4.3.2.4 朴素贝叶斯分类器总结

朴素贝叶斯分类器是一个很容易理解的模型，对它的总结如下：

1. 朴素贝叶斯分类器**原理简单**，**预测准确度较好**，故常被使用。
2. 朴素贝叶斯分类器的使用前提是假设所有属性之间是**条件独立**的。这虽然是一个比较苛刻的假设，很多时候并不符合实际情况，但却为后续计算大幅度**节省了成本**。

4.3.3 朴素贝叶斯分类器在文本分类中的应用

文本分类（text categorization）是自然语言处理中相当经典的问题，在生活中也非常常见。文本分类会对输入的一大串文本进行特征提取，判定其属于哪一类，比如对垃圾邮件的判定，对文章种类的判定等。文本分类的主要步骤，包括对输入的一大串文本进行特征提取，对这些特征进行表示，并把提取的特征映射到类别。通过这几步，能够从文本数据获取最具代表性且更易处理的特征，并完成分类任务。

下面介绍的文本分类方法主要适用于英文文本分类。对于文本分类的特征提取步骤，将介绍一种常用的方法：词袋法。对于文本分类的其他步骤，将介绍两种基于不同概率分布的朴素贝叶斯分类器。

4.3.3.1 词袋法

词袋法通常用于文本分类的特征提取。顾名思义，“词袋”即“词语的口袋”。词袋法实质上可以看做是 $N=1$ 时的 N-gram 模型[9]，**其忽略文本自身的语法和语序等因素，将文本看作是多个词的集合，而词之间是相互独立的。**当外界输入一个文本的模型时，模型首先对其进行预处理，进行比如分词、去停用词等操作，从而将一大段文本转为词汇的集合 $\mathcal{S}$。接着，模型进行特征提取步骤。我们需要预先准备一个包含大量词汇的字典（dictionary），每个词汇就相当于一个特征。模型将对照字典中的每个特征，判定集合 $\mathcal{S}$ 中是否也包含它，并将结果记录下来。为了记录此种信息，研究者们提出了向量表示方法。常见的两种方法分别为布尔值表示法和词频表示法。

布尔值表示法　根据词是否出现进行表示，1 表示出现，0 表示未出现，最后得到的向量形如 $[1,0,1,1,0,\cdots,1,0]$，如图4.10。

对于在集合 S 中出现而字典中没有出现的词，可以在字典中设置一个并不是词的特征，即 <UNK>（unknown），在词频法中，所有未知词的个数即是该特征的值，而在用布尔值表示的方法中，一旦出现未知词，该特征就会被赋予 1 的值。

词频表示法　词频表示法根据词出现的频率进行表示，最后得到的向量形如 $[2,2,1,1,0,\cdots,0,1]$，该向量的每一个元素对应一个词在文本中出现的频率，如图4.11。

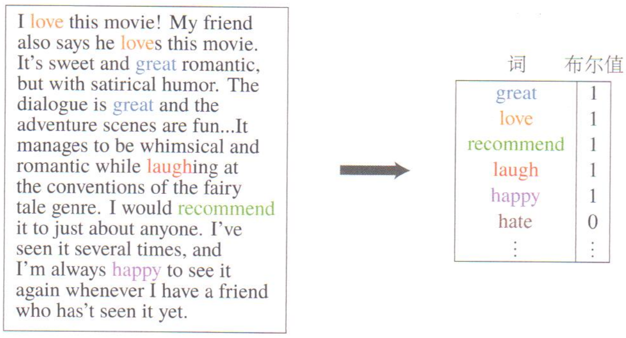

图 4.10 布尔值表示法

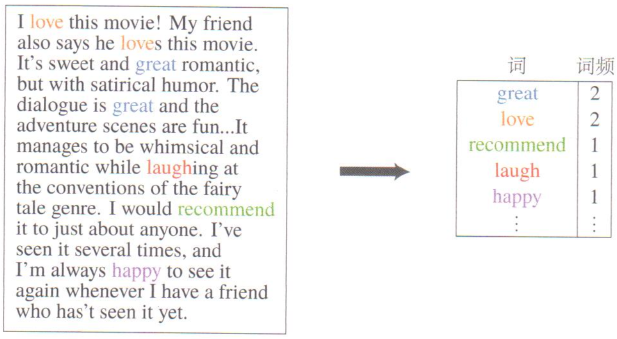

图 4.11 词频表示法

一般来说，为了保证所有词都能囊括在字典里，字典的规模会特别大。另外，由于很多词语未必会出现，字典内存在大量的 0，这就造成了词袋模型存在两个问题——高维度性和高稀疏性。正是这两种特性，导致词袋模型不仅在存储上具有极大的空间复杂度，同时计算的时间复杂度也不小。同时，由于词袋模型忽略了上下文关系，故会造成部分信息的丢失，这会对预测的准确性造成影响。

