**决策树算法实验指导书**

**一、实验目的**

1. 了解决策树模型的相关知识；
2. 熟悉如何通过Python实现ID3和CART决策树算法；
3. 练习分类模型的评估方法。

**二、实验环境**

系统：windows10/linux

软件：Python

**三、背景知识**

1. 决策树的基本概念

机器学习中，决策树是一个树型结构预测模型，它代表的是对象属性值与对象值之间的一种映射关系。树中每个节点表示某个对象，每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应具有上述属性值的子对象。决策树方法的基本思想是：利用训练集数据自动地构造决策树，然后根据这个决策树对任意实例进行判定。决策树归纳的学习和分类步骤是简单和快速的，推理过程可以表示成If Then 形式，并且具有很好的准确率；获取的知识用树的形式表示是直观的，并且容易被人理解。因而，决策树归纳分类是目前应用最广泛的归纳推理算法之一，在数据挖掘中受到研究者的广泛关注。

2. ID3算法

ID3算法最早是由罗斯昆（J. Ross Quinlan）于1975年在悉尼大学提出的一种分类预测算法，算法的核心是“信息熵”。ID3算法通过计算每个属性的信息增益，认为信息增益高的是好属性，每次划分选取信息增益最高的属性为划分标准，重复这个过程，直至生成一个能完美分类训练样例的决策树。该算法的基本策略如下：

（1）选择一个属性放置在根节点，为每个可能的属性值产生一个分支；

（2）根据属性值将样本划分成多个子集，一个子集对应于一个分支；

（3）在每个分支上递归地重复这个过程，仅使用真正到达这个分支的样本；

（4）如果在一个节点上的所有样本拥有相同的类别，即停止该部分树的扩展。

**四、实验内容**

1. 基于西瓜数据集2.0的ID3决策树算法实现

数据集xiguadata.csv,来自教材中的西瓜数据集2.0，共有17个样本数据。实验中，选取其中的15个样本构成训练集，其余2个样本构成测试集。基于ID3算法采用训练样本构造决策树，并简单绘出生成的决策树。最后，测试该决策树对测试样本的类别划分情况。

2.基于Wifi\_Localization/iris/breast数据集分析CART决策树算法实现（要求两种实现方式）

文本

描述已自动生成

1. 基于西瓜数据集2.0(xiguadata.csv)的CART决策树算法实现，要求使用两种实现方式（sklearn实现时请注意数据格式的转换）
2. 分析所实现代码和Sklearn第三方库结果的差异
3. 使用PCA对breast数据进行降维表示，并可视化
4. 通过可视化决策树分析实验结果差异（选做）

**五、实验步骤**

1. 见参考代码1\_id3.py

2. 见参考代码 2\_cart\_tree.py, 2\_cart\_use\_sklearn.py（两种实现方式）

3. 需自行实现（要求两种实现方式）

4. 分析实验结果差异

5. 需自行实现

**六、实验报告提交**

[发送电子版至TA邮箱iamironmancc@qq.com，格式：2024](mailto:发送电子版至TA邮箱1121296671@qq.com，格式：2024)机器学习第三次实验报告暨期中测试-姓名，截止时间2025年12月29日11:59pm。

附：

**关于决策树可视化（选做）**

使用sklearn的决策树可视化生成dot文件，运行dot文件需要安装Graphviz软件，步骤较为复杂，不在本课程的教学范围内，具体可参考[链接](https://blog.csdn.net/lanchunhui/article/details/49472949?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522163749352616780271929401%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=163749352616780271929401&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduend~default-2-49472949.first_rank_v2_pc_rank_v29&utm_term=windows%E4%B8%8BGraphviz%E5%AE%89%E8%A3%85%E5%8F%8A%E5%85%A5%E9%97%A8%E6%95%99%E7%A8%8B&spm=1018.2226.3001.4187)。

简单的决策树可视化，可在**PyCharm**中安装***dotplugin***插件，安装完成后在**PyCharm**中打开dot文件即实现可视化。（**VSCode**中安装***Graphviz (dot) language support for Visual Studio Code***插件，使用方式请查看插件说明）

**关于数据集**

**“Iris”**

鸢尾花数据集包含4种特征，以及3种鸢尾花类别。

详情请看sklearn官方文档 [链接](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris)

**“Breast”**

乳腺癌数据包含30个特征，以及患病、不患病两个类别。

详情请看sklearn官方文档 [链接](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer)

**“wifi”**

WiFi定位数据集包含7个特征，以及4个位置类别。

详情请看uci机器学习数据库 [链接](https://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization)