机器学习算法--贝叶斯

# 课程大纲

|  |  |
| --- | --- |
| 朴素贝叶斯算法原理 | Bayes算法概述 |
| Bayes算法思想 |
| Bayes算法要点 |
| 朴素贝叶斯算法案例1 | 需求 |
| Python实现 |
| 朴素贝叶斯算法案例2 | 需求 |
| Python实现 |

课程目标：

1. 理解朴素贝叶斯算法的核心思想
2. 理解朴素贝叶斯算法的代码实现
3. 掌握朴素贝叶斯算法的应用步骤：数据处理、建模、运算和结果判定

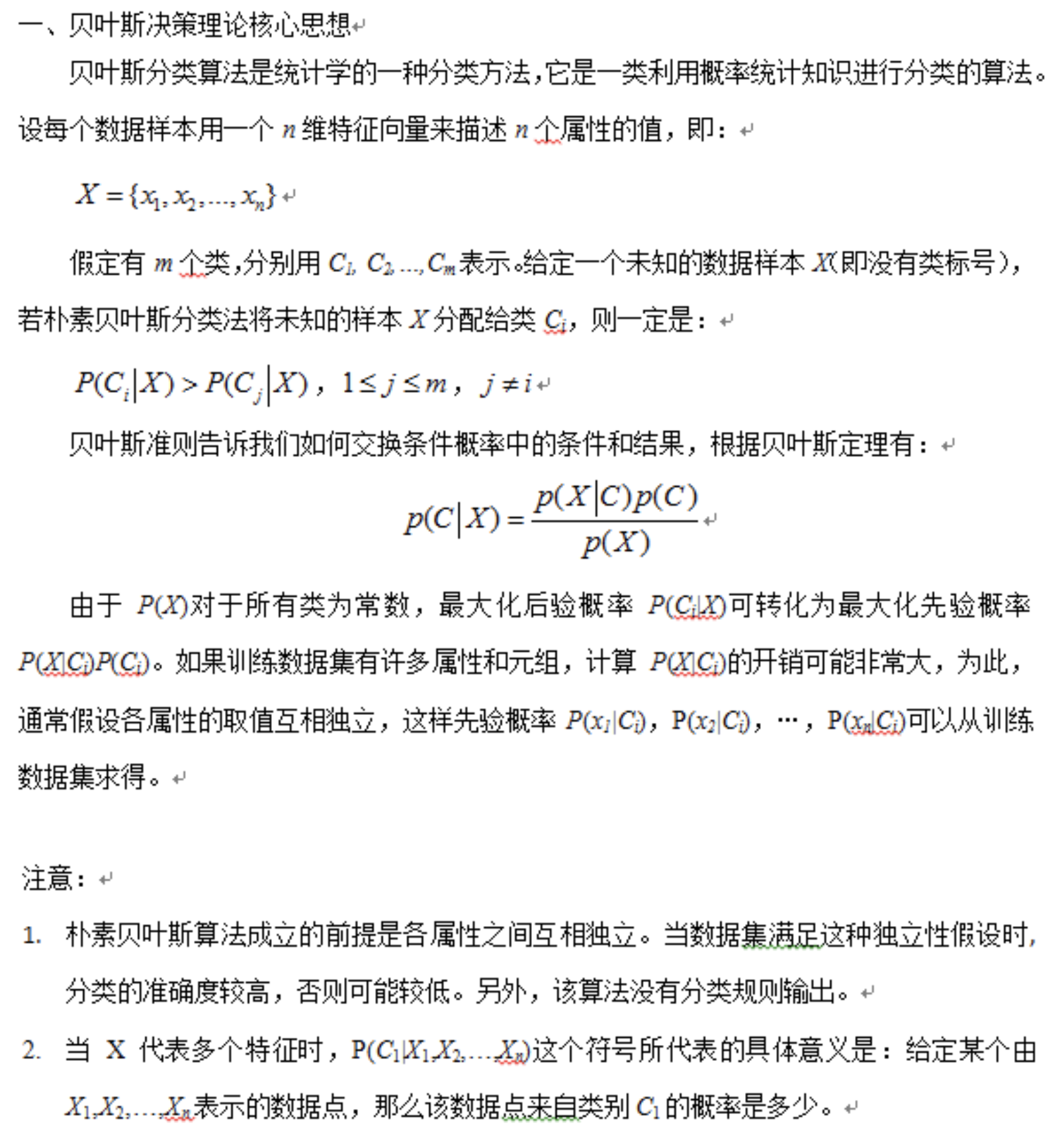
# 1. 朴素贝叶斯分类算法原理

## 1.1 概述

朴素贝叶斯分类算法是贝叶斯分类算法中最简单的一种

*注：朴素的意思是条件概率独立性，也就是我们的特征之间互相独立。*

## 1.2 算法思想



再扩展一下，假如在街上看到一个黑人讲英语，那我们是怎么去判断他来自于哪里？

提取特征：

肤色： 黑

语言： 英语

黑色人种来自非洲的概率： 80%

黑色人种来自于美国的概率：20%

讲英语的人来自于非洲的概率：10%

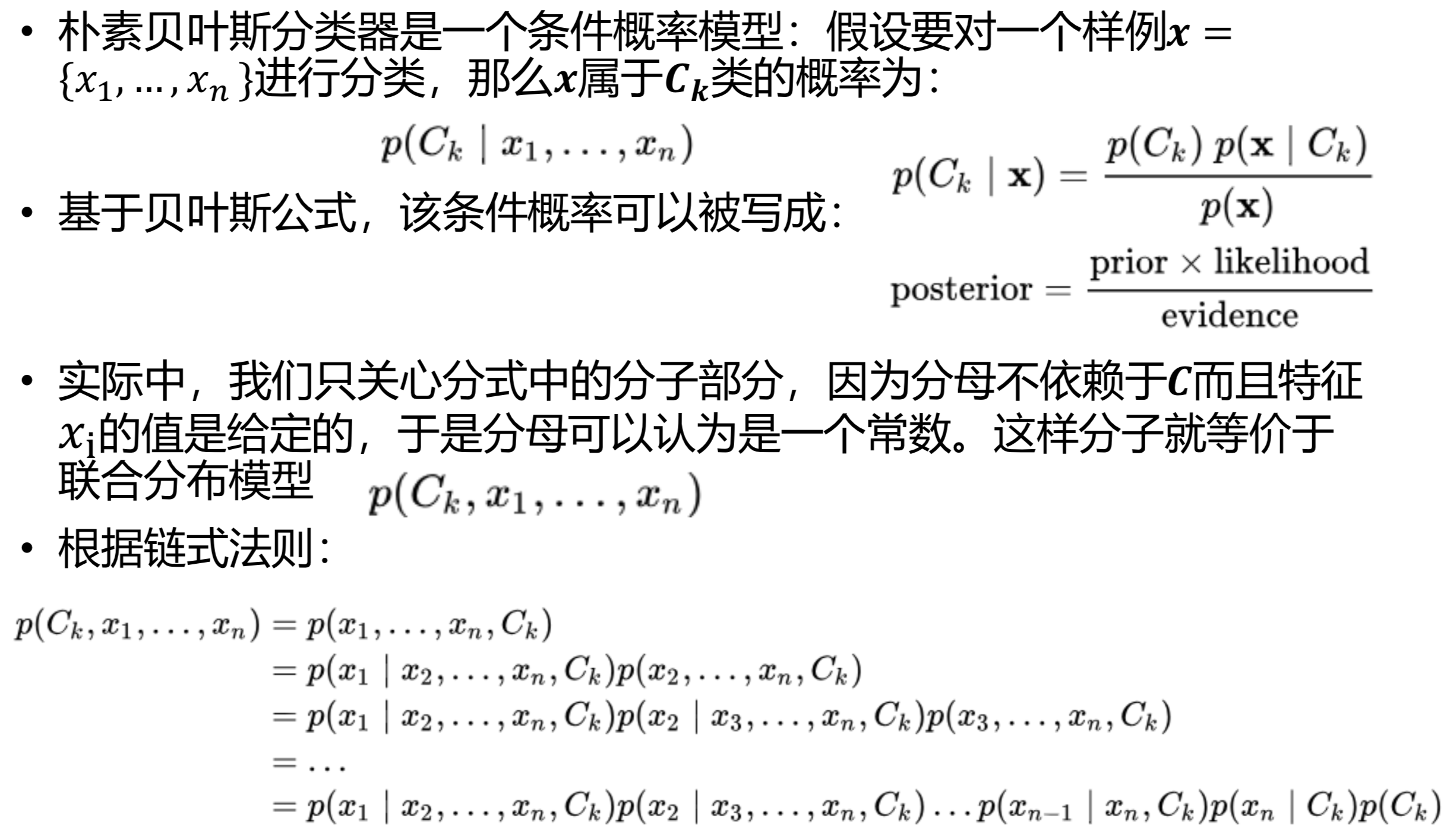
