文本分类作业报告

王翔宇22200215

阙杭明

1. 实验概述

本实验的目标是通过文本分类任务，掌握文本分类的基本概念、数据预处理、模型训练、评估的关键步骤。文本分类是自然语言处理（NLP）中的重要应用之一，它可以用于垃圾邮件过滤、情感分析、主题分类等多个场景。本实验选择了电影评论（MR）数据集，并实现了两种经典的文本分类算法：Naive Bayes和支持向量机（SVM）。通过对比这两种算法的性能，我们可以深入理解它们的优缺点及适用场景。

1. 实验相关准备

实验开始之前进行了相关软件包的下载：

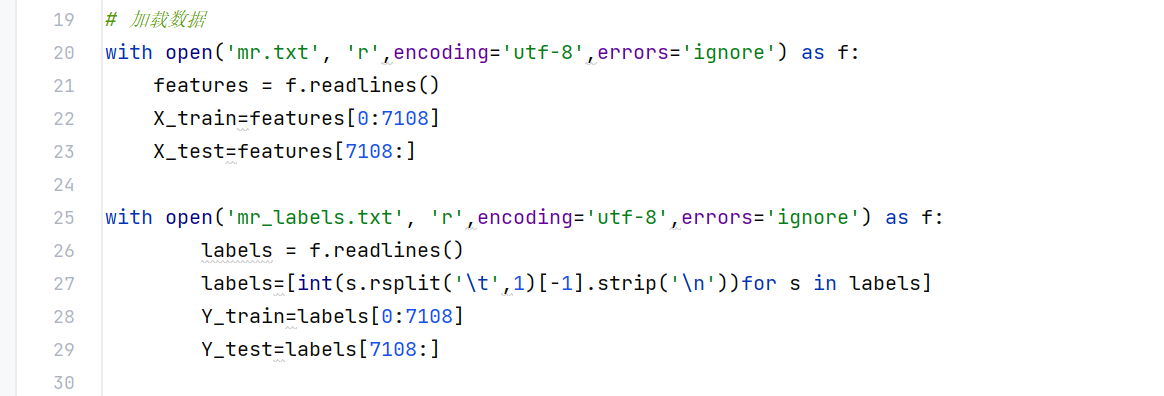


1. 数据收集与预处理
2. 数据集的选择

本实验选择了MR（Movie Reviews）数据集，MR数据集包含10662条电影评论，每条评论都标注了对应的情感标签（正面或负面）。其中，共有7108条训练集，有3554条测试集评论。数据集中，评论文本文件为mr.txt，对应的标签文件为mr\_labels.txt。

1. 数据预处理
2. 加载数据

首先，我们将数据集加载到内存中。MR数据集包含的电影评论被存储在文本文件中，每一行代表一条评论。对应的标签文件中，每一行包含一个标签，表示评论的情感分类（0为负面，1为正面）。

