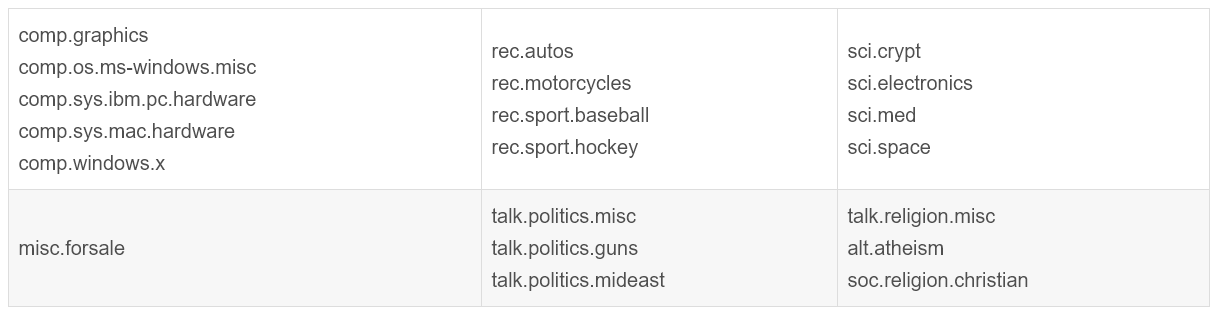
高级机器学习——分类实验报告

吕芳锐 2020101273

1. 数据集

实验的数据集选用20newsgroups数据集。20newsgroups数据集是用于文本分类、文本挖据和信息检索研究的国际标准数据集之一。数据集收集了大约20,000左右的新闻组文档，均匀分为20个不同主题的新闻组集合。一些新闻组的主题特别相似(e.g. comp.sys.ibm.pc.hardware/ comp.sys.mac.hardware)，还有一些却完全不相关 (e.g misc.forsale /soc.religion.christian)。其类别如下表所示：



在sklearn中，该模型有两种装载方式，第一种是fetch\_20newsgroups，返回一个可以被文本特征提取器，是可以自定义参数提取特征的原始文本序列；第二种是fetch\_20newsgroups\_vectorized，返回一个已提取特征的文本序列，即不需要使用特征提取器。

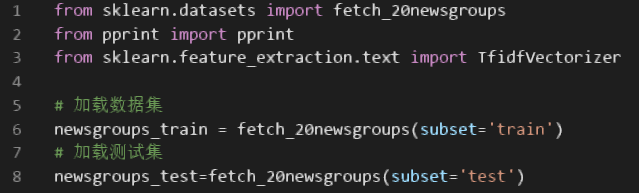
本次实验选取全部的原始文本序列，全部的20个类别，再使用TF-IDF统计词频，抽取特征。

TF-IDF是一种用于信息检索与文本挖掘的常用加权技术。TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF的主要思想是：如果某个单词在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

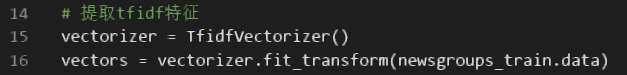
再抽取特征后使用具体的机器学习方法进行分类。

下面简单介绍处理数据集的流程：

首先加载数据集：



再提取TF-IDF特征向量：

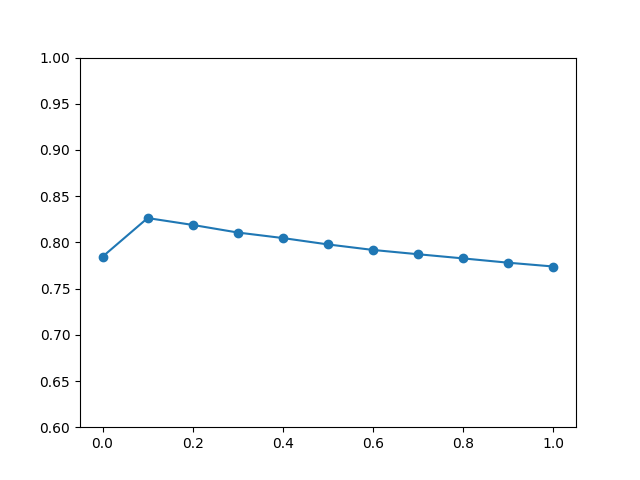


1. 先验为多项式分布的朴素贝叶斯

首先第一个实验使用的方法是贝叶斯方法。贝叶斯方法是以贝叶斯原理为基础，使用概率统计的知识对样本数据集进行分类。由于其有着坚实的数学基础，贝叶斯分类算法的误判率是很低的。贝叶斯方法的特点是结合先验概率和后验概率，即避免了只使用先验概率的主观偏见，也避免了单独使用样本信息的过拟合现象。贝叶斯分类算法在数据集较大的情况下表现出较高的准确率，同时算法本身也比较简单。

而朴素贝叶斯方法是在[贝叶斯](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF/1405899)算法的基础上进行了相应的简化，即假定给定目标值时属性之间相互条件独立。也就是说没有哪个属性变量对于决策结果来说占有着较大的比重，也没有哪个属性变量对于决策结果占有着较小的比重。虽然这个简化方式在一定程度上降低了贝叶斯分类算法的分类效果，但是在实际的应用场景中，极大地简化了贝叶斯方法的复杂性。

本次实验采用的是先验为多项式分布的朴素贝叶斯方法。在sklearn中，先验为多项式分布的朴素贝叶斯方法的函数为MultinomialNB(alpha=1.0，fit\_prior=True，class\_prior=None)。这个函数中有三个主要参数，分别代表：添加拉普拉斯平滑，是否要考虑先验概率，以及修改先验概率。在实验中发现第一个参数alpha对实验结果影响比较明显，下面给出算法准确率随alpha的变化：



可以看到alpha取到0.1时算法可以得到最好的效果，得到效果如下：



下面给出实验的代码：

