

1 Classification and Regression Trees

The CART algorithm generates a classification tree by sequentially doing binary splits on the data. The simplest case is when splits are made on individual variables rather than combinations of variables.

For example, in the olive oils data, to separate the 3 regions CART would first split on eicosenoic acid, then linoleic acid. Amazingly it only uses these two variables:

```
If eicosenoic acid  $\geq$  7 then the region is South (1)
Else
  If linoleic  $\geq$  1048.5 then the region is Sardinia (2)
  Else the region is North (3)
```

The missclassification rate, including training and test data is $1/572 = 0.00175$. (Figure 1 shows the tree and a scatterplot illustrating the split.)

```
> olive.area<-factor(d.olive.train[,1])
> olived<-data.frame(d.olive.train[,-c(1,2)])
> olive.tree<-tree(olive.area~.,olived)
> summary(olive.tree)

Classification tree:
tree(formula = olive.area ~ ., data = olived)
Variables actually used in tree construction:
[1] "eicosenoic" "linoleic"
Number of terminal nodes: 3
Residual mean deviance: 0 = 0 / 433
Misclassification error rate: 0 = 0 / 436

> olive.tree

node), split, n, deviance, yval, (yprob)
* denotes terminal node

1) root 436 851.2 1 ( 0.5642 0.1697 0.2661 )
 2) eicosenoic<7 190 254.0 3 ( 0.0000 0.3895 0.6105 )
   4) linoleic<1048.5 116 0.0 3 ( 0.0000 0.0000 1.0000 ) *
   5) linoleic>1048.5 74 0.0 2 ( 0.0000 1.0000 0.0000 ) *
 3) eicosenoic>7 246 0.0 1 ( 1.0000 0.0000 0.0000 ) *

> par(mfrow=c(1,2), pty="s")
> plot(olive.tree)
> text(olive.tree, all=T)
> plot(d.olive.train[,10],d.olive.train[,7],xlab="Eicosenoic",ylab="Linoleic",type="n")
> text(d.olive.train[,10],d.olive.train[,7],d.olive.train[,1])
> text(d.olive.test[,10],d.olive.test[,7],d.olive.train[,1])
> abline(v=7)
> lines(c(0,7),c(1048.5,1048.5))
> table(d.olive.train[,1],predict(olive.tree,olived,type="class"))
 1  2  3
1 246  0  0
2  0 74  0
3  0  0 116
> olivet<-data.frame(d.olive.test[,-c(1,2)])
> table(d.olive.test[,1],predict(olive.tree,olivet,type="class"))
 1  2  3
1 77  0  0
2  0 24  0
3  0  1 34
```