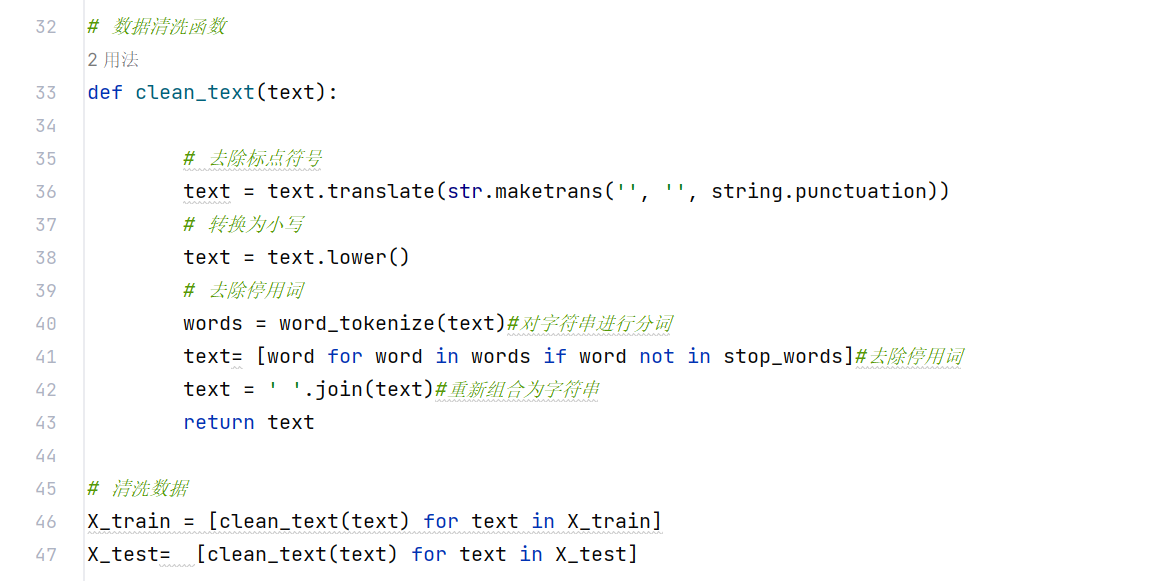


在加载数据集的过程中，手动将训练集和测试集进行分类。由于标签文件包含其他无用数据，因此调用函数对提取出来的标签数据进行了有用数据提取。

1. 数据清洗

数据清洗是预处理过程中的重要步骤，它包括去除无关字符、转换大小写、分词、去除停用词等操作。以下是具体步骤：

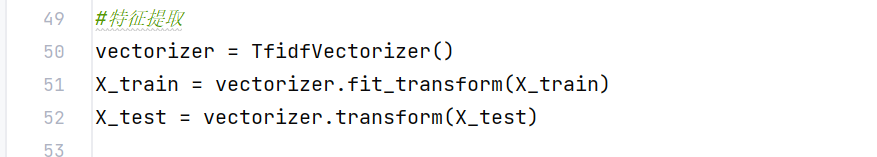
* 去除标点符号：标点符号对文本分类没有实质性帮助，通常需要去除。
* 转换为小写：将所有文本转换为小写，避免同一个单词因大小写不同被当作不同的特征。
* 分词：将文本拆分为单独的单词。
* 去除停用词：去除像“the”、“and”等高频但无实际意义的词。



分别对训练集和测试集进行了数据清洗。

1. 特征提取

在完成数据清洗后，需要将文本转换为机器学习模型可以处理的数值特征。这里我们使用TF-IDF（Term Frequency-Inverse Document Frequency）方法，它能够衡量一个词在文档中的重要性。



通过TF-IDF，我们将每篇文本转换为特征向量，向量的每一维度对应一个词的权重，从而使得文本可以输入到机器学习模型中进行训练。

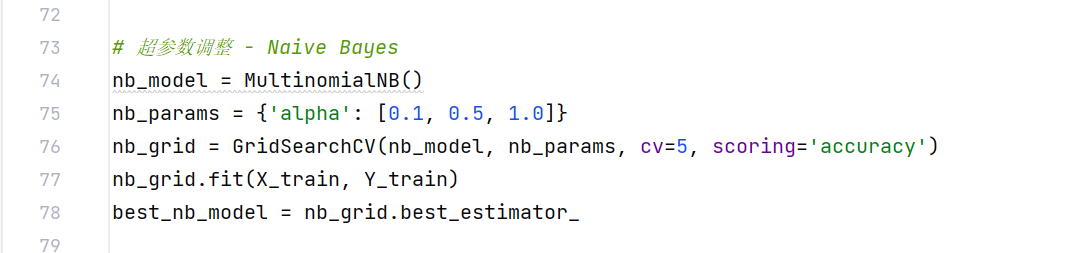
1. 模型训练

选择算法：

我们选择了Naive Bayes和支持向量机（SVM）两种经典的文本分类算法进行实验。Naive Bayes算法假设特征之间相互独立，适用于高维数据。SVM通过最大化分类间隔来实现分类效果，适用于复杂的分类任务。

**Naive Bayes**：

Naive Bayes模型是一种基于贝叶斯定理的概率分类方法。我们使用MultinomialNB，并通过网格搜索（GridSearchCV）来调整模型的超参数alpha。