讲英语的人来自于美国的概率：90%

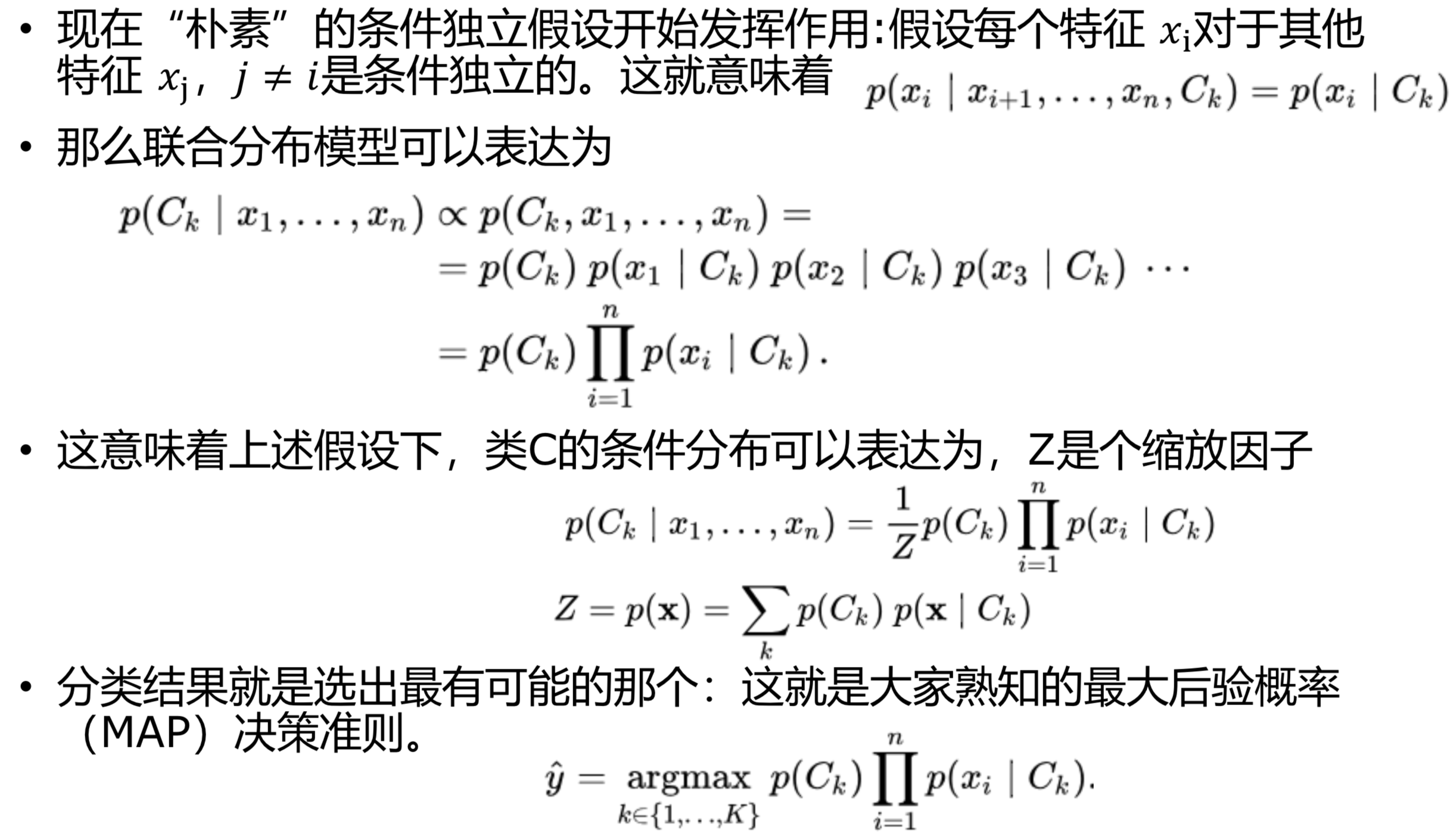
在我们的自然思维方式中，就会这样判断：

这个人来自非洲的概率：80% \* 10% = 0.08

这个人来自美国的概率：20% \* 90% =0.18

我们的判断结果就是：此人来自美国！

****

****

## 1.3 算法要点

### 1.3.1 算法步骤

1. 分解各类先验样本数据中的特征
2. 计算各类数据中，各特征的条件概率

*（比如：特征1出现的情况下，属于A类的概率p(A|特征1)，属于B类的概率p(B|特征1)，属于C类的概率p(C|特征1)......）*

1. 分解待分类数据中的特征（特征1、特征2、特征3、特征4......）
2. 计算各特征的各条件概率的乘积，如下所示：

判断为A类的概率：*p(A|特征1)\*p(A|特征2)\*p(A|特征3)\*p(A|特征4).....*

判断为B类的概率：*p(B|特征1)\*p(B|特征2)\*p(B|特征3)\*p(B|特征4).....*

判断为C类的概率：*p(C|特征1)\*p(C|特征2)\*p(C|特征3)\*p(C|特征4).....*

