On word frequency information and negative evidence in Naïve Bayes text classification

Authors

Karl-Michael Schneider

Abstract

Naïve Bayes有很多版本。其中一个版本叫Multi-variate Bernoulli NB(也叫做 binary independence model)，因为它用的是binary value嘛，就是看这个词有无出现。还有一个版本是叫multinomial NB，用的是TF来做values，就是说看这个词出现了多少次。经过对比，multinomial NB的性能比较好，有些人认为性能好的原因是用了word frequency的原因。作者认为不是这样子的。他们把word frequency information移除之后，发现multinomial NB性能更好了。作者认为出现这个差异的原因是negative evidence incorporate到model中。因此呢，这篇论文就是为了帮助我们理解这两个NB版本之间的差异。

1 Introduction

在text classification任务上，NB是一个很流行的方法，因为简单且有效。NB有很多不同的版本，主要是看这个document或者说是message是怎么表示的，是TF呢，还是binary。如果这个document是用binary的，就是我们选定vocabulary(假设选了3000个词之类的)，然后就看这些词有没有在document中出现过，出现了，那么就是1，没有出现就是0，使用的model是multi-variate Bernoulli NB。因为这个document的组成可以看做是多次的Bernoulli trails的组成。另外一个版本呢，就是document用TF来组成，就是看某个词出现了多少次。然后就用multinomial NB。因为这个document就是相当于多项式分布嘛。

之前的文献中就有人发现multinomial NB的效果会比multi-variate Bernoulli NB好，分类准确率高。很多人认为multinomial NB效果较好的原因是document使用的是TF来表示的，使用TF的话会capture到更多的information，而使用binary value capture到的information不是很多。所以multi-variate Bernoulli NB性能一般。

这篇论文主要是argue multinomial NB性能好的原因不是因为用了TF作为attributes的值导致的。作者呢，把word frequency的information给移除了，multinomial NB性能非但没有降，还提升了。作者呢，还argue说这两个版本之所以性能不一样，是因为处理negative evidence的方式不一样，negative evidence的意思就是document里没有出现的词(假设fixed attributes有3000词，不是说这所有的3000词都在document中有出现的)