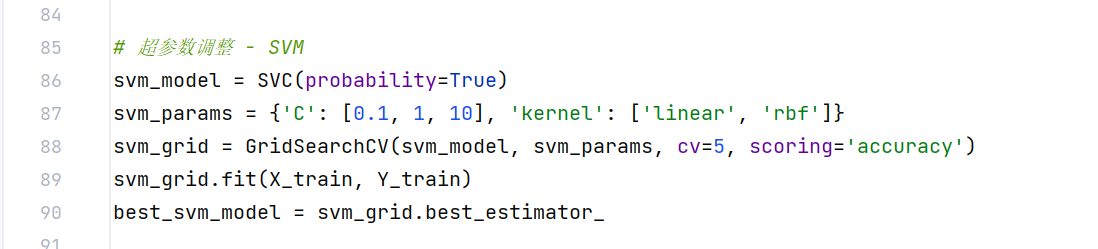


这段代码执行了以下操作：

1. 定义了一个朴素贝叶斯模型，这里使用的是MultinomialNB，适用于多项式分布的朴素贝叶斯模型。
2. 定义了一个参数字典nb\_params，其中包含了要尝试的不同参数值的范围，这里只包含了一个参数alpha的值列表。
3. 使用GridSearchCV函数进行网格搜索，通过传入模型、参数范围、交叉验证的折数（cv=5）以及评分指标（scoring='accuracy'），来寻找最佳的参数组合。
4. 使用训练数据X\_train和对应的标签Y\_train来拟合GridSearchCV对象，它将尝试所有参数组合，并通过交叉验证来评估每个组合的性能。
5. 通过nb\_grid.best\_estimator\_来获取最佳的模型，这是在网格搜索过程中表现最好的模型，其参数已经在交叉验证中进行了优化调整。

**SVM**：

支持向量机（SVM）是一种通过寻找最优超平面来实现分类的模型。我们使用SVM，并通过网格搜索调整C和kernel参数。参数C决定了误分类的惩罚度，kernel定义了核函数类型。



