## 第四章 决策树

4.1 什么是熵？

参考答案：1948年，香农Claude E. Shannon引入信息（熵），将其定义为离散随机事件的出现概率。一个系统越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。所以说，信息熵可以被认为是系统有序化程度的一个度量。

4.2 对于树形结构为什么不需要归一化？

　　参考答案：数值缩放，不影响分裂点位置。因为第一步都是按照特征值进行排序的，排序的顺序不变，那么所属的分支以及分裂点就不会有不同。对于线性模型，比如说LR，我有两个特征，一个是(0,1)的，一个是(0,10000)的，这样运用梯度下降时候，损失等高线是一个椭圆的形状，这样我想迭代到最优点，就需要很多次迭代，但是如果进行了归一化，那么等高线就是圆形的，那么SGD就会往原点迭代，需要的迭代次数较少。

另外，注意树模型是不能进行梯度下降的，因为树模型是阶跃的，阶跃点是不可导的，并且求导没意义，所以树模型（回归树）寻找最优点事通过寻找最优分裂点完成的。

4.3 决策树学习算法如何处理“过拟合”？

参考答案：剪枝处理。

4.4 决策树如何进行“剪枝”处理？

参考答案：预剪枝是指在决策树生成过程中，对每个结点在划分前先进行估计，若当前结点的划分不能带来决策树泛化性能提升，则停止划分并将当前结点标记为叶结点。

后剪枝则是先从训练集生成一棵完整的决策树，然后自底向上地对非叶结点进行考察，若将该结点对应的子树替换为叶结点能带来决策树泛化性能提升，则将该子树替换为叶结点.

4.5试证明对于不含冲突数据（即特征向量完全相同但标记不同）的训练集，必存在与训练集一致（即训练误差为0）的决策树。

参考答案：假设不存在与训练集一致的决策树，那么训练集训练得到的决策树至少有一个节点上存在无法划分的多个数据（若节点上没有冲突数据，那么总是能够将数据分开的）。这与前提-不含冲突数据矛盾，因此必存在与训练集一致的决策树。

4.6试析使用“最小训练误差”作为决策树划分选择的缺陷。

参考答案：若以最小训练误差作为决策树划分的依据，由于训练集和真是情况总是会存在一定偏差，这使得这样得到的决策树会存在过拟合的情况，对于未知的数据的泛化能力较差。因此最小训练误差不适合用来作为决策树划分的依据。

