Classification Trees

Data Mining 2015: assignment 1

Sebastiaan Jong (5546303) & Bas Geerts (5568978)

1 Introduction

This report is written for the first assignment of the Data Mining (2015) course at Utrecht University. The goal of this assignment was to write a function in the R programming language that constructs a classification tree on a certain dataset, and to figure out efficient parameters for this tree.

2 Data

For this assignment we used the Heart Disease dataset from the University of California, Irvine machine learning repository. The preprocessed version for this assignment contains 297 instances and 14 attributes. The class label for this assignment is AHD, which indicates whether the patient has been diagnosed with a heart disease, the preprocessed dataset contains 137 instances where this is the case. Used attributes:

1. Age, numeric attribute.
2. Sex, categorical attribute.
3. ChestPain, categorical attribute.
4. RestBP, numeric attribute, resting bloodpressure.
5. Chol, numeric attribute, serum cholesterol.
6. Fbs, categorical attribute, fasting bloodsugar test result.
7. RestECG, categorical attribute, resting electrocardiographic results.
8. MaxHR, numeric attribute, maximum heart rate achieved.
9. ExAng, categorical attribute, exercise induced angina.
10. Oldpeak, numeric attribute, exercise induced ST-depression.
11. Slope, categorical attribute, slope of the peak exercise ST segment.
12. Ca, numeric attribute, number of major vessels colored by flourosopy.
13. Thal, categorical attribute, thallium heart scan.

3 Experiments

The size of the classification tree is controlled by the *nmin* (minimum internal node size) and *minleaf* (minimum leaf size) parameters. Instead of brute forcing all possible values it is better to only try some sensible values and narrow down the optimal settings from there. It is possible to make a few observations regarding sensible parameter values (where n is the dataset size):

- Both $nmin$ and $minleaf$ should not be larger than $n/2$.

- The $minleaf$ parameter should not be larger than $nmin/2$.

To quickly find an approximation of good parameter values, we tried all likely values for $nmin$ between 1 and 100 with a certain ratio to $minleaf$. The results are plotted in Figure 1 below, results are obtained by performing a 10-fold cross validation 5 times for each different parameter. For example, the ratio 3:1 means that at $nmin = 45$, we tried $minleaf = 15$. The results from Figure 1 indicate that a well performing value for $nmin$ might be found between $10 - 20$, since the error rate is consistently low at these points.