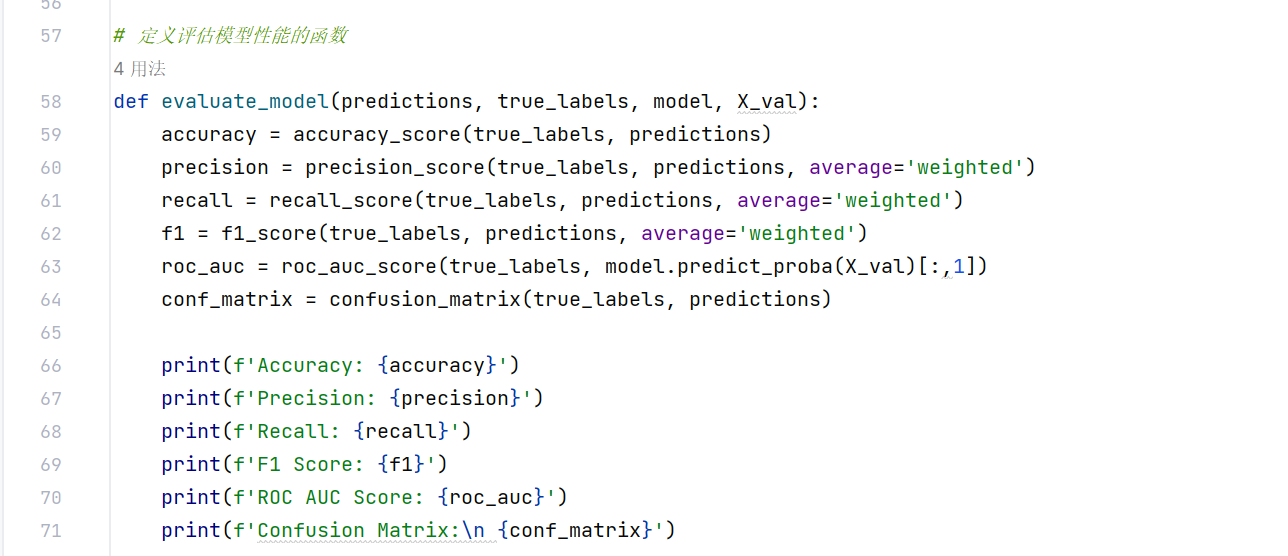
这段代码执行了以下操作：

1. 定义了一个支持向量机（SVM）模型，这里使用的是SVC(probability=True)，其中probability=True表示希望模型能够输出类别的概率。
2. 定义了一个参数字典svm\_params，其中包含了要尝试的不同参数值的范围，包括C（C是一个重要的超参数，用于控制模型的复杂度和拟合程度。C的作用在于平衡模型在训练数据上的拟合程度和对误分类样本的惩罚程度。C是正则化参数，它的倒数控制着正则化强度。较大的C值表示较小的正则化强度，模型更倾向于在训练数据上达到更高的准确度，即更好地拟合训练数据，但这可能会导致模型对噪声或异常值过度敏感，从而导致过拟合。相反，较小的C值表示较大的正则化强度，模型更倾向于寻找一个更简单的决策边界，以更好地泛化到未见过的数据，但可能会导致在训练集上的准确度较低。）和kernel（核函数，核函数是一种用来处理非线性可分问题的技术。它将原始的输入特征空间映射到一个更高维的特征空间，使得数据在新的空间中更容易被线性分离。我们采用了线性核函数和高斯径向基核函数）。
3. 使用GridSearchCV函数进行网格搜索，通过传入模型、参数范围、交叉验证的折数（cv=5）以及评分指标（scoring='accuracy'），来寻找最佳的参数组合。
4. 使用训练数据X\_train和对应的标签Y\_train来拟合GridSearchCV对象，它将尝试所有参数组合，并通过交叉验证来评估每个组合的性能。
5. 通过svm\_grid.best\_estimator\_来获取最佳的模型，这是在网格搜索过程中表现最好的模型，其参数已经在交叉验证中进行了优化调整。