2 Naïve Bayes

2.1 Multi-variate Bernoulli Model

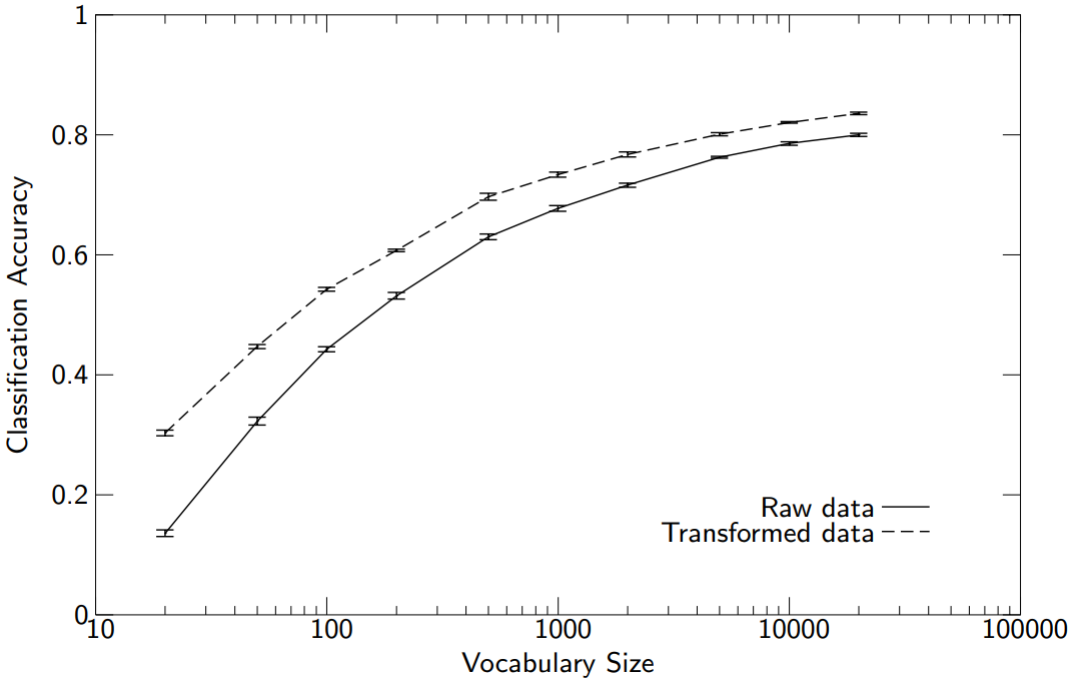
2.2 Multinomial Model

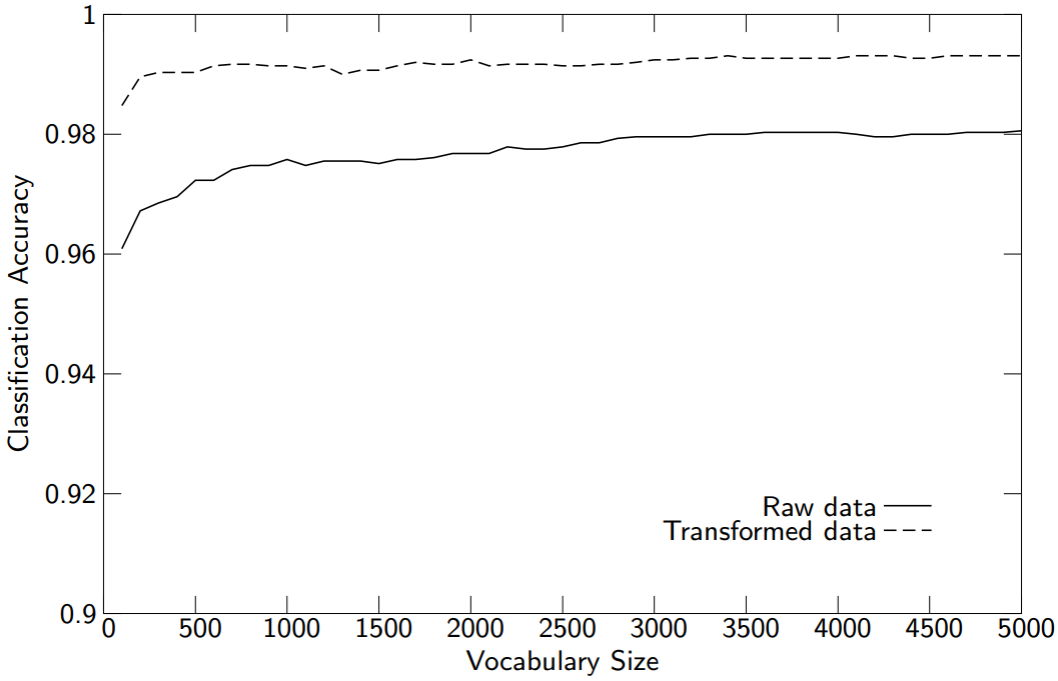
3 Word frequency information

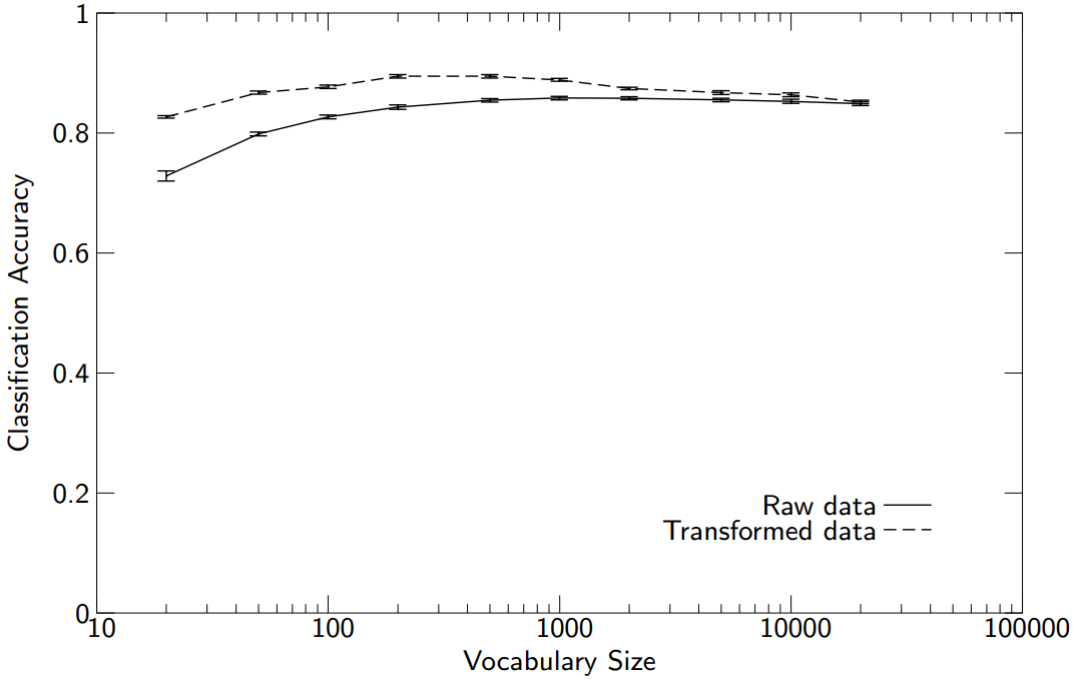
https://www.bilibili.com/video/BV16Z4y1V7uE/

