**ML\_决策树 集成学习**

**简答题**

1. 根据下表所给的训练数据集，利用信息增益比（C4.5算法）生成决策树。
2. 对下表所给的训练数据集，根据信息增益准则选择最优特征。
3. 根据下表所给的训练数据集，利用ID3算法生成决策树。
4. 根据下表所给的训练数据集，利用CART算法生成决策树。

贷款申请样本数据表

| **ID** | **年龄** | **有工作** | **有自己的房子** | **信贷情况** | **类别** |
| --- | --- | --- | --- | --- | --- |
| 1 | 青年 | 否 | 否 | 一般 | 否 |
| 2 | 青年 | 否 | 否 | 好 | 否 |
| 3 | 青年 | 是 | 否 | 好 | 是 |
| 4 | 青年 | 是 | 是 | 一般 | 是 |
| 5 | 青年 | 否 | 否 | 一般 | 否 |
| 6 | 中年 | 否 | 否 | 一般 | 否 |
| 7 | 中年 | 否 | 否 | 好 | 否 |
| 8 | 中年 | 是 | 是 | 好 | 是 |
| 9 | 中年 | 否 | 是 | 非常好 | 是 |
| 10 | 中年 | 否 | 是 | 非常好 | 是 |
| 11 | 老年 | 否 | 是 | 非常好 | 是 |
| 12 | 老年 | 否 | 是 | 好 | 是 |
| 13 | 老年 | 是 | 否 | 好 | 是 |
| 14 | 老年 | 是 | 否 | 非常好 | 是 |
| 15 | 老年 | 否 | 否 | 一般 | 否 |

1. 已知如下表所示的训练数据，试用平方误差损失准则生成一个二叉回归树。

| **x** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| y | 4.50 | 4.75 | 4.91 | 5.34 | 5.80 | 7.05 | 7.90 | 8.23 | 8.70 | 9.00 |

1. 试析使用“最小训练误差”作为决策树划分选择准则的缺陷。
2. 试将4.4.2节对缺失值的处理机制推广到基尼指数的计算中去。
3. 某公司招聘职员考查身体、业务能力、发展潜力这3项。身体分为合格1、 不合格0两级，业务能力和发展潜力分为上1、中2、下3三级。分类为合格1、不合格-1两类。已知10个人的数据，如表所示。假设弱分类器为决策树桩。试用 AdaBoost 算法学习一个强分类器。

应聘人员情况数据表

|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 身体 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| 业务能力 | 1 | 3 | 2 | 1 | 2 | 1 | 1 | 1 | 3 | 2 |
| 发展潜力 | 3 | 1 | 2 | 3 | 3 | 2 | 2 | 1 | 1 | 1 |
| 分类 | -1 | -1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | -1 |

1. 给定如表所示训练数据。假设弱分类器由 x＜v或 x＞v产生，其阈值v使该分类器在训练数据集上分类误差率最低。试用 AdaBoost 算法学习一个强分类器。

| **序号** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| y | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 |

1. 试析随机森林为何比决策树 Bagging 集成的训练速度更快。
2. MultiBoosting 算法［Webb，2000］将AdaBoost 作为Bagging 的基学习器。Iterative Bagging 算法［Breiman，2001b］则是将 Bagging 作次 AdaBoost 的基学习器。试比较二者的优缺点。
3. GradientBoosting［Friedman，2001］是一种常用的Boosting 算法，试析其与 AdaBoost 的异同。
4. 试析 Bagging 通常为何难以提升朴素贝叶斯分类器的性能。
5. 简述如何增强个体学习器的多样性。