**......**

1. 结果中的最大值就是该样本所属的类别

### 1.3.2 算法应用举例

大众点评、淘宝等电商上都会有大量的用户评论，比如：

|  |  |
| --- | --- |
| *1、衣服质量太差了！！！！颜色根本不纯！！！*  *2、我有一有种上当受骗的感觉！！！！*  *3、质量太差，衣服拿到手感觉像旧货！！！*  *4、上身漂亮，合身，很帅，给卖家点赞*  *5、穿上衣服帅呆了，给点一万个赞*  *6、我在他家买了三件衣服！！！！质量都很差！* | *0*  *0*  *0*  *1*  *1*  *0* |

其中1/2/3/6是差评，4/5是好评

现在需要使用朴素贝叶斯分类算法来自动分类其他的评论，比如：

|  |
| --- |
| *a、这么差的衣服以后再也不买了*  *b、帅，有逼格*  *……* |

目的：是已经这句话求出好评和差评的概率，看哪个大。

步骤：求出差评的概率，求出好评的概率，求出在差评（好评）情况下，出现这句话的概率，这句话可以用多个单词计算，这样会有多个单词的联合概率，根据上面的链式法则，假设单词之间是互相独立的，则只需要求出差评（好评）中出现这个词的概率即可。这样根据全概率公式就求出结果了。

