



哈尔滨工业大学（深圳）

Harbin Institute of Technology, Shenzhen

机器学习概论 实验（三）任务书

题 目：基于朴素贝叶斯（或其它方法）的文本分类算法

院 （系） 计算机科学与技术

日 期 2025 年 10 月 24 日

实验背景

文本分类是一项重要的自然语言处理任务，旨在将文本按照其内容或主题分为不同的类别。在本实验中，需使用朴素贝叶斯算法或其他相关方法，进行文本分类实验。

本次实验为**新闻标题文本分类**，根据给出的新闻标题文本和标签训练一个分类模型，然后对测试集的新闻标题文本进行分类。

实验目的

1. 理解文本分类的基本概念和任务；
2. 学习如何使用朴素贝叶斯算法或其他方法进行文本分类；
3. 实现一个文本分类模型，能够对给定的新闻标题文本进行分类；
4. 评估分类模型的性能，并进行结果分析和讨论。

数据集介绍

本实验使用的是处理后的 THUCNews 数据集。THUCNews 是根据新浪新闻 RSS 订阅频道 2005~2011 年间的历史数据筛选过滤生成，包含 74 万篇新闻文档，均为 UTF-8 纯文本格式。在原始新浪新闻分类体系的基础上，重新整合划分出 14 个候选分类类别：**财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐**。

数据集内含训练集、验证集和测试集三个文件，文件均为 txt 格式。训练集包含 678998 条数据，验证集包含 75636 条数据，测试集包含 77837 条数据。文本格式为：**新闻标题文本 类别**。数据内容示例如下：

巴萨 1 年前地狱重现这次却是天堂 再赴魔鬼客场必翻盘 体育
美国称支持向朝鲜提供紧急人道主义援助 时政
增资交银康联 交行夺参股险商首单 股票
夏日大学游园会 诺基亚 E66 红黑独家对比 科技

实验步骤

1. 数据准备与数据预处理
 - a. 加载和预处理数据集，确保数据格式正确并进行必要的预处理步骤，如分词等；
 - b. 将文本数据转换为机器学习算法可以处理的特征表示形式；
2. 构建网络模型，进行模型训练
 - a. 需要选择使用朴素贝叶斯算法或其他相关方法进行文本分类；
 - b. 使用训练集数据训练分类模型，并更新模型的参数以获得较好的性能；
 - c. 设置合适的训练轮数和批量大小，控制模型的训练过程。
3. 模型评估
 - a. 使用验证集对训练好的模型进行评估，选择合适的评估指标评估模型；
 - b. 分析评估结果，探究模型的性能和潜在改进方向。
4. 撰写实验报告
 - a. 展示实验结果，包括模型的准确率和其他评估指标，撰写实验报告。
 - b. 探讨实验中遇到的问题、挑战和解决方案。

实验报告要求

实验报告要求包含以下内容：

1. 实验目的和背景介绍；
2. 数据集的描述和处理方式；
3. 实验步骤、训练过程和关键代码的展示；
4. 实验结果的展示和分析、对模型性能的讨论；
5. 对实验过程中遇到的问题和解决方案、实验心得体会等。

其他注意事项

1. 本次实验可进行组队，每组人数为 1~2 人；
2. 为体现工作量，推荐自己实现代码，而非直接使用现成的框架；
3. 实验报告和代码具体提交方式后续通知；

4. 如果组队，需在实验报告内写清组队信息，以及自己在组内负责的工作；
5. 实验报告和代码提交截止日期：2025 年 10 月 31 日 24:00；
6. 实验报告或代码有抄袭行为，按 0 分处理。