# CIS 635 Data Mining

Homework 4

## Description

In this homework you will be going through the process of building a decision tree. The process is different for nominal vs numeric data so there will be two tasks, one for each. There are several splitting metrics but you will use Gini for this homework.

## Instructions

### Part 1 – manual calculations

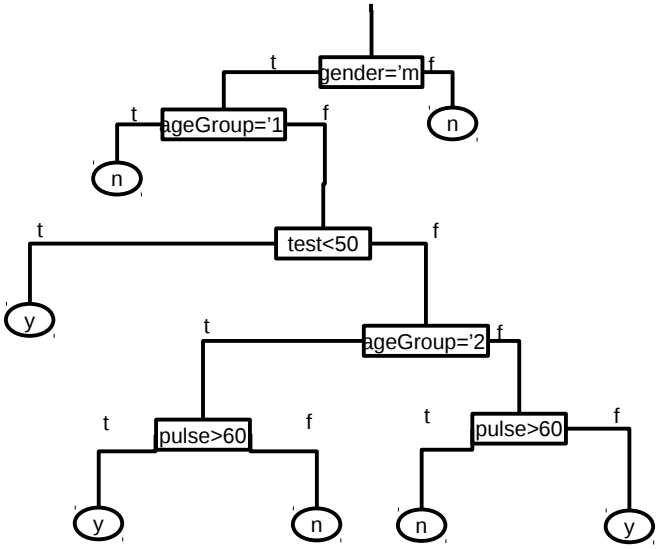
The table to the right contains information about patients in a study for a particular virus. The attributes consist of id, gender, age group, pulse/heart rate, test and the class, virus. The test is the results of a medical test that can range from 1 to 100. It is thought that gender, ageGroup, pulse and/or test may be predictive of the presence of the virus. Do the analysis to answer the questions below.

| id | gender | ageGroup | pulse | test | virus |
|---|---|---|---|---|---|
| 101 | M | 1 | 67 | 45 | Y |
| 102 | M | 3 | 68 | 58 | N |
| 103 | F | 2 | 73 | 72 | N |
| 104 | M | 2 | 77 | 67 | N |
| 105 | F | 1 | 62 | 39 | Y |
| 106 | F | 3 | 81 | 61 | N |
| 107 | M | 1 | 70 | 73 | N |
| 108 | F | 2 | 81 | 52 | Y |
| 109 | M | 3 | 105 | 66 | N |
| 110 | F | 2 | 63 | 47 | Y |

1. calculate the gini scores for splitting on
    a) gender
    b) ageGroup (2 different splits)
2. calculate the gini scores for splitting on
    a) pulse
3. what is the best split of all the attributes?
4. after that attribute is selected, describe *briefly* in your own words, what the next step is.

## Part 2 – classifying instances

In this part, we are using a different data set from the one above with more instances (thousands) but the same attributes. It is also not the same as the data set used in the next exercise below. You are not allowed to see the data but you can see the decision tree that was built from the data set to the right.




Using just the tree, do the following:

1. categorize the following instances:
   a) {2031, f, 2, 66, 50 ,?}
   b) {2032, m, 1, 90 ,81 ,?}
   c) {2033, m, 2, 75 ,45 ,?}
   d) {2034, m, 3, 90 ,55 ,?}
   e) {2035, m, 2, 78 ,67 ,?}

2. characterize the patients with the virus by *briefly* describing their characteristics.

3. if there was a middle aged male, with a test score > 50, what could he do to try to avoid the virus?

## Part 3 – building a model in R

You have been provided three data sets with many instances (same attributes as above but different data). The data sets are hw04a.txt, hw04b.txt and hw04c.txt.

Follow these directions to process the data, create models and examine the results:

1. for each data set, perform the following actions
   a) read the data from the file into a data matrix
   b) use rpart to build a decision tree model using virus as the class and all of the attributes
   c) plot the decision tree using rpart.plot
   d) identify the pure leaves and the impure leaves
   e) suggest any pruning (choose one branch to remove – answer can be none)

2. determine which data set, hw04a or hw04b, appears to be more suited to analysis by decision tree and support your answer

3. explain why the tree for data set hw04c is so simple