

CIS 635 Data Mining

Homework 3

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

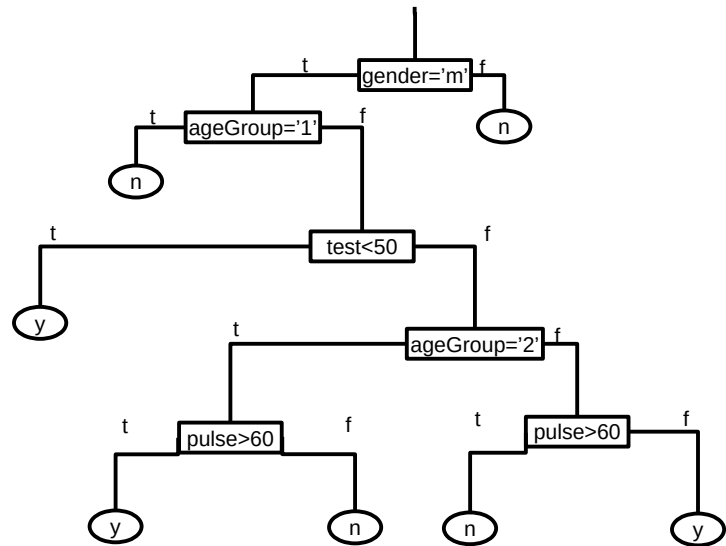
The table to the right contains information about patients in a study for a particular virus. The attributes consist of id, gender, age group, pulse/heart rate, test and the class, virus. The test is the results of a medical test that can range from 1 to 100. It is thought that gender, ageGroup, pulse 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	2	77	45	n
102	m	1	72	58	n
103	f	2	79	72	n
104	f	1	67	67	y
105	m	2	64	39	n
106	f	2	58	61	y
107	m	3	63	73	y
108	m	3	65	52	n
109	f	1	52	66	y
110	m	2	64	47	n

1. calculate the gini scores for splitting on
 - a) gender
 - b) ageGroup (split by 1 on one side and 2 & 3 on the other)
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, 1, 66, 50 ,?}
- b) {2032, m, 3, 90 ,81 ,?}
- c) {2033, m, 2, 75 ,45 ,?}

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 a data set that consists of data for the 100 US senators, as of summer 2020. The attributes consist of name, political party (democrat, republican or independent), age, eight binary attributes for their previous career (law, military, social or non-profit, etc.), the number of years in office and finally, the class which is 0/1. The class=0 means their voters don't approve of them and class=1 means that the voters do approve of them. This was taken from a website of senator rankings.

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

1. 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 approval as the class and all of the attributes except name
 - c) plot the decision tree using `rpart.plot`
2. What is one of the main determiners of voter approval?
3. Is a senator's party important in predicting their approval rating? Explain briefly.