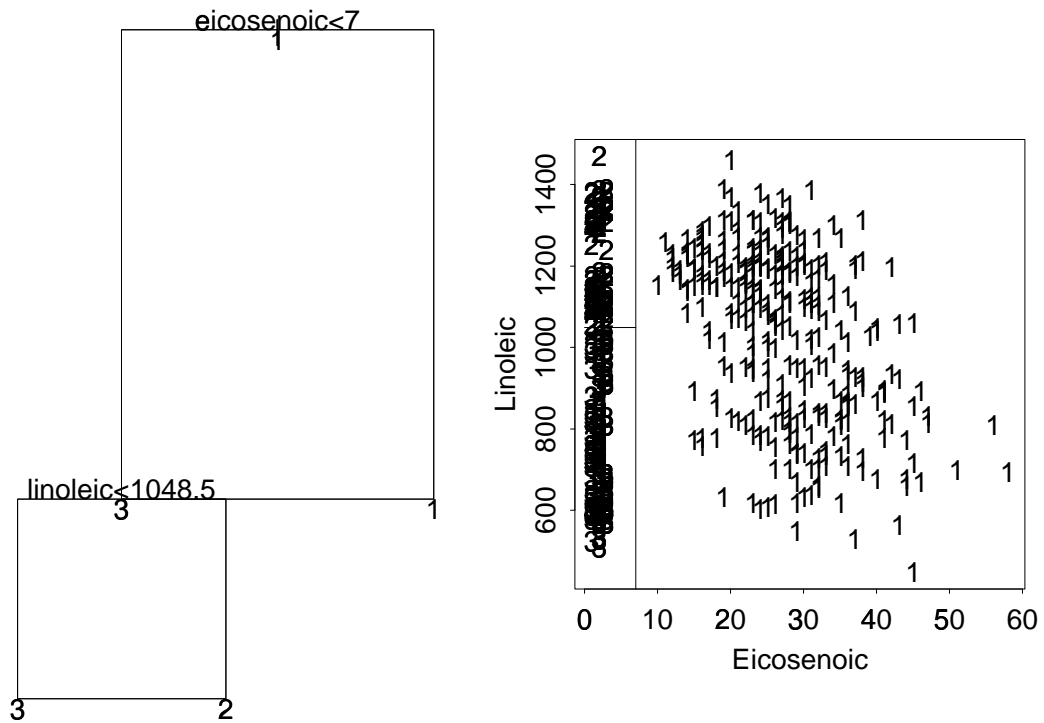


Figure 1: CART on Regions: (left) tree view, (right) data view of boundaries.

How the splits are decided

Consider that each case in the training sample is assigned to a leaf, so at each leaf we have a random sample size n_{ik} from a multinomial distribution specified by p_{ik} . The conditional likelihood (given the training sample) is proportional to

$$\prod_{i=1}^{\#leaves} \prod_{j=1}^{\#classes} p_{ik}^{n_{ik}}$$

Then a deviance measure (how impure/matches true classification) is given by (take negative twice the log likelihood)

$$D = -2 \sum_{i=1}^{\#leaves} \sum_{j=1}^{\#classes} n_{ik} \log p_{ik}$$

Estimate p_{ik} by $\hat{p}_{ik} = n_{ik}/n_i$ (maximum likelihood estimator). The splitting strategy is to choose the attribute which minimizes this deviance.

(Algorithm in S-Plus is due to Clark and Pregibon, 1992.)

In the olive oil data the initial stage has all observations in one group, then

$$n_{1,1} = 246, n_{1,2} = 74, n_{1,3} = 116, n_1 = 436$$

and

$$\hat{p}_{1,1} = 246/436 = 0.5642, \hat{p}_{1,2} = 74/436 = 0.1697, \hat{p}_{1,3} = 116/436 = 0.2661$$

and $D = 851.2$.

At the first split,

$$\begin{aligned} n_{1,1} &= 246, n_{1,2} = 0, n_{1,3} = 0, n_1 = 246, \\ n_{2,1} &= 0, n_{2,2} = 74, n_{2,3} = 116, n_2 = 190 \end{aligned}$$

and

$$\begin{aligned} \hat{p}_{1,1} &= 246/246 = 1, \hat{p}_{1,2} = 0/246 = 0, \hat{p}_{1,3} = 0/246 = 0, \\ \hat{p}_{2,1} &= 0/190 = 0, \hat{p}_{2,2} = 74/190 = 0.3895, \hat{p}_{2,3} = 116/190 = 0.6105 \end{aligned}$$

and $D = 254.0$ (taking $0 \log(0) = 0$).

At the second split (in the group where *linoleic* > 1048.5),

$$\begin{aligned} n_{1,1} &= 246, n_{1,2} = 0, n_{1,3} = 0, n_1 = 246, \\ n_{2,1} &= 0, n_{2,2} = 74, n_{2,3} = 0, n_2 = 74 \\ n_{3,1} &= 0, n_{3,2} = 0, n_{3,3} = 116, n_3 = 116 \end{aligned}$$

$$\begin{aligned}\hat{p}_{1,1} &= 1, \hat{p}_{1,2} = 0, \hat{p}_{1,3} = 0, \\ \hat{p}_{2,1} &= 0, \hat{p}_{2,2} = 1, \hat{p}_{2,3} = 0, \\ \hat{p}_{3,1} &= 0, \hat{p}_{3,2} = 0, \hat{p}_{3,3} = 1\end{aligned}$$

so $D = 0$ (taking $0 \log(0) = 0$).