4.7 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

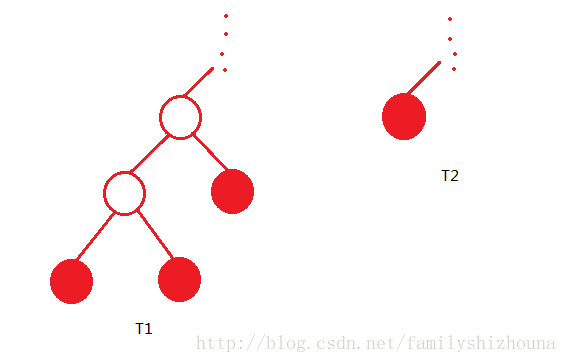
参考答案：把信息增益的公式换成基尼指数就行，包括扩展到连续参数，缺失参数。

4.8 证明CART剪枝算法中，当α确定时，存在唯一的最小子树Tα使损失函数Cα(T)最小。

证明：

假设当α确定时，存在两颗子树T1,T2都使损失函数Cα最小。

可以先看T1,T2不同的一个分支。如下图所示，T1,T2有相同的最小损失函数。但T2比较小，所以T1不是最小子树。



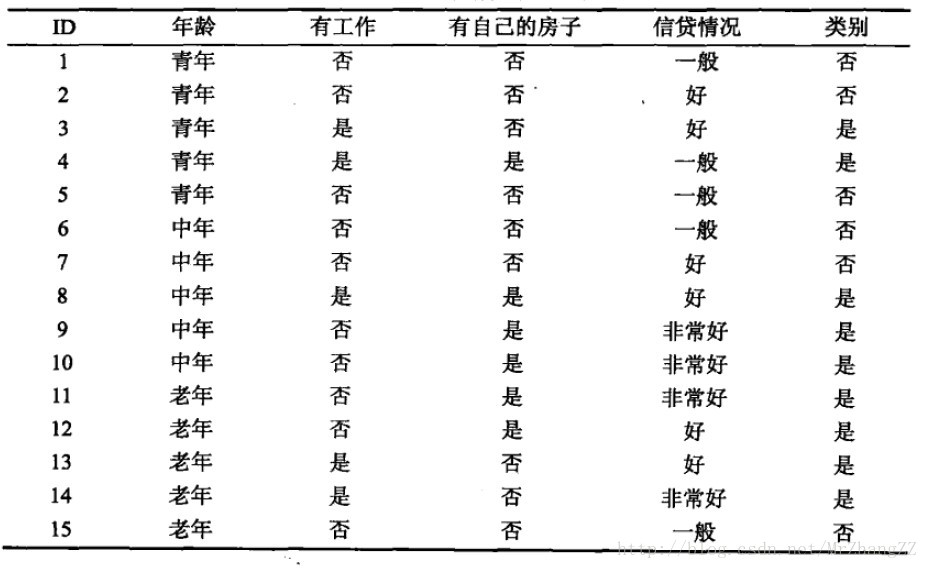
4.9 根据表4-1的训练数据集D，根据信息增益准则选择最优特征。  


表4-1−−贷款申请样本数据表

参考答案：

首先，计算经验熵

然后计算各个特征对数据集D的信息增益，分别以,,,表示年龄、有工作、有自己的房子和信贷情况4个特征，则

(1)

这里,,分别是D中（年龄）取值为青年、中年和老年的样本子集。类似的，

(2)

(3)

(4)

由于特征（有自己的房子）的信息增益值最大，所以选择特征作为最优特征.

4.10 对表1中的训练数据集，利用ID3算法建立决策树

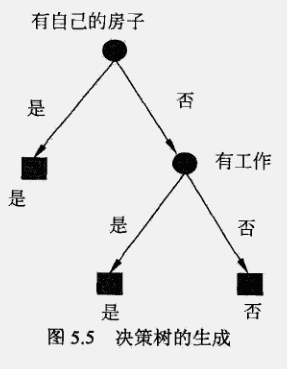
参考答案：

利用上题的结果，由于特征 (有自己的房子)的信息增益值最大，所以选择特征作为根节点的特征。它将训练数据集D划分为两个子集（取值为“是”）和 (取值为“否”)。由于只有同一类的样本点，所以它成为一个叶结点，结点标记为“是”。

对则需从特征（年龄），（有工作）和 (信贷情况)中选择新的特征，计算各个特征的信息增益：

选择信息增益最大的特征(有工作)作为结点的特征。由于有两个可能取值，从这一结点引出两个子结点：一个对应“是”（有工作）的子结点，包含3个样本，且都属于同一类别“是”（允许贷款），所以这是一个叶结点，标记为“是”；另一个是对应“否”（无工作）的子结点，包含6个样本，它们同属于一类“否”（“不允许贷款”），所以这也是一个叶结点，类标记为“否”。

这样生成一个如下图所示的决策树。该决策树只用了两个特征（有两个内部结点）



4.11对表1所给的训练数据集，应用CART算法生成决策树

参考答案：

首先计算各特征的基尼系数，选择最优特征以及其最优切分点。分别以,,,表示年龄、有工作、有自己的房子和信贷情况4个特征，并以1，2，3表示年龄的值为青年、中年和老年，以1，2表示有工作和有自己的房子的值为是和否，以1，2，3表示信贷情况的值为非常好、好和一般。

求特征A1的基尼指数：

由于和相等，且最小，所以和都可以选作的最优切分点.

求特征和的基尼指数：

由于和只有一个切分点，所以它们就是最优切分点.

求特征的基尼指数：

最小，所以为的最优切分点.

