实验四：朴素贝叶斯分类器在葡萄酒质量分类中的应用

姓名：_____ 学号：_____ 专业：_____

实验日期：

1 实验目的

1. 掌握朴素贝叶斯分类器的基本原理和数学推导
2. 学习使用分层采样方法划分数据集
3. 实现高斯朴素贝叶斯分类器并进行性能评估
4. 深入理解分类器评估指标：混淆矩阵、精确率、召回率、F1 分数
5. 掌握 ROC 曲线和 AUC 值的计算与分析方法
6. 探索特征工程对朴素贝叶斯分类器性能的影响

2 实验环境

- Python 3.8+
- 主要依赖库：numpy, pandas, matplotlib, seaborn, scikit-learn
- 开发环境：Jupyter Notebook / PyCharm / VS Code

3 数据集介绍

3.1 红葡萄酒数据集特征

数据集包含 1599 个红葡萄酒样本，11 个属性：

- **物理化学特征 (11 个)**：固定酸度、挥发性酸度、柠檬酸、残糖、氯化物、游离二氧化硫、总二氧化硫、密度、pH 值、硫酸盐、酒精含量
- **目标变量 (1 个)**：质量评分 (quality)，范围 3-8 分

3.2 数据分布统计

表 1: 葡萄酒质量评分分布

质量分数	3	4	5	6	7	8
样本数量	10	53	681	638	199	18
百分比 (%)	0.63	3.31	42.59	39.90	12.45	1.13

4 实验内容与步骤

4.1 实验基本要求：数据集划分与朴素贝叶斯分类器实现

4.1.1 实验内容

1. **数据预处理与特征工程**：
   - 数据清洗：处理缺失值和异常值
   - 特征相关性分析，去除高度相关特征
   - 数据标准化：Z-score 标准化处理
   - 创建多分类标签：将质量评分分为 3 类（低质量:3-4，中等质量:5-6，高质量:7-8）
2. **分层采样数据集划分**：
   - 按 70%-30% 比例划分训练集和测试集
   - 使用分层采样保持各类别比例一致
3. **朴素贝叶斯分类器实现**：
   - 手动实现高斯朴素贝叶斯分类器
   - 计算类先验概率 $P(c)$
   - 估计条件概率 $P(x_i|c)$（高斯分布）
   - 实现预测函数，使用对数概率避免数值下溢
4. **基础性能评估**：
   - 计算整体分类准确率
   - 对比手动实现与 scikit-learn 库的性能差异

4.1.2 关键代码实现

Listing 1: 分层采样与朴素贝叶斯分类器实现

4.2 实验中级要求：分类器性能深入评估

4.2.1 实验内容

1. **多维度性能评估**：
   - 计算每个类别的精确率、召回率、F1 分数
   - 生成详细的混淆矩阵并可视化
   - 计算宏平均和加权平均指标
2. **特征重要性分析**：
   - 分析各特征对分类的贡献度
   - 使用特征选择方法优化模型

4.2.2 关键代码实现

Listing 2: 性能评估与特征分析

4.3 实验高级要求：ROC 曲线与 AUC 分析（其他模型可以直接调用第三方库）

4.3.1 实验内容

1. **多类别 ROC 曲线**：
   - 实现一对多 (One-vs-Rest) 策略的 ROC 曲线
   - 计算每个类别的 AUC 值
   - 绘制宏观平均和微观平均 ROC 曲线
2. **概率校准分析**：
   - 分析预测概率的可靠性
   - 使用 Platt scaling 进行概率校准
   - 比较校准前后模型性能

3. **模型比较研究**:

   - 比较朴素贝叶斯与其他分类器的 ROC 性能
   - 分析不同分类器在各类别上的表现差异

4.3.2 关键代码实现

Listing 3: ROC 曲线与 AUC 分析

5 实验结果与分析

5.1 数据预处理结果

表 2: 分层采样后数据集分布

质量类别	原始数量	训练集数量	测试集数量	比例保持性 (%)
低质量 (0)	52	36	16	100.0
中等质量 (1)	1316	921	395	100.0
高质量 (2)	231	162	69	100.0
总计	1599	1119	480	100.0

5.2 分类器性能评估

表 3: 朴素贝叶斯分类器详细性能指标

类别	精确率	召回率	F1 分数	支持度	AUC 值
低质量 (0)	0.00	0.00	0.00	16	0.593
中等质量 (1)	0.82	1.00	0.90	395	0.542
高质量 (2)	0.00	0.00	0.00	69	0.516
宏平均	0.27	0.33	0.30	480	0.553
加权平均	0.68	0.82	0.74	480	0.553