### 1.3.3 算法应用流程

1. 分解出先验数据中的各特征

*（即分词，比如“衣服”“质量太差”“差”“不纯”“帅”“漂亮”，“赞”……）*

1. 计算各类别（好评、差评）中，各特征的条件概率

（比如 p(“衣服”|差评)、p(“衣服”|好评)、p(“差”|好评) 、p(“差”|差评)……）

1. 分解出待分类样本的各特征

（比如分解a： “差” “衣服” ……）

1. 计算类别概率

P(好评|句子) = p(单词1|好评) \*p(单词2|好评)\*……p(好评)/p(句子)

P(差评|句子) = p(单词1|差评) \*p(单词2|差评)\*……p(差评)/p(句子)

其中句子是多个单词组成的，都是分母可以省略

5、显然P(差评|句子)的结果值更大，因此a被判别为“差评”

# 2. 朴素贝叶斯分类算法案例1

## 2.1 需求

**客户评论分类：**

以在线社区的留言板为例。为了不影响社区的发展，我们要屏蔽侮辱性的言论，所以要构建一个快速过滤器，如果某条留言使用了负面或者侮辱性的语言，那么就将该留言标识为内容不当。过滤这类内容是一个很常见的需求。对此问题建立两个类别：侮辱类和非侮辱类，使用1和0分别标识。

