作业 2：决策树构建、模型评估

1. 为什么不能用训练集中的样本来测试训练好的模型？请给出留出法和 K 折交叉验证法的具体步骤。

   答：模型测试是为了评估模型的泛化能力。用训练集中的样本进行测试得到的是训练误差，训练误差小可能出现过拟合，因此训练误差不能有效评估模型的泛化能力。

   留出法：

   一般采用分层采样。从每个类别中按照给定的比例（无放回）随机采样一部分样本作为测试集，每个类别中未被采样到的样本作为训练集。一般用多次留出法的结果取平均。

   K 折交叉验证法：

   把训练集随机分成 K 个大小一样不相交的子集，每次分别用其中 1 个子集作为测试集，剩下的 K-1 个子集中的所有样本作为训练集，总共训练 K 次，对 K 个结果取平均。

2. 假设有一个测试集，其真实标签和模型预测出来的标签分别如下，请给出混淆矩阵，并计算查准率(precision)、查全率(recall)、和 F1 度量(F1-measure)。

编号	真实标签	预测标签
1	是	否
2	否	否
3	否	否
4	是	是
5	是	否
6	否	否
7	否	是
8	是	是

混淆矩阵：

	预测正例	预测反例
真实正例	TP=|{4,8}|=2	FN=|{1，5}|=2
真实反例	FP=|{7}|=1	TN=|{2，3，6}|=3

$$Precision = \frac{TP}{TP + FP} = \frac{2}{2 + 1} = 0.67$$

$$Recall = \frac{TP}{TP+FN} = \frac{2}{2+2} = 0.5$$

$$F1 - measure = \frac{2*Presicion*Recall}{resicion+Recall} = \frac{2*0.67*0.5}{0.67+0.5} = 0.57$$

3. 基于顾客的“年龄”、“收入”、“学生”这三个属性构造决策树用于预测一个学生是否会买电脑。其中三个属性的取值范围分别为：“年龄”={青年、中年、老年}，“收入”={高、中、低}，“学生”={是、否}。具体要求如下：

   a.用以下训练集基于信息增益构造决策树，给出中间步骤，并画出决策树。

   b.基于上面构造的决策树，对测试集中样本进行预测，即顾客是否会买电脑， 并计算 accuracy。

   *方便大家计算，现给出以下对数的具体数值：$\log_2(3) = 1.5850$，$\log_2(5) = 2.3219$

训练集

编号	**类别:** 是否买电脑	年龄	收入	学生
1	否	青年	高	否
2	否	青年	高	否
3	是	中年	高	否
4	是	老年	中	否
5	是	老年	中	是
6	否	老年	低	是
7	否	青年	中	否
8	是	老年	中	是

测试集

编号	**类别:** 是否买电脑	年龄	收入	学生
1	是	中年	低	是
2	是	青年	中	是
3	否	老年	低	否

参考答案

a.以下给出具体的构造过程。注意：该题目中**第一列为要预测的类别**。

划分前数据集：D={1,2,3,4,5,6,7,8}，其信息熵$Ent(D) = -\sum_{i=1}^{n} p_i\, log_2(p_i) = -\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right) = 1$。

根据$Ent(D,a) = \sum_{v=1}^{V} \frac{|D_v|}{|D|} Ent(D_v)$，以及当前可用属性集$A = \{$*年龄*、*收入*、*学生*$\}$，计算 A 中每个特征划分后的信息熵如下：

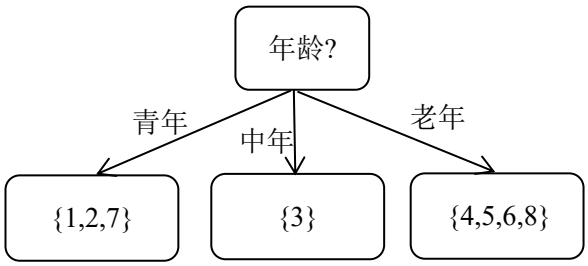
$$Ent(D, 年龄) = \frac{3}{8}\left(-\frac{3}{3}log_2\left(\frac{3}{3}\right)\right) + \frac{1}{8}\left(-\frac{1}{1}log_2\left(\frac{1}{1}\right)\right) + \frac{4}{8}\left(-\frac{1}{4}log_2\left(\frac{1}{4}\right) - \frac{3}{4}log_2\left(\frac{3}{4}\right)\right) = 0.4056$$

$$Ent(D, 收入) = \frac{1}{8}\left(-\frac{1}{1}log_2\left(\frac{1}{1}\right)\right) + \frac{4}{8}\left(-\frac{1}{4}log_2\left(\frac{1}{4}\right) - \frac{3}{4}log_2\left(\frac{3}{4}\right)\right) + \frac{3}{8}\left(-\frac{2}{3}log_2\left(\frac{2}{3}\right) - \frac{1}{3}log_2\left(\frac{1}{3}\right)\right) = 0.75$$