Southern Oils

```
> olive.area<-factor(d.olive.train[d.olive.train[,1]==1.2])
> olived<-data.frame(d.olive.train[d.olive.train[,1]==1,-c(1,2)])
> olive.tree<-tree(olive.area~.,olived)
> summary(olive.tree)

Classification tree:
tree(formula = olive.area ~ ., data = olived)
Variables actually used in tree construction:
[1] "linoleic" "palmitoleic" "stearic" "palmitic" "linolenic"
[6] "arachidic" "oleic"
Number of terminal nodes: 13
Residual mean deviance: 0.2131 = 49.66 / 233
Misclassification error rate: 0.04472 = 11 / 246

> olive.tree
node), split, n, deviance, yval, (yprob)
* denotes terminal node

1) root 246 505.000 3 ( 0.077240 0.17070 0.64230 0.10980 )
 2) linoleic<949 89 215.200 2 ( 0.202200 0.47190 0.06742 0.25840 )
 4) palmitoleic<95.5 36 57.630 1 ( 0.500000 0.02778 0.00000 0.47220 )
 8) stearic<262 19 12.790 1 ( 0.894700 0.00000 0.00000 0.10530 )
 16) stearic<246.5 14 0.000 1 ( 1.000000 0.00000 0.00000 0.00000 ) *
 17) stearic>246.5 5 6.730 1 ( 0.600000 0.00000 0.00000 0.40000 ) *
 9) stearic>262 17 15.090 4 ( 0.058820 0.05882 0.00000 0.88240 )
 18) palmitic<980 5 5.004 4 ( 0.200000 0.00000 0.00000 0.80000 ) *
 19) palmitic>980 12 6.884 4 ( 0.000000 0.08333 0.00000 0.91670 ) *
 38) stearic<285 5 5.004 4 ( 0.000000 0.20000 0.00000 0.80000 ) *
 39) stearic>285 7 0.000 4 ( 0.000000 0.00000 0.00000 1.00000 ) *
 5) palmitoleic>95.5 53 73.340 2 ( 0.000000 0.77360 0.11320 0.11320 )
 10) linolenic<41.5 21 45.320 2 ( 0.000000 0.42860 0.28570 0.28570 )
 20) arachidic<68.5 12 15.280 2 ( 0.000000 0.66670 0.33330 0.00000 )
 40) palmitoleic<152.5 7 0.000 2 ( 0.000000 1.00000 0.00000 0.00000 ) *
 41) palmitoleic>152.5 5 5.004 3 ( 0.000000 0.20000 0.80000 0.00000 ) *
 21) arachidic>68.5 9 15.280 4 ( 0.000000 0.11110 0.22220 0.66670 ) *
 11) linolenic>41.5 32 0.000 2 ( 0.000000 1.00000 0.00000 0.00000 ) *
 3) linoleic>949 157 49.310 3 ( 0.006369 0.00000 0.96820 0.02548 )
 6) arachidic<76.5 146 11.960 3 ( 0.006849 0.00000 0.99320 0.00000 )
 12) palmitoleic<138.5 5 5.004 3 ( 0.200000 0.00000 0.80000 0.00000 ) *
 13) palmitoleic>138.5 141 0.000 3 ( 0.000000 0.00000 1.00000 0.00000 ) *
 7) arachidic>76.5 11 14.420 3 ( 0.000000 0.00000 0.63640 0.36360 )
 14) oleic<6770.5 5 0.000 3 ( 0.000000 0.00000 1.00000 0.00000 ) *
 15) oleic>6770.5 6 7.638 4 ( 0.000000 0.00000 0.33330 0.66670 ) *

>
> par(mfrow=c(1,2), pty="c")
> plot(olive.tree)
> text(olive.tree, all=T)
> par(pty="s")
> plot(d.olive.train[d.olive.train[,1]==1.7],d.olive.train[d.olive.train[,1]==1.4],
      ylab="Palmitoleic",xlab="Linoleic",type="n")
> text(d.olive.train[d.olive.train[,1]==1.7],d.olive.train[d.olive.train[,1]==1.4],
      d.olive.train[d.olive.train[,1]==1.2])
> text(d.olive.test[d.olive.test[,1]==1.7],d.olive.test[d.olive.test[,1]==1.4],
      d.olive.test[d.olive.test[,1]==1.2])
> abline(v=949)
```

```

> lines(c(350,949),c(95.5,95.5))
> table(d.olive.train[d.olive.train[,1]==1,2].predict(olive.tree,olived,type="class"))
  1  2  3  4
1 17  0  1  1
2  0 39  1  2
3  0  0 154  4
4  2  0  0 25
> olivet<-data.frame(d.olive.test[d.olive.test[,1]==1,-c(1,2)])
> table(d.olive.test[d.olive.test[,1]==1,2].predict(olive.tree,olivet,type="class"))
  1  2  3  4
1  3  1  0  2
2  0  9  2  3
3  0  1 46  1
4  0  1  1  7

```

There are 11 errors in the training sample, and 12 errors in the test sample, giving an error rate of $23/572 = 0.040$. The variables used were palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, all variables except eicosenoic! (Figure 2 shows the tree and scatterplot for the first 2 splits.)