4.3.3.2 多元伯努利分布和多项式分布

现在，我们已经完成了预处理和特征提取这两步，接下来需要构建分类器，完成从词向量到类别标签的映射。

由于在词袋法中，可以将词语所映射的特征视为随机变量，那么自然可以使用适当的概率分布假设来完成分类。在布尔值表示法中，使用布尔值来表示词向量，单个词则可以假设其符合伯努利分布，对于整个词向量，可以假设其符合**多元伯努利分布**（**multivariate Bernoulli distribution**）；词频表示法使用词频来表示词向量，作为多元伯努利的推广，可以假设其符合**多项式分布**（**multinomial distribution**）。下面将对这两种分布进行具体介绍。

多元伯努利分布 在了解多元伯努利分布之前，先来回顾一下伯努利分布。

伯努利分布，又称为两点分布或者 0-1 分布，是一种离散型的概率分布，若随机变量 X 服从伯努利分布，则该随机变量的取值只有两种可能，用 0 和 1 来表示这两种可能，随机变量取值为 1 的概率为 $p(0 < p < 1)$，它的概率表达式为

$$\mathbb{P}(X = x) = p^x(1 - p)^{1-x} \tag{4.31}$$

多元伯努利分布，即同时进行多个不同的伯努利实验，若发生了 n 次独立的伯努利实验，每个伯努利实验的概率参数为 p_i，那么 n 次独立伯努利实验的概率表达式为

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_{i=1}^{n} p_i^{x_i}(1 - p_i)^{1-x_i} \tag{4.32}$$

多项式分布 多项式分布其实就是伯努利分布的推广。在一次实验中，对于随机变量 X，设其有 d 种状态（当 $d = 2$ 时，多项式分布本质上就是伯努利分布），可以将其表示为一个 d 维的向量，每一维代表一种状态，随机变量可以表示为 $X = (X_1, X_2, X_3, \cdots, X_d)$，且 $X_i \in \{0, 1\}$。假设 $X_i = 1$ 的概率为 μ_i，且 $\sum\limits_{i=1}^{d} \mu_i = 1$，该随机变量的概率表达式为

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_d = x_d) = \prod_{i=1}^{d} \mu_i^{x_i} \tag{4.33}$$

若发生了 n 次独立实验，假设出现了 m_i 次 $X_i = 1$ 的情况，且 $\sum\limits_{i=1}^{d} m_i = n$，那么概率可表示为

$$\mathbb{P}(X_1 = m_1, X_2 = m_2, \cdots, X_d = m_d) = \frac{n!}{m_1! m_2! \cdots m_d!} \prod_{i=1}^{d} \mu_i^{m_i} \tag{4.34}$$

4.3.3.3 基于伯努利分布的朴素贝叶斯

考虑单词的出现服从伯努利分布，即一个单词在文档中会被标记为出现（True）或者不出现（False）这两种可能。具体来说，用 W_i，$i = 1, 2, \cdots, k$ 表示字典中的每一个词语，其中字典中总词数为 k。对于一篇需要分类的文档 D，$W_i = \text{True}$ 当且仅当单词 W_i 出现在 D 中，否则 $W_i = \text{False}$。因此，对于某一篇文档 D，单词出现的概率可以表示为

$$\mathbb{P}(W_1 = \text{True}, W_2 = \text{False}, \cdots, W_k = \text{True} \mid C = c) \tag{4.35}$$

其中 C 为一随机变量，表示文档 D 所属类别。除此之外，基于朴素贝叶斯分类器的假设，给定文档的类别后，每一个词语之间应该是独立的，所以有

$$\begin{aligned} &\mathbb{P}(W_1 = \text{True}, W_2 = \text{False}, \cdots, W_k = \text{True} \mid C = c) \\ &= \mathbb{P}(W_1 = \text{True} \mid C = c) \times \cdots \times \mathbb{P}(W_k = \text{True} \mid C = c) \end{aligned} \tag{4.36}$$

其中每一个 $\mathbb{P}(W_i = \text{True} \mid C = c)$ 都服从伯努利分布。这种基于伯努利分布的分类器很适合具有二值的变量 (binary variable)。而且对于每个单词，只需要计算 $\mathbb{P}(W_i = \text{True} \mid C =$

c)，因为

$$\mathbb{P}(W_i = \text{False} \mid C = c) = 1 - \mathbb{P}(W_i = \text{True} \mid C = c) \tag{4.37}$$

对于基于伯努利分布的朴素贝叶斯分类器，概率值的计算和朴素贝叶斯模型相似，用频率来代替概率，即

$$\mathbb{P}(W_i = \text{True} \mid C = c) = \frac{N(W_i = \text{True}, C = c)}{N(C = c)} \tag{4.38}$$

直接的解释是所有类别为 c 的文本中，出现单词 W_i 的文本的比例。先验概率 $\mathbb{P}(C = c)$ 为

$$\mathbb{P}(C = c) = \frac{N(C = c)}{N(D)} \tag{4.39}$$

其中 $N(D)$ 为文档的总数量。先验概率即为所有文档中类别为 c 的文档比例。

有了 $\mathbb{P}(W_i = \text{True} \mid C = c)$ 和先验概率 $\mathbb{P}(C = c)$，就可以通过计算 $\mathbb{P}(C = c \mid W)$ 来对文本进行分类了。为了处理未出现的值，同样可以用 §4.3.2.1中介绍的平滑化方法来避免概率为 0 的值出现。

