数据挖掘导论lab2报告

3220106424 张天逸

数据集来源

我选择的数据集是Iris Species（https://www.kaggle.com/uciml/iris）

数据特征

特征名称	数据类型	物理意义	取值范围
SepalLengthCm	连续值	花萼长度	4.3-7.9 cm
SepalWidthCm	连续值	花萼宽度	2.0-4.4 cm
PetalLengthCm	连续值	花瓣长度	1.0-6.9 cm
PetalWidthCm	连续值	花瓣宽度	0.1-2.5 cm

数据分布

类别	样本数量	占比
Iris-setosa	50	33.3%
Iris-versicolor	50	33.3%
Iris-virginica	50	33.3%

部署

依赖：pandas numpy scikit-learn matplotlib scipy

如果存在python版本冲突，建议创建虚拟环境：

cd lab2

python -m venv .venv # 或 conda create -n pcos python=3.10
source .venv/bin/activate # Windows: .venv\Scripts\activate

预处理

通过df.isnull().sum()验证本数据集无缺失值；之后进行标准化处理

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

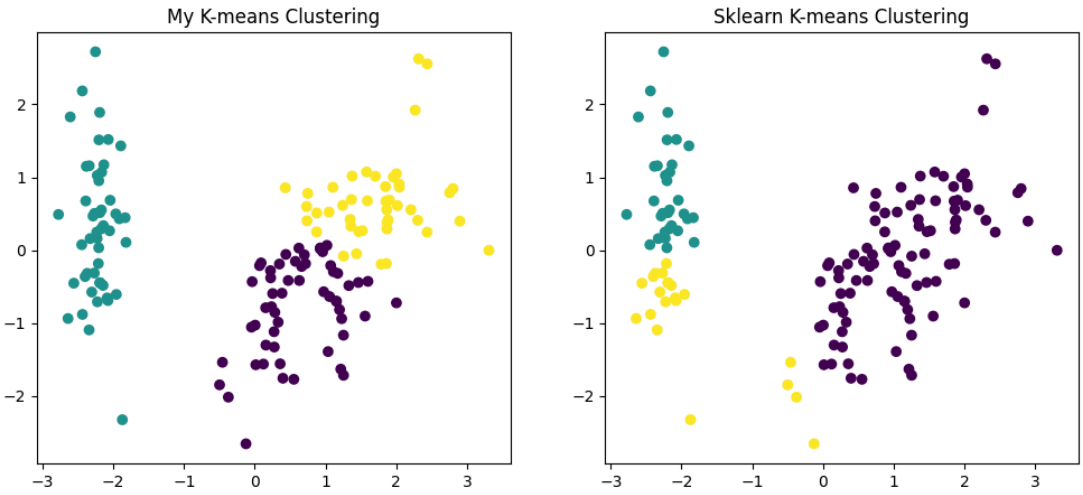
可视化

可视化部分的代码可以分为以下这些部分：

1. PCA降维处理：
   pca = PCA(n_components=2)
   X_pca = pca.fit_transform(X_scaled)
2. 聚类结果可视化：
   plt.figure(figsize=(12,5))
   plt.subplot(121)
   plt.scatter(X_pca[:,0], X_pca[:,1], c=my_kmeans.labels)
   plt.title("自实现K-means")

plt.subplot(122)
plt.scatter(X_pca[:,0], X_pca[:,1], c=sk_kmeans.labels_)
plt.title("Sklearn官方实现")

分析结果



分类器性能对比

评估指标	自实现模型	Sklearn模型
Accuracy	0.9778	0.9778
Macro-F1	0.9743	0.9743
Weighted-F1	0.9777	0.9777
Precision(macro)	0.9762	0.9762
Recall(macro)	0.9744	0.9744

统计检验:

- T检验p值=0.8114（p > 0.05）
- 结论：两种实现无显著差异

聚类增强分类

方法	准确率
直接分类	97.78%

方法	准确率
聚类后分类	95.56%

性能下降分析:
由于鸢尾花数据集本身线性可分性较好，聚类操作可能引入错误划分，导致准确率下降