Ivy Plan 第一次作业报告

问题描述

给定「20 News Group」数据集，希望使用朴素贝叶斯方法，对数据集进行分类预测。

数据集探索

「20 News Group」数据集按文档内容划分为20组，每组包含文档 1000 份左右，绝大多数文档的词量在 100 ～ 500 之间，文档内容为英文。

技术实现

1.样本划分

将数据集随机划分为80%的训练集和 20%的测试集。

2.数据预处理

对文本进行下列处理：去掉标点符号->转为小写->分词->提取词干->过滤停用词，生成每一篇文章的词表。

3.计算 tf-idf

tf-idf 是重要性调整系数，衡量一个词是不是常见词。如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性，因此会给予这个词更高的权重。
使用训练集：

3.1 计算每一个 News Group 的词频（tf）。
3.2 计算每一个词的逆文档频率（idf）。
3.3 将 tf 中的每一个词频乘上词的 idf，即得到 News Group 的 tf-idf。

4. 通过 NB 的 likelihood 进行分类

给定待分类文档，分别计算其在每个 News Group 下的 likelihood，取最大值的作为模型的分类结果。

5. 验证

依次对测试集里的数据进行分类，统计准确率。

6. 重复验证

重复上述5个步骤多次，取平均准确率作为该模型的分类准确率。

结论

基于上述步骤，通过朴素贝叶斯模型对「20 News Group」数据分类预测。**对每一个 News Group 取 500 个文档作为训练集，100 个文档作为测试集，进行分类的准确率为 85.45%**

思考和改进

1. 实验最初并没有用 tf-idf 作为词的权重，后来从词频改为 tf-idf 后准确率有少许提升，但是对提升值没有做多次的验证和量化。
2. 可以尝试对数据预处理进行不同的尝试，看是否有更优的方案。
3. 实验的过程和中间数据应当尽可能的保留，方便做对照和可视化。
4. 可以尝试一些其他的机器学习模型，进行对比。

参考资料

[1] 20 Newsgroups 文档分类
[2] 自然语言处理 文档归类
[3] Multinomial Naive Bayes Classifier for Text Analysis
[4] Using Naive Bayes algorithm To Classify Text
[5] GitHub:/Loc-Tran/NaiveBayes20NewsGroup
[6] TF-IDF与余弦相似性的应用