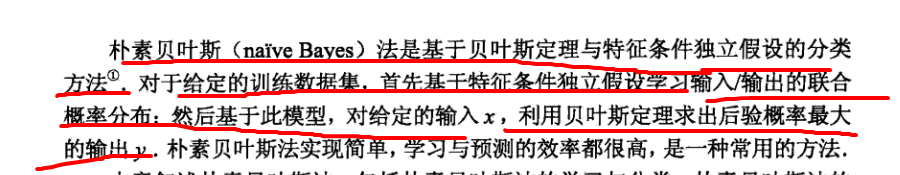
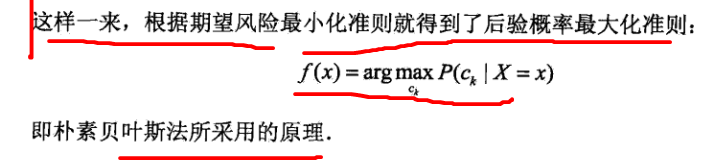
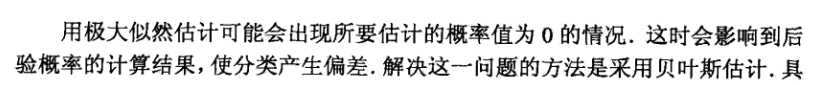
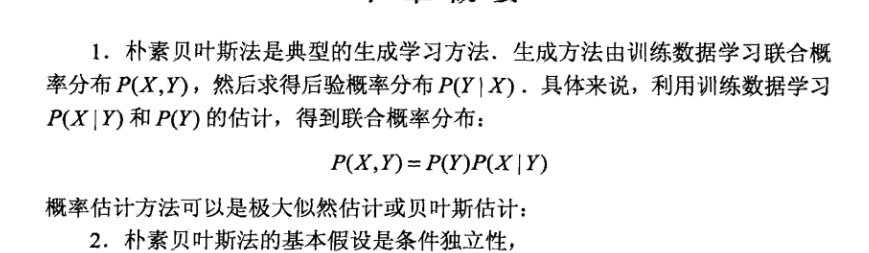
朴素贝叶斯









朴素贝叶斯分类及模型：

<https://blog.csdn.net/zhihaoma/article/details/51052064>

朴素贝叶斯分类器是一种有监督学习，常见有两种模型，多项式模型(multinomial model)和伯努利模型(Bernoulli model)。

**1.多项式模型**

在多项式模型中， 设某文档d=(t1,t2,…,tk)，tk是该文档中出现过的单词，允许重复，则

先验概率P(c)= 类c下单词总数/整个训练样本的单词总数

类条件概率P(tk|c)=(类c下单词tk在各个文档中出现过的次数之和+1)/(类c下单词总数+|V|)

V是训练样本的单词表（即抽取单词，单词出现多次，只算一个），|V|则表示训练样本包含多少种单词。

**2.伯努利模型**

P(c)= 类c下文件总数/整个训练样本的文件总数

P(tk|c)=(类c下包含单词tk的文件数+1)/(类c的文档总数+2)

有了以上类条件概率，开始计算后验概率，

P(yes | d)=P(yes)×P(Chinese|yes) ×P(Japan|yes) ×P(Tokyo|yes)×(1-P(Beijing|yes)) ×(1-P(Shanghai|yes))×(1-P(Macao|yes))

**3.两个模型的区别**

二者的计算粒度不一样，多项式模型以单词为粒度，伯努利模型以文件为粒度，因此二者的先验概率和类条件概率的计算方法都不同。

计算后验概率时，对于一个文档d，多项式模型中，只有在d中出现过的单词，才会参与后验概率计算，伯努利模型中，没有在d中出现，但是在全局单词表中出现的单词，也会参与计算，不过是作为“反方”参与的。

<https://blog.csdn.net/qq_38688564/article/details/79521544>  
<https://blog.csdn.net/longxinchen_ml/article/details/50597149>

朴素贝叶斯总结：

1. 属性可以离散、 也可以连续

2. 数学基础扎实、 分类效率稳定

3. **对缺失值和噪声不太敏感**

4. **属性如果不相关， 分类效果很好；** 如果相关， 分类效果不低于决策树

5. **面对孤立的噪声点，朴素贝叶斯分类器是健壮的。**因为在从数据中估计条件概率时，这些点会被平均。

6. 面对无关属性， 该分类器时健壮的， 如果一个 X 时无关属性， 那么 P(x|Y)几乎变成了均匀分布， X 条件下概率不会对总的后验概率的计算产生影响。

7. **相关属性会降低朴素贝叶斯的分类的性能，因为这些属性，条件独立的假设不成立。**

**平滑技术：在增加未出现词的次数减少已出现词的次数。**

工程上的一些tricks:

1. 取对数：使得判断所需要的计算时间被转移到了训练阶段，实时运行速度快的多
2. 转换为权重：这样可以根据权重的大小来评估和筛选特征
3. 选取top k的关键词，需要依靠经验选定一个正数的阈值

