机器学习讲义(L17-C)：朴素贝叶斯分类在文本分类中的应用

授课教师：王贝伦 / 助教：张嘉琦，黄旭，谈笑，徐浩卿

1 基于字典的对文本的向量空间表示

文本分类（Text categoriation）是自然语言处理中相当经典的问题，在我们生活中也非常常见，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，根据视频的标题为其打上“娱乐”“体育”等这样的标签。
问题在于，我们怎么对输入的一大串文本进行特征提取，怎么去表示这些特征，又怎么把提取的特征映射到类别。
我们这里介绍的文本分类主要适用于英文文本分类。

1.1 词袋法（Bags of words, BoW）

词袋法实质上可以看做是 $N=1$ 时的 N-gram 模型，**其忽略文本自身的语法和语序等因素，将文本看做是多个词的集合，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、去停用词的操作，将一大段文本转为词的集合 S。其次进行特征提取，我们要预先准备一个字典（dictionary），其中包含大量词汇，这就相当于特征，对照字典中的每个特征，判定 S 中是否也包含，并记录下来。由此，为了记录此种信息，出现了两种向量表示方法，一是词频表示法，即根据词出现的频率进行表示，最后得到的向量形如 $[2,2,1,1,0,...,0,1]$，如图 1；二是布尔值表示法，根据词是否出现进行表示，1 表示 Yes，0 表示 No，最后得到的向量形如 $[1,0,1,1,0,...,1,0]$，如图 2。

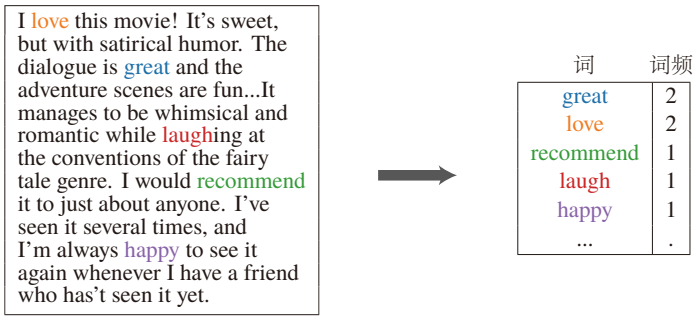

Figure 1: 词频表示法

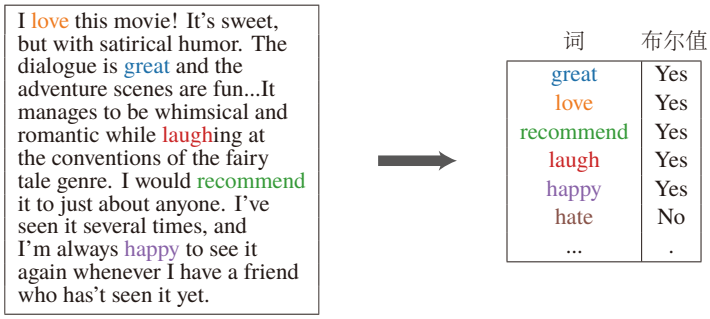

Figure 2: 布尔值表示法

而对于在 S 中出现，而字典中没有出现的词，我们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 Yes 的值。
一般来说，为了保证所有词都能囊括在字典里，不出现未知的词，字典的规模会特别大，这就造成了词袋这种模型本身最大的两个问题，也就是——高维度性和高稀疏性。正是这两种特性，导致其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。另一方面，由于词袋模型忽略了上下文关系，丢失了这部分信息，也会对预测的准确性造成影响。

2 多元伯努利分布和多项式分布

现在我们已经结束了预处理和特征提取这两步，那我们接下来该怎么构建分类器，完成从向量到类别的映射呢？
与高斯朴素贝叶斯相似，我们可以基于其他概率分布的假设来完成分类。在这里我们介绍两种分布，多元伯努利分布（Multivariate

Bernoulli Distribution）和多项式分布（Multinomial Distribution）。

2.1 多元伯努利分布

在了解多元伯努利分布之前，我们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，是一个离散型的概率分布，若随机变量 X 服从伯努利分布，对于它的两种状态，我们可以用 0 和 1 表示，已知取值为 1 的概率为 $p(0<p<1)$，其概率表示的式子为

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

多元伯努利分布，通俗说就是同时进行多个不同的伯努利实验，若发生了 n 次独立伯努利实验，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

$$P(X_1=x_1,X_2=x_2,...,X_n=x_n)=\prod_i p_i^{x_i}(1-p_i)^{(1-x_i)} \tag{2}$$

我们可以将基于该分布的概率模型套用至词袋的布尔值表示法，（注意：这里并没有做出如式子 (2) 中独立性假设）那么文本分类问题就转化为

$$\begin{gathered}\operatorname{argmax} P(W=w|C=c)P(C=c)\\ P(W=w|C=c)=P(W_1=w_1,W_2=w_2,...,W_k=w_k|C=c)\\ w_i\in\{0,1\}\end{gathered}$$

2.2 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，我们可以将其表达为一个 d 维的向量，每一维代表一种状态，我们表示为 $X=(X_1,X_2,X_3,...,X_d)$，且 $X_i\in\{0,1\}$，假设 $X_i=1$ 的概率是 μ_i，且 $\sum_{i=1}^{d}\mu_i=1$，那么概率表示的式子为

$$P(X_1=x_1,X_2=x_2,...,X_i=x_i,...,X_d=x_d)=\prod_{i=1}^{d}\mu_i^{x_i} \tag{3}$$

由此扩展到 n 次独立实验，假设出现了 m_i 次 $X_i=1$ 的情况，且 $\sum_i m_i=n$，那么概率可表示为

$$\begin{aligned}&P(X_1=m_1,X_2=m_2,...,X_i=m_i,...,X_d=m_d)\\ &=\frac{n!}{m_1!m_2!...m_d!}\prod_{i=1}^{d}\mu_i^{m_i}\end{aligned} \tag{4}$$

我们可以将基于该分布的概率模型套用至词袋的词频表示法，那么文本分类问题就转化为

$$\begin{gathered}\operatorname{argmax} P(W=w|C=c)P(C=c)\\ P(W=w|C=c)=P(W_1=n_1,W_2=n_2,...,W_k=m_k|C=c)\end{gathered}$$

3 基于伯努利分布的朴素贝叶斯

在上次的讲义中，我们介绍了基于高斯分布的朴素贝叶斯分类器，在这一节我们借助文本分类这个例子来介绍另外两个使用不同概率分布的朴素贝叶斯分类器。我们先考虑当我们假设单词的出现是属于伯努利分布（Bernoulli distribution）的，即一个单词在文档中会被表示为出现（True）或者不出现（False）这两种可能。具体来说，我们用 $W_i,\ i=1,2,\cdots,k$ 表示字典中的每一个词语，其中字典中总词数为 k。对于一篇我们需要分类的文档 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，它出现的概率可以表示为

$$P(W_1=\text{True}\ W_2=\text{False},\cdots,W_k=\text{True}\mid C=c) \tag{5}$$

其中 C 表示文档 d 的类别。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，所以我们有

$$\begin{aligned}&P(W_1=\text{True},W_2=\text{False},\cdots,W_k=\text{True}\mid C=c)\\ &=P(W_1=\text{True}\mid C=c)\times\cdots\times P(W_k=\text{True}\mid C=c)\end{aligned} \tag{6}$$

其中每一个 $P(W_i=\text{True}\mid C=c)$ 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，我们只需要计算 $P(W_i=\text{True}\mid C=c)$ 因为

$$P(W_i=\text{False}\mid C=c)=1-P(W_i=\text{True}\mid C=c) \tag{7}$$

对于基于伯努利分布的朴素贝叶斯，概率值的计算和我们最开始介绍的朴素贝叶斯模型相似，用频率来代替概率，即

$$P(W_i = \text{True} \mid C = c) = \frac{N(W_i = \text{True}, C = c)}{N(C = c)} \tag{8}$$

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 $P(C = c)$ 为

$$P(C = c) = \frac{N(C = c)}{N(d)} \tag{9}$$

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 $P(W_i = \text{True} \mid C = c)$ 和先验概率 $P(C = c)$ 我们就可以通过计算 $P(C = c \mid W)$ 来对文本进行分类了。为了处理未出现的值，我们同样可以用平滑的方式来避免概率为 0 的值出现。

在用朴素贝叶斯分类器时，还存在概率值计算**下溢出**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，造成下溢出。这时计算结果就变成了 0。因此为了解决这种问题，我们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。而且由于 $\log$ 函数是单调函数，这种操作并不会影响分类结果。具体来说，对于我们这一节考虑的文本分类问题，我们需要计算

$$\begin{aligned} \widehat{c} &= \operatorname*{argmax}_c P(W_1, \cdots, W_k \mid c) P(c) \\ &= \operatorname*{argmax}_c \log P(W_1, \cdots, W_k \mid c) P(c) \\ &= \operatorname{argmax} \log P(c) + \sum_{i=1}^{k} \log P(W_i \mid c) \end{aligned} \tag{10}$$

