**朴素贝叶斯算法（分类器）**

1. **朴素贝叶斯算法的优缺点：**

优点： 在数据较少的情况下任然有效，可以处理多类问题

缺点：对于输入数据的准备方式较为敏感

适用数据类型：标称型数据

1. **核心思想:** 通过概率来进行分类
2. **朴素贝叶斯的一般过程：**
3. 收集数据：可以使用任何方法。
4. 准备数据： 需要数字型或布尔型数据
5. 分析数据：有大量特征时，绘制特征作用不大，此时使用直方图效果更好。
6. 训练算法：计算不同独立特征的条件概率
7. 使用算法：一个常见的贝叶斯应用是文档分类。可以在任意分类场景中使用朴素贝叶斯分类器，不一定非要文本。

**4.贝叶斯分类器**

朴素贝叶斯分类器是一个概率分类器。假设现有的类别C={c1，c2，……cm}。给定一篇文档d，文档d最有可能属于哪个类呢？这个问题用数学公式表示如下：

IMG_256（公式一）

c^就是：在所有的类别C={c1，c2，……cm} 中，使得：条件概率P(c|d)取最大值的类别。使用贝叶斯公式，将（公式一）转换成如下形式：

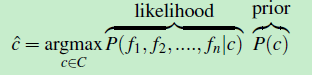
IMG_257(公式二)

对类别C中的每个类型，计算 [p(d|c)\*p(c)]/p(d) 的值，然后选取最大值对应的那个类型ci ，该ci就是最优解c^，因此，可以忽略掉分母 p(d)，（公式二）变成如下形式：

IMG_258(公式三)

这个公式由两部分组成，前面那部分P(d|c) 称为似然函数，后面那部分P(c) 称为先验概率。

前面提到使用词袋模型来表示 文档d，文档d的每个特征表示为：d={f1,f2,f3……fn}，那么这里的特征fi其实就是单词wi出现的频率（次数），公式三转化成如下形式：

(公式四)

对文档d 做个假设：假设各个特征之间是相互独立的。那么p(f1,f2……fn|c)=p(f1|c)\*p(f2|c)\*……\*p(fn|c)，公式四转化成如下形式：

IMG_260（公式五）

由于每个概率值很小（比如0.0001）若干个很小的概率值直接相乘，得到的结果会越来越小。为了避免计算过程出现下溢(underflower)，引入对数函数Log，在 log space中进行计算。然后使用词袋模型的每个单词wi 出现频率作为特征，得到如下公式

IMG_261（公式六）

**五，训练朴素贝叶斯分类器**

训练朴素贝叶斯的过程其实就是计算先验概率和似然函数的过程。

**①先验概率P(c)的计算**

P(c)的意思是：在所有的文档中，类别为c的文档出现的概率有多大？假设训练数据中一共有Ndoc篇文档，只要数一下类别c的文档有多少个就能计算p(c)了，类别c的文档共有Nc篇，先验概率的计算公式如下：

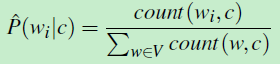
IMG_262(公式七)

先验概率 其实就是 准备干一件事情时，目前已经掌握了哪些信息了

**②似然函数P(wi|c)的计算**

由于是用词袋模型表示一篇文档d，对于文档d中的每个单词wi，找到训练数据集中所有类别为c的文档，数一数 单词wi在这些文档（类别为c）中出现的次数：count(wi,c)

然后，再数一数训练数据集中类别为c的文档一共有多少个单词IMG_263。计算 二者之间的比值，就是似然函数的值。似然函数计算公式如下：

（公式八）

其中V，就是词库。（有些单词在词库中，但是不属于类别C，那么 count(w,c)=0）

从上面计算似然函数的过程来看，其实相当于一个发掘（统计）潜藏规律的过程。

**六，unknow words的情形**

假设只考虑文本二分类：将文档分成 positve类别，或者negative类别，C={positive, negative}

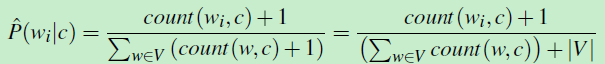
在训练数据集中，类别为positive的所有文档 都没有 包含 单词wi = fantastic（fantastic可能出现在类别为negative的文档中）

那么 count(wi=fantastic，ci=positive)=0 。那么：

IMG_265

而注意到前面公式五中的累乘，整篇文档的似然函数值为0，也就是说：如果文档d中有个单词fantastic在类别为c的训练数据集文档中从未出现过，那文档d被分类到类别c的概率为0，尽管文档d中还有一些其他单词（特征），而这些单词所代表的特征认为文档d应该被分类 到 类别c中

解决方案就是**add-one smoothing**。（不介绍了），其实就是将“出现次数加1”。似然函数公式变成如下形式：

（公式九）

其中|V|是词库中所有单词的个数。

（以上为摘要前人的劳动成果）

**6.具体实例，以及Python实现**

先说：在这里我是假设你已经了解了一些基本的概念，比如：概率论里面的 条件概率，贝叶斯，以及先验概率和似然函数等等。如果你还不了解这些，就先去学一下。朴实贝叶斯分类的核心就是通过概率来对我们的目标进行分类。

1.<通过朴素贝叶斯对文档分类 实例>

假设我们现在要给王者荣耀的聊天系统做一个过滤器，这个过滤器的目的是过滤掉那些给玩家带来不良影响的言辞。

那么这个时候我们就可以把所有的两天记录分为两类，一类是对用户有影响的不良言辞，一类是没有有影响的言辞。即：C={C0, C1}（C0为不会造成不良影响的，C1是会造成不良影响的）

这里我们收集了10条用户数据（当然这是远远不够的，要想使我们的算法模型在进行工作的时候准确性越高，那么我们就得拿跟多的数据来训练我们的模型，这里我只是作为一个简单的例子在说。）当热既然是拿来训练的数据，那么我们肯定是知道他的分类的。一下是我们的数据：

1. 玩的真他妈菜垃圾（C1）
2. 小哥哥玩的真好爱你哟（C0）
3. 原来帅哥是个大神啊（C0）
4. 这他妈还能玩吗傻逼（C1）
5. 今天我就日了狗了遇到你们这些傻逼（C1）
6. 大神带我装逼带我飞啊（C0）
7. 遇到一个百年打野的大傻逼草尼玛的（C1）
8. 帅哥你玩的这么好能带带小妹吗（C0）
9. 这操作6的飞起啊（C0）
10. 相信我能把你们带飞的（C0）

好了，现在我们的训练的数据样本是有了，

1. 小哥哥玩的好牛逼啊

现在我们要让我们的算法模型知道11这句话到底是属于C0,还是C1呢，好了现在我们用python语言来构建我们的算法模型。