Figure 1. *Parameter experiment results*

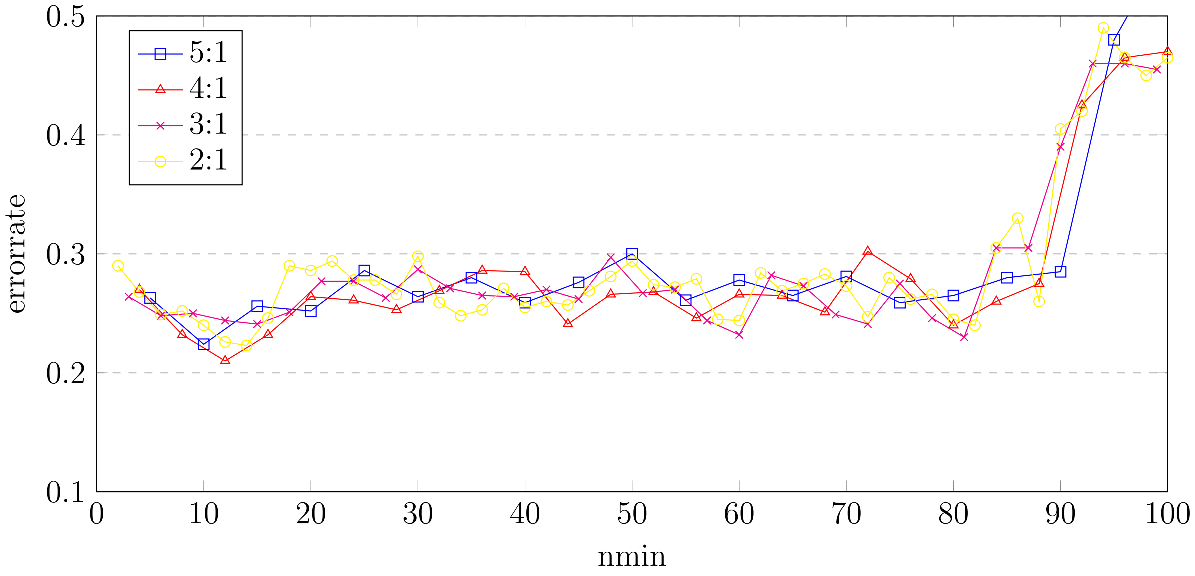
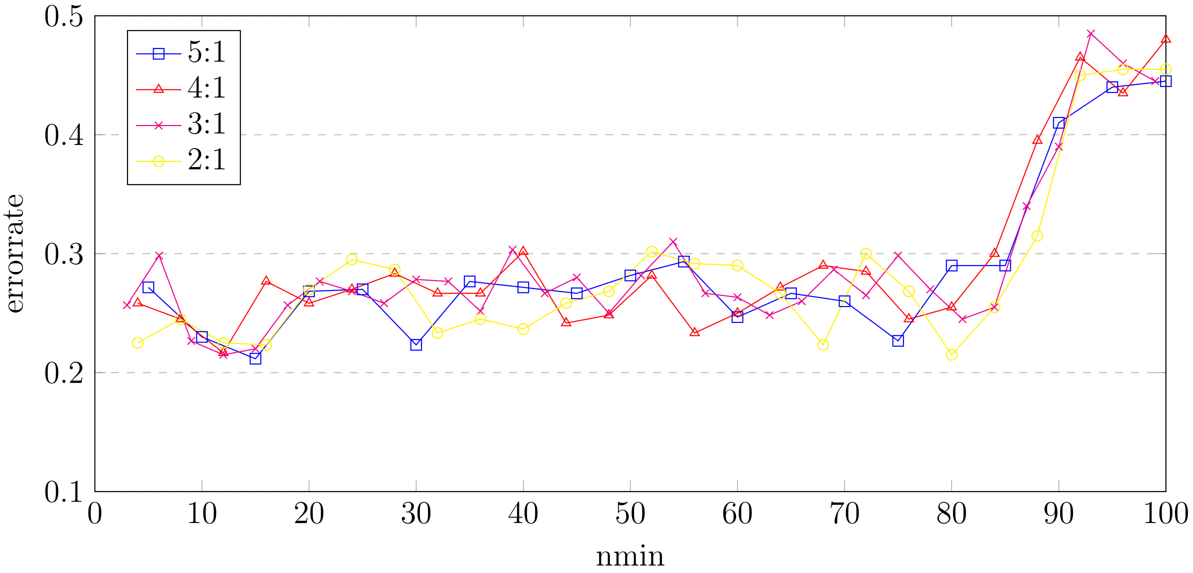
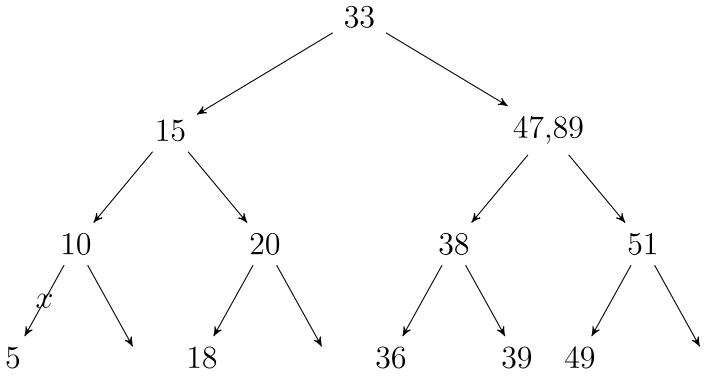


Figure 1. *Parameter experiment results*


errorrate
nmin
5:1
4:1
3:1
2:1



33
15
47,89
10
20
38
51
x
5
18
36
39
49