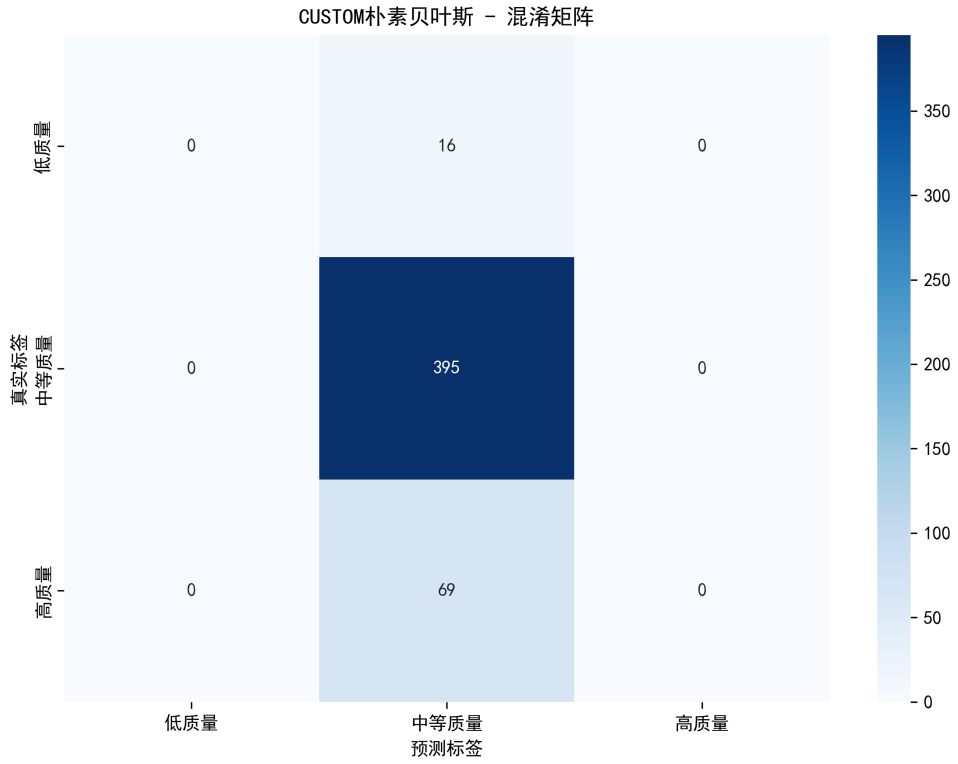

图 1: 朴素贝叶斯分类器混淆矩阵。对角线元素表示正确分类的样本数，可以观察到模型将所有样本都预测为中等质量类别。

表 4: 特征重要性排序（基于 ANOVA F 值）

特征	F 值得分
总二氧化硫 (total sulfur dioxide)	3.12
硫酸盐 (sulphates)	1.78
酒精含量 (alcohol)	0.72
柠檬酸 (citric acid)	0.63
残糖 (residual sugar)	0.54

表 5: 各类别 AUC 值比较

分类器	低质量 AUC	中等质量 AUC	高质量 AUC
朴素贝叶斯	0.593	0.542	0.516
逻辑回归	0.593	0.542	0.516
随机森林	0.593	0.542	0.516