在之前的文献中，就发现multinomial NB的性能要比Multi-variate Bernoulli NB要好，特别是当vocabulary size变得更加大的时候。所以很多人就认为multinomial NB性能好的原因是attributes用了TF。但是作者认为这个word frequency information不是导致multinomial NB性能好的主要原因

于是作者就在三个public的数据集上做了实验，然后把attributes的TF value transform 成 $x'_t = min\{x_t, 1\}$，其实就是变成了binary values。我们来看看效果如何，因为有三个不同的数据集，所以就为我们展示了三幅图





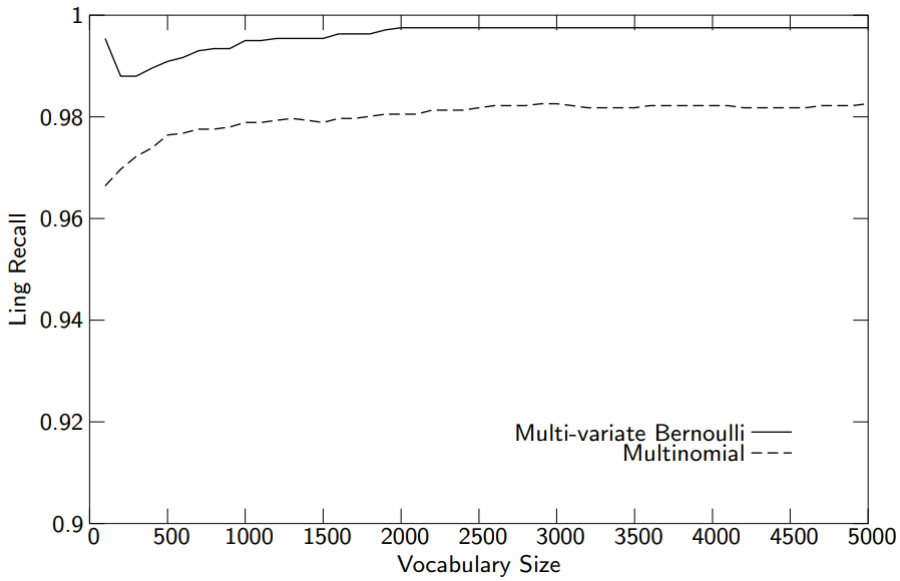


我们会发现，把TF换成binary value之后，multinomial NB效果更佳。作者选择vocabulary size的方法是通过mutual information来选择的(这里我们当做information gain来看，差不多的意思)。我们发现vocabulary size其实也是有一定的影响的。然后我们发现，两个model的差距在vocabulary size小的时候，差距更大。

