机器学习课程　第4次作业

黄昊　　*20204205*

4.1　显然成立：构造这样一颗决策树：第一层判断特征向量的第一个分量，第二层判断第二个... 以此类推。由于数据各不相同，故这样构造出来1决策树，必然能分到一个叶节点，且只有一个数据符合。根据这个构造方法，每个数据到达叶节点的路径各不相同，且一定完全符合（因为各不冲突），故训练误差为0.

4.2　把训练误差作为训练准则容易出现泛化能力差的问题。

4.3

4.4

4.8　算法见下页。如果属性取值较多但属性少，BFS比DFS空间消耗更大；若属性多但属性值少，则DFS比BFS空间消耗更大，DFS有爆栈的风险。

Algorithm 1: 决策树生成算法——基于广度优先搜索

Data: 训练集D=$\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
属性集A=$\{a_1, a_2, \ldots, a_d\}$
最大高度 MaxDepth

Result: 决策树T

生成节点N，节点信息包括数据集D，属性集A，高度信息h；
记录决策树T的根为N；
生成节点队列Q；
将N压入队列Q的队尾；
while 节点队列Q非空 **do**
    从节点队列Q中取出队首节点N；
    if 节点$N.D$中样本全属于同一类别C **then**
        将N标记为C类叶节点；**continue**；
    end
    if 节点$N.h$已达到$MaxDepth$ ***OR*** $N.A = \emptyset$ ***OR*** $N.D$中样本在$N.A$上的取值相同 **then**
        将N标记为叶节点，其类别标记为N.D中样本最多的类；**continue**；
    end
    从N.A中选择最优划分属性a_*；
    for a_*的每一个值a_*^v **do**
        为N生成一个分支；令D_v为D中在a_*上取值为a_*^v的样本子集；
        if D_v为空 **then**
            将分支节点标记为叶节点，其类别表及为D中样本最多的类；continue；
        else
            生成节点N_s，节点信息包括数据集D_v，属性集$A \backslash \{a_*\}$，高度信息$N.h + 1$；
            将N_s压入节点队列Q
        end
    end
end
return 决策树T