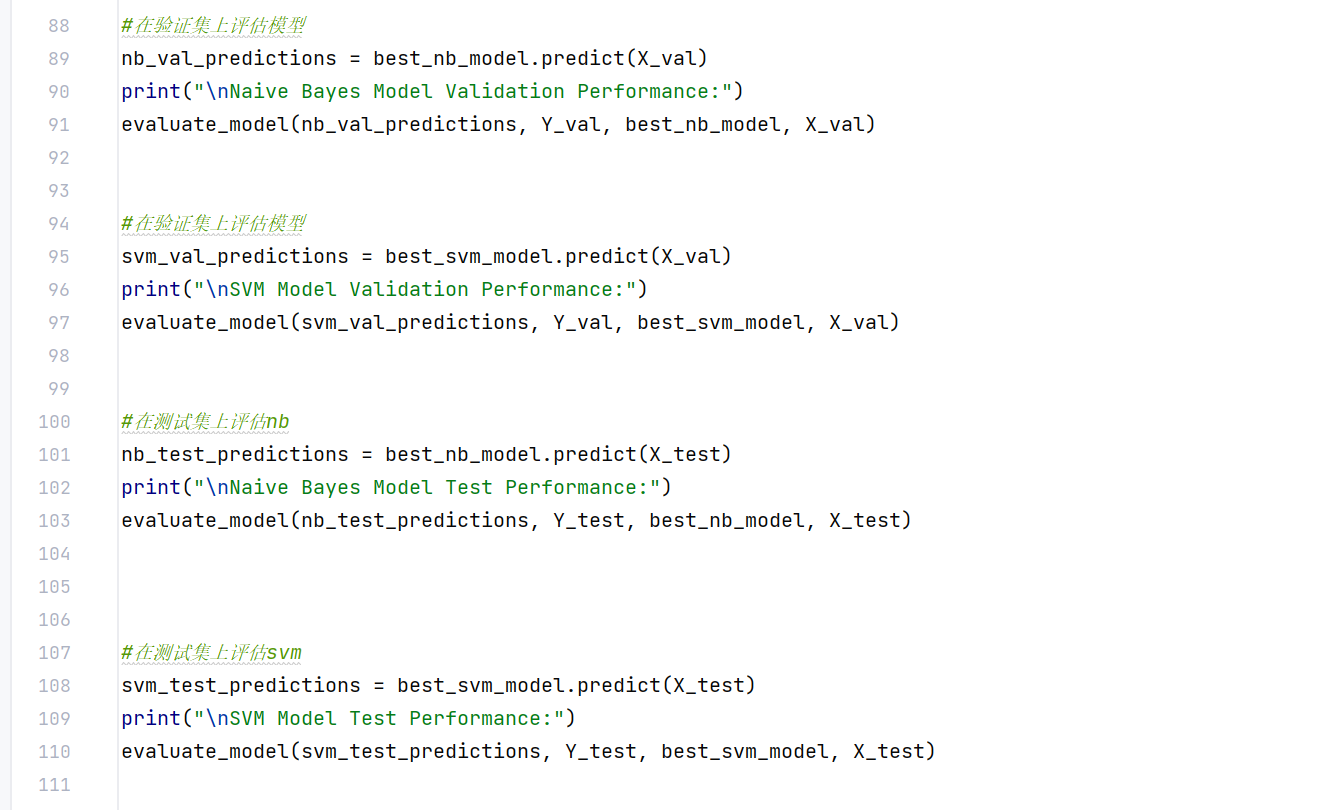
模型评估

模型训练完成后，需要对其进行评估。我们使用以下指标来评估模型的性能：

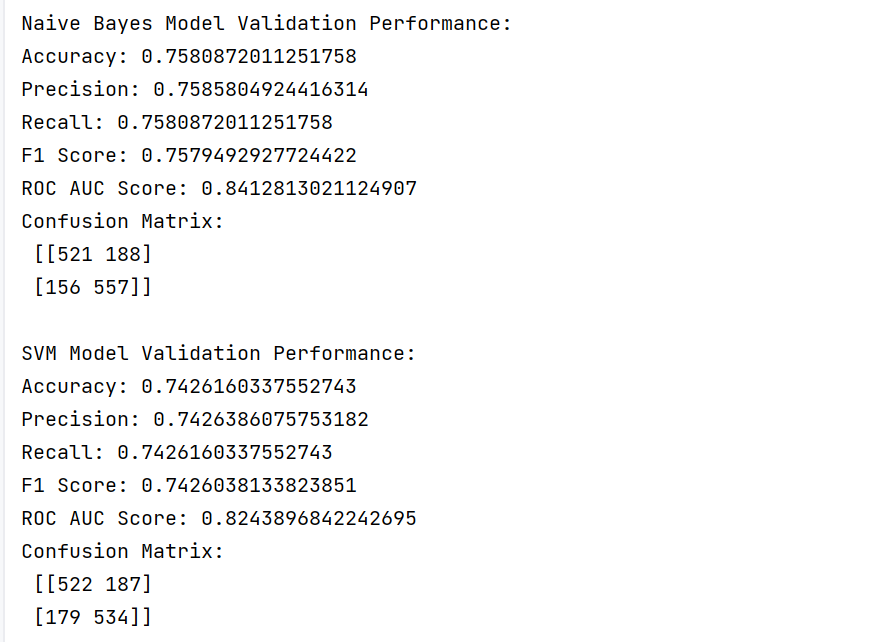
* 准确率（Accuracy）：正确分类的样本数占总样本数的比例。
* 精确率（Precision）：被正确分类为正类的样本数占被分类为正类的样本数的比例。
* 召回率（Recall）：被正确分类为正类的样本数占实际为正类的样本数的比例。
* F1分数（F1 Score）：精确率和召回率的调和平均数。
* ROC曲线下面积（AUC-ROC）：衡量模型分类能力的综合指标。
* 混淆矩阵（Confusion Matrix）：详细展示分类结果，包括真正类（TP）、假正类（FP）、真负类（TN）和假负类（FN）。

