数据归约技术实验报告：基于 Wine Quality 数据集的 PCA 与 LDA 分析

GitHub 仓库：https://github.com/djj316/Data-Reduction

附上二维码：

最后更新：2025年4月13日

目录

1. 实验目的

本实验旨在探讨数据归约中两种经典降维技术——主成分分析（Principal Component Analysis, PCA）与线性判别分析（Linear Discriminant Analysis, LDA）在葡萄酒质量分类任务中的应用效果，具体目标如下：

- 比较不同降维方法对分类模型性能的影响；
- 可视化高维数据在低维空间中的分布特征；
- 分析方差保留率与维度压缩之间的权衡关系；
- 评估数据规约对后续学习任务的作用与影响。

2. 数据集

数据来源

实验所使用的数据集来源于 UCI 机器学习仓库：
👉 Wine Quality Dataset (ID:186)

特征描述

特征类别	数量	示例特征
理化指标	11	酸度、pH值、酒精浓度等
目标变量	1	质量评分（范围 3~9）

数据预处理

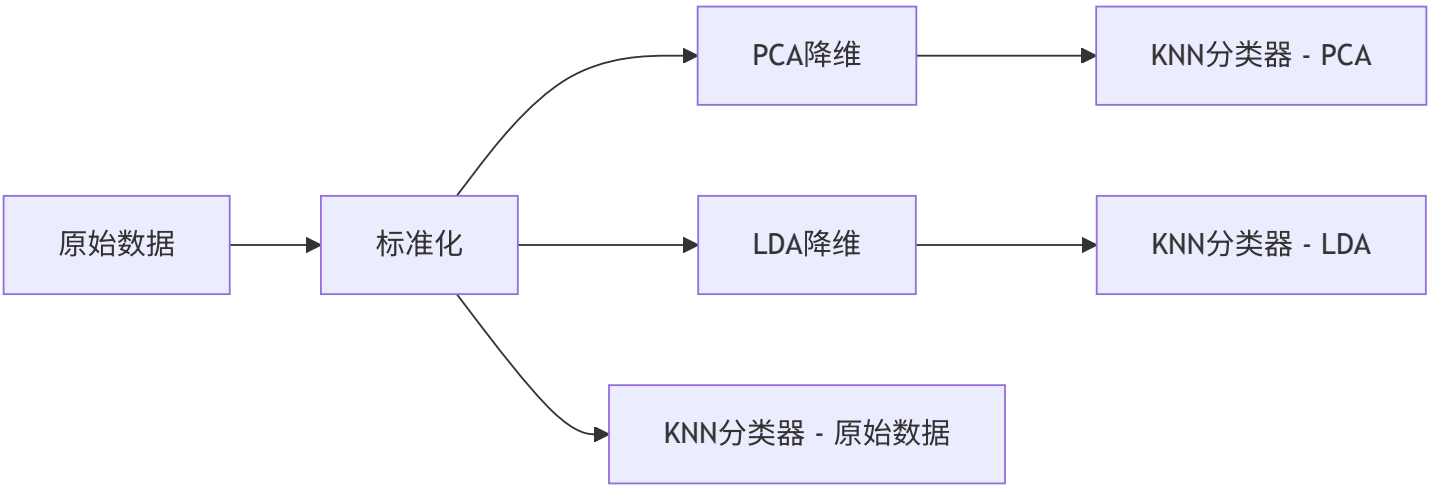
为简化分类任务，将原始的葡萄酒质量评分离散化为三个等级（低、中、高）：

```
# 将评分按阈值进行分箱，得到三类标签
y = np.digitize(y, bins=[3, 6], right=True) - 1
```

此外，数据集按照 7:3 的比例划分为训练集和测试集，并对特征进行了标准化处理以适应 PCA 处理要求。

3. 方法

技术流程概述



本实验采用 KNN（K-近邻）作为统一的分类器，对原始特征、PCA 降维后特征以及 LDA 降维后特征分别进行分类评估。PCA 为无监督降维方法，主要基于数据方差；而 LDA 属于监督式方法，目标是最大化类间距离与最小化类内距离。

