深 圳 大 学 实 验 报 告

课程名称：Python 程序设计

项目名称：鸢尾花数据分类与可视化

学　　院：人工智能学院

专　　业：计算机科学与技术 (IEEE 荣誉班)

指导教师：舒婷，樊超

报告人：姜顺元　　学号：2024401029

实验时间：2025 年 12 月 14 日

提交时间：2025 年 12 月 14 日

教务处制

1 实验目的

本实验通过 Python 编程实现对经典 Iris（鸢尾花）数据集的探索性数据分析（EDA）和多分类器可视化，主要目标包括：

- 掌握使用 pandas、seaborn、matplotlib 和 plotly 等库进行数据预处理、统计描述与可视化；
- 理解鸢尾花数据集的特征分布与类别可分性；
- 实现多种分类器（逻辑回归、线性 SVM、决策树）的训练与决策边界可视化；
- 掌握交互式 2D 和 3D 可视化技术，提升数据分析结果的表达效果。

2 实验概述

实验以 Iris 数据集为对象，主要流程如下：

1. 数据加载与预处理：使用 sklearn 加载数据集，划分训练集与测试集；
2. 探索性数据分析：统计特征分布、绘制箱线图与散点图矩阵，观察各类别间的差异；
3. 分类器训练：在最具区分度的特征子集上训练多种分类器；
4. 可视化实现：使用 matplotlib/seaborn 进行静态分析图，使用 plotly 生成交互式决策边界、概率热图及 3D 边界/概率体积渲染，使用 HTML 展示结果。**网页可以交互，具有更好的展示效果**。

3 实验实现

实验分为四个独立 Python 脚本（task1.py、task2.py、task3.py、task4.py），分别实现不同维度的分类器可视化。所有脚本均使用 scikit-learn 训练模型、numpy 构建网格、plotly 生成交互式 HTML 输出，便于动态观察决策边界与概率分布。

3.1 task1.py：可视化不同分类器的结果

该脚本实现三种分类器在 petal length 与 petal width 两个特征上的决策边界与概率对比：

- 训练 Logistic Regression、Linear SVM(启用 probability)和 Decision Tree(depth=5)；
- 构建高分辨率 2D 网格（500×500），预测硬决策区域与后验概率；
- 使用 make_subplots 创建 3 行 5 列布局：每行对应一个分类器，左侧与右侧显示决策区域（彩色填充 + 黑边真实标签点），中间三列为各物种概率热图（白 → 该类颜色渐变）；
- Decision Tree 无概率时显示说明文字；
- 输出 task1.html。

3.2 task2.py：可视化 3D Boundary

该脚本针对二分类（versicolor vs virginica）与三特征（sepal length, petal length, petal width）实现 3D 决策超平面：

- 使用线性 SVM 训练，计算决策超平面方程；
- 构建 3D 网格，绘制决策平面（Surface）与彩色数据点；
- 输出 task2.html。

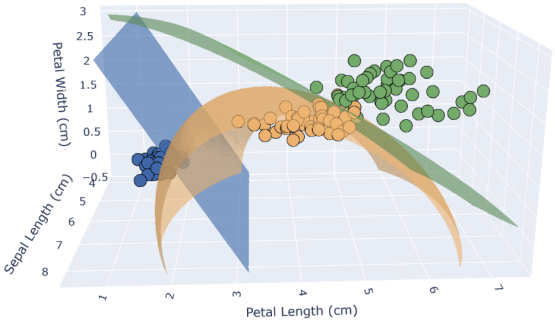

Figure 1: task2.py 生成 html 页面的效果

3.3 task3.py：可视化 3D Probability Map

该脚本在相同二分类与三特征设置下，实现概率分布的 3D 体积渲染：

- 使用 RBF 核 SVM（启用 probability）训练；
- 构建 3D 网格，预测概率；
- 使用 Volume 渲染概率体积（透明度表示概率强度），Isosurface 绘制 P=0.5 决策面；
- 输出 task3.html。

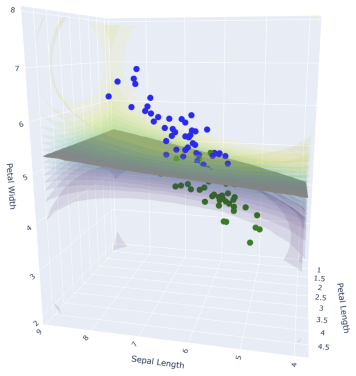

Figure 2: task3.py 生成 html 页面的效果

3.4 task4.py：高级 3D 多分类可视化

该脚本扩展到完整三分类，使用 Gaussian Process Classifier 实现非线性边界：

- 构建 35×35×35 网格，预测平滑概率；
- 三子图布局： 每个类 P=0.5 等值面（Isosurface）； 概率体积渲染（Volume，以 virginica 为例）； 交互 3D 散点图（图例开关类别）。
- 输出 task4.html。

通过以上四个任务，逐步从 2D 多分类对比深化到 3D 非线性概率可视化，全面掌握了 Python 在数据分析与机器学习可视化中的应用。

4 实验结果

- 探索性分析结果显示：setosa 类在所有特征上与其他两类有明显分离；versicolor 与 virginica 在 petal 特征上高度可分，但在 sepal 特征上重叠较多。
- 2D 可视化（task1.html）清晰对比了三种分类器的决策风格：逻辑回归与线性 SVM 边界平滑且提供可靠概率；决策树边界呈阶梯状，无概率输出。
- 3D 可视化（task2.html、task3.html、task4.html）揭示了二维投影隐藏的复杂非线性结构，概率等值面与体积渲染直观展示了模型不确定性区域。

交互式 HTML 结果显著优于静态图片，便于动态观察数据分布与模型行为。

5 讨论与分析

- petal length 与 petal width 是 Iris 数据集最具区分度的特征组合，适合用于 2D 决策边界演示；
- 不同分类器在该数据集上均取得高准确率，但决策边界形态差异显著：软分类器（逻辑回归、SVM）更适合需要概率输出的场景；决策树边界更易解释但易过拟合；
- 交互式可视化（plotly）相比传统 matplotlib/seaborn 具有明显优势，可 hover 查看数值、旋转 3D 视图，大幅提升分析效率与报告表现力；
- 3D 可视化进一步证明了高维特征空间中非线性模型的优势，同时也暴露了可视化计算开销较大的问题（网格分辨率需权衡）；
- 本实验证明，python 在数据分析和可视化方面有非常大的优势，海量支持库和工具也可以为 python 语言提供和其他工具（如 html）交互的能力，这说明 python 作为“胶水语言”，易用性和全面性都非常出色。此外，HTML 展示效果良好但占用较大，这是该方案的不足之处。

指导教师批阅意见：

成绩评定：

指导教师签字：

备注：

- 报告内的项目或内容设置，可根据实际情况加以调整和补充。
- 教师批改学生实验报告时间应在学生提交实验报告时间后 10 日内。