实验二 机器学习

问题描述

鸢尾花（Iris）数据集是机器学习领域中一个非常著名且常用的数据集：它包含了分别来自三种不同的鸢尾属植物的 150 个样本，即 Iris Setosa、Iris Versicolour 和 Iris Virginica。其中每种植物各收集了 50 个样本。每个样本测量了 4 个特征，分别是：

1. 花萼长度（sepal length）
2. 花萼宽度（sepal width）
3. 花瓣长度（petal length）
4. 花瓣宽度（petal width）

所有特征的单位都是厘米（cm）。鸢尾花数据集最初由英国统计学家和生物学家 Ronald Fisher 在 1936 年的论文《The use of multiple measurements in taxonomic problems》中引入，用以介绍线性判别分析。鸢尾花数据集因其简单性和特征的清晰分离而受到青睐，这使得它成为测试新机器学习模型的基准。

图 1 鸢尾花

实验要求

分别使用聚类算法和分类算法对鸢尾花数据集（https://archive.ics.uci.edu/dataset/53/iris，Iris - UCI Machine Learning Repository）进行分析。分类时需要按 7：3 随机划分训练集和测试集。设置分类和聚类的评价指标，使用两种不同的聚类/分类算法（建议使用 k-means、层次聚类，逻辑回归、决策树分类算法）实现。比较不同的方法，分析参数选择与性能之间的关系。

可以自行实现模型，也可以使用 Python 或 R 语言中相应的支持。需要以可视化方式呈现结果。

实验报告

不要简单地拷贝代码。通过观察与定量分析，得出一般性的结论。
请交叉检查，并评价其他同学的实现方法，并在对方实验报告上署名签字。