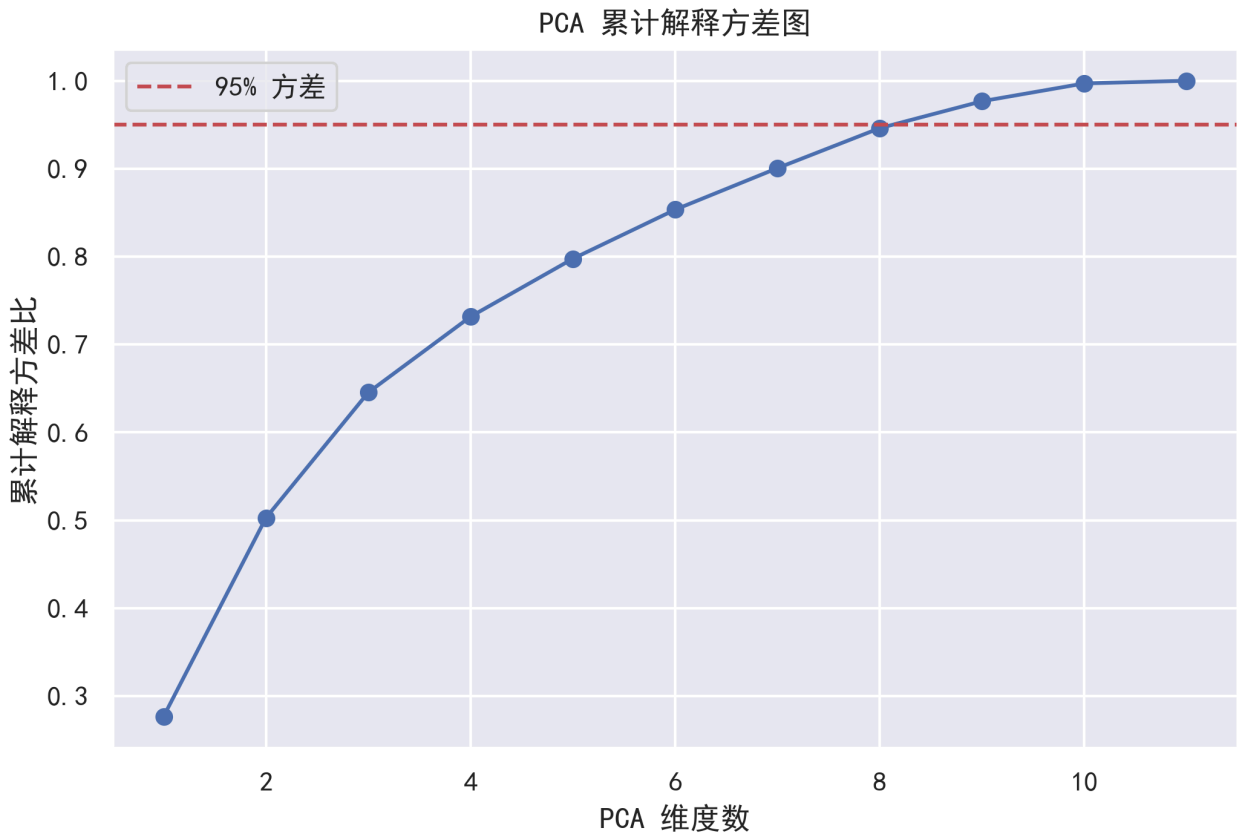
4. 实验结果

4.1 分类准确率比较

