Naive Bayes
朴素贝叶斯

王 博

2015.5.20

简介

贝叶斯定理是18世纪英国数学家托马斯·贝叶斯（Thomas Bayes）提出得重要概率论理论。以下摘一段 wikipedia 上的简介：

> *所谓的贝叶斯定理源于他生前为解决一个“逆概”问题写的一篇文章，而这篇文章是在他死后才由他的一位朋友发表出来的。在贝叶斯写这篇文章之前，人们已经能 够计算“正向概率”，如“假设袋子里面有N 个白球，M 个黑球，你伸手进去摸一把，摸出黑球的概率是多大”。而一个自然而然的问题是反过来：“如果我们事先并不知道袋子里面黑白球的比例，而是闭着眼睛摸出一个（或好几个）球，观察这些取出来的球的颜色之后，那么我们可以就此对袋子里面的黑白球的比例作出什么样的推测”。这个问题，就是所谓的逆向概率问题。*

推导

我们可以从条件概率的定义推导出贝叶斯定理。

根据条件概率的定义，在事件 B 发生的条件下事件 A 发生的概率为：$P(A|B) = \frac{P(A\cap B)}{P(B)}$

同样地，在事件 A 发生的条件下事件 B 发生的概率为：$P(B|A) = \frac{P(A\cap B)}{P(A)}.$

结合这两个方程式，我们可以得到：$P(A|B)\,P(B) = P(A\cap B) = P(B|A)\,P(A).$

这个引理有时称作概率乘法规则。上式两边同除以 P(A)，若P(A)非零，可以得到贝叶斯定理:

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}.$$

理解

通常，事件 A 在事件 B 发生的条件下的概率，与事件 B 在事件 A 发生的条件下的概率是不一样的；然而，这两者是有确定关系的，贝叶斯定理就是这种关系的陈述。它能够告诉我们如何利用新证据修改已有的看法。

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}.$$

我们把P(B)称为“先验概率”，即在A事件发生之前，我们对B事件概率的一个判断。P(B|A)称为“后验概率”，即在事件A发生之后，我们对B事件的重新评估。P(A|B)/P(A)称为“可能性函数”，这是一个调整因子，使得预估概率更接近真实概率。

后验概率 = 先验概率 X 调整因子

> *在西门那边公交车上下来一个人，不考虑其他因素，他是华师大的学生的概率是多少。*
> *现在你看到他下车后从华师大西门进入了华师大，这时，他是华师大的学生的概率又是多少。*

举例

事件B：明天下雨　　　事件A：今晚多云

如果你今天晚上看到多云了，那么明天下雨的概率是多大呢？

$\mathrm{P}(B|A)$?

每一天下雨的概率：$\mathrm{P}(B)$

每晚多云的概率：$\mathrm{P}(A)$

如果某天下雨了，那么前一晚多云的概率：$\mathrm{P}(A|B)$

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}.$$

朴素贝叶斯

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法，对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。朴素贝叶斯法实现简单，学习与预测的效率都很高，是一种常用的方法。

邮件分类器

一封邮件，朴素贝叶斯法如何来判定它呢，我们用S来表示垃圾邮件，H表示正常邮件。
随便给你一封邮件，它是垃圾邮件的概率为：$P(S)$，它是正常邮件的概率为：$P(H)$

邮件D = $\{W_1, W_2, \ldots, W_n\}$

该邮件是垃圾邮件的概率：

$$P(S|D) = \frac{P(S)P(D|S)}{P(D)}$$

该邮件是正常邮件的概率：

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

如果P(S|D) > P(H|D) ,这封邮件是垃圾邮件。
如果P(S|D) < P(H|D) ,这封邮件是正常邮件。

先验概率$P(S)$

条件概率$P(D|S)$

极大似然估计

在不考虑其他情况下，从网上随便拿出一封邮件，它为垃圾邮件的概率记为P(S),那么它为正常邮件的概率记为P(H),由经验可知: P(S)+P(H)=1
从网上所有邮件中取出了m+n封邮件，其中m封垃圾邮件，n封正常邮件的概率为：

$$P = P(S)^m P(H)^n = P(S)^m (1 - P(S))^n$$

现有100封邮件的训练样本，其中30封垃圾邮件，70封正常邮件，它是由上面的概率模型产生的，那么我们可以依靠这个样本来估计参数P(S)，这个估计基于这样的思想：我们所估计的模型参数，要使得产生这个给定样本的可能性最大。
即我们现在要求出P(S)，使得下面的P最大

$$P = P(S)^{30} (1 - P(S))^{70}$$

对上式求导，并令其等于零得：

$$30P(S)^{29}(1 - P(S))^{70} - 70(1 - P(S))^{69} P(S)^{30} = 0$$

$$30P(S)^{29}(1 - P(S))^{70} = 70(1 - P(S))^{69} P(S)^{30}$$

$$\frac{1 - P(S)}{P(S)} = \frac{7}{3} \Rightarrow P(S) = 0.3 = \frac{\text{垃圾邮件的个数}}{\text{邮件总数}}$$

特征条件独立假设

邮件： $\mathrm{D}=\{W_1, W_2, \ldots, W_n\}$ $P(D|S)$？

$$P(D|S)=P(W_1, W_2, \ldots, W_n|S)$$

$$P(D|S)=P(W_1|S)P(W_2|S, W_1)P(W_3|S, W_1, W_2) \ldots P(W_n|S, W_1, \ldots, W_{n-1})$$

假设W_i, W_{i-1}在邮件中相互独立

$$P(D|S)=P(W_1|S)P(W_2|S)P(W_3|S) \ldots P(W_n|S)$$

$$P(W_i|S)=\frac{P(W_i, S)}{P(S)}$$

$$P(S)=\frac{\text{垃圾邮件的个数}}{\text{邮件总数}} \quad P(W_i, S)=\frac{\text{包含词}W_i\text{的垃圾邮件个数}}{\text{邮件总数}}$$

$$P(W_i|S)=\frac{\text{包含词}W_i\text{的垃圾邮件个数}}{\text{垃圾邮件的个数}}$$

拉普拉斯平滑

$$D = \{W_1, W_2, W_k\}(k \neq 1 \ldots n)$$

$$P(W_k|S) = \frac{包含词w_k的垃圾邮件个数}{垃圾邮件的个数}=0$$

$$P(W_k|H) = \frac{包含词w_k的正常邮件个数}{正常邮件的个数}=0$$

$$P(S|D) = \frac{P(S)P(D|S)}{P(D)} = \frac{P(S)P(W_1|S)P(W_2|S)P(W_k|S)}{P(D)} = 0$$

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} = \frac{P(H)P(W_1|H)P(W_2|H)P(W_k|H)}{P(D)} = 0$$

> *因为我们要将邮件分为两类，一类垃圾邮件，一类正常邮件，那么我们设想在训练集之外还存在两封邮件，一封为垃圾邮件，另一封为正常邮件，这两封邮件都包含了词典中的所有词。*

$$P(W_k|S) = \frac{包含词W_k的垃圾邮件个数+1}{垃圾邮件的个数+1} \neq 0 \qquad P(W_k|H) = \frac{包含词W_k的正常邮件个数+1}{正常邮件的个数+1} \neq 0$$

作业

训练集
4000封 spam_train.txt

测试集
1000封 spam_test.txt

使用训练集上训练朴素贝叶斯分类器，对测试集分类并给出准确率。
1-2人一组，10月27日晚上8点前把结果发至邮箱bwang@ica.stc.sh.cn。

PS：朴素贝叶斯是生成模型，逻辑回归是判别模型，有兴趣的同学在使用朴素贝叶斯算法分类后，可以使用上次的逻辑回归模型在这个数据集上做下实验，对比下这两个分类模型的效果。

参考资料

斯坦福机器学习公开课**Naive Bayes**讲义

《统计学习方法》

贝叶斯推断及其互联网应用（二）：过滤垃圾邮件

http://blog.chinaunix.net/uid-26548237-id-3853480.html

数学之美番外篇：平凡而又神奇的贝叶斯方法

http://mindhacks.cn/2008/09/21/the-magical-bayesian-method/

分类算法之朴素贝叶斯分类**(Naive Bayesian classification)**

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