Sardinia

```

> olive.area<-factor(d.olive.train[d.olive.train[,1]==2,2])
> olived<-data.frame(d.olive.train[d.olive.train[,1]==2,-c(1,2)])
> olive.tree<-tree(olive.area~,olived)
> summary(olive.tree)
Classification tree:
tree(formula = olive.area ~ ., data = olived)
Variables actually used in tree construction:
[1] "linoleic"
Number of terminal nodes:  2
Residual mean deviance:  0 = 0 / 72
Misclassification error rate: 0 = 0 / 74

> olive.tree
node), split, n, deviance, yval, (yprob)
* denotes terminal node

1) root 74 94.66 5 ( 0.6622 0.3378 )
 2) linoleic<1246.5 49  0.00 5 ( 1.0000 0.0000 ) *
 3) linoleic>1246.5 25  0.00 6 ( 0.0000 1.0000 ) *

> table(d.olive.train[d.olive.train[,1]==2,2].predict(olive.tree,olived,type="class"))
  1  2
5 49  0
6  0 25
> olivet<-data.frame(d.olive.test[d.olive.test[,1]==2,-c(1,2)])
> table(d.olive.test[d.olive.test[,1]==2,2].predict(olive.tree,olivet,type="class"))
  1  2
5 16  0
6  0  8

```

There are no errors in classification here for training and test samples, and only one variable is used: linoleic acid. However the separation between the two groups is very small. (Figure 3 has the tree and a histogram illustrating the split.)

North

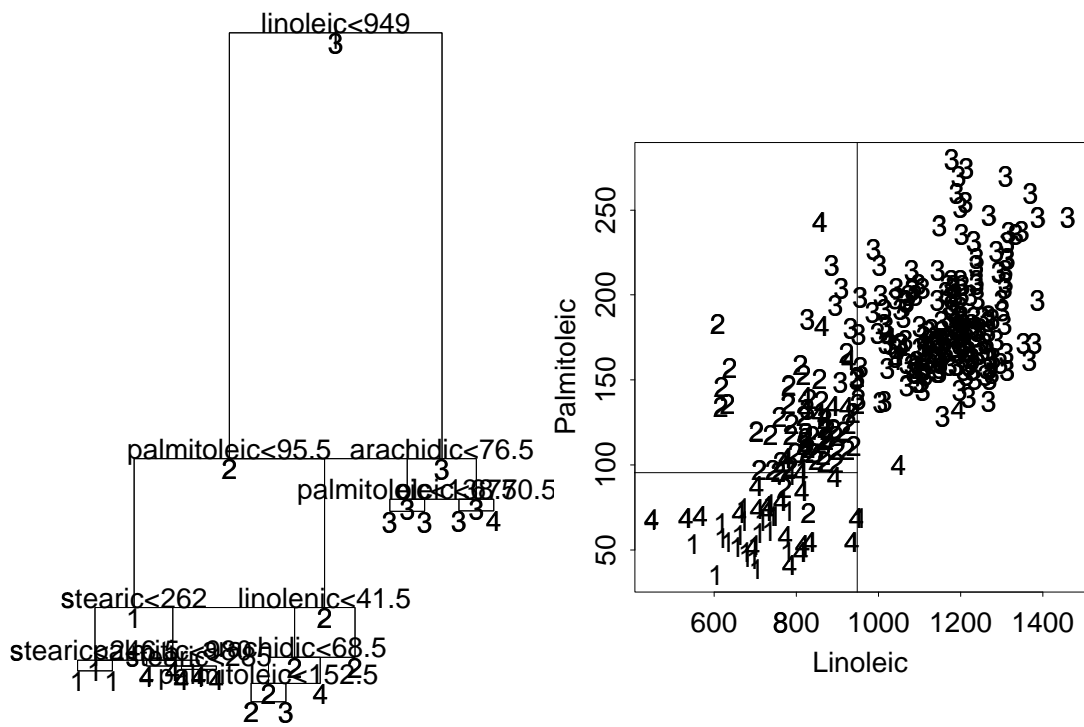


Figure 2: CART on South: (left) tree view, (right) data view of first split boundary.