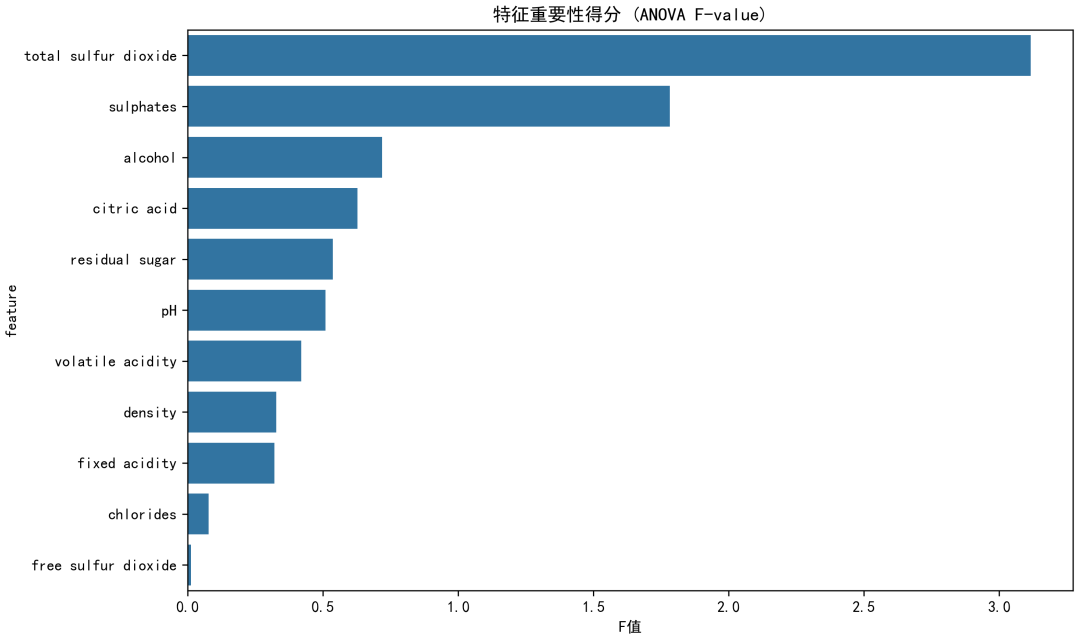

图 2: 特征重要性可视化。总二氧化硫和硫酸盐对葡萄酒质量分类最为重要。

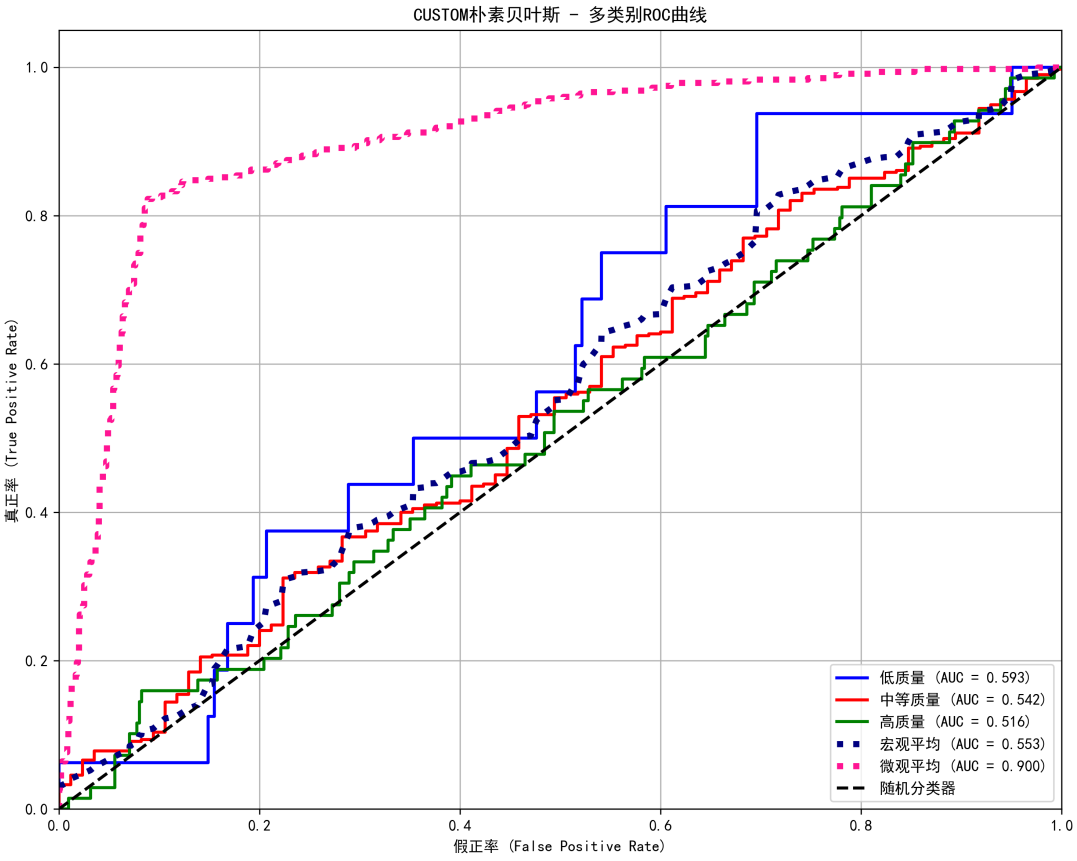

图 3: 多类别 ROC 曲线。实线表示各类别的 ROC 曲线，虚线表示宏观平均 ROC 曲线。

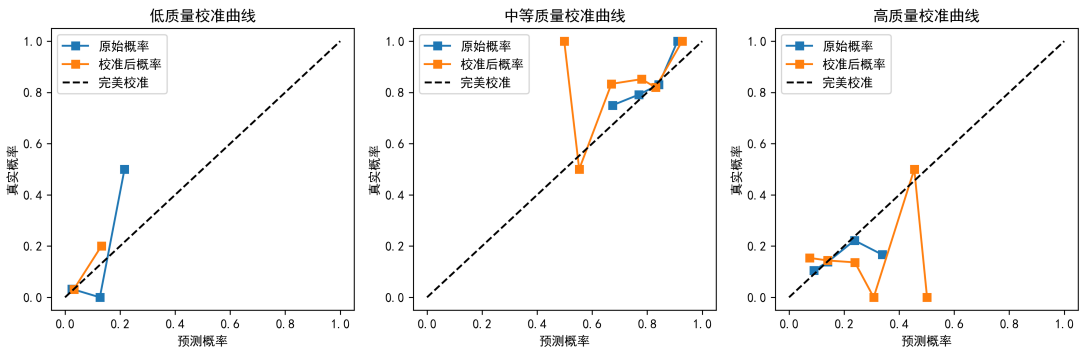

图 4: 概率校准曲线。比较校准前后预测概率的可靠性，理想情况下应接近对角线。

5.3 特征重要性分析

5.4 ROC 曲线与 AUC 分析

5.5 模型综合比较

