Datamining – Recursive partitioning trees

Søren Højsgaard

Department of Mathematical Sciences

Aalborg University, Denmark

August 22, 2012

Contents

1 Introduction

Data mining is an umbrella for a wide variety of techniques for exploring data.

We illustrate one particular technique: Recursive partitioning trees.

2 Example - wine data

The wine data has measurements on the chemical composition of samples of 3 different cultivars (varieties) of wine.

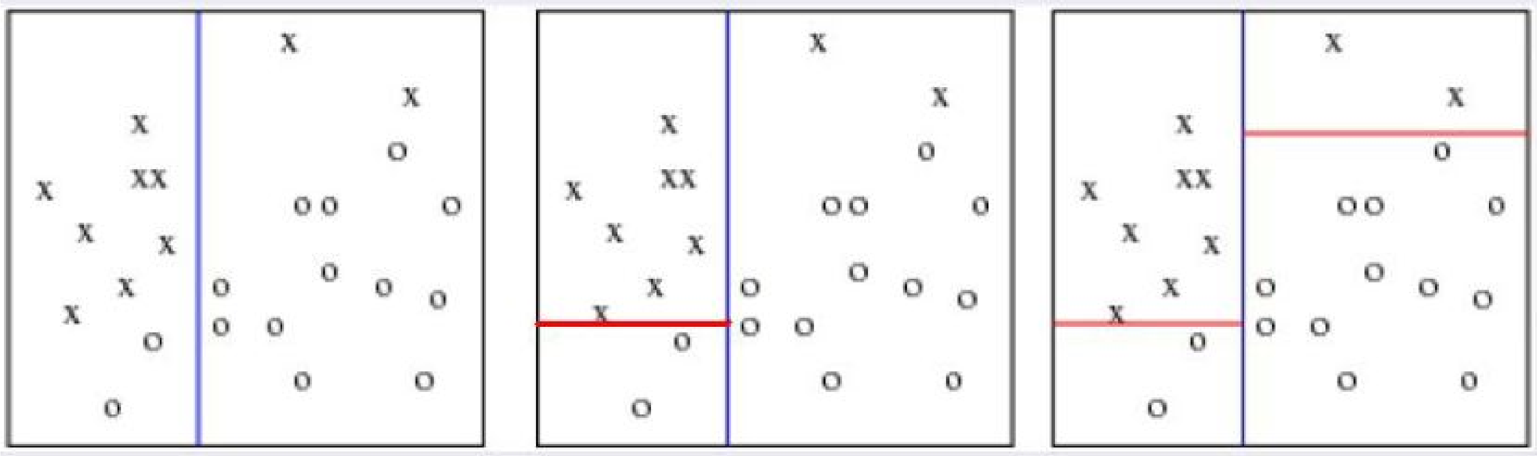
```
 data(wine, package="gRbase")
 head(wine)

  Cult  Alch Mlca  Ash Aloa Mgns Ttlp Flvn Nnfp Prnt Clri  Hue Oodw Prln
1   v1 14.23 1.71 2.43 15.6  127 2.80 3.06 0.28 2.29 5.64 1.04 3.92 1065
2   v1 13.20 1.78 2.14 11.2  100 2.65 2.76 0.26 1.28 4.38 1.05 3.40 1050
3   v1 13.16 2.36 2.67 18.6  101 2.80 3.24 0.30 2.81 5.68 1.03 3.17 1185
4   v1 14.37 1.95 2.50 16.8  113 3.85 3.49 0.24 2.18 7.80 0.86 3.45 1480
5   v1 13.24 2.59 2.87 21.0  118 2.80 2.69 0.39 1.82 4.32 1.04 2.93  735
6   v1 14.20 1.76 2.45 15.2  112 3.27 3.39 0.34 1.97 6.75 1.05 2.85 1450

 table(wine$Cult)

v1 v2 v3
59 71 48
```

Question: Can we construct a model that will be good at classifying the variety from the chemical measurements.

The general picture: We have a categorical response variable y (3 levels for the wine data) and a number of predictor variables $x_1, \ldots x_p$ (13 predictors for the wine data).

Idea:

- Split data into two subgroups according to the values of one of the predictors, say x_1.
- Split the first subgroup according to the values of one of the other predictors, say x_2.
- Split the second subgroup according to the values of one of the other predictors, say x_3 (or possibly also x_2).
- and so on...



