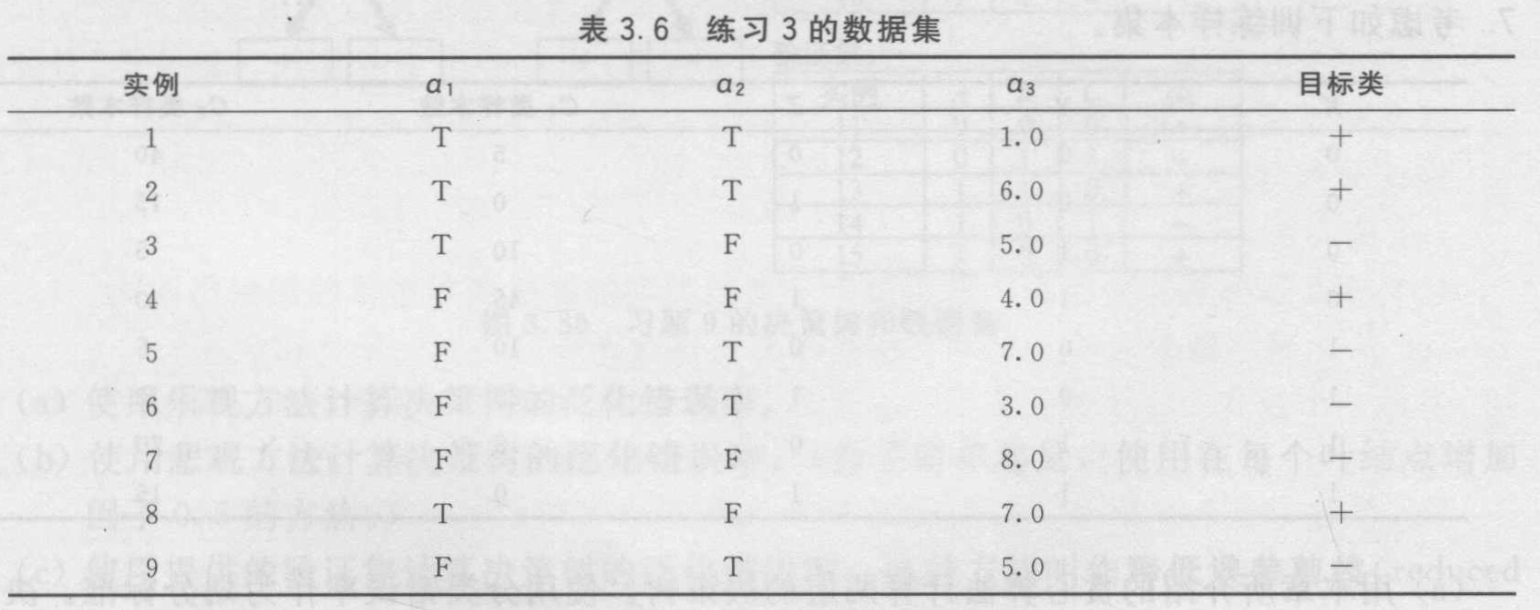
# 练习四

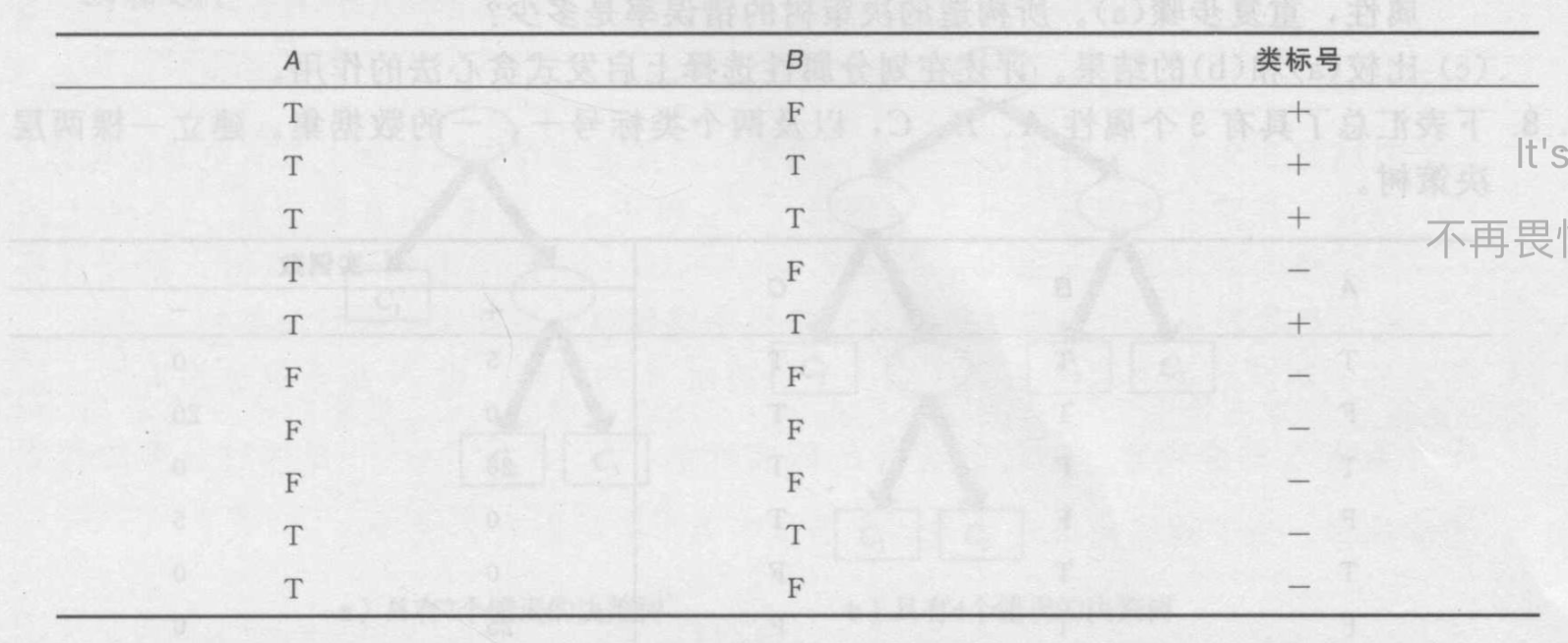