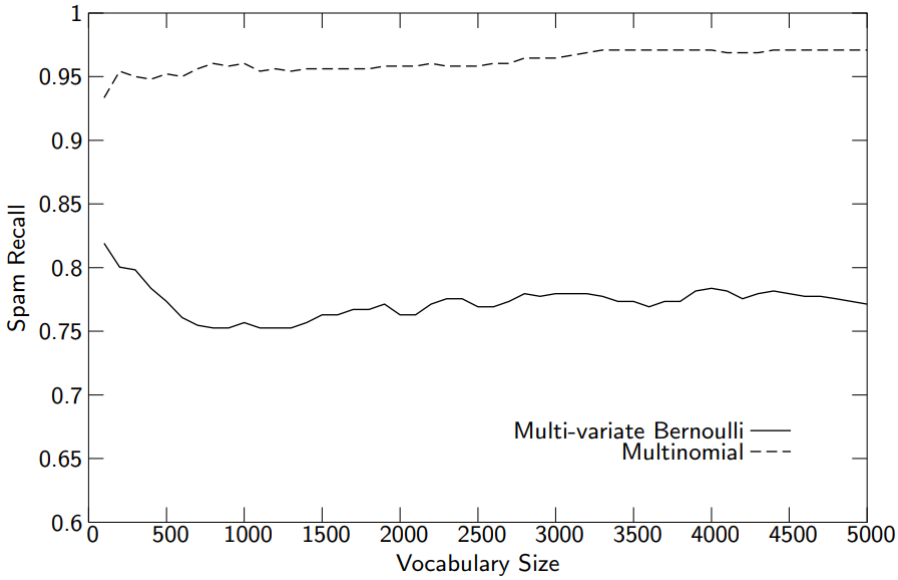
4 Negative evidence

为什么multinomial NB会比multi-variate Bernoulli NB效果要好呢？作者用其中一个数据集(ling-spam)作为学习参考。作者呢，把ling class recall和spam class recall给我们画出来了，我们看看，如下图

这个是ling-class recall，recall就是求全率，希望把所有的ling-class给找出来，这里假设spam是positive，毕竟是要找spam email。所以这里的ling-class recall就是 $\frac{TN}{TN+FP}$.



然后我们再看spam-class recall。我们希望把所有的spam-class给找出来，所以就是 $\frac{TP}{TP+FN}$.



我们对比会发现Multi-variate Bernoulli NB有着较高的ling-class recall但是spam recall却很低。而multinomial NB却相对比较平衡，两者都差不多。作者认为造成multi-variate Bernoulli NB相差比较大的一个原因是数据集的问题。我们来看下ling-spam dataset的构成，我们可以看到，spam email占的比例较低。然后我们再看下vocabulary，只有8.3%的词是没有出现在Ling class的，然而spam email中，却有81.2%的词是没出现的。

	Total	Ling	Spam
Documents	2893	2412 (83.4%)	481 (16.6%)
Vocabulary	59,829	54,860 (91.7%)	11,250 (18.8%)

我们回忆一下multi-variate Bernoulli distribution，它的公式如下

$$p(\overrightarrow{x}|c) = \prod_{i=1}^{m} p(t_i|c)^{x_i}(1 - p(t_i|c))^{(1-x_i)}$$

我们可以看到每一个词都是要去算概率的，不管这个词有无出现。如果出现了(positive evidence),其概率为$p(t_i|c)^{x_i}$, 如果没有出现(negative evidence) 其概率为$(1 - p(t_i|c))^{(1-x_i)}$.

我们还是得看下ling-spam dataset的average distribution，如下表所示，vocabulary 那一列指的是vocabulary size。这张表为我们展示的是平均下来每个document有多少个distinct word和一个单词平均出现在多少个document中

Vocabulary	Total		Ling		Spam	
	Words	Documents	Words	Documents	Words	Documents
Full	226.5	11.0	226.9	9.1	224.5	1.8
MI 5000	138.5	80.2	133.8	64.5	162.5	15.6
MI 500	44.0	254.5	39.6	190.9	66.2	63.7

我们可以看到平均每个文档，大约有226.5个distinct words，大约占总词汇的0.38%。每个词平均出现在11个document中。如果仅仅只是用mutual information 选出来的5000个词，那么平均下来每个document大约有138.5个词，约占2.77%。然后每个词也平均出现在80.2个document中。如果我们再减少到500个词，distinct word的比例增加到了8.8%, (44 out of 500). 然后，我们仍旧可以知道，还是有大部分的词是没有出现在一个document中的。也就是说不管是TF还是binary value，很多位置的值还是0的。

这个发现表明，$p(\overrightarrow{x}|c)$的概率极大可能是基于没有出现的词的概率的，也就是说，这个文档的分类要很大地依赖于negative evidence，也就是文档中没有出现的词。如果一篇空文档会出现什么样的概率呢？如下表所示