在用朴素贝叶斯分类器时，还存在概率值计算**算术下溢**（**arithmetic underflow**）的问题。因为概率值都是属于 $[0, 1]$ 范围的数，多个概率值相乘之后，计算结果可能会超过计算机内存所能表示的范围，造成算术下溢，这时计算结果就变成了 0。为了解决这种问题，并不直接对概率值做乘法，而是通过 log 函数来将乘法转换成加法。由于 log 函数是单调函数，这种操作并不会影响分类结果。具体来说，对于本节考虑的文本分类问题，需要计算

$$\begin{aligned}
\widehat{c} &= \arg\max_{c} \ \mathbb{P}(W_1, \cdots, W_k \mid c)\mathbb{P}(c) \\
&= \arg\max_{c} \ \log \mathbb{P}(W_1, \cdots, W_k \mid c)\mathbb{P}(c) \\
&= \arg\max_{c} \ \log \mathbb{P}(c) + \sum_{i=1}^{k} \log \mathbb{P}(W_i \mid c)
\end{aligned} \tag{4.40}$$

这样，通过将乘法运算转换成了加法运算，进而解决了算术下溢的问题。

4.3.3.4 基于多项式分布的朴素贝叶斯

下面，再介绍另外一种基于多项式分布的模型。一个多项式分布由这个参数决定：实验重复的次数 n 以及每次实验成功的概率 $p_1, p_2, \cdots, p_n$。在文本分类问题中，假设每个单词 W_i 的出现次数 n_i 服从一个多项式分布。这时，对于某一篇文档 D，它出现的概率可以表示为

$$\begin{aligned}
&\mathbb{P}(W_1 = n_1, W_2 = n_2, \cdots, W_k = n_k \mid C = c, N, p_{1,c}, \cdots, p_{k,c}) \\
&= \frac{N!}{n_1! n_2! \cdots n_k!} \cdot p_{1,c}^{n_1} p_{2,c}^{n_2} \cdots p_{k,c}^{n_k}
\end{aligned} \tag{4.41}$$

这里 N 是文档 D 中的单词总数，$p_{i,c}$ 表示对于类别为 c 的文档，单词 W_i 出现一次的概率。特别地，有

$$\sum_{i=1}^{k} n_i = N, \quad \sum_{i=1}^{k} p_{i,c} = 1 \tag{4.42}$$

注意到公式（4.41）第一项与分类实际上无关，只需要计算

$$\widehat{c} = \arg\max_{c} \mathbb{P}(C=c) \cdot p_{1,c}^{n_1} p_{2,c}^{n_2} \cdots p_{k,c}^{n_k} \tag{4.43}$$

对于基于多项式分布的朴素贝叶斯分类器，先验概率的计算应该为

$$\mathbb{P}(C=c) = \frac{N(C=c)}{N(D)} \tag{4.44}$$

与其他几种朴素贝叶斯分类器相同。而条件概率的计算有所不同

$$\mathbb{P}(W_i = n_i \mid C=c) = \frac{n_{i,c}}{n_c} \tag{4.45}$$

这里 $n_{i,c}$ 是所有类别为 c 的文本中单词 W_i 出现的次数，n_c 是所有类别为 c 的文档的总单词数。同样地，可以使用 §4.3.2.1中介绍的平滑化方法来避免零概率出现，比如

$$\mathbb{P}(W_i = n_i \mid C=c) = \frac{n_{i,c}+1}{n_c+k} \tag{4.46}$$

这里 k 是词典的大小（即词典中单词数）。

4.3.3.5 两种朴素贝叶斯方法在文本分类上的效果

这里，依据论文[10]，在数据集 WebKB 4 上简单比较一下两种朴素贝叶斯方法在文本分类上的效果。由图4.12可见，当字典规模不是很大时，采用基于多元伯努利分布的概率模型（布尔值表示法）的准确率比采用基于多项式分布的概率模型（词频表示法）高。但是随着字典规模逐渐变大，后者的准确率大于前者，而前者的准确率持续走低。

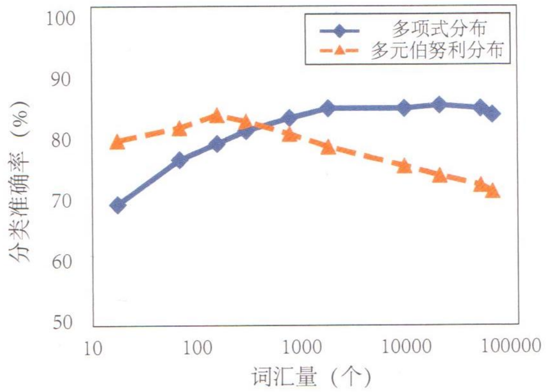

图 4.12　基于多元伯努利分布的概率模型 VS. 基于多项式分布的概率模型