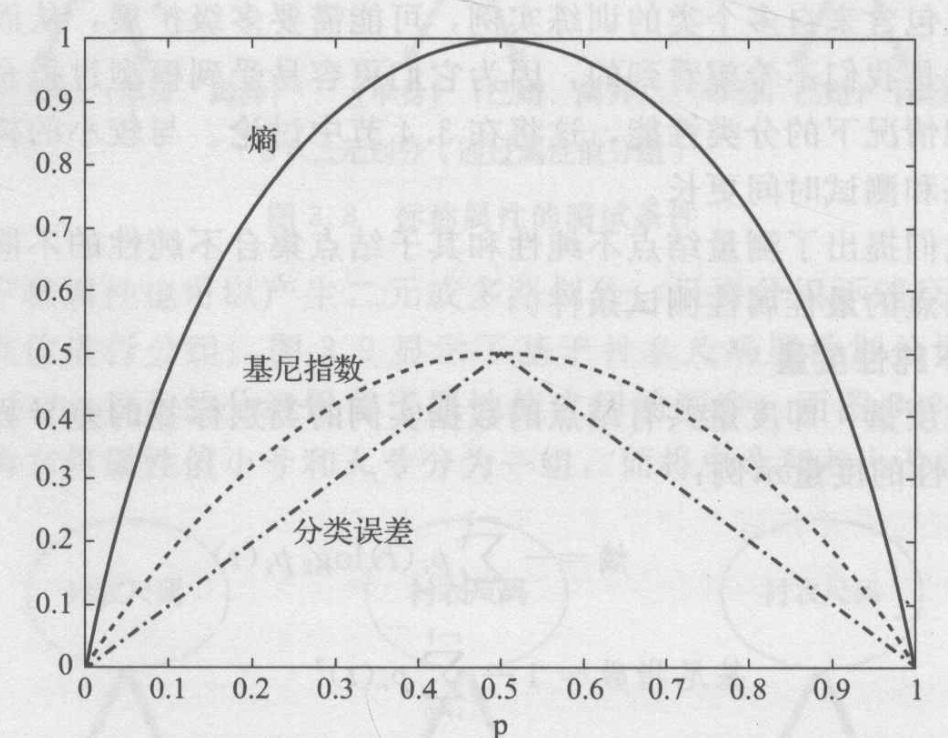
1. 考虑表3.5中二分类问题的训练样本。