在,,,几个特征中， 最小，所以选择特征为最优特征，为其最优切分点。于是根据根结点生成两个子结点，一个是叶结点。对另一个结点继续使用以上方法在,,中选择最优特征及其最优切分点，结果是.依此计算得知，所得结点都是叶结点。

对于本问题，按照CART算法所生成的决策树与按照ID3算法所生成的决策树完全一致.

4.12 考虑表4-2中二元分类问题的训练样本。

(a)计算整个训练样本集的Gini指标值。

(b)计算属性顾客ID的Gini指标值。

(c)计算属性性别的Gini指标值。

(d)计算使用多路划分属性车型的Gini指标值。

(e)计算使用多路划分属性衬衣尺码的Gini指标值。

(f)下面哪个属性更好，性别、车型还是衬衣尺码？

(g)解释为什么属性顾客ID的Gini值最低，但是不能作为属性测试条件。



表4-2 习题4.12的数据集

参考答案：

(a)

(b)每个顾客ID的Gini为0，故顾客ID的总Gini基尼系数为0

(c)男性的Gini系数值为，女性也为0.5.

因此性别属性的Gini指标

(d)

(e)

(f) 车型最好，因为该属性的Gini最小.

(g) 因为顾客ID对于新顾客没有预测作用.

4.13 考虑表4-3中的二元分类问题的训练样本集。

(a) 整个训练样本集关于类属性的熵是多少？

(b) 关于这些训练样本，和的信息增益是多少？

(c) 对于连续属性，计算所有可能的划分的信息增益。

(d) 根据信息增益，哪个是最佳划分（在，和中）？

(e) 根据分类错误率，哪个是最佳划分（在中）？

(f) 根据Gini指标，哪个是最佳划分（在中）？

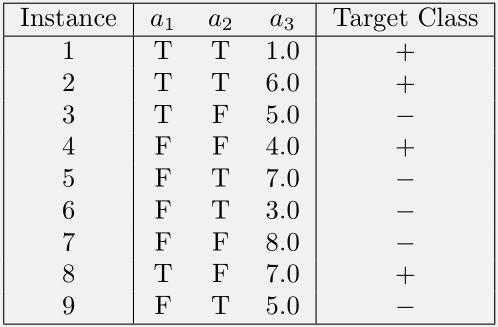
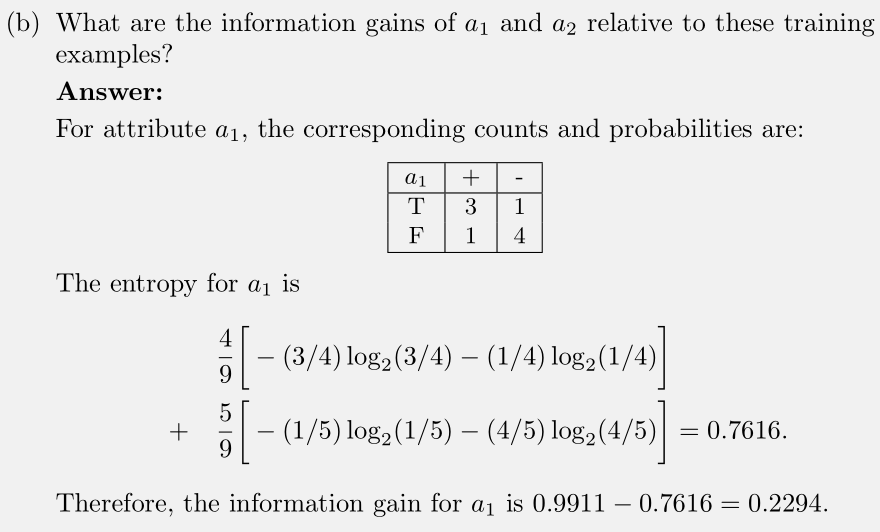


表4-3 习题4.13的数据集

参考答案：

(a)

