用于学习和分类文本的朴素贝叶斯算法

- Learn_Naive_Bayes_Text(Examples, V)

  Examples为一组文本文档以及它们的目标值。V为所有可能目标值的集合。此函数作用是学习概率项$P(w_k|v_j)$和$P(v_j)$。
  - 收集Examples中所有的单词、标点符号以及其他记号
    - Vocabulary←在Examples中任意文本文档中出现的所有单词及记号的集合
  - 计算所需要的概率项$P(v_j)$和$P(w_k|v_j)$
    - 对V中每个目标值v_j
      - $docs_j$←Examples中目标值为v_j的文档子集
      - $P(v_j)$←$|docs_j|$ / |Examples|
      - $Text_j$←将$docs_j$中所有成员连接起来建立的单个文档
      - n←在$Text_j$中不同单词位置的总数
      - 对Vocabulary中每个单词w_k
        - n_k←单词w_k出现在$Text_j$中的次数
        - $P(w_k|v_j)$←(n_k+1) / (n+|Vocabulary|)

把属于vj类的所有文章合成一篇文章

Textj中的总位置数

- 分类状态得分（CSV, Categorization Status Value）：用于描述将文档归于某个类别下有多大的可信度。
- 准确率（Precision）：在所有被判断为正确的文档中，有多大比例是确实正确的。
- 召回率（Recall）：在所有确实正确的文档中，有多大比例被我们判为正确。
- 假设：计算机对训练集背后的真实模型（真实的分类规则）的猜测称为假设。可以把真实的分类规则想像为一个目标函数，我们的假设则是另一个函数，假设函数在所有的训练数据上都得出与真实函数相同（或足够接近）的结果。
- 泛化性：一个假设能够正确分类训练集之外数据（即新的，未知的数据）的能力称为该假设的泛化性。

输入原训练集

➢概率估计

$$v_{NB} = \arg\max_{v_j \in \{like, dislike\}} P(v_j) \prod_{i=1}^{19} P(a_i \mid v_j)$$

$$= \arg\max_{v_j \in \{like, dislike\}} P(v_j) P(a_1 = "this" \mid v_j) \ldots P(a_{19} = "sentences" \mid v_j)$$

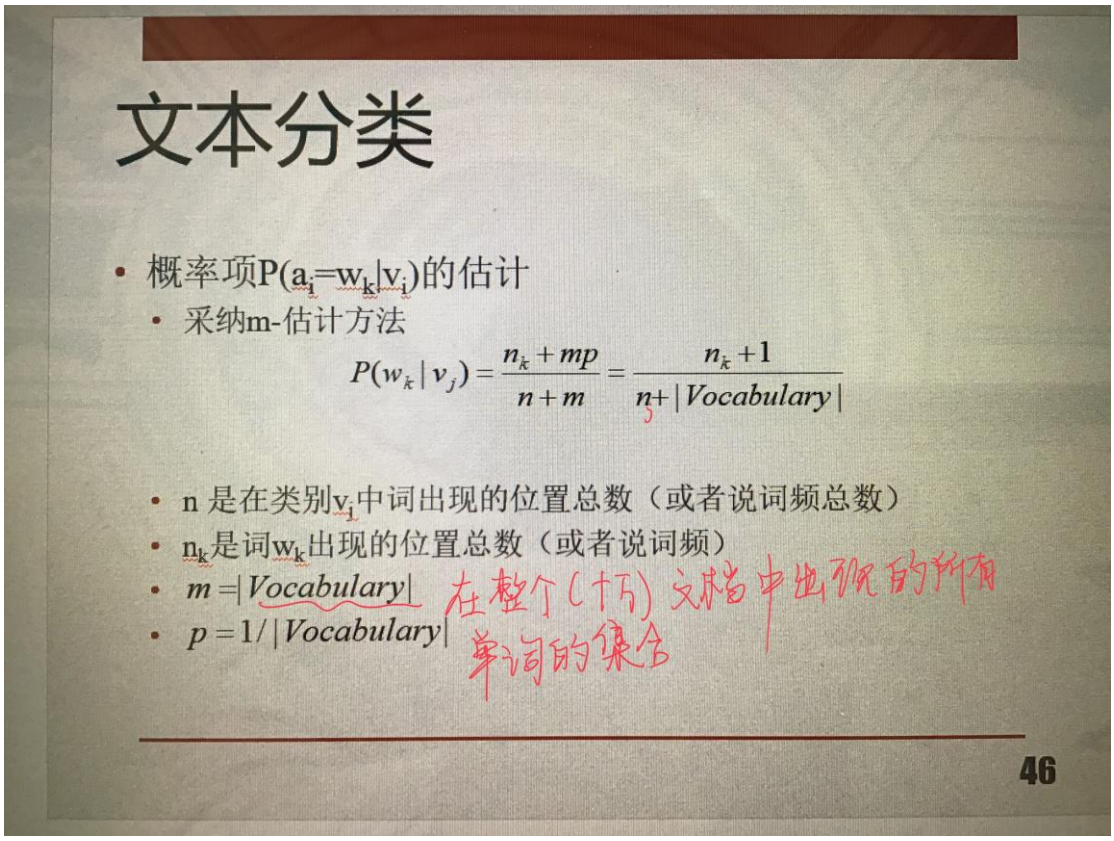
- 此处贝叶斯分类器隐含的独立性假设并不成立。通常，某个位置上出现某个单词的概率与前后位置上出现的单词是相关的
- 但是在实践中，朴素贝叶斯学习器在许多文本分类问题中性能非常好

