

人工智能引论

《决策树分类实验报告》

2025年05月01日

1. 实验概括

本次实验基于 sklearn 提供的鸢尾花（Iris）数据集，采用决策树分类器对数据进行建模和评估。我们通过手动划分训练集和测试集，对决策树模型进行训练，使用可视化工具生成树结构图，并评估模型在测试集上的准确率、精确率和召回率，验证其在该数据集上的分类效果。

1. 小组成员

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 学号 | 班级 | 专业 |
| 冯汉禹 | 2024210097 | 2024211104 | 通信工程 |
| 雷皓 | 2024210104 | 2024211104 | 通信工程 |
|  |  |  |  |
|  |  |  |  |

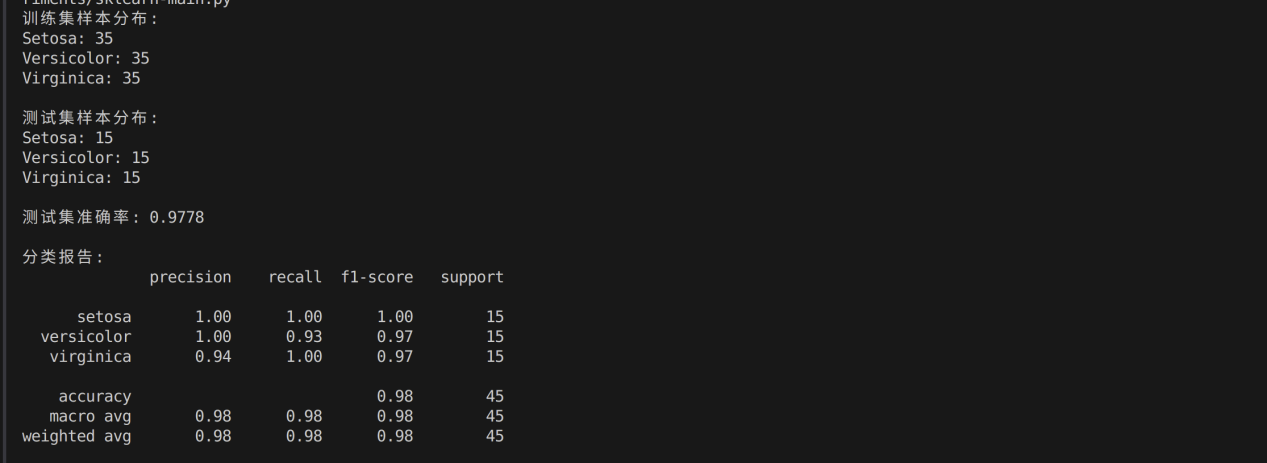
1. 实验环境

软件：venv python3.12.3

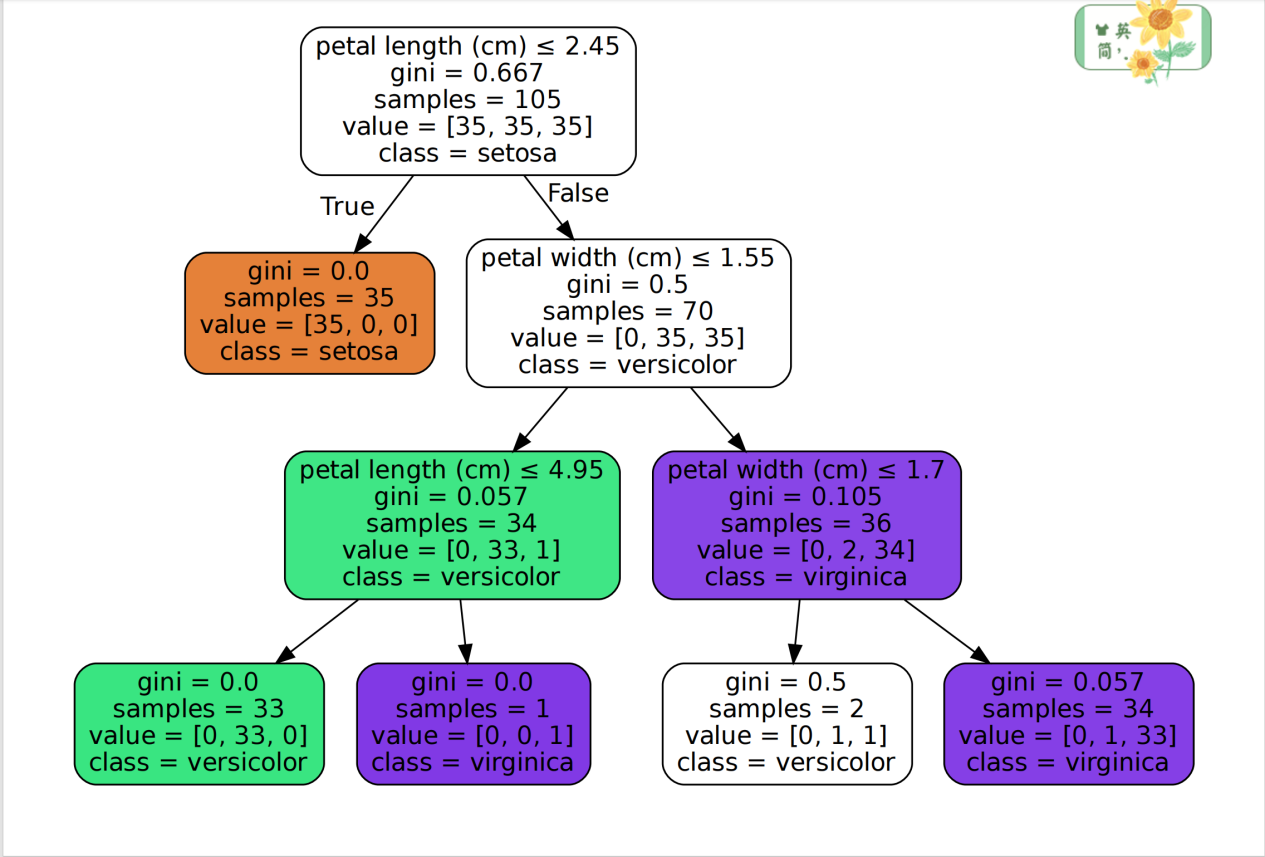
硬件：ubuntu-24.04

1. 主要实验设计
2. 本实验使用的主要工具为 scikit-learn 中的 DecisionTreeClassifier 类，其中包含的库包括numpy matplotlib graphviz
3. 决策树是一种常用的监督学习算法，广泛应用于分类和回归问题。它通过模拟人类“如果-那么”的推理过程，将决策过程表示为一棵树形结构。在决策树中，每一个内部节点表示对某个特征的判断，每条分支代表判断的结果，每个叶子节点对应一个类别或预测值。构建决策树的基本过程包括选择最优特征对数据进行划分，通常依据信息增益、信息增益比或基尼指数等准则进行判断；然后将样本递归划分形成子树，直至满足停止条件，如样本纯净或特征用尽。决策树的主要优点在于结构清晰、易于理解和可视化，能够处理数值型和分类型特征，对数据的预处理要求较低。然而，其缺点也较明显，如容易过拟合、对噪声敏感、泛化能力有限等。在实际应用中，决策树被广泛用于客户分类、医疗诊断、信用评估等领域，尤其适合需要可解释性的任务场景。
4. 具体实验过程
5. 使用virtualenv创建虚拟环境，使用pip来下载所需依赖
6. 编写决策树程序并得出相应的数据