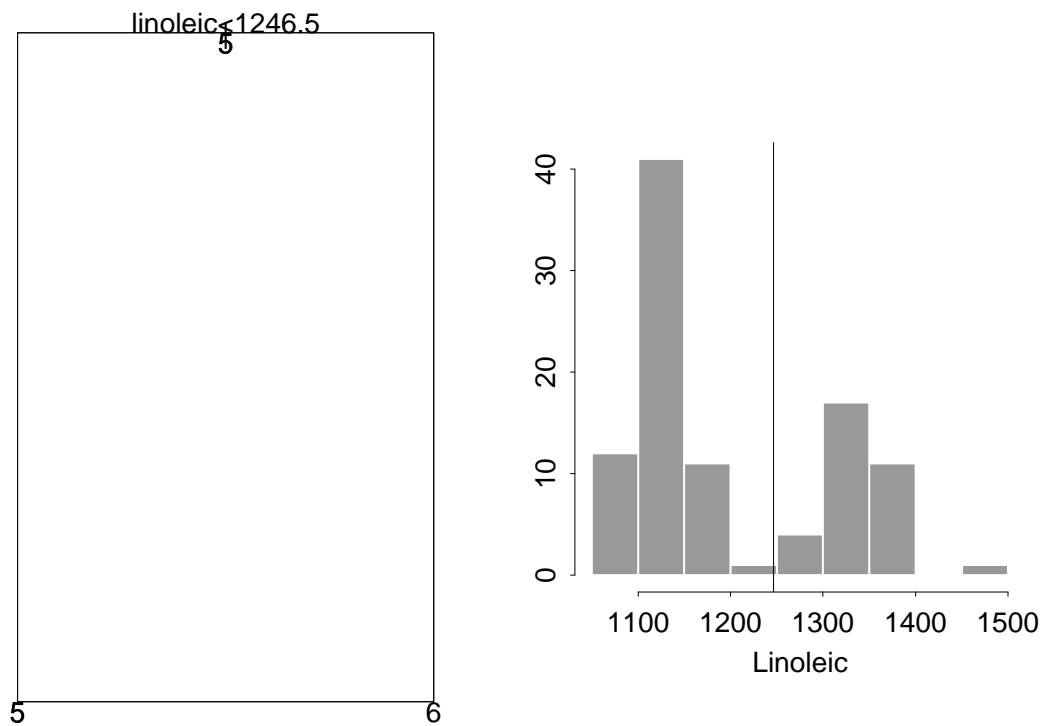


Figure 3: CART on Sardinia: (left) tree view, (right) data view of boundary.

```

> olive.area<-factor(d.olive.train[d.olive.train[,1]==3.2])
> olived<-data.frame(d.olive.train[d.olive.train[,1]==3.-c(1,2)])
> olive.tree<-tree(olive.area~.,olived)
> summary(olive.tree)
Classification tree:
tree(formula = olive.area ~ ., data = olived)
Variables actually used in tree construction:
[1] "linoleic" "oleic" "arachidic"
Number of terminal nodes: 6
Residual mean deviance: 0.1365 = 15.01 / 110
Misclassification error rate: 0.02586 = 3 / 116

> olive.tree
node), split, n, deviance, yval, (yprob)
* denotes terminal node

1) root 116 254.800 9 ( 0.32760 0.32760 0.3448 )
 2) linoleic<628 44 26.810 9 ( 0.09091 0.00000 0.9091 )
   4) oleic<7905 5 5.004 7 ( 0.80000 0.00000 0.2000 ) *

```

```

5) oleic>7905 39 0.000 9 ( 0.00000 0.00000 1.0000 ) *
3) linoleic>628 72 99.590 8 ( 0.47220 0.52780 0.0000 )
6) arachidic<15 33 0.000 8 ( 0.00000 1.00000 0.0000 ) *
7) arachidic>15 39 29.870 7 ( 0.87180 0.12820 0.0000 )
14) linoleic<810 34 9.023 7 ( 0.97060 0.02941 0.0000 )
28) oleic<7845 29 0.000 7 ( 1.00000 0.00000 0.0000 ) *
29) oleic>7845 5 5.004 7 ( 0.80000 0.20000 0.0000 ) *
15) linoleic>810 5 5.004 8 ( 0.20000 0.80000 0.0000 ) *

> par(mfrow=c(1,2), pty="c")
> plot(olive.tree)
> text(olive.tree, all=T)
> par(pty="s")
> plot(d.olive.train[d.olive.train[,1]==3.6],d.olive.train[d.olive.train[,1]==3.7],
xlab="Oleic",ylab="Linoleic",type="n")
> text(d.olive.train[d.olive.train[,1]==3.6],d.olive.train[d.olive.train[,1]==3.7],
d.olive.train[d.olive.train[,1]==3.2])
> text(d.olive.test[d.olive.test[,1]==3.6],d.olive.test[d.olive.test[,1]==3.7],
d.olive.test[d.olive.test[,1]==3.2])
> abline(h=628)
> lines(c(7905,7905),c(498,628))
> table(d.olive.train[d.olive.train[,1]==3.2],predict(olive.tree,olived,type="class"))
 1 2 3
7 37 1 0
8 1 37 0
9 1 0 39
> olivet<-data.frame(d.olive.test[d.olive.test[,1]==3,-c(1,2)])
> table(d.olive.test[d.olive.test[,1]==3.2],predict(olive.tree,olivet,type="class"))
 1 2 3
7 11 0 1
8 0 12 0
9 2 0 9