当前待分类文本的向量表示
所有类的向量表示

文本分类

举例：学习分类文本（2）

- 应用朴素贝叶斯分类器的两个主要设计问题：
  - 怎样将任意文档表示为属性值的形式
  - 如何估计朴素贝叶斯分类器所需的概率
- 表示文档的方法
  - 给定一个文本文档，对每个单词的位置定义一个属性，该属性的值为在此位置上找到的英文单词
- 假定我们共有1000个训练文档，其中700个分类为dislike，300个分类为like，现在要对下面的新文档进行分类：
  - This is an example document for the naive Bayes classifier. This document contains only one paragraph, or two sentences.



评估指标

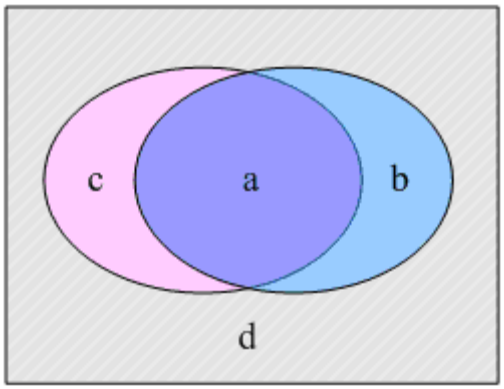
人们根据不同的文本分类应用背景提出了多种评估分类系统性能的标准。常用的评估标准：召回率(Recall)、准确率(Precision)、F1-评测值(F1-measure)、微平均(Micro-average)和宏平均(Macro-average)。另外一些使用较少的评估方法包括平衡点(break-even point)、11 点平均正确率(11-point average precision)等。本文中所涉及到的“精度”(Accuracy)一般指广义精度，可以代表召回率、精确率、F1-评测值(简记：F1 值)、微平均和宏平均等评价指标。

假设一个文本分类系统针对类别 c_i 的分类标注结果统计如表所示：

文本与类别的实际关系 / 分类器的分类判断	属于	不属于
标记为“是”	a	b
标记为“否”	c	d

表 1 分类结果邻接表

或者用等价的集合描述如图所示:



分类结果集合示意图

图左侧椭圆表示实际测试集类别标注，右侧椭圆是经过分类器分类后标注的分类结果。上图与表的中符号的意义如下：

1) a 表示正确地标注测试集文本为类别 c_i 的文本数量；

2) b 表示错误地标注测试集文本为类别 c_i 的文本数量；

3) c 表示错误地排除测试集文本在类别 c_i 之外的文本数量；

4) d 表示正确地排除测试集文本在类别 c_i 之外的文本数量。

召回率与准确率

a) 分类器在类别 c_i 上的召回率(又称查全率)定义如式：

$$recall_i = \frac{a}{a+c} \times 100\%$$

分母为实际=5000

b) 分类器在类别 c_i 上的准确率(又称查准率)定义如式：

$$precision_i = \frac{a}{a+b} \times 100\%$$

分母为分完类后被分到这类下面的文本数

F1-评测值

c) 分类器在类别 c_i 上的 F1 值定义如式：

$$F_{1i} = \frac{2 \times precision_i \times recall_i}{precision_i + recall_i}$$

召回率和准确率分别从两个方面考察分类器的分类性能。召回率过高可能导致准确率过低，反之亦然。所以综合考虑分类结果召回率和准确率的平衡，采用F1-评测值比较合理。