* 
* （a）计算整个训练样本集的基尼指数值。
* （b）计算属性顾客ID的基尼指数值。
* （c）计算属性性别的基尼指数值。
* （d）计算使用多路划分属性车型的基尼指数值。
* （e）计算使用多路划分属性衬衣尺码的基尼指数值。
* （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的基尼指数值最低，但却不能作为属性测试条件。

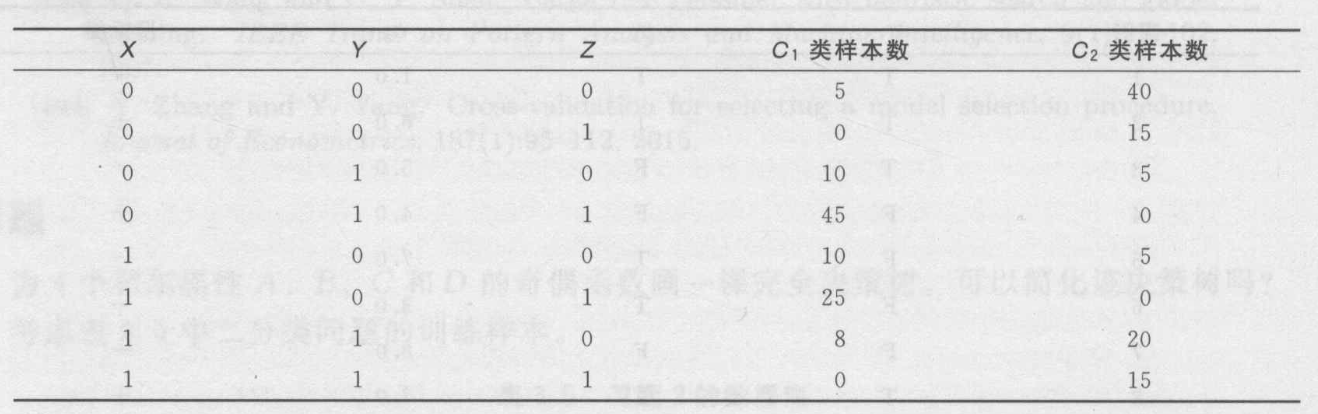
1. 考虑表3.6中的二分类问题的训练样本集。

* 
* （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（在和中）？
* （f）根据基尼指数，哪个是最佳划分（在和中）？