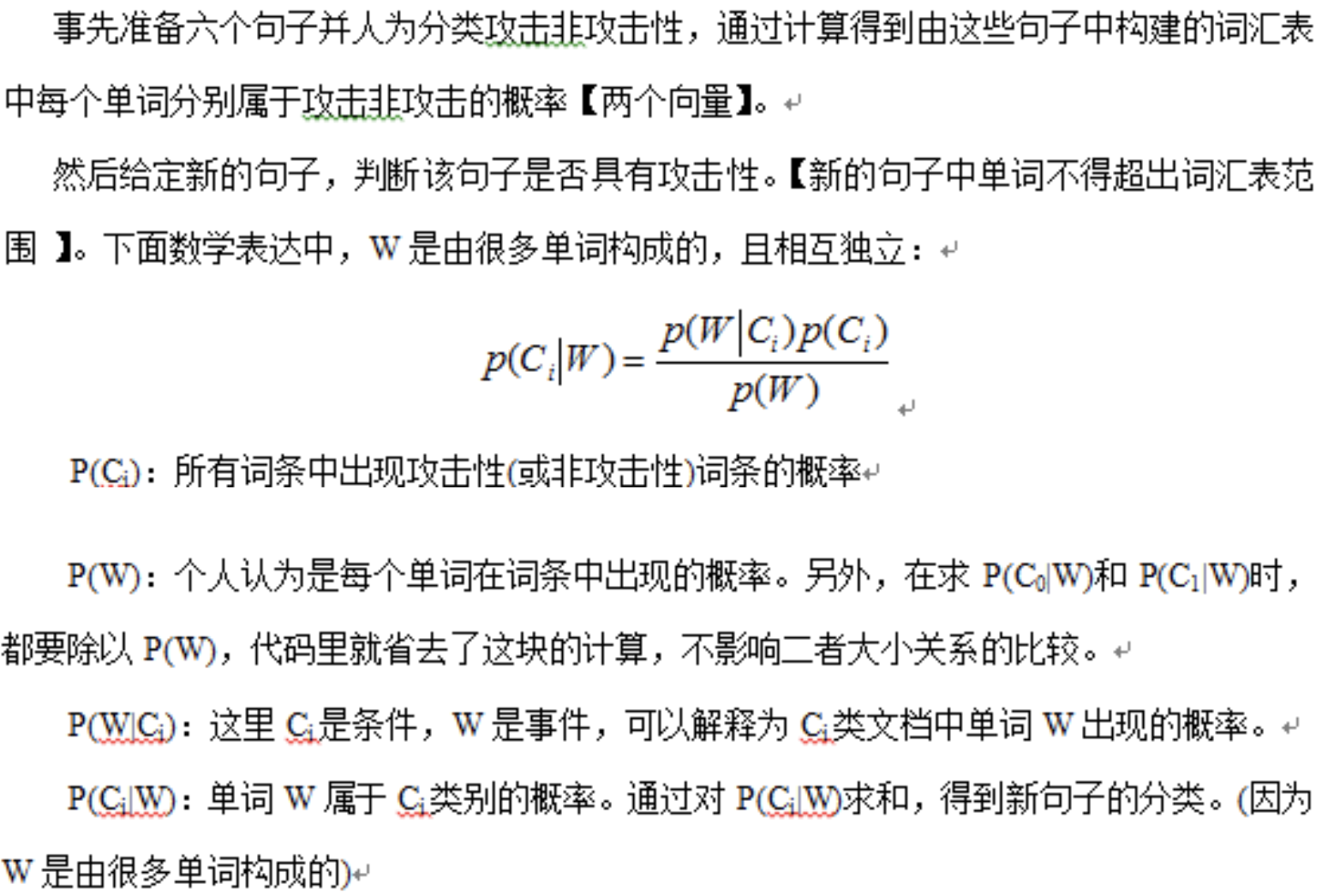
有以下先验数据，使用bayes算法对未知类别数据分类

|  |  |
| --- | --- |
| 帖子内容 | 类别 |
| 'my','dog','has','flea','problems','help','please' | 0 |
| 'maybe','not','take','him','to','dog','park','stupid' | 1 |
| 'my','dalmation','is','so','cute','I','love','him' | 0 |
| 'stop','posting','stupid','worthless','garbage' | 1 |
| 'mr','licks','ate','my','steak','how','to','stop','him' | 0 |
| 'quit','buying','worthless','dog','food','stupid' | 1 |