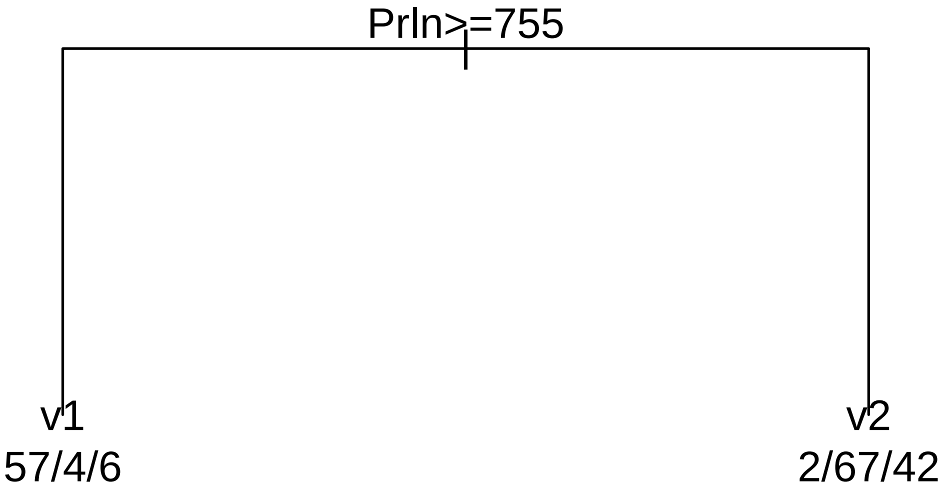
To get this to work we need

- Some rule for deciding on which variable to split
- A rule for deciding when to stop splitting

This is implemented in the `rpart()` function in the **rpart** package.

A simple usage where we allow one split only:

```
library(rpart)
f1<-rpart(Cult~., data=wine, control=rpart.control(maxdepth=1))
plot(f1, uniform=T,margin=0.2)
text(f1, use.n=TRUE)
```

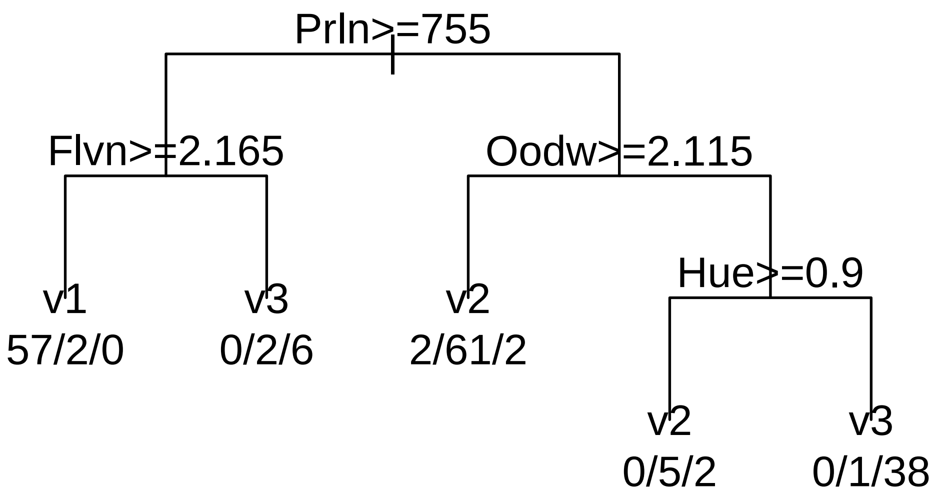


Read this as:

- Split on whether `Prln` ≥ 755. “Yes” is to the left, “no” to the right.
- $57 + 4 + 6 = 67$ cases appear on the leaf to the left. These cases are all given the label `v1`;
- 57 cases have variety `v1`, 4 are of variety `v2` and 6 are of variety `v3`.

Alternatively, we can leave it to data to suggest the number of splits

```
f2<-rpart(Cult~., data=wine)
plot(f2, uniform=T,margin=0.2)
text(f2, use.n=TRUE)
```



Having done so, at natural question is to ask how good our classification is:

```
table(wine$Cult, predict(f1, type="class"))

    v1 v2 v3
 v1 57  2  0
 v2  4 67  0
 v3  6 42  0

table(wine$Cult, predict(f2, type="class"))

    v1 v2 v3
 v1 57  2  0
 v2  2 66  3
 v3  0  4 44
```