作业答案

1.课上所有的案例都要自己手动完成

略

2.下面以常用的贷款申请样本数据表为样本集，计算信息增益计算过程，并构建ID3决策树。

ID	年龄	有工作	有房子	信贷情况	类别
1	青年	否	否	一般	拒绝
2	青年	否	否	好	拒绝
3	青年	是	否	好	同意
4	青年	是	是	一般	同意
5	青年	否	否	一般	拒绝
6	中年	否	否	一般	拒绝
7	中年	否	否	好	拒绝
8	中年	是	是	好	同意
9	中年	否	是	非常好	同意
10	中年	否	是	非常好	同意
11	老年	否	是	非常好	同意
12	老年	否	是	好	同意
13	老年	是	否	好	同意
14	老年	是	否	非常好	同意
15	老年	否	否	一般	拒绝

Step1 计算经验熵

类别一共是两个拒绝/同意，数量分别是6和9，根据熵定义可得：

$$H(D)=-\frac{9}{15}\log_2\frac{9}{15}-\frac{6}{15}\log_2\frac{6}{15}=0.971$$

Step2 各特征的条件熵

将各特征分别记为 A_1,A_2,A_3,A_4 ， 分别代表年龄、有无工作、有无房子和信贷情况，那么

$$H(D\mid A_1)=\frac{5}{15}(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5})+\frac{5}{15}(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5})+\frac{5}{15}(-\frac{4}{5}\log_2\frac{4}{5}-\frac{1}{5}\log_2\frac{1}{5})$$
$$H(D\mid A_2)=\frac{5}{15}(-\frac{5}{5}\log_2\frac{5}{5})+\frac{10}{15}(-\frac{4}{10}\log_2\frac{4}{10}-\frac{6}{10}\log_2\frac{6}{10})=0.647$$
$$H(D\mid A_3)=0.551$$
$$H(D\mid A_4)=0.608$$

Step3 计算增益

$$g(D,A_1)=H(D)-H(D\mid A_1)=0.083$$
$$g(D,A_2)=H(D)-H(D\mid A_2)=0.324$$
$$g(D,A_3)=H(D)-H(D\mid A_3)=0.420$$
$$g(D,A_4)=H(D)-H(D\mid A_4)=0.363$$

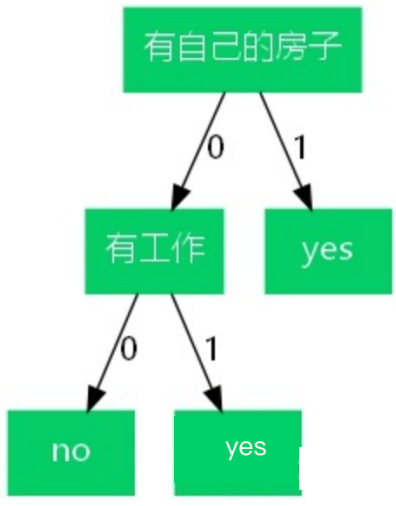
根据计算所得的信息增益，选取最大的A_3 作为根节点的特征。它将训练集 D 划分为两个子集D_1（取值为“是”）和D_2（取值为“否”）。由于D_1只有同一类的样本点，所以成为一个叶节点，节点标记为“是”。

对于D_2需从特征A_1,A_2,A_4中选择新的特征。计算各个特征的信息增益

$$g(D_2, A_1) = 0.918 - 0.668 = 0.251$$
$$g(D_2, A_1) = 0.918$$
$$g(D_2, A_1) = 0.474$$

选择信息增益最大的特征A_2作为节点的特征。由于A_2有两个可能取值，一个是“是”的子节点，有三个样本，且为同一类，所以是一个叶节点，类标记为“是”；另一个是“否”的子节点，包含6个样本，也属同一类，所以也是一个叶节点，类别标记为“否”。

最终构建的决策树如下：



3.说明CART分类树和回归树的特点，并说明构建过程

- CART分类树
  - 采用基尼指数，计算量减小，一定是二叉树，预测输出的是一个离散值，使用叶子节点多数类别作为预测类别
  - 构建过程：各个特征先分类，计算基尼值，计算基尼指数。如果是多个值，将值排序，以相邻中间值作为待确定分裂点，计算出两部分的基尼指数，比较出最小的基尼指数，为该特征的基尼指数。比较各个特征的基尼指数，优先选择最小的特征。
- CART回归树：
  - 预测输出的是一个连续值，使用平方损失作为划分、构建树的依据，采用叶子节点里均值作为预测输出
  - 构建过程：将特征值排序，以相邻中间值作为待划分点，根据划分点，将数据集分为两部分，两部分平方损失相加作为该切分点平方损失，取最小的平方损失的划分点，作为当前特征的划分点，依次类推，计算所以特征的最优划分点和对应损失值，比较所有特征的最小平方损失的划分点，作为当前树的分裂点

4.说明决策树剪枝的方法有哪些？及各自的特点是什么？

- 预剪枝：（边构建树边剪枝）

使决策树的很多分支没有展开，不单降低了过拟合风险，还显著减少了决策树的训练、测试时间开销但是有些分支的当前划分虽不能提升泛化性能，但后续划分却有可能导致性能的显著提高；预剪枝决策树也带来了欠拟合的风险

- 后剪枝：（先构建树，然后自底向上剪枝）

比预剪枝保留了更多的分支。一般情况下，后剪枝决策树的欠拟合风险很小，泛化性能往往优于预剪枝但是训练时间开销比未剪枝的决策树和预剪枝的决策树都要大得多。