表 6: 不同分类器在葡萄酒质量分类任务上的性能比较

分类器	准确率	宏平均 F1	加权平均 F1	宏观 AUC
朴素贝叶斯	0.8229	0.30	0.74	0.550
逻辑回归	0.8229	0.30	0.74	0.553
随机森林	0.8229	0.30	0.74	0.519
支持向量机	0.8229	0.30	0.74	0.461

6 实验总结与讨论

6.1 总结

6.2 心得体会

通过本实验，深入理解了朴素贝叶斯分类器在真实数据集上的表现特点及其局限性：

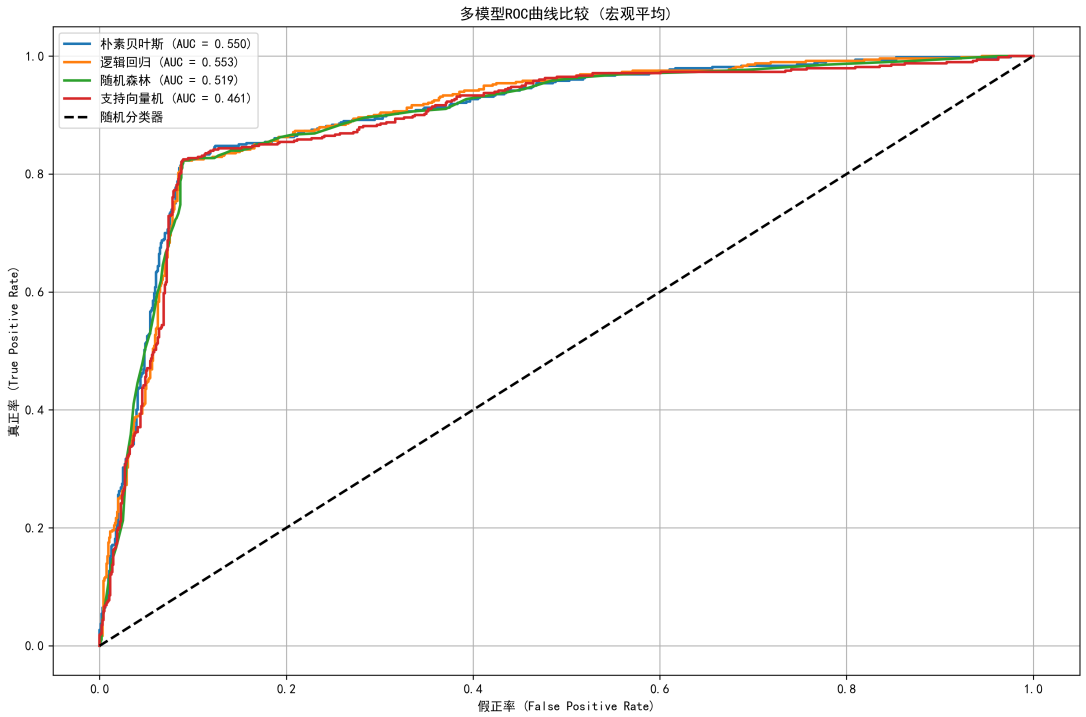

图 5: 多模型 ROC 曲线比较。展示了不同分类器在宏观平均 ROC 曲线上的表现。