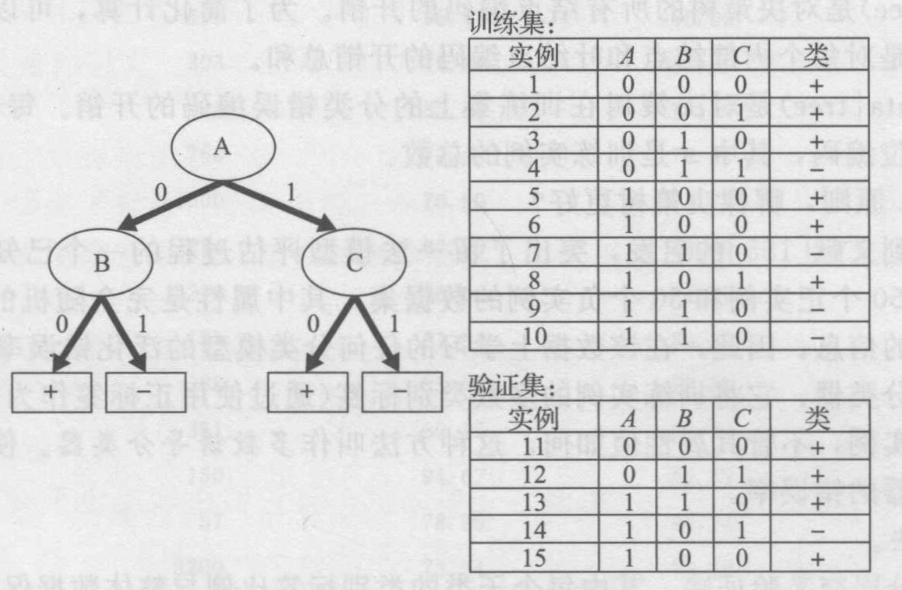
1. 考虑如下二分类问题的数据集。

* 
* （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划分时的基尼指数。决策树归纳算法将会选择哪个属性？
* （c）从图3.11可以看出熵和基尼指数在区间[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基尼指数增益支持不同的属性？解释你的理由。
* 

1. 考虑如下训练样本集。

* 
* （a）用本章所介绍的贪心算法计算两层的决策树。使用分类错误率作为划分标准。决策树的总错误率是多少？
* （b）使用作为第一个划分属性，两个后继结点分别在剩余的属性中选择最佳的划分属性，重复步骤（a）。所构造的决策树的错误率是多少？
* （c）比较（a）和（b）的结果。评述在划分属性选择上启发式贪心法的作用。

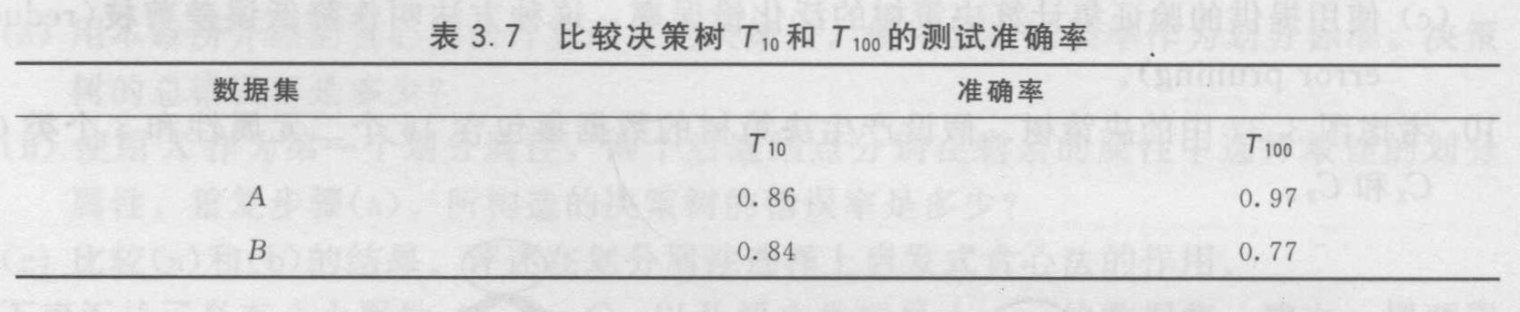
1. 考虑图3.36中的决策树。

* 
* （a）使用乐观方法计算决策树的泛化错误率。
* （b）使用悲观方法计算决策树的泛化错误率。（为了简单起见，使用在每个叶结点增加因子的方法。）
* （c）使用提供的验证集计算决策树的泛化错误率。这种方法叫作降低误差剪枝（reduced error pruning）。

1. 考虑图3.37中的决策树。假设产生决策树的数据集包含16个二元属性和3个类、和。根据最小描述长度原则计算每棵决策树的总描述长度。

* 
  + 树的整体描述长度由下式给出：
  + 树的每个内部结点用划分属性的ID进行编码。如果有个属性，为每个属性编码的代价是个二进位。
  + 每个叶结点使用与之相关联的类的ID编码。如果有个类，为每个类编码的代价是个二进位。
  + 是对决策树的所有结点编码的开销。为了简化计算，可以假设决策树的总开销是对每个内部结点和叶结点编码的开销总和。
  + 是对决策树在训练集上的分类错误编码的开销。每个错误用个二进位编码，其中是训练实例的总数。
* 根据MDL原则，哪棵决策树更好？