这样我们将乘法运算转换成了加法运算解决了运算下溢出的问题。

4 基于多项式分布的朴素贝叶斯

我们再介绍另外一种基于多项式分布（multinomial distribution）[2] 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，我们假设每个单词 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

$$\begin{aligned} &P(W_1 = n_1\ W_2 = n_2, \cdots, W_k = n_k \mid C = c, N, p_{1,c}, \cdots, p_{k,c}) \\ &= \frac{N!}{n_1! n_2! \cdots n_k!} \cdot p_{1,c}^{n_1} p_{2,c}^{n_2} \cdots p_{k,c}^{n_k} \end{aligned} \tag{11}$$

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，我们有

$$\sum_{i=1}^{k} n_i = N \qquad \sum_{i=1}^{k} p_{i,c} = 1 \tag{12}$$

注意到公式（11）第一项与分类实际上无关，我们只需要计算

$$\widehat{c} = \operatorname*{argmax}_c\ P(C = c) \cdot p_{1,c}^{n_1} p_{2,c}^{n_2} \cdots p_{k,c}^{n_k} \tag{13}$$

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

$$P(C = c) = \frac{N(C = c)}{N(d)} \tag{14}$$

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

$$P(W_i = n_i \mid C = c) = \frac{n_{i,c}}{n_c} \tag{15}$$

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 的文档的总单词书数。同样地，可以使用平滑方法来避免零概率出现，比如

$$P(W_i = n_i \mid C = c) = \frac{n_{i,c} + 1}{n_c + k} \tag{16}$$

这里 k 是词典的大小（即词典中单词数）。

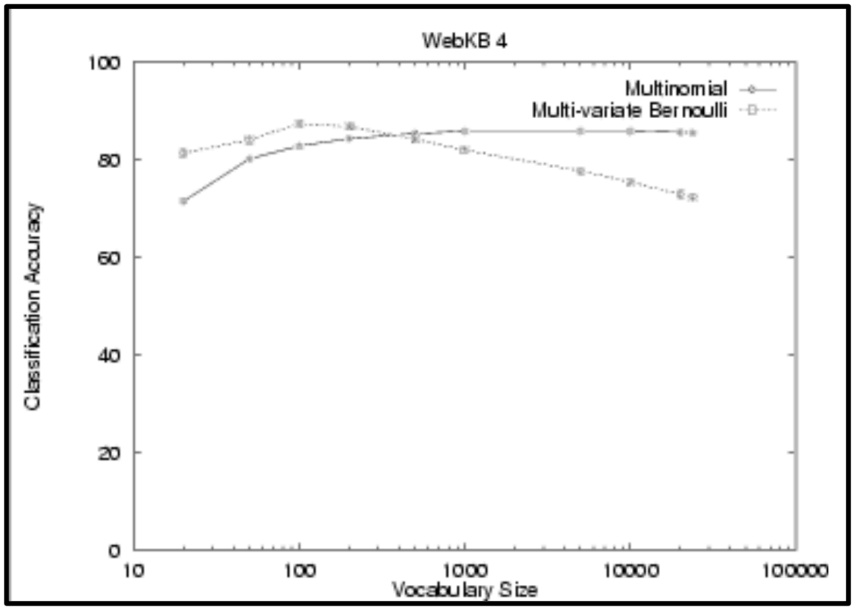

Figure 3: 词频表示法 Vs 布尔值表示法

我们这里简单比较一下两种朴素贝叶斯方法在文本分类上的效果。由上图可见，在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。

5 朴素贝叶斯总结

朴素贝叶斯分类器虽然非常易于理解且容易实现，它在分类任务上的表现并没有想象中那么差。在 1997 KDD CUP 中，朴素贝叶斯分类器的分类结果获得了前两名。朴素贝叶斯的有点在于对分类结果无影响的属性会被模型自动忽略。而我们之后将要介绍的决策树分类器对于无关属性很敏感。而朴素贝叶斯在排除无关属性影响的情况下还能获得与决策树相差不大的分类效果。在很多分类问题中，尤其是文本分类中，研究者通常会先用朴素贝叶斯分类器做一次分类，将分类准确率作为基线（baseline），然后在这个基础上去提升模型效果。朴素贝叶斯分类器潜在的好处还有运算的快速以及很小的内存需求。

引用

[1] Bernoulli Distribution `https://en.wikipedia.org/wiki/Bernoulli_distribution`

[2] Multinomial Distribution: `https://en.wikipedia.org/wiki/Multinomial_distribution`

[3] `https://zhuanlan.zhihu.com/p/53302305`