Data mining exercise 1

Jaap van der Plas 3998312
Linus Oleander F120180
Daniel Tell F120181

Introduction

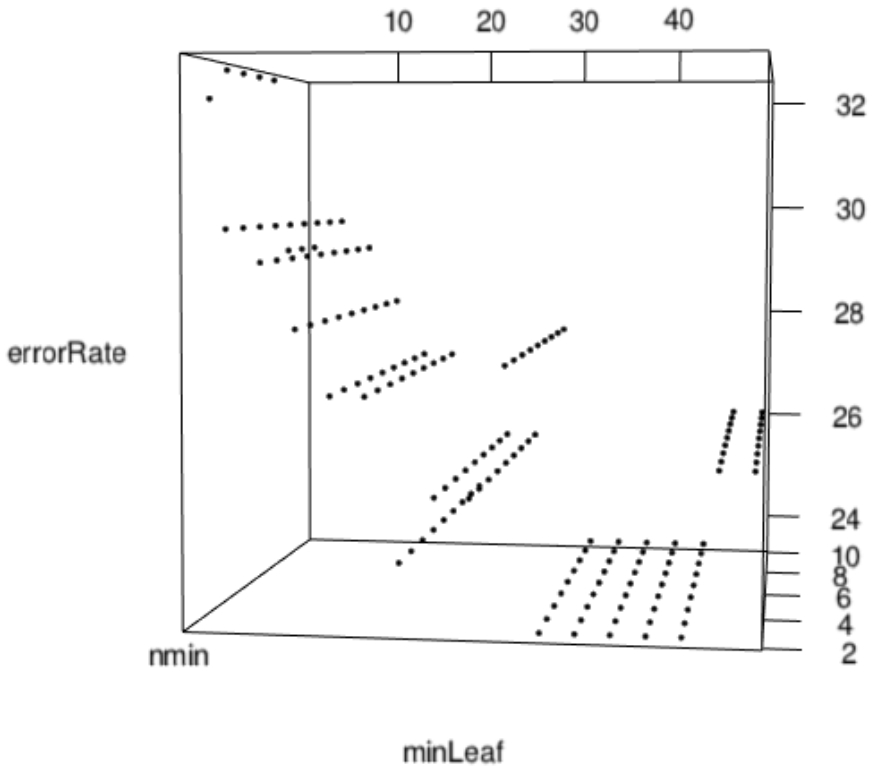
We analysed the “Pima Indians diabetes” data set from the UC Irvine Machine Learning Repository. The data set contains 768 rows. Each row consists of eight numerical attributes, and a binary class.

Parameters

We wrote a function which took one sample and tried many different values for the parameters (*tree.analyse*) in order to identify the optimal values. We then selected the best parameters and wrote a function to try it on 100 different samples (*tree.sample*). We subsequently used this function to calculate the average error rate of our tree. The parameters *nmin = 2* and *minleaf = 31* were found to give good results, resulting in an average error rate of 25%.

A tree that is too simple cannot take all important attributes into consideration when classifying data. A tree that is too complex on the other hand will suffer from overfitting, which means that it will attempt to make predictions based on random noise.

Both parameters limit the size of the tree, however the *minleaf* parameter proved to be most effective at providing a compact tree with a low error rate. Based on our simulations we conclude that given *minleaf*, error rate is independent of *nmin*. As shown in the below 3d plot of error rates, the data points end up on parallel, straight lines.



Figure

To more easily assess the quality of the tree and get a better understanding of the output generated by *tree.analyse* we used a graphical library called *rgl* for visualisation. Figure 1 is the result of *nmin* range from 2 to 10 with increment 1 and *minleaf* range from 1 to 50 with increment 3. The image displays the curve that was discussed on during the lectures. Here, a *minleaf* value equal to 28 to 42 results in the lowest error rate (x axis). We can also see that *nmin* (z axis) has a minimal impact on the quality of the tree.

Classification tree

We used 75 percent of the data set for training and the remaining 25 percent for testing. To ensure the quality and to counteract any structures in the original data, we chose our sample randomly.

Our program chose *Plasma glucose concentration* as the root attribute for our tree, hence deeming it the most important factor. Splitting at 127,5 gave the highest reduction in impurity. From a 500/268 distribution root node this split created 2 nodes, the largest having a 391/94 distribution.

Other attributes that showed up in the tree were *Body mass index*, *Age* and *Diabetes pedigree function*. The attributes *Diastolic blood pressure*, *Triceps skin fold thickness* and *2-Hour serum insulin* on the other hand, did not seem to influence the probability of developing diabetes. In total, the tree has 16 nodes and 17 leaves.

Appendix: the classification tree

Generated by the *tree.print* function.

```
Node: (500|268) { bestI: 2, bestS: 127.5 }
  Node: (391|94) { bestI: 8, bestS: 28.5 }
    Node: (248|23) { bestI: 6, bestS: 30.9 }
      Node: (149|2) { bestI: 1, bestS: 2.5 }
        Leaf: (113|0)
        Leaf: (36|2)
      Node: (99|21) { bestI: 7, bestS: 0.483 }
        Node: (62|8) { bestI: 7, bestS: 0.248 }
          Leaf: (26|7)
          Leaf: (36|1)
        Leaf: (37|13)
    Node: (143|71) { bestI: 6, bestS: 26.2 }
      Leaf: (39|2)
      Node: (104|69) { bestI: 2, bestS: 99.5 }
        Leaf: (45|10)
        Node: (59|59) { bestI: 7, bestS: 0.557 }
          Node: (50|34) { bestI: 6, bestS: 34.6 }
            Leaf: (27|26)
            Leaf: (23|8)
          Leaf: (9|25)
  Node: (109|174) { bestI: 6, bestS: 29.9 }
    Node: (52|24) { bestI: 2, bestS: 145.5 }
      Leaf: (35|6)
      Leaf: (17|18)
    Node: (57|150) { bestI: 2, bestS: 157.5 }
      Node: (45|70) { bestI: 8, bestS: 30.5 }
        Leaf: (27|23)
        Node: (18|47) { bestI: 7, bestS: 0.443 }
          Leaf: (13|18)
          Leaf: (5|29)
      Node: (12|80) { bestI: 7, bestS: 0.341 }
        Leaf: (7|25)
        Leaf: (5|55)
```