1. 考虑一个包含100个数据实例的标记数据集，该数据实例被随机分为A和B两组，每组包含50个实例。使用A作为训练集来学习两个决策树，具有10个叶结点，具有100个叶结点。数据集A和B的两个决策树有相应准确率数据。

* 
* （a）根据给定的准确率，你认为哪个分类模型在未知实例上具有更好的性能？
* （b）现在，在整个数据集上测试了和，发现在数据集上的分类准确率为0.85，而在数据集上的分类准确率为0.87。根据这些新信息和之前的观察结果，你最终会选择哪个分类模型进行分类？

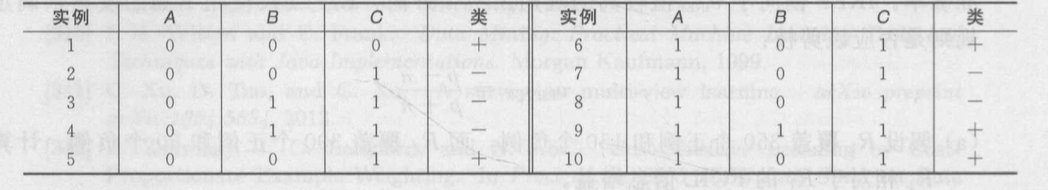
1. 考虑一个二分类问题，属性集和属性值如下：
   * 空调 = {可用，不可用}。
   * 引擎 = {好，差}。
   * 行车里程 = {高，中，低}。
   * 生锈 = {是，否}。

* 假设一个基于规则的分类器产生的规则集如下：
  + 行车里程 = 高 → 价值 = 低
  + 行车里程 = 低 → 价值 = 高
  + 空调 = 可用，引擎 = 好 → 价值 = 高
  + 空调 = 可用，引擎 = 差 → 价值 = 低
  + 空调 = 不可用 → 价值 = 低  
    （a）这些规则是互斥的吗？  
    （b）这些规则集是完全的吗？  
    （c）规则需要排序吗？  
    （d）规则集需要默认类吗？

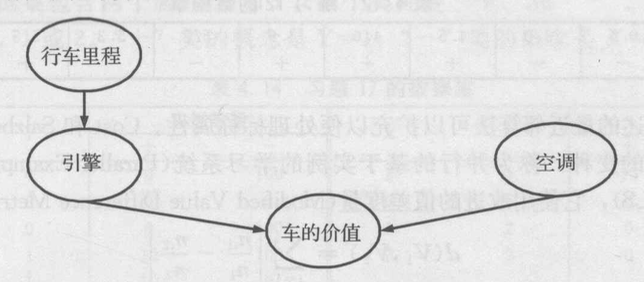
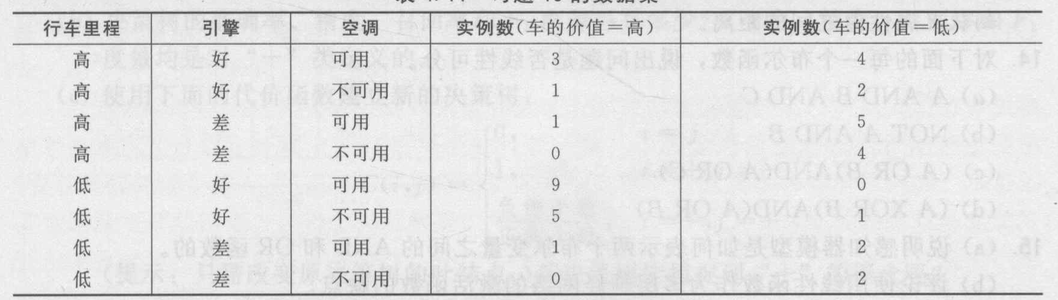
1. 考虑一个训练集，包含100个正例和400个负例。对于下面的候选规则：
   * ：（覆盖4个正例和1个负例）
   * ：（覆盖30个正例和10个负例）
   * ：（覆盖100个正例和90个负例）

