贝叶斯概率计算公式的应用

本文主要说明一下贝叶斯公式的相关数学推导和应用场景。贝叶斯概率公式主要是逆向概率的计算，通过已知的先验概率来推测后验概率。我们先给出贝叶斯的计算公式，如下所示：

$$\boldsymbol{P(B|A) = \frac{p(A|B)P(B)}{P(A)}}$$

那么这个公式是如何得来的呢，这就要用到条件概率了，条件概率的计算公式如下：

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

那么我们将这个条件概率计算公式的分母乘到左边，我们可以得到以下公式：

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

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

这样我们就依靠条件概率推导出了贝叶斯概率计算公式。

下面我们再来说明一下如何利用贝叶斯公式来解决实际问题。经常使用的是单词拼写错误检查，比如我输出了一个错误的单词 the，程序自动帮助我们修正为 the 或者 they，需要根据山下文来考虑，这里我想到会用到深度学习的循环神经网络来处理这种依赖上下文的情况。拼写错误方面的网上有先关例子，大家可以查阅。

还有一种使用贝叶斯概率计算的经典例子就是垃圾邮件过滤，这方面的思路网上很多，我主要是讲解一些里面的一个计算公式的推导，网上文章都没有说明这个问题。首先也简要说明一下垃圾邮件过滤问题。我们要判断一封邮件是否是

垃圾邮件，我们用h^+表示为垃圾邮件事件，w表示为接受到的邮件。那么我们要判断的就是$P(h^+|w)$,即是接收到的邮件是垃圾邮件的概率，根据概率来判断是否是垃圾邮件。直接去求解这个概率比较困难，我们就想到了贝叶斯方法。通过先验概率来推测后验概率。我们就可以写成如下计算公式：

$$\boldsymbol{P(h^+|w) = \frac{P(w|h^+)P(h^+)}{P(w)}}$$

因为$P(w)$对我们判断是垃圾邮件还是正常邮件，不会产生任何影响，它就是个客观事实，因此我可以把$P(w)$去掉。仅仅考虑$\boldsymbol{p(w|h^+)P(h^+)}$这个式子。$\boldsymbol{P(h^+)}$就是先验概率，我们可以统计接收到的邮件中哪些是垃圾邮件哪些是正常的，就可以知道$\boldsymbol{P(h^+)}$的值了。问题的关键是$\boldsymbol{p(w|h^+)}$。它表示的是如果垃圾邮件，那么内容和w一样的概率。这可很难求解啊，所以我们需要对这个概率做一下处理。我们设置w是由单词$\{w_1, w_2, w_3 \dots\dots w_n\}$组成。那么我们可得到下下面式子：

$$\boldsymbol{P(w|h^+) = P(\{w_1, w_2, w_3 \dots\dots w_n\}|h^+)}$$

$$\boldsymbol{P(\{w_1, w_2, w_3 \dots\dots w_n\}|h^+) = \frac{P(\{w_1, w_2, w_3 \dots\dots w_n\}, h^+)}{P(h^+)}}$$

$$\boldsymbol{P(\{w_1, w_2, w_3 \dots\dots w_n\}|h^+)P(h^+) = P(\{w_1, w_2, w_3 \dots\dots w_n\}, h^+)}$$

我们大学学习概率论数理统计的时候接触过这样一个公式，我们一般叫做乘法规则，该计算公式如下：

$$\boldsymbol{P(E_1E_2E_3 \dots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1, E_2) \dots P(E_n|E_1E_2E_3 \dots E_{n-1})}$$

证明这个公式很简单啊，只要对等式右边的这些条件概率表达式，用条件概率计算公式重新表达就可以,即下面这个例子：

$$\boldsymbol{P(E_1E_2E_3\dots E_n)=P(E_1)\frac{P(E_1,E_2)}{P(E_1)}\frac{P(E_1,E_2,E_3)}{P(E_1,E_2)}\dots\frac{P(E_1,E_2,E_3\dots E_{n-1}E_n)}{P(E_1,E_2,E_3\dots E_{n-1})}}$$

分母都约掉，发现两边式子就相等了，证明完毕了。我们就要利用这个乘法规则来帮助我们解决之前$\boldsymbol{P(\{w_1,w_2,w_3\dots\dots w_n\},h^+)}$这个如何计算的问题。那么利用乘法规则，可得如下式子：

$$\boldsymbol{P(\{w_1,w_2,w_3\dots w_n\},h^+)}$$

$$\boldsymbol{=P(h^+)P(w_1|h^+)P(w_2|w_1,h^+)\dots P(w_n|w_1,w_2\dots w_n,h^+)}$$

又因为

$$\boldsymbol{P(\{w_1,w_2,w_3\dots\dots w_n\}|h^+)P(h^+)=P(\{w_1,w_2,w_3\dots\dots w_n\},h^+)}$$

$$\boldsymbol{P(\{w_1,w_2,w_3\dots\dots w_n\}|h^+)P(h^+)}$$

$$\boldsymbol{=P(h^+)P(w_1|h^+)P(w_2|w_1,h^+)\dots P(w_n|w_1,w_2\dots w_n,h^+)}$$

等式两边约掉$\boldsymbol{P(h^+)}$,就得到如下式子：

$$\boldsymbol{P(\{w_1,w_2,w_3\dots\dots w_n\}|h^+)}$$

$$\boldsymbol{=P(w_1|h^+)P(w_2|w_1,h^+)\dots P(w_n|w_1,w_2\dots w_n,h^+)}$$

这里我们假设$\boldsymbol{w_1,w_2,w_3\dots\dots w_n}$是彼此独立的，这就是朴素贝叶斯。我们之所以这样假设，也是为了我们可以计算概率。这样我们就可以得到关于判断是否是垃圾邮件的最终概率计算公式：

$$\boldsymbol{P(w_1|h^+)P(w_2|w_1,h^+)\dots P(w_n|w_1,w_2\dots w_n,h^+)}$$

$$\boldsymbol{=P(w_1|h^+)P(w_2|h^+)P(w_3|h^+)\dots P(w_n|h^+)}$$

这样就可以很容易计算出来了，只要在垃圾邮件中统计$\boldsymbol{w_1,w_2,w_3\dots\dots w_n}$这些词分别出现的频率，就可以计算出$P(w_1|h^+),P(w_2|h^+),P(w_3|h^+),\dots P(w_n|h^+)$这些概率了。这就是关于贝叶斯相关的数学推导过程。