待分类数据：

|  |  |
| --- | --- |
| 'love','my','dalmation' | ? |
| 'stupid','garbage' | ? |

## 2.2 模型分析



## 2.2 Python实现

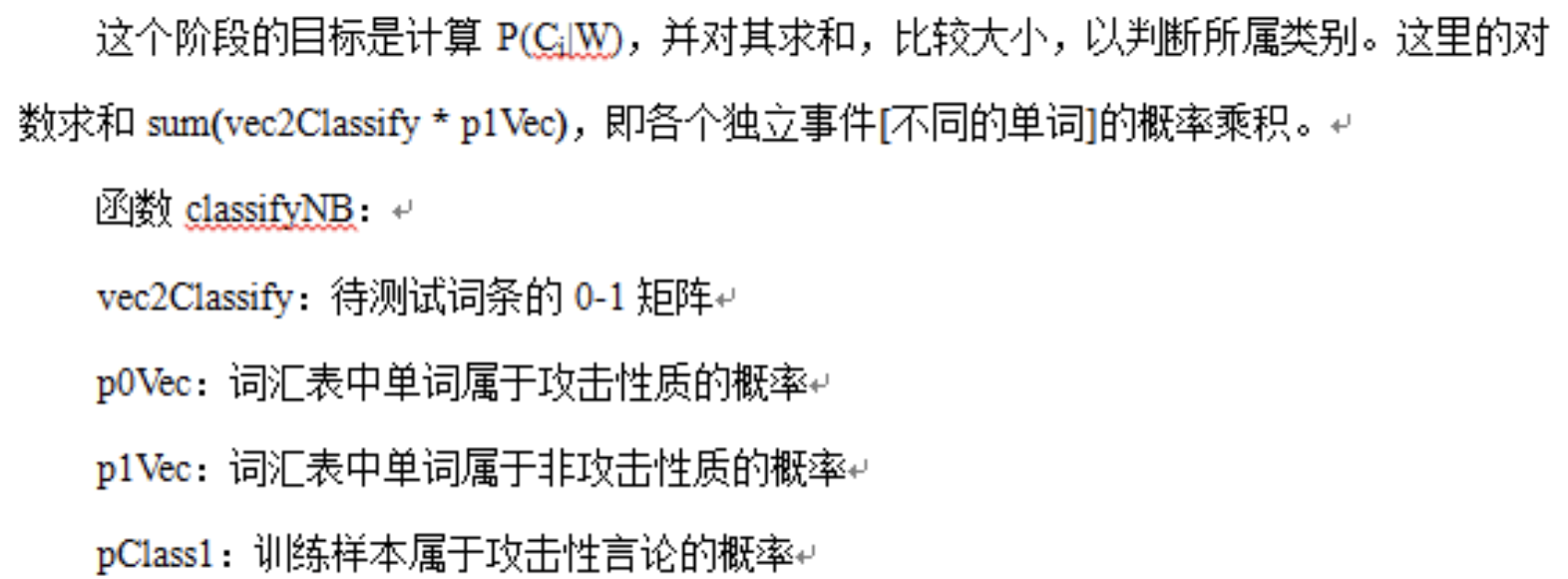
#### (1) 准备数据，构建词条向量

|  |
| --- |
| from numpy import \*  #过滤网站的恶意留言    def loadDataSet():  postingList = [['my','dog','has','flea','problems','help','please'],  ['maybe','not','take','him','to','dog','park','stupid'],  ['my','dalmation','is','so','cute','I','love','him'],  ['stop','posting','stupid','worthless','garbage'],  ['mr','licks','ate','my','steak','how','to','stop','him'],  ['quit','buying','worthless','dog','food','stupid']]  classVec = [0,1,0,1,0,1]  return postingList, classVec  # 创建一个包含在所有文档中出现的不重复的词汇表（列表）  def createVocabList(dataSet):  vocabSet = set([]) #创建一个空集  for document in dataSet:  vocabSet = vocabSet | set(document) #创建两个集合的并集  return list(vocabSet)  #将文档词条转换成词向量    def setOfWords2Vec(vocabList, inputSet):  returnVec = [0]\*len(vocabList) #创建一个其中所含元素都为0的向量  for word in inputSet:  if word in vocabList:  #returnVec[vocabList.index(word)] = 1 #index函数在字符串里找到字符第一次出现的位置 词集模型  returnVec[vocabList.index(word)] += 1 #文档的词袋模型 每个单词可以出现多次  else: print "the word: %s is not in my Vocabulary!" % word  return returnVec |

#### 训练算法

|  |
| --- |
| def trainNB0(trainMatrix, trainCategory):  numTrainDocs = len(trainMatrix)  numWords = len(trainMatrix[0])  pAbusive = sum(trainCategory)/float(numTrainDocs)  # p0Num = zeros(numWords); p1Num = zeros(numWords)  #p0Denom = 0.0; p1Denom = 0.0  p0Num = ones(numWords); #避免一个概率值为0,最后的乘积也为0  p1Num = ones(numWords); #用来统计两类数据中，各词的词频  p0Denom = 2.0; #用于统计0类中的总数  p1Denom = 2.0 #用于统计1类中的总数  for i in range(numTrainDocs):  if trainCategory[i] == 1:  p1Num += trainMatrix[i]  p1Denom += sum(trainMatrix[i])  else:  p0Num += trainMatrix[i]  p0Denom += sum(trainMatrix[i])  # p1Vect = p1Num / p1Denom  #p0Vect = p0Num / p0Denom  p1Vect = log(p1Num / p1Denom) #在类1中，每个次的发生概率  p0Vect = log(p0Num / p0Denom) #避免下溢出或者浮点数舍入导致的错误 下溢出是由太多很小的数相乘得到的  return p0Vect, p1Vect, pAbusive |

