯过滤算法举例

- 计算得在本表中：
- “法”出现的概率为 0.3
- “轮”出现的概率为 0.3
- “功”出现的概率为 0.3

贝叶斯过滤算法举例

- 根据邮件B生成hashtable_good，该哈希表中的记录为：
- 法： 1 次
- 律： 1 次
- 计算得在本表中：
- “法”出现的概率为 0.5
- “律”出现的概率为 0.5

贝叶斯过滤算法举例

- 综合考虑两个哈希表，共有四个
TOKEN 串：法、轮、功、律
- 当邮件中出现“法”时，该邮件为垃圾邮件的概率为：
- $P = 0.3 / (0.3 + 0.5) = 0.375$
- 出现“轮”时，该邮件为垃圾邮件的概率为：
- $P = 0.3 / (0.3 + 0) = 1$

贝叶斯过滤算法举例

- 出现“功”时，该邮件为垃圾邮件的概率为：
- $P = 0.3 / (0.3 + 0) = 1$
- 出现“律”时，该邮件为垃圾邮件的概率为：
- $P = 0 / (0 + 0.5) = 0$

贝叶斯过滤算法举例

- 由此可得第三个哈希表
hashtable_probability，其数据为：
- 法： 0.375
- 轮： 1
- 功： 1
- 律： 0

贝叶斯过滤算法举例


- 现在新到一封含有“功律”的邮件，我们可得到两个TOKEN串：功、律
查询哈希表 hashtable_probability 可得：
- $P(\text{垃圾邮件} | \text{功}) = 1$
- $P(\text{垃圾邮件} | \text{律}) = 0$
- 此时该邮件为垃圾邮件的可能性为：
- $P = (0 * 1) / [0 * 1 + (1 - 0) * (1 - 1)] = 0$
- 由此可推出该邮件为非垃圾邮件。


贝叶斯算法的应用


- 目前，贝叶斯过滤技术被广泛应用于各种反垃圾邮件的应用程序中。
- 如：卡巴斯基反垃圾邮件程序
- foxmail、gmail、outlook的反垃圾邮件程序等

总结


2003年5月BBC专题报道称，贝叶斯可以达到99.7%的垃圾邮件识别率，同时误判率极低。是目前最有效的反垃圾邮件技术。

- 
- 贝叶斯过滤技术对邮件的所有内容进行分析，不仅仅是其中的某个关键词，而且他能判别邮件是垃圾邮件还是正常邮件。
 - 可以说，贝叶斯具有一定的智能，它对邮件中的关键词汇能综合的进行评判，可以把握“好”与“坏”之间的平衡。显然，这种技术远远高于非1即0的静态过滤技术。

- 
- 贝叶斯过滤技术具备自适应功能——通过学习新的垃圾邮件及正常邮件样本，贝叶斯将能对抗最新的垃圾邮件。并且对变体字有奇效。
 - 比如，垃圾邮件中用5ex代替sex，贝叶斯也推算出他是垃圾邮件的可能性也急剧增加。

- 
- 贝叶斯过滤技术更加个性化。他能学习并理解用户对邮件的偏好。
 - 贝叶斯过滤技术支持多语种或者说与编码无关。对于贝叶斯而言，他分析的是字串，无论他是字、词、符号、还是别的什么，当然更与语言无关。

- 贝叶斯过滤器很难被欺骗。垃圾邮件发送高手通常通过减少垃圾词汇（如free、viagra、发票）或者在信中多掺一些好的词汇（如合同、文件）来绕过检查一般的邮件内容检查，但由于贝叶斯具有的个性化色彩，要想成功的绕过贝叶斯的检查，他就不得不对每个收件人的偏好进行研究，这简直是“不可能完成的任务”。垃圾邮件发送者无法容忍的。若采用变化字，则如前所述贝叶斯判断其为垃圾邮件的可能性反而增加。

- 
- 尽管贝叶斯过滤器非常有效，但它仍需要进行优化才能真正完美。比如它可以结合“白名单”降低误报率，结合“黑名单”降低漏过率，还可以利用其他技术如源址认证使其成为更加精确的垃圾邮件过滤器。



- The End~~~

- Thank you!