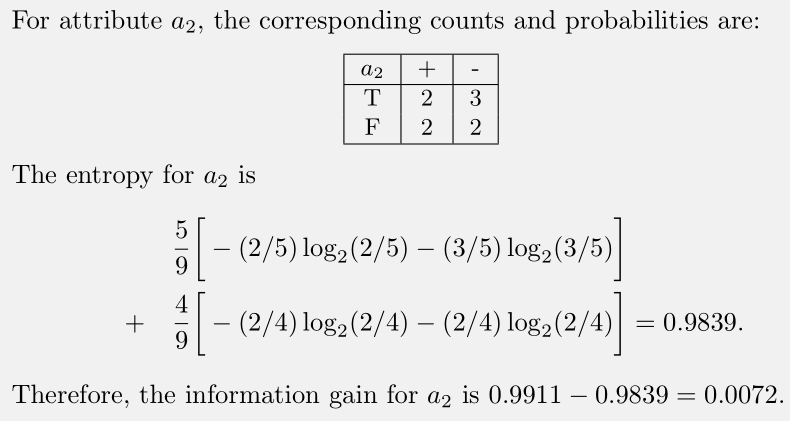
(b) 对于属性，相应的统计为



因此，的熵为

因此，的信息增益为 0.9911-0.7616=0.2294.

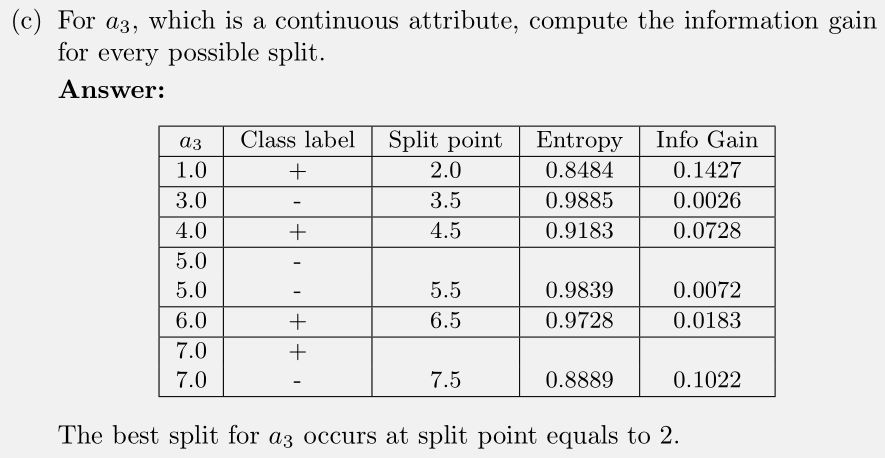
对于属性，相应的统计为



的熵为

因此，的信息增益为 0.9911-0.9839=0.0072

(c)



最好的划分点在等于2的位置.

(d)