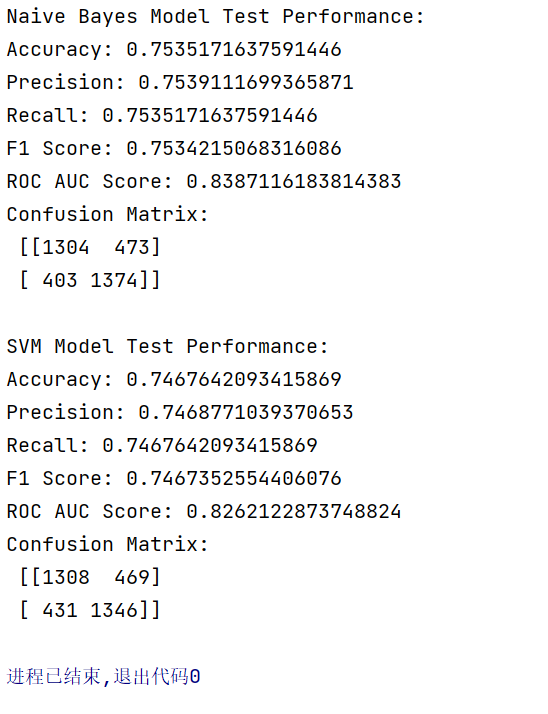


在验证集和测试集上分别评估模型



实验结果如下：





1. 总结与感悟

实验总结

本次实验旨在比较两种常见的机器学习算法——朴素贝叶斯（Naive Bayes）和支持向量机（SVM）在文本分类任务中的性能。通过特征提取、模型训练、超参数调整及性能评估等步骤，得出了以下主要结论：

1. 朴素贝叶斯模型表现优异:
   * 在验证集和测试集上的各项性能指标（准确率、精确率、召回率、F1得分、ROC AUC得分）均略优于SVM模型。
   * 朴素贝叶斯模型的准确率、精确率、召回率和F1得分均在0.75左右，ROC AUC得分在0.84左右。
2. SVM模型表现略逊:
   * 虽然SVM模型也表现良好，但各项性能指标稍逊于朴素贝叶斯模型，特别是ROC AUC得分稍低。
   * SVM模型的准确率、精确率、召回率和F1得分均在0.74左右，ROC AUC得分在0.83左右。
3. 模型的稳定性:
   * 两种模型在验证集和测试集上的表现均较为一致，表明模型具有较好的泛化能力。
   * 朴素贝叶斯模型在验证集和测试集上的表现尤其稳定，说明其在当前任务中的适用性较好。

实验感悟

1. 模型选择的重要性:
   * 通过本次实验，可以清晰地看到不同模型在相同任务上的表现差异。选择合适的模型对于提升任务性能至关重要。
   * 朴素贝叶斯模型由于其简单、有效，在文本分类任务中表现优异，适合快速实现和部署。
2. 超参数调整的意义:
   * 超参数调整是提升模型性能的重要手段。在实验中，通过网格搜索对朴素贝叶斯和SVM模型的超参数进行了调整，显著提升了模型的性能。
   * 特别是在SVM模型中，不同的超参数组合（例如不同的C值和kernel类型）对模型性能的影响较大，合理调整超参数是发挥模型潜力的关键。
3. 特征提取和数据预处理的关键性:
   * 实验中，使用TF-IDF进行特征提取，有效地提升了文本分类的准确性。
   * 数据清洗和预处理（如去除停用词、标点符号等）对模型性能有显著影响，是实验成功的基础。
4. 验证集的重要性:
   * 验证集在超参数调整和模型评估中起到了重要作用。通过划分验证集，可以在模型训练过程中进行实时评估，防止过拟合。
   * 本实验中，将训练集的一部分划分为验证集，确保了超参数调整的有效性和模型的泛化能力。
5. 持续学习与改进:
   * 实验过程中遇到了一些挑战，如特征提取方法的选择、模型的参数调整等。这些挑战促使我们不断学习和改进方法。
   * 未来可以尝试更多高级特征提取方法（如词嵌入）、更复杂的模型（如深度学习模型）以及更丰富的超参数调整策略，以进一步提升模型性能。

结论

通过本次实验，深入了解了朴素贝叶斯和SVM模型在文本分类任务中的表现差异和各自优缺点。实验结果表明，朴素贝叶斯模型在当前任务中表现优异，且具有较好的稳定性和泛化能力。支持向量机模型在进一步调优后，也展现了较强的竞争力。

本次实验不仅提升了对不同机器学习模型的理解，也加深了对特征提取、数据预处理及超参数调整等关键步骤的认识。未来在实际应用中，可以根据具体任务需求选择合适的模型和方法，不断优化和提升模型性能。