explanation

olxh

22 .5 .3

1 朴素贝叶斯算法

朴素贝叶斯算法是一种有监督的机器学习，其优点是简单易懂易上手。

1.1 概率与条件概率

我们可以计算出事件A发生的概率P0，以及事件A不发生的概率P1。我们计算出这两个概率P0和P1，将他们进行比较，如果P0大于P1,我们可以认为事件A发生，否则我们认为事件A不发生。

且我们有条件概率的计算公式：

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

稍作变形可得：

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

其中，我们把$P(A)$叫做这件事发生的先验概率，即可以看作我们在平时生活中估计所得出的概率。$P(A|B)$叫做这件事发生的后验概率，即在一定条件下索发生的概率。 $\frac{P(B|A)}{P(B)}$ 叫做可能性函数，我们可以通过调整它，使得先验概率通过一系列计算得到我们所需要的后验概率。

在实际计算的过程中，我们可以所需要得到的是两个后验概率。因为它们的分母都是相同的，我们可以只通过比较它们的分子来得到哪件事情发生的概率更大，从而计算出结论。这叫做贝叶斯决策理论。

1.2 朴素贝叶斯推断

朴素贝叶斯推断则把我们的条件的概率分布做了条件独立性的假设，使得我们更加方便对其进行计算。

假设一个事件X有n个特征$X_i(1 \leq i \leq n \& i \in Z)$，也就是说我们有这个公式：

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)} = \frac{P(x_1, x_2, x_3 \dots x_n|A)P(A)}{P(X)} = \frac{P(A)\sum_{i=1}^{n} P(x_i|A)}{P(X)}$$

因为我们需要比较的两个概率的分母都相等，所以我们只要关心分子。

1.3 拉普拉斯平滑

其实我们这么简单粗暴地进行计算是存在一定问题的，我们需要对这种算法进行改进。

我们可以观察上面的式子，当$P(x_i|A)$这一项里面有任意一个是0，那么最后我们求出的结果就都为0。这样显然是不合理的。为了降低这种影响，假如这个单词出现过，我们把它的贡献记作$\frac{1}{2}$而不是0。在实际实现的过程中，我们可以直接把分子分母都加上1来解决这个问题。这种做法就叫做拉普拉斯平滑，又叫做加1平滑，用来解决这种概率为0的问题。

同时，当我们将很多个大于0小于1的数字相乘的时候，可能最后会造成小数精度不够的问题（下溢出）。为了解决这个问题，我们可以对需要求的答案进行取对数的操作，从而解决精度不够的问题。需要注意的是，我们取对数后，一开始条件概率公式里面的乘积会变成加法。

2 具体实现过程

通过两个文件，$text_for_negative.txt$与$text_for_positive.txt$对读入的文件进行处理，分别代表了具有侮辱性的语句和不具有侮辱性的语句。

接着，我们可以对样本集合进行处理，得到一张单词表（去重过）以及各个单词出现的频率进行统计。

最后我们可以对一组词通过上面的公式得到它是否为侮辱性词汇，最后进行输出。结果输出：

p0: -8.558

p1: -8.580

['handsome','rich','happy'] is not offensive

p0: -7.131

p1: -4.678

['stupid','pig'] is offensive

注意到上面$P0$和$P1$均为负数，因为我们对最终答案取了对数。

3 不足及改进

首先，运用机器学习找出不符合社区环境的词语具有一定的局限性。在各类语境下面，运用一些否定词或者否定词的前缀会使得原先的一些词改变它本身的意思，从而干扰机器学习的正确性与否。我们可以通过过滤一些含有否定词的语句或者运用语法逻辑改变一些语句是否具有冒犯性来解决这方面的问题。

其次，样本集合过于小且具有主观性，可以通过下载网上的各类资料来减少这方面所带来的误差，使得结果更为精准可靠。

同时，我们估计得到的两个概率P0和P1相差比较小的时候，其实是不能精确地确定它是否为我们所需要查找到的关键词（误差可能会比较大），这时候我们就需要通过人工筛查或者规定一个区间范围从而进行筛查。