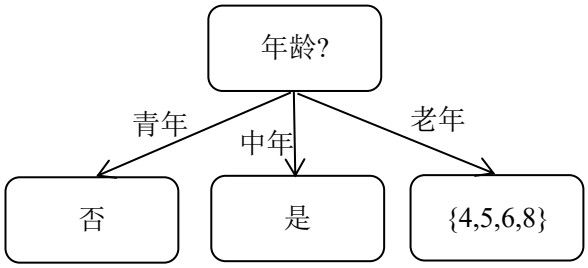
$$Ent(D, 学生) = \frac{5}{8}\left(-\frac{3}{5}log_2\left(\frac{3}{5}\right) - \frac{2}{5}log_2\left(\frac{2}{5}\right)\right) + \frac{3}{8}\left(-\frac{1}{3}log_2\left(\frac{1}{3}\right) - \frac{2}{3}log_2\left(\frac{2}{3}\right)\right) = 0.9512$$

根据信息增益：$\text{Gain}(D,a) = \text{Ent}(D)\text{-Ent}(D,a)$，得到每个特征的信息增益为：$\text{Gain}(D, 年龄) = 1 - 0.4056 = 0.5944$，$\text{Gain}(D, 收入) = 1 - 0.75 = 0.25$，$G\text{ain}(D, 学生) = 1 - 0.9512 = 0.0488$

“年龄”属性的信息增益最高，所以作为当前分裂属性，得到



其中，“青年”和“中年”两个节点样本类别已经一致，所以作为叶子节点，得到：



对“老年”节点再划分，此时 D={4,5,6,8}，$\mathrm{Ent}(D) = -\frac{1}{4}log_2\left(\frac{1}{4}\right) - \frac{3}{4}log_2\left(\frac{3}{4}\right) = 0.8113$。

当前可用属性集更新为$A = \{收入、学生\}$，计算 A 中每个特征划分后的信息熵如下：

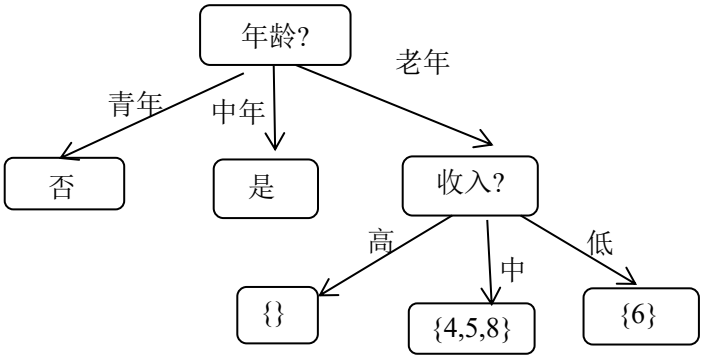
$$Ent(D, 收入) = \frac{1}{4}\left(-\frac{1}{1}log_2\left(\frac{1}{1}\right)\right) + \frac{3}{4}\left(-\frac{3}{3}log_2\left(\frac{3}{3}\right)\right) = 0$$

$$Ent(D, 学生) = \frac{1}{4}\left(-\frac{1}{1}log_2\left(\frac{1}{1}\right)\right) + \frac{3}{4}\left(-\frac{1}{3}log_2\left(\frac{1}{3}\right) - \frac{2}{3}log_2\left(\frac{2}{3}\right)\right) = 0.6887$$

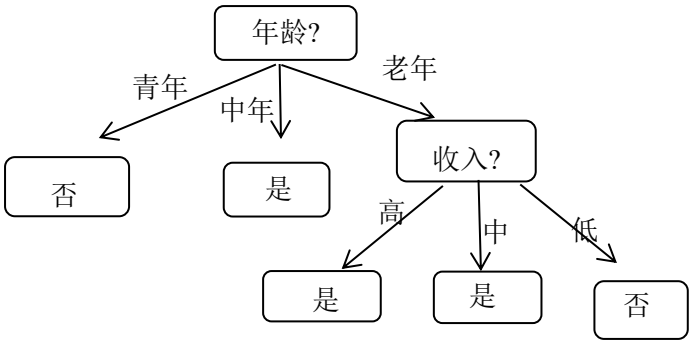
得到每个特征的信息增益为：

$\mathrm{Gain}(D, 收入) = 0.8113 - 0 = 0.8113$，$\mathrm{Gain}(D, 学生) = 0.8113 - 0.6887 = 0.1226$

“收入”属性的信息增益最高，所以作为当前分裂属性，决策树更新为：



其中，“老年->收入=低”和“老年->收入=中”两个节点样本的类别一致，所以作为叶子节点，“老年-收入高”对应的训练样本数目为 0，此时用父亲节点“老年”的样本{4,5,6,8}中数目最多的类别“是”来标记。最后得到的决策树为：



由于当前所有节点均为叶子节点，决策树构造完成。

b.测试：

根据以上所构造决策树，从上至下对各个特征遍历，得到

序号 9：年龄“中年”，预测“是”，实际“是”，正确；

序号 10：年龄“青年”，预测“否”，实际“是”，错误；

序号 11：年龄“老年”、收入“低”，预测“否”，实际“否”，正确；

Accuracy=$\frac{2}{3}$=0.67。