Vocabulary	Total	Ling	Spam
Full	3.21e-137	1.29e-131	5.2e-174
MI 5000	6.44e-78	8.4e-76	1.45e-96
MI 500	5.21e-24	1.41e-22	3.59e-37

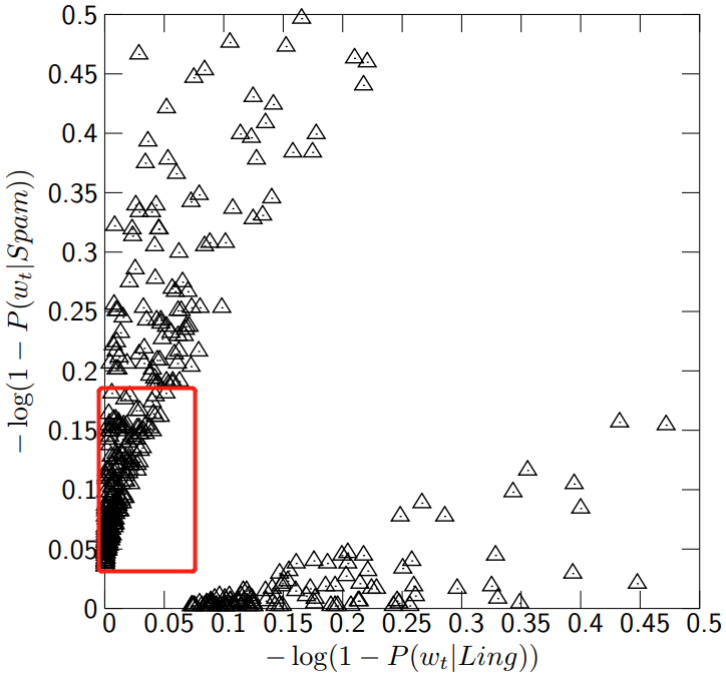
我们可以看到，如果是空文档，那么是spam的概率很低。都是划分到ling-class中。这可以有如下解释，ling-class占的词比较多，毕竟占80%多呢。但是，ling-class的distinct word却不比spam-class 的distinct占比高，特别是当vocabulary size减到5000,500的时候。因此呢，在ling-class的每一个词的概率都是低于spam class的。因为Multi-variate Bernoulli NB中的$p(t|c)$的计算方式如下

$$p(t|c) = \frac{1 + M_{t,c}}{2 + M_c}$$

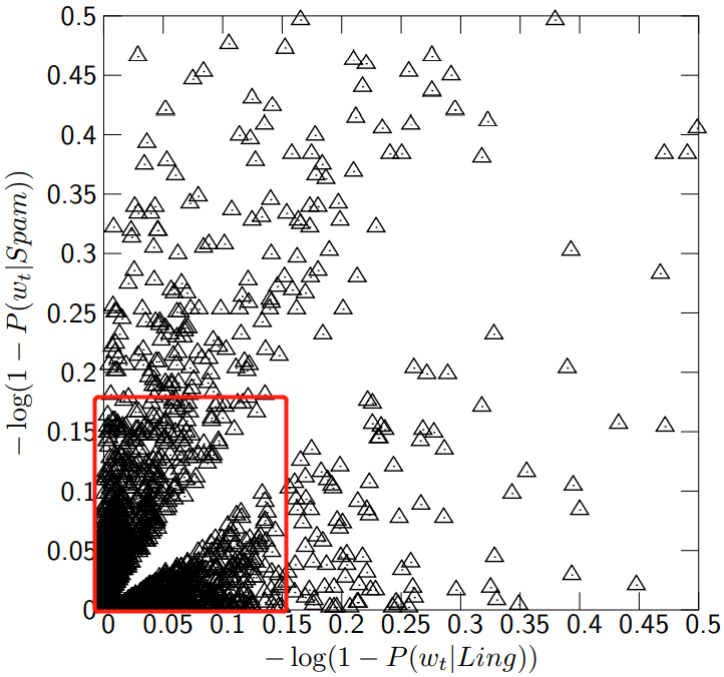
这里，$M_{t,c}$就是指类别c中包含token t 的message有多少个。M_c代表的意思就是类别c一共有多少个。在ling-class中，M_c比较大，但是$M_{t,c}$比较小。在一个document在被划分的时候，也就是算概率的时候，大多数的词都是negative evidence的，都是没在这个document出现过的。因此呢，划分到ling-class的时候，negative evidence会有一定的主导，因为条件概率会更低一些。即使$p(t|c)$这个先验概率可以忽略不计，但是很多的词连乘之后，效果就显而易见了。

我们可以可视化negative evidence的impact，其实就是可视化每个词的weight，如下图所示，分别可视化当选了500个词和5000个词后的情况。x轴是spam class的attributes，这个attributes是没有出现在spam document中的，然后把每一个词的weight都给可视化出来。y轴就是ling-class 的attributes，也是没有出现在ling document中的。至于计算weight的方法是 $p(t|c) = \frac{1+M_{t,c}}{2+M_c}$.

