研究生算法课课堂笔记

上课日期：2016 年 9 月 26 日　　第(2)节课

组长学号及姓名：李念语 1601111282
组员学号及姓名：梁晶晶 1601111285
组员学号及姓名：娄一翎 1601111287

一、内容概要：

本节课内容主要包括以下几点：

1、第二次作业习题讲解

2、决策树

二、详细内容：

1、第二次作业习题讲解：

a) 棋盘问题

解题方法：类似于八皇后，采用递归方法

b) Sorting It All Out

解题方法：采用拓扑排序。

需要注意的几个情况：

如果所有顶点都已经存在于拓扑排序中，那么忽略后续边的关系，直接输出结果（针对后续有环的情况）。

示例输入 2 中，只读入第一条边时，输出结果是无法确定，因为一共有三个顶点。读入第二条边后，就发现了环。

c) Gone Fishing

解题方法：

可理解为：有 n 个有序数组，要求从中取出前 k 个最大数。

方法：构建最大堆，每次取根节点即为当前最大值，然后调整堆，直到取出 k 个数，时间复杂度为：O(klgn)

或者采用循环，效率为 O(kn)

现增加要求：按照数组顺序输出。

方法：记录每个数组从中取值的个数。最后依次输出。

2、决策树

a) 决策树的定义

决策树（decision tree）是一个预测模型，是对象属性与对象值之间的一种映射关系。

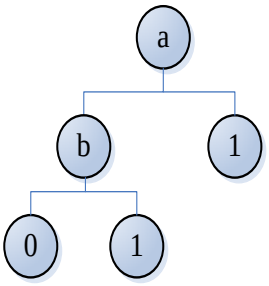
i. 每一个非叶子节点（内部节点）代表了一种属性

ii. 每一个分支代表该属性中的测试输出

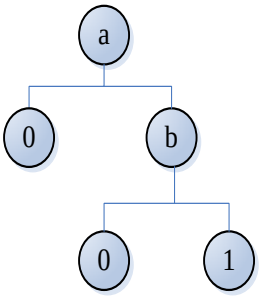
iii. 每一个叶子节点代表一种类别

b) 决策树的表达能力

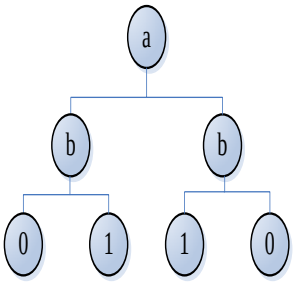
或关系 $a \vee b$



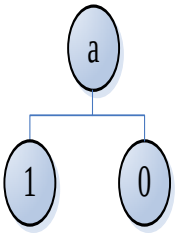
与关系 $a \wedge b$



异或关系 XOR



非关系 $\neg a$



注：上述关系中假设左分支属性值为 0，右分支属性值为 1
由上述基本关系可以表示出更复杂的表达式

e. g. : $(A \wedge B) \vee (C \wedge \neg D \wedge E)$

注：表示同一函数或者表达式的决策树并不是唯一的，对于某一表示式，给予不同的变量（属性）顺序可以产生不同的决策树。

c) 决策树适用范围：

i. 测试样本是（属性-值） 对
ii. 对象值是不连续，离散的
iii. 需要不连续的假设情形
iv. 训练样本中存在噪音的数据集

d) 根据训练数据构建决策树方法：构建决策树算法

i. 选取目前最优属性 A
ii. 将 A 作为下一节点的决策属性
iii. 对于 A 属性的不同值，分别创建不同的子节点
iv. 将训练集根据属性 A 的不同值进行分类
v. 如果子训练集分类完成，循环停止，否则对子节点从 i 进行操作

e) 如何确定最优的属性？

$$H(A) = \sum_{i=1}^{k} P_i \log \frac{1}{p_i}$$

通过 Entropy 熵来确定最优属性

Information Gain: 熵减少的量即为信息增量，即父节点的熵与子节点加权平均后熵的差值为信息增量，使得信息增量最大的属性即为当前推理中最优的属性。