(e) 属性的分类错误率=2/9，属性的分类错误率=4/9

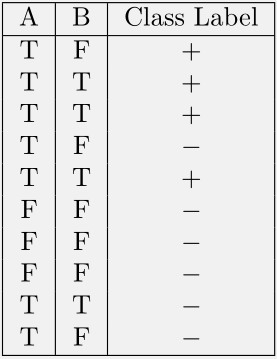
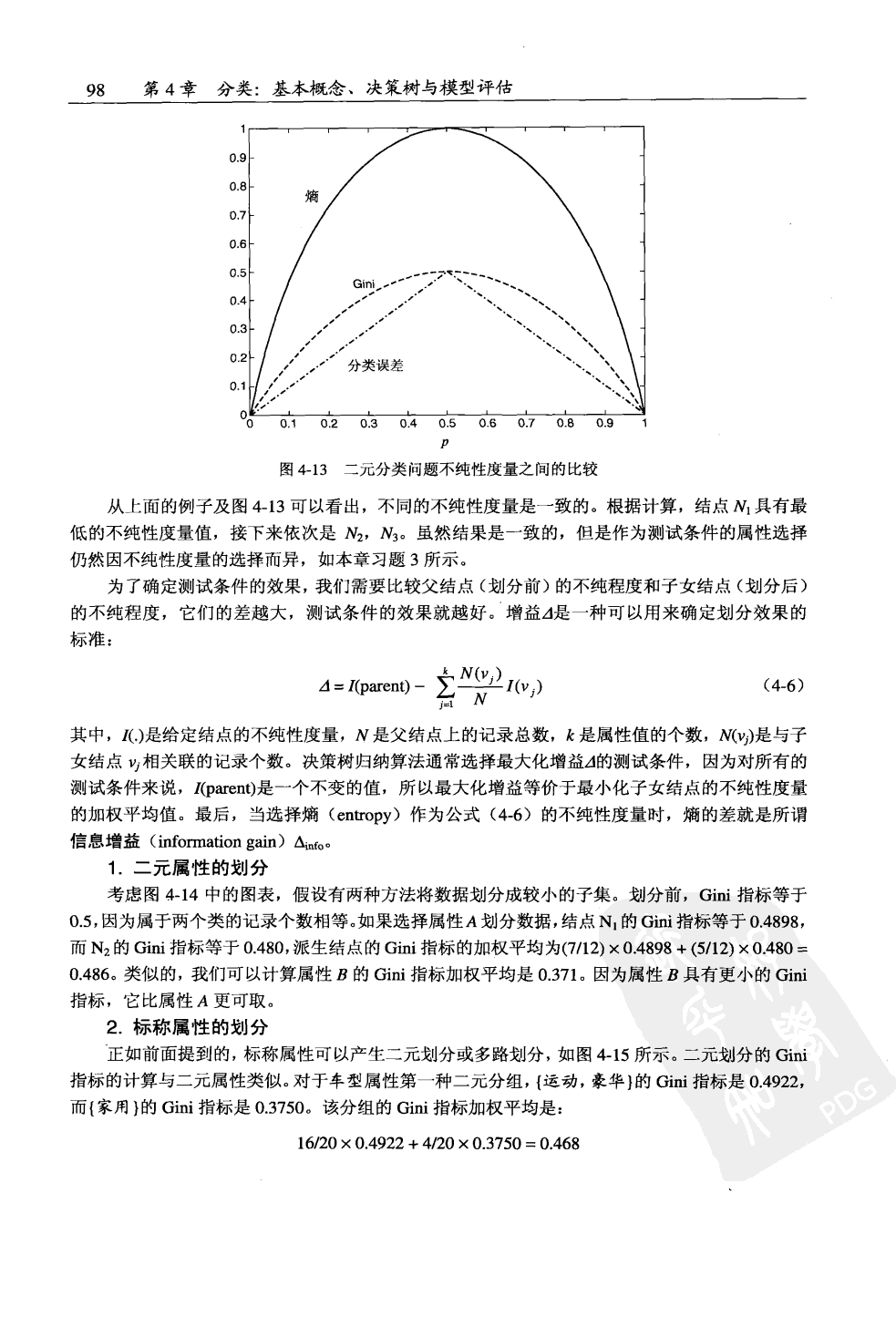
因此最佳划分属性为

