1、决策树是一类常见的机器学习方法，是基于树结构进行决策的。一般的，一棵决策树包含两类结点：内部节点和叶结点，其中内部节点表示表示一个特征或属性，叶结点表示\_决策结果\_\_\_\_\_\_\_\_\_\_\_\_\_。

2、在决策树学习中，一般情况下，属性的信息增益越大，则意味着使用属性来进行划分获得的\_\_\_“纯度提升”越大\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_。

3、信息增益准则对\_\_可取数值数目较多\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_的属性有所偏好，增益率准则对\_\_\_可取数值数目较少\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_的属性有所偏好。

4、在决策树学习中，C4.5决策树算法中采用\_二分法\_对连续属性进行离散化处理。

5、决策树学习算法包括3部分：特征选择、树的生成和树的剪枝。特征选择的目的在于选择对训练数据能够分类的特征。特征选择的关键是其准则，常用的准则有哪些，请简单描述。

信息增益：信息增益是指假定采用a来对样本集进行划分，设有n个取值，第v个取值为x，把样本属性a取值为x的样本拿出来计算信息熵，附上权重后与原集合的信息熵做差即为用属性a划分的信息增益；

增益率：增益率准则对可取数值数目较少的属性有所偏好，C4.5算法并不是直接选择增益率最大的属性进行划分，而是使用了一个启发式：先从候选划分属性中找出信息增益高于平均水平的属性，再从其中找出增益率最高的，这样中和了两种准测对样本数目的偏好；

基尼指数：基尼指数反映了从数据集D中随机抽取两个样本，其类别标记不一致的概率，因此基尼指数越小，则数据集D的纯度越高。

6、目标变量在训练集上的 10 个实际值 [0,0,0,0,1,1,1,1,1,1]，则目标变量的熵是\_\_\_-3/5 \* log3/5\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_。

7、C4.5决策树算法中采用\_\_\_\_二分法\_\_\_\_\_\_\_\_\_\_\_\_\_\_对连续属性进行处理。

8、常用的决策树学习算法有ID3、C4.5和CART，介绍它们采用的特征选择准则是什么？

ID3采用的是信息增益作为其特征选择的标准，C4.5采用增益率，CART采用基尼指数。

9、简述决策树生成与决策树剪枝。

决策树生成：产生一个泛化能力强，即处理未见示例能力强的决策树，其基本流程遵循简单且直观的“分而治之”策略；

决策树剪枝：剪枝是决策树学习算法对付“过拟合”的主要手段，基本策略有“预剪枝”和“后剪枝”，都是为了提高决策树的泛化能力。

10、决策树剪枝的基本策略有预剪枝和后剪枝，请简述并分析两种剪枝策略。

预剪枝是指在决策树生成的过程中，对每个结点在划分前后先进行估计，若当前结点的划分不能带来决策泛化性能提升，则停止划分并把当前节点划分为叶子节点；

后剪枝则是先从训练集生成一颗完整的决策树，然后自底向上地对非叶子结点进行考察，若将该节点对应的子树替换成一个叶节点能提升决策树泛化性能，则将该子树替换成叶节点。

11、根据表4.1中的西瓜数据集，计算属性“纹理”的信息增益。