下图是500词的



下图是5000词的



这两幅图给我们展示了multi-variate Bernoulli NB给negative evidence多少weight。我们可以发现所有选择出来的词通常要么是属于Ling-class的negative evidence，要么是属于Spam-class 的negative evidence，(这里的意思是更偏向于出现在哪一个class)，因为要么在对角线以上(更偏向出现在spam-class)，要么就在对角线以下(更偏向出现在ling-class)。这张图呢，越小的value表示这个evidence越强烈，也就表示几乎没有出现在document中。我们可以看到对角线以上的数量比较多，说明很多词就不怎么出现在Ling-class document中。

5 Discussion

之前就有文献证明multinomial NB其实是Naïve Bayes Poisson model改编过来的，需要assume 这个document length跟class无关，也就是相互独立。在Naïve Bayes Poisson model中，每一个词,token t_i都会用一个值来表示，取非负数值来表示这个词在document出现的次数，因此是直接把词频incorporate到model中，但是这个token，也就是这个特征里所有的值，得服从Poisson distribution。当然了，token与token之间也是independence的，这是NB的正常的assumption啦。但是，之前的文献中就有发现NB Poisson model并没有表现地比multinomial NB要好。这个multinomial NB也是assume 这个词频是服从Poisson distribution的。

那为啥multinomial NB把词频的attributes换成binary之后，性能提升了呢？之前就有文献讨论documents的词的分布，发现呢，一个词在同一篇document中第二次出现的概率会大于一个词只出现一次的概率。Poisson distribution不能在这种情况中表现好。还有其他文献把更为复杂的分布(mixture of Poisson distribution)给应用在模型中，这样更贴合词语在document中的分布。然而，也有文献表示把词频做一个简单的transformation，就是$x'_t = log(d + x_t)$，就足以提升multinomial NB的分类性能。这种简单的变换能够把大的词频变小，因此在multinomial NB中，就是documents中如果出现很多相同的词，这个document得到的probability也会高一些(因为小数的次方，次方越大，值越小。把这个文档中的经常出现的词给降低了，整体的值就不会变得很小，所以才说是higher probability)。

作者把TF改成binary之后，分类性能也提升上去了。使用TF的话，主要是不服从Poisson distribution，导致性能变差。multinomial TF需要assume每个词的出现是independence的，但这是不可能的。就像刚刚说的，一个词在同一篇document中第二次出现的概率会大于一个词只出现一次的概率。就是这个词出现了一次，就还有可能继续出现，且概率很大。因此呢，multinomial TF对于处理这样的data效果是很差的。但是改成binary之后，会减小这种影响。

那么，multi-variate Bernoulli NB和multinomial NB的区别在哪呢？Multi-variate对positive 和negative evidence都是一样处理，可以说是没有区别对待。但是multinomial对每个词，positive还是negative的对待方式是不一样的，根据词频变化而变化。如果一个词没有出现的话，它是没有直接contribute 到 $p(\overrightarrow{x}|c)$,而且，它是0次方，不怎么有影响。这样子计算的话，negative evidence的影响

会小于在multi-variate Bernoulli NB中negative evidence的影响。

6 Conclusion

这篇论文是主要是对比两个model，使得我们更加理解这两个model。multinomial NB会比multivariate Bernoulli NB性能好不是因为TF的原因，而是因为对negative evidence的处理方式不同。事实上，只要经过一些小小的变化，比如说加log，或者变成binary，都能提升性能。

我们发现multi-variate Bernoulli NB中大部分的evidence都是negative evidence，而且，不同的class，negative evidence的分布也很不均衡，这也就导致multivariate Bernoulli NB更容易倾向于某一个class因为这个model给negative evidence的weight太高，导致正确分类率不高。