(f) 属性的Gini指标为：

属性的Gini指标为：

因此最佳划分属性为

4.14 考虑如下二元分类问题的数据集。

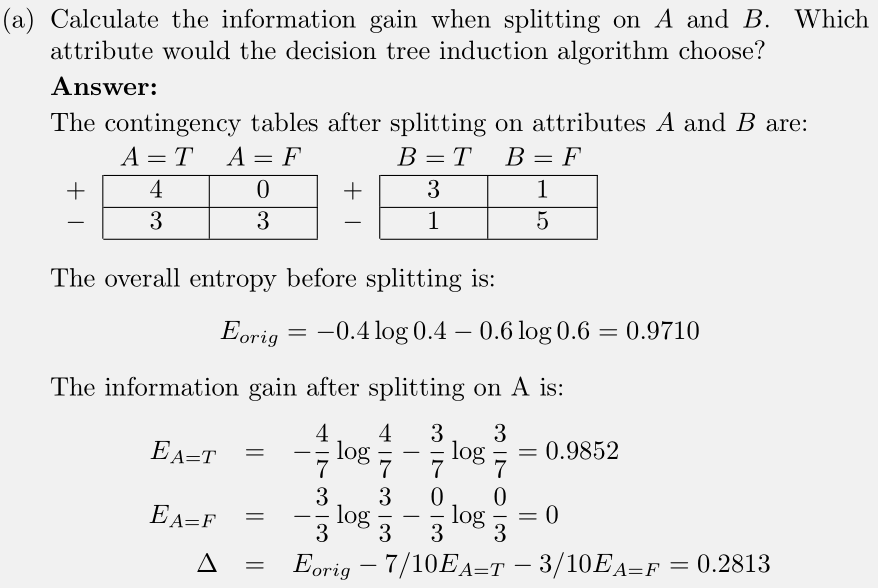
(a) 计算按照属性A和B划分时的信息增益。决策树归纳算法将会选择哪个属性？

(b) 计算按照属性A和B划分时的Gini指标。决策树归纳算法将会选择哪个属性？

(c) 从图4-13可以看出熵和Gini指标在区间[0,0.5]都是单调递增的，而在区间[0.5,1]都是单调递减的。有没有可能信息增益和Gini指标增益支持不同的属性？为什么？

参考答案：

(a)



按照属性A划分时的信息增益为

按照属性B划分时的信息增益为

因此，最佳划分属性为A.

(b)

总体Gini指标为

按照属性A划分时的Gini指标:

按照属性B划分时的Gini指标:

因此，最佳划分属性为B.

(c)有可能。虽然这两种测量方式有相同的区间和变化趋势，但是最佳划分属性是按照相对的增益选择的，在相对增益上，它们并不一定具有相同的趋势，如(a)、(b)所示.

4.15 某公司为适应市场的需要，准备扩大生产能力，有两种方案可供选择：

方案一：建大厂

方案二：先建小厂，后考虑扩建。

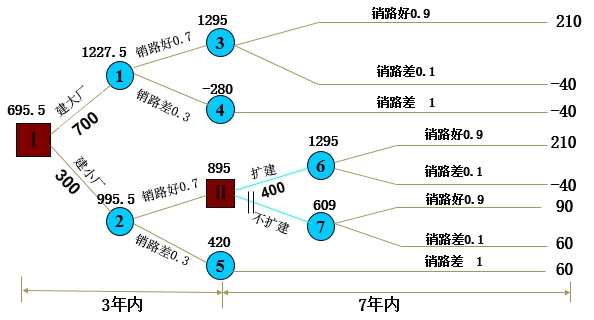
如建大厂，需投资700万元，在市场销路好时，每年收益210万元，销路差时，每年亏损40万元。在第二方案中，先建小厂，如销路好，3年后进行扩建。建小厂的投资为300万元，在市场销路好时，每年收益90万元，销路差时，每年收益60万元，如果3年后再扩建，扩建投资为400万元，收益情况同第一方案一致。

未来市场销路好的概率为0.7，销路差的概率为0.3；如果前3年销路好，则后7年销路好的概率为0.9，销路差的概率为0.1。如前3年销路差，则后7年的销路差为100%。无论选用何种方案，使用期均为10年。

试通过决策树法作出分析。

参考答案：

1、画出决策树图

[](http://www.cnitpm.com/upload/img/201509/20150922082649927284.jpg)

2、从右向左计算各点的期望收益值

点6：210×0.9×7-40×0.1×7-400=895（万元）

点7：90×0.9×7+60×0.1×7=609（万元）

点Ⅱ是个决策点，比较点6和点7的期望值，选择扩建。

点3：210×0.9×7-40×0.1×7=1295（万元）

点4：-40×1×7=-280（万元）

点5：60×1×7=420（万元）

点1：1295×0.7+210×0.7×3-280×0.3-40×0.3×3=1227.5（万元）

1227.5-700=527.5(万元)

点2：895×0.7+90×0.7×3+420×0.3+60×0.3×3=995.5（万元）

995.5-300=695.5(万元)

3、进行决策

∵　695.5 ＞ 527.5

∴　建小厂方案为优

比较点1和点2的期望值，点2期望值较大，可见，最优方案是先建小厂，如果销路好，3年以后再进行扩建。