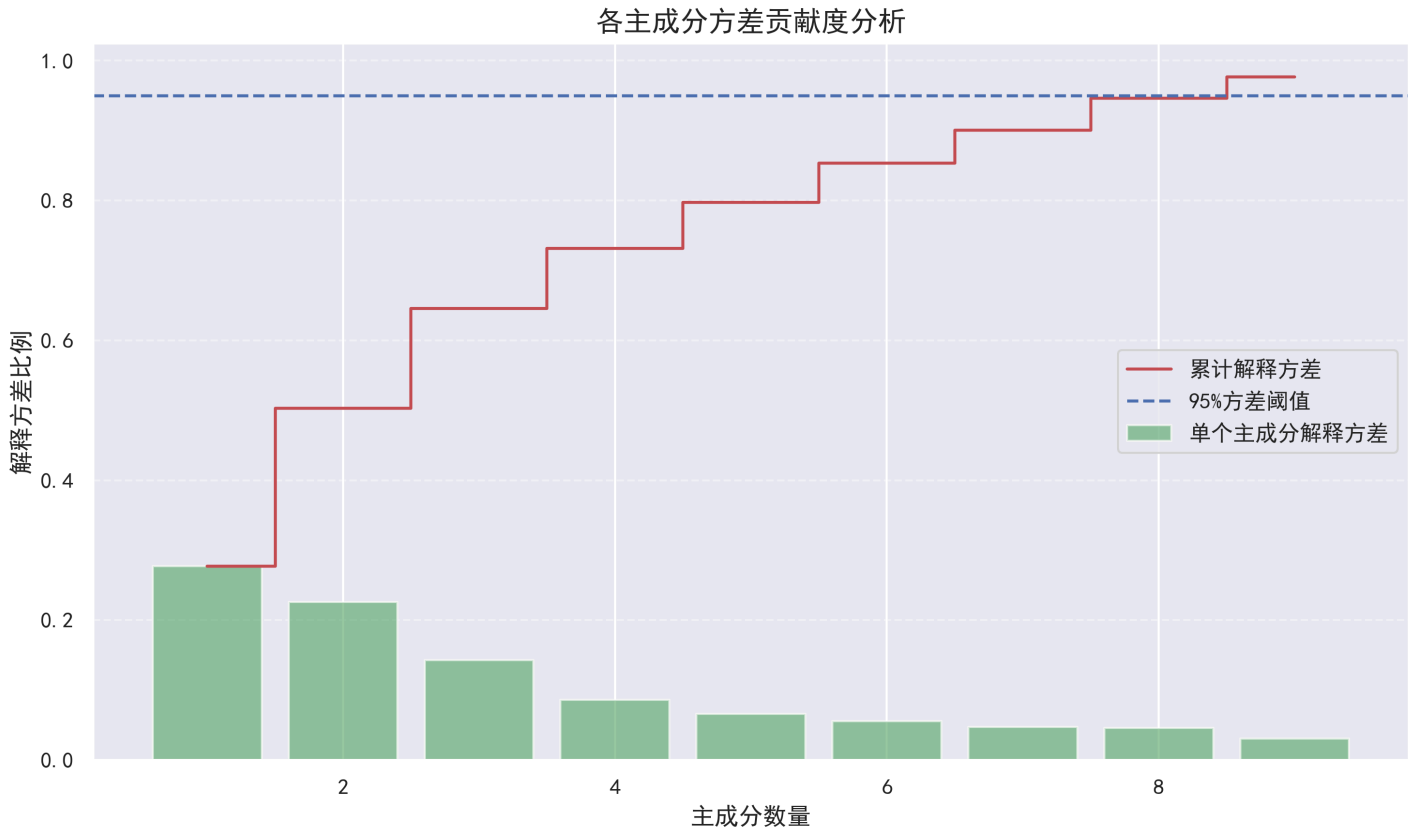
方法	测试准确率	降维后维度
原始特征集	0.80	11
PCA 降维	0.84	9
LDA 降维	0.82	2