```

There are 3 errors in the training sample and 3 errors in the test sample. Hence the error rate is $6/151 = 0.0397$. (Figure 4 has the tree and scatterplot for the first 2 splits.)

Summary

Overall the missclassification rate is $30/572 = 0.0524$.

Resources

Ripley (1999) *Pattern Recognition and Neural Networks and Venables and Ripley Modern Applied Statistics with S-Plus (3rd ed)*

Ripley's web site is www.stats.ox.ac.uk, which has the tree code used here, and also `rpart` code written by Terry Therneau and Elizabeth Atkinson at the Mayo Clinic.

Quinlan,J.R.: C4.5: Programs for Machine Learning Morgan Kauffman, 1993 (A tutorial on using C4.5 is at <http://www.cs.uregina.ca/dbd/cs831/notes/ml/dtrees/>)

Original reference is: Breiman,Friedman,Olshen,Stone: Classification and Decision Trees Wadsworth, 1984

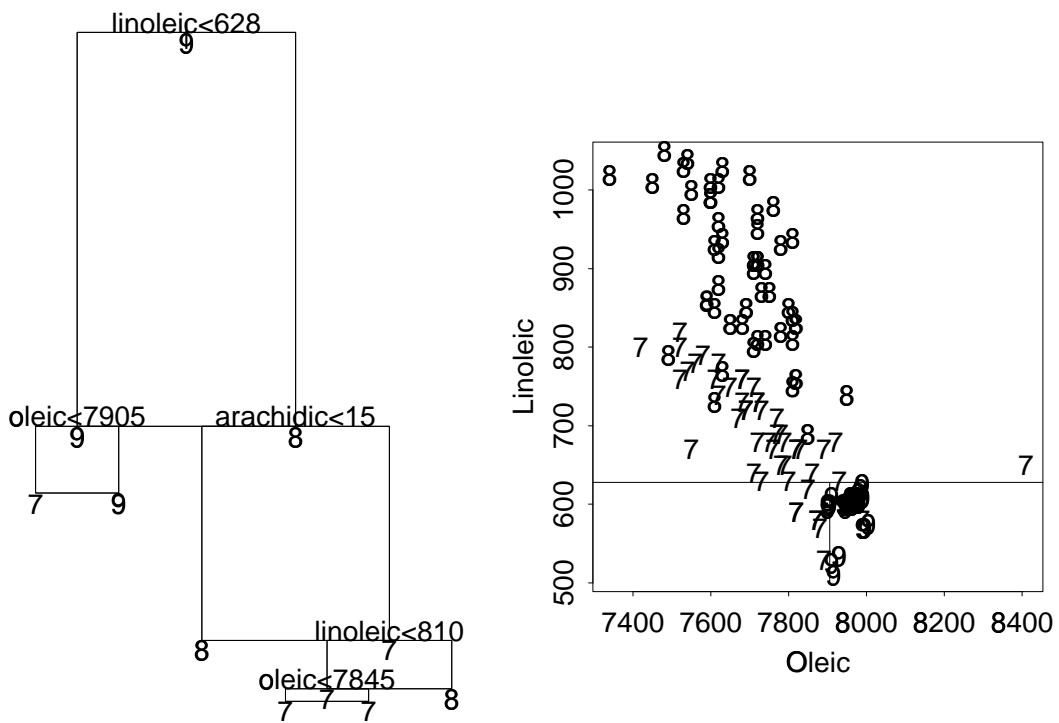


Figure 4: CART on North: (left) tree view, (right) data view of boundary.