##### QG工作室数据挖掘小组实验报告

|  |  |  |
| --- | --- | --- |
| 实习生：李炜乐 | 导师：张润鹏 | 日期：2019 年 7 月 20日 |

|  |
| --- |
| **实验名称：CART分类树和回归树的实验** |
| **已完成内容：**   1. **完成了CART分类树的算法** 2. **完成了CART回归树的算法** |
| **未完成内容：暂无** |
| **未完成原因：暂无** |
| **需要帮助：暂无** |

|  |  |
| --- | --- |
| **实验总结** | |
| **知识点总结：**   1. **CART分类树包括计算基尼系数、判断基尼系数的是大小，递归生成决策树三个步骤。其中算法停止的条件：（1）结点中地样本个数小于预定阈值（2）样本基地基尼指数小于预定阈值（3）没有更多地特征** 2. **CART分类树的工作原理是，通过基尼公式计算基尼系数，选取基尼系数最少的特征作为最优点，进行分割。运行时间为0.06968551s** 3. ****CART回归算法正好适用于连续型特征。CART算法使用二元切分法来处理连续型变量。而使用二元切分法则易于对树构建过程进行调整以处理连续型特征。具体的处理方法是：如果特征值大于给定值就走左子树，否则就走右子树。**** 4. **CART回归算法一共分为两步。（1）：决策树生成：递归地构建二叉决策树的过程，基于训练数据集生成决策树，生成的决策树要尽量大；自上而下从根开始建立节点，在每个节点处要选择一个最好的属性来分裂，使得子节点中的训练集尽量的纯。不同的算法使用不同的指标来定义"最好"。（2）决策树剪枝：用验证数据集对已生成的树进行剪枝并选择最优子树，这时损失函数最小作为剪枝的标准。** | |
| **遇到问题**：在处理数据集的时候发现数据集中较多的数据都存在着缺失值，这个会一定程度影响决策树的判断 | **解决过程：使用出现次数最多的哪个数字进行填补** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **导师评价** | | | | |
| **实验分数** | **知识掌握情况** | **代码编写能力** | **建议** | **评价日期** |
|  |  |  |  |  |