# 贝叶斯算法

## 1 垃圾邮件处理模型

以预测垃圾邮件为例子，介绍贝叶斯算法。词库中有5000个单词。我们有输入x，，是一个50000维的向量，其中为1表示词库中第i个单次在该邮件中出现，为0表示词库中第i个单次在该邮件中没有出现。

已经邮件是否为邮件,样本出现的概率为，可以定义为下式:



注: 上式很好理解。右侧第一个式子为已知y，x1出现的概率。第二个为一直y和x1出现x2的概率。因此两个式子相乘表示一直y初选(x1,x2)的概率。依次类推，可以得到这个公式。

接下来我们做个假设，即表示每个词在邮件中出现不影响其他词在邮件中出现，这些时间是独立的。这样就可以得到下式:



2 公式推导

这里求最大释然函数，我们设,,。









然后我们对求偏导数：



所以有:



然后我们对





所以有:



如果一个单词不再训练样本中，就会出现0/0的现象。避免这个事件发生，引入拉普拉斯平滑。

同理有:



注:试分析这样的公式，就是总的垃圾邮件数目除以从样本数。就是所有垃圾邮件中出现对应的单词的数目除以所有垃圾邮件的数目。实际上这是个很容理解的道理。当然这就是数学的魅力，即便很简单的公司都是有着其理论依据的。

## 2 另一种垃圾邮件的模型

下面采用另外一种方式进行建模。对于。其中代表邮件的第一个单词在词库中的索引为1。仍然有如下公式



我们设置，。

然后求最大释然函数









我们对求偏导数，与之前相同，这里直接写出：



我们对求偏导数，如下：



其中有：





所以有：



所以最后得到:



同理有:

