2023-2024 学年秋季学期《模式识别与机器学习》课程第 3 章作业题

一、填空题

1. 构建决策树时，选择最优划分属性的准则有_____。

2. 在决策树中，信息熵越大，证明样本集合的纯度越_____（高/低）。

3. 构建决策树的信息增益准则对取值数目更_____的属性有利。

二、选择题

1. （**单选题**）以下关于决策树的说法错误的是（　　）

A．决策树容易过拟合，可以使用预剪枝或后剪枝进行处理

B．相比于预剪枝，后剪枝更容易得到更简单的决策树

C．后剪枝需要判断根结点的剪枝与否

D．决策树既可用于分类问题，也可用于回归问题

2. （**多选题**）关于决策树方法，以下说法正确的是（　　）

A. 多变量决策树能得到与坐标轴呈一定夹角的分类面

B. 增益率准则对取值数目多的属性有所偏好

C. 基尼指数衡量了数据在类别上的不确定性的减小程度

D. 预剪枝和后剪枝策略都可以用于预防过拟合

三、计算题

下表是由 15 个样本组成的贷款申请数据集，数据包括贷款申请人的年龄、收入情况、是否有车、信用情况四项属性，最后一列为是否同意贷款，作为我们的预测结果：

序号	年龄	是否有车	收入情况	信用情况	是否同意贷款
1	19	否	一般	一般	否
2	22	否	一般	好	否
3	75	否	一般	一般	否
4	21	否	一般	一般	否
5	36	否	一般	一般	否
6	40	否	一般	好	否
7	69	是	一般	好	是
8	45	是	良好	好	是
9	52	是	一般	非常好	是
10	66	是	一般	非常好	是

11	25	否	良好	好	是
12	42	是	一般	非常好	是
13	59	否	良好	好	是
14	61	否	良好	非常好	是
15	29	是	良好	一般	是

1. 假如我们按照小于 30 岁、30 到 60 岁、大于 60 岁将申请人的年龄分为三组，并将{5，6，14，15}号样本作为验证集，其余样本作为训练集。请利用增益率和后剪枝策略建立决策树，写出决策树的建立过程并画出最终的决策树结构。

2. 将 1 中的增益率准则替换为基尼指数准则，后剪枝策略替换为预剪枝策略，请重新建立决策树，写出决策树的建立过程并画出最终的决策树结构。