实验课题2-基于决策树的个人年收入分类任务

介绍

- 在本课题任务中，基于Adult数据集，应用决策树分类模型，根据各种人口统计学和就业相关特征预测个人年收入是否超过5万美元。具体任务包括数据预处理，使用交叉验证调整超参数，分析数据，可视化决策树并解释实验结果。

实验目的

- 掌握使用决策树进行分类任务的实践经验。
- 学习处理缺失值和特征变量的数据预处理技术。
- 学会使用交叉验证进行超参数的调优。
- 分析决策树分类器的性能并解释结果。

数据

- Adult数据集包含个人相关信息、包括年龄、教育程度、职业和国籍等特征。目标变量是收入情况，代表个人年收入是否超过5万美元。
- 数据集来源：Adult dataset.
- 数据格式与特征说明，请查看Adult dataset website.

分类任务

1. **数据预处理：**

- 处理缺失值：删除或归因缺失值(Missing values)。通过程序读出Adult数据为Pandas的DataFrame格式，可查看其中一些数据的特征是缺失的（NaN）,因此在使用前，需要先删除这些具有缺失值的数据样本。
- 编码分类变量：使用One-hot encoding将分类变量(Category variables)转换为数值变量(Numerical variables)，或者使用LabelEncoder(标签编码)转换分类变量为数值变量。
- 在构建模型之前，我们需要完成一些预处理步骤。首先，请注意，数据中的特征有分类变量（Category variables）和数值变量（Numerical variables）。在线性和逻辑回归等模型中，可以为分类变量创建One-hot encoding，因为这些模型（数学方程）只能处理数值变量。但在决策树中不需要，因为它们可以轻松处理分类变量。然而，我们仍然需要将分类变量编码为标准格式，以便sklearn能够理解它们并构建决策树。我们可使用sklearn.preprocessing附带的LabelEncoder()类来做到标签编码，可以在此处阅读LabelEncoder的文档，其对目标标签进行编码，其值在0到n_classes-1之间，其中n_classes代表类别个数。

2. **决策树分类器：**

- 将数据集拆分为训练和测试集。
- 构建决策树分类器。
- 可视化决策树，以了解模型的决策过程。

3. **超参数调优：**

- 使用交叉验证（CV）来调整决策树分类器的超参数。
- 尝试max_depth、min_samples_split和min_samples_leaf等超参数的不同值。
- 尝试使用不同的编码方法（One-hot encoding或者标签编码），观察不同编码方法的分类性能，并分析原因。

4. **数据分析：**

- 分析数据集中特征的分布。
- 探索特征和目标变量之间的关系。
- 讨论数据中观察到的规律或潜在模式。

5. **结果分析：**

- 使用准确性、精度、召回和F1得分来评估决策树分类器的性能。
- 比较模型在超参数调优之前和之后的性能。
- 解释结果并讨论模型的优势和局限性。

6. **代码编写：**

- 建议采用Python语言。
- 附件提供起始代码文件dt_tree_starter.py, 可在此文件基础上开始编写代码。
- 需要在`# write your code`位置补充代码，完成实验。
- 其他部分，可自行修改，也可不修改。

提交内容：

- 代码：包含代码实现的Jupyter笔记本或Python源代码。
- 实验报告：报告记录数据预处理步骤、超参数调优过程、数据分析结果和结果分析、决策树的可视化、讨论从分析中得出的见解和结论。
- 实验报告模版：模版单独提供，可在智慧树网站的实验任务附件里找到。

提交地址：

- 智慧树网站.
- 迟交地址(会有扣分)：我的邮箱（zpshao@hunnu.edu.cn）

资源:

- Scikit-learn Documentation
- Pandas Documentation
- Pandas 中文资料
- Matplotlib Documentation

附加说明:

- 可以与同学讨论并合作，但确保提交的工作是你自己的。
- 如果您在作业期间遇到任何困难，请向老师或同学寻求帮助。