4.2 可视化结果

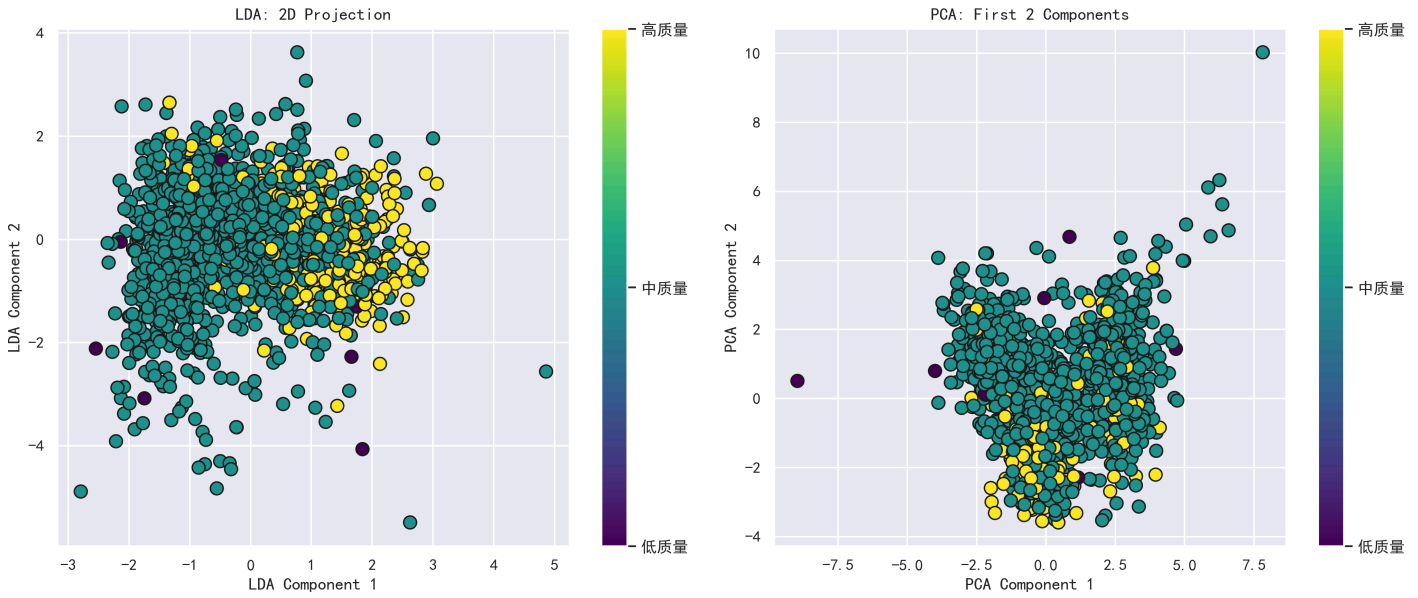
PCA 累计方差解释率



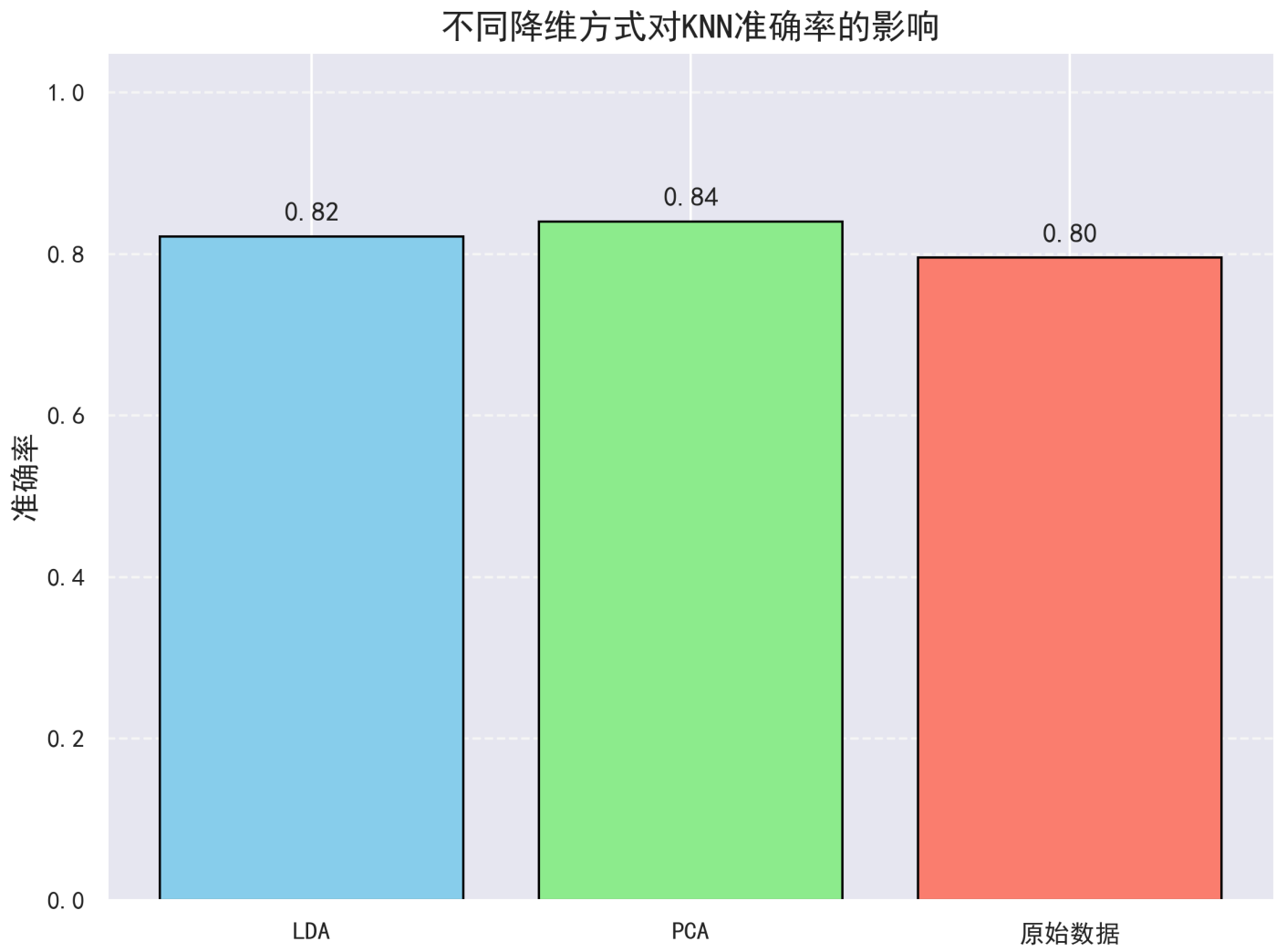
主成分方差贡献度分析



LDA 与 PCA 二维投影对比



分类准确率对比图



5. 结果分析与讨论

PCA 方法分析

- 前两个主成分共解释约 50% 的方差，说明数据在前两个维度上仍存在大量信息损失；
- 在保留 95% 方差的前提下，可将维度从 11 降至 9，降维效果显著；
- 由于 PCA 为无监督方法，其低维投影可能未能有效突出类别间差异，因此在分类任务中表现略逊于 LDA；
- 适合用于探索性数据分析与可视化。

LDA 方法分析

- 尽管被约束至二维空间，LDA 仍能维持较高的分类准确率，展示出良好的类别判别能力；
- LDA 通过监督学习显式最大化类间距离，提升了低维空间的可分性；
- 理论上，LDA 的投影维度不超过类别数减一（C-1），本实验中为 2 维，限制了降维灵活性；
- 更适用于有监督的降维与可视化场景。

6. 结论

结合实验结果，得出以下结论：

1. 在保留 95% 总方差的前提下，PCA 可有效将原始特征维度从 11 降至 9，维度压缩率为 18.2%；
2. LDA 虽仅保留两个维度，但其监督性质使得在分类准确率上优于原始特征，且与 PCA 表现接近；
3. **实用建议**：
    - 进行特征探索或可视化时，推荐优先使用 PCA；
    - 若目标为提升分类性能，且具有可靠标签信息，则建议使用 LDA；
    - 在建模过程中可结合两者进行综合评估与选择。

7. 附录

实验环境

```
Python 版本：3.8+
依赖库：
- numpy >= 1.21
- scikit-learn >= 1.0
- matplotlib >= 3.5
```

完整源代码

详见本文末尾代码块，或访问 GitHub 仓库获取：Data-Reduction

▶ 点击展开完整源代码

报告作者:

zyh