7. 决策树相关数据

测试集和验证集划分方式及每类样本的数量



决策树视图及测试集上面的准确度



相关指标

1. 准确率 97.78%
2. 加权平均精准率 98%
3. 加权平均召回率 98%

决策树在数据集上的表现

决策树在鸢尾花数据集上表现出色，测试准确率达到97.78%，仅有一个样本被误分类。模型主要依赖花瓣长度和宽度这两个关键特征，这与植物学上区分鸢尾花品种的实际依据高度吻合。具体而言，决策树首先通过花瓣长度是否小于等于2.45厘米这一判断条件完美区分出了Setosa品种，随后再根据花瓣宽度等特征进一步细分Versicolor和Virginica品种。这种分类逻辑不仅准确率高，而且具有很强的可解释性，每个判断条件都清晰明了。不过，模型在Versicolor和Virginica两类交界处的样本（花瓣长度在4.6-5.0厘米范围内）仍存在少量误判，这表明决策树对边界样本的区分能力还有提升空间。总体而言，决策树算法在这个特征明确、类别可分的数据集上展现出了优异的性能，既保证了分类准确率，又保持了模型的简洁性和可解释性，非常适合作为鸢尾花分类的基准模型。若需进一步提升性能，可以考虑引入更复杂的特征工程或尝试集成学习方法，但这可能会以牺牲部分模型可解释性为代价。

1. 实验总结

通过本次鸢尾花分类实验，我对决策树算法有了更深入的认识。该算法在特征明确的分类任务中展现出显著优势，仅通过花瓣长度和宽度两个关键特征就实现了97.78%的高准确率，特别是对Setosa类别的完美区分验证了决策树在规则明确场景下的强大性能。实验过程中，决策树可视化提供的直观分类逻辑让我深刻体会到机器学习模型可解释性的重要价值。同时，Versicolor和Virginica两类在特征空间的重叠区域出现的个别误判案例，也让我意识到算法在边界样本处理上的局限性。基于实验发现，我认为可以从三个维度进行改进：首先优化模型参数和采用集成学习来提升边界样本识别能力；其次通过构造长宽比等衍生特征增强特征表达能力；最后引入交叉验证等更严谨的评估方法。这些改进方向既着眼于性能提升，又兼顾了实际应用中对模型可解释性的需求。这次实验不仅验证了决策树算法的适用性，更启发我在未来工作中要更加注重算法特性与数据特征的匹配，以及模型性能与可解释性的平衡。

附录

试验源码