#### 测试分类器



注意：主要从以下两点对分类器进行修改

1. 贝叶斯概率需要计算多个概率的乘积以获得文档属于某个类别的概率，即计算p(w0|1)p(w1|1)p(w2|1)。如果其中一个概率值为0，那么最后的乘积也为0
2. 第二个问题就是下溢出，这是由于太多过小的数相乘造成的。由于大部分因子都非常小，所以程序会下溢出或者得不到正确的答案。解决办法是对乘积取自然对数这样可以避免下溢出或者浮点数舍入导致的错误。
3. 每个单词的出现与否作为一个特征，被称为词集模型；在词袋模型中，每个单词可以出现多次。

|  |
| --- |
| #测试分类器  def classifyNB(vec2Classify, p0Vec, p1Vec, pClass1):  p1 = sum(vec2Classify\*p1Vec) + log(pClass1)  p0 = sum(vec2Classify\*p0Vec) + log(1.0-pClass1)  if p1 > p0:  return 1  else: return 0  下面是完整的测试流程  def testingNB():  listOPosts, listClasses = loadDataSet()  myVocabList = createVocabList(listOPosts)  trainMat = []  for postinDoc in listOPosts:  trainMat.append(setOfWords2Vec(myVocabList, postinDoc))  p0V, p1V, pAb = trainNB0(array(trainMat), array(listClasses))  testEntry = ['love','my','dalmation']  thisDoc = array(setOfWords2Vec(myVocabList, testEntry))  print testEntry, 'classified as: ', classifyNB(thisDoc, p0V, p1V, pAb)  testEntry = ['stupid','garbage']  thisDoc = array(setOfWords2Vec(myVocabList, testEntry))  print testEntry, 'classified as: ', classifyNB(thisDoc, p0V, p1V, pAb) |

#### 运行测试

|  |
| --- |
| >>>testingNB()  ['love','my','dalmation'] classified as: 0  ['stupid','garbage'] classified as: 1 |

# 朴素贝叶斯分类算法案例2

## 3.1 需求

利用大量邮件先验数据，使用朴素贝叶斯分类算法来自动识别垃圾邮件

## 3.2 python实现

|  |
| --- |
| #过滤垃圾邮件  def textParse(bigString): #正则表达式进行文本解析  import re  listOfTokens = re.split(r'\W\*',bigString)  return [tok.lower() for tok in listOfTokens if len(tok) > 2]  def spamTest():  docList = []; classList = []; fullText = []  for i in range(1,26): #导入并解析文本文件  wordList = textParse(open('email/spam/%d.txt' % i).read())  docList.append(wordList)  fullText.extend(wordList)  classList.append(1)  wordList = textParse(open('email/ham/%d.txt' % i).read())  docList.append(wordList)  fullText.extend(wordList)  classList.append(0)  vocabList = createVocabList(docList)  trainingSet = range(50);testSet = []  for i in range(10): #随机构建训练集  randIndex = int(random.uniform(0,len(trainingSet)))  testSet.append(trainingSet[randIndex]) #随机挑选一个文档索引号放入测试集  del(trainingSet[randIndex]) #将该文档索引号从训练集中剔除  trainMat = []; trainClasses = []  for docIndex in trainingSet:  trainMat.append(setOfWords2Vec(vocabList, docList[docIndex]))  trainClasses.append(classList[docIndex])  p0V, p1V, pSpam = trainNB0(array(trainMat), array(trainClasses))  errorCount = 0  for docIndex in testSet: #对测试集进行分类  wordVector = setOfWords2Vec(vocabList, docList[docIndex])  if classifyNB(array(wordVector), p0V, p1V, pSpam) != classList[docIndex]:  errorCount += 1  print 'the error rate is: ', float(errorCount)/len(testSet) |