* 根据下面的度量，确定最好规则和最差规则。
* （a）规则准确率。
* （b）FOIL信息增益。
* （c）似然比统计量。
* （d）拉普拉斯度量。
* （e）度量（且 ）。

1. 考虑表4.9中的数据集。

* 
* （a）估计条件概率 、、、、和 。
* （b）根据（a）中的条件概率，使用朴素贝叶斯方法预测测试样本的类标签。
* （c）使用估计方法（且）估计条件概率。
* （d）同（b），使用（c）中的条件概率。
* （e）比较估计概率的两种方法。哪一种更好？为什么？

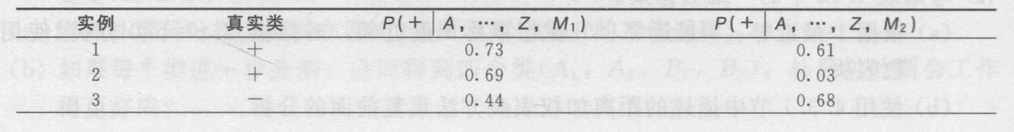
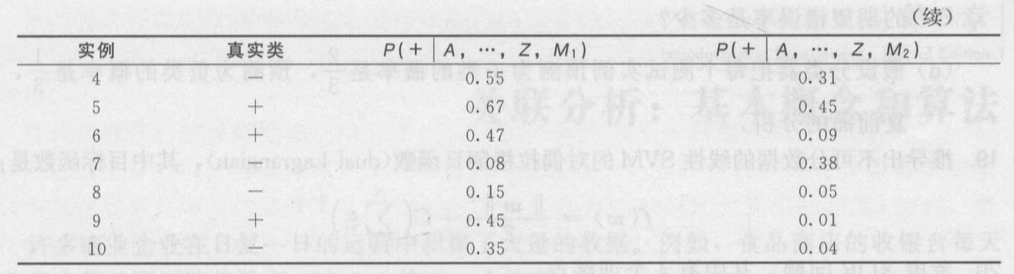
1. 图4.57给出表4.11中数据集对应的贝叶斯网络（假设所有属性都是二元的）。

* 
* 
* （a）画出网络中每个结点对应的概率表。
* （b）使用贝叶斯网络计算 。

1. 对下面的每一个布尔函数，说出问题是否线性可分。

* （a）
* （b）
* （c）
* （d）

1. 请评价两个分类模型和的性能。所选择的测试集包含26个二元属性，记作 。表 4.13 是模型应用到测试集时得到的后验概率（图中只显示正类的后验概率）。因为这是二分类问题，所以，。假设希望从正类中检测实例。

* 
* 
* （a）画出曲线和的ROC曲线（画在同一幅图中）。哪个模型更好？给出理由。  
  （b）对模型，假设截止阈值。即任何后验概率大于的测试实例都被看作正例。计算模型在此阈值下的精度、召回率和度量。  
  （c）对模型使用相同的截止阈值重复（b）的分析。比较两个模型的度量，哪个模型更好？所得结果和从ROC曲线中得到的结论一致吗？   
  （d）使用阈值对模型重复（b）的分析。和哪个阈值更好？该结果与从ROC曲线中得到的结果一致吗？

1. 阅读建议: 数据挖掘导论(原书第2版),[[美] 陈封能](https://book.douban.com/author/1248168) [[美]迈克尔·斯坦巴赫](https://book.douban.com/search/迈克尔·斯坦巴赫) / [[美]阿努吉·卡帕](https://book.douban.com/search/阿努吉·卡帕坦)，[机械工业出版社](https://book.douban.com/press/2793)，出版年: 2019-8-1，ISBN: 9787111631620第四章节内容