Titre

Sébastien Guyader

2018 November 28

Iris data

The iris data set gives data on the dimensions of sepals and petals measured on 50 samples of three different species of iris (setosa, versicolor and virginica).

```
library(partykit)
```

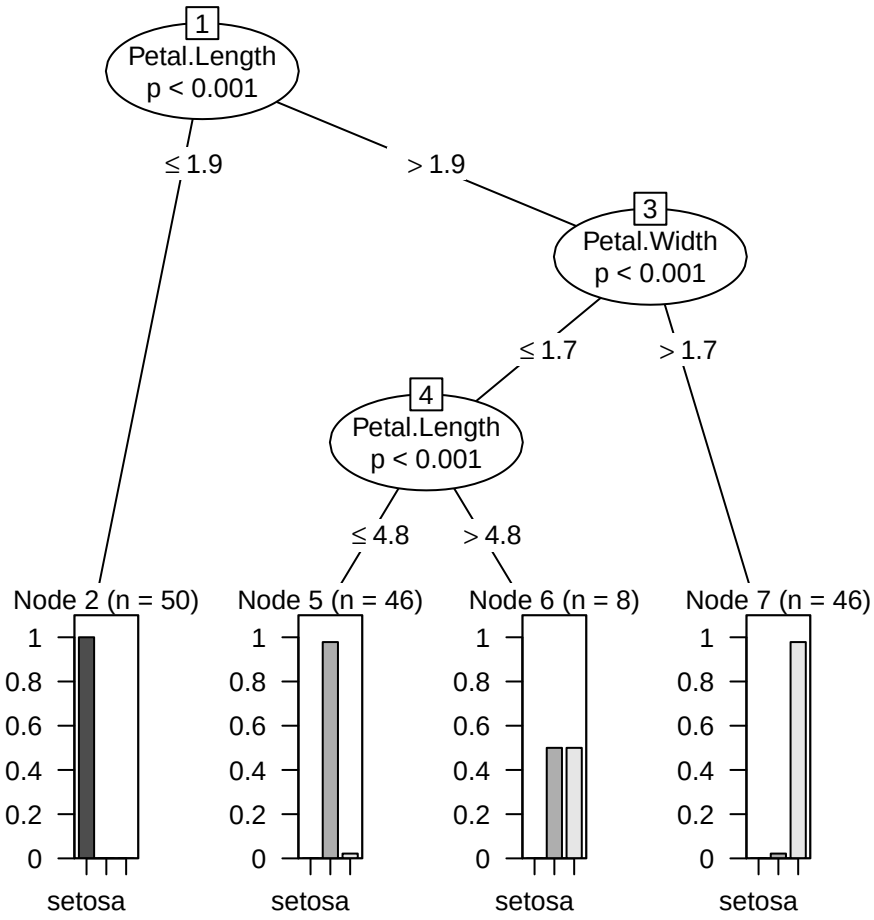
```
## Loading required package: grid
```

```
## Loading required package: libcoin
```

```
## Loading required package: mvtnorm
```

We will construct a model of iris species as a function of the other covariates.

```
iris.ct <- ctree(Species ~ .,data = iris)
plot(iris.ct)
```



Get the node p-values.

```
nodeapply(iris.ct, ids = nodeids(iris.ct),
  FUN = function(n) info_node(n)$p.value
  )
```

```
## $`1`
## Petal.Length
## 1.393271e-30
##
## $`2`
## NULL
##
## $`3`
##  Petal.Width
## 6.900972e-16
##
## $`4`
## Petal.Length
## 0.0007854878
##
```

```
## $`5`
## Petal.Width
##   0.1115842
##
## $`6`
## NULL
##
## $`7`
## Petal.Length
##    0.4961648
```

The structure of the tree is essentially the same. Only the representation of the nodes differs because, whereas ozone was a continuous numerical variable, iris species is a categorical variable. The nodes are thus represented as bar plots. Node 2 is predominantly setosa, node 5 is mostly versicolor and node 7 is almost all viriginica. Node 6 is half versicolor and half virginica and corresponds to a category with long, narrow petals. It is interesting to note that the model depends only on the dimensions of the petals and not on those of the sepals.

We can assess the quality of the model by constructing a confusion matrix. This shows that the model performs perfectly for setosa irises. For versicolor it also performs very well, only classifying one sample incorrectly as a virginica. For virginica it fails to correctly classify 5 samples. The model seems to perform well overall, however, this is based on the training data, so it is not really an objective assessment!

```
table(iris$Species, predict(iris.ct), dnn = c("Actual species",
                                              "Predicted species"))
```

```
##                Predicted species
## Actual species setosa versicolor virginica
##     setosa         50          0         0
##     versicolor      0         49         1
##     virginica       0          5        45
```

Finally, we can use the model to predict the species for new data (no need to specify sepal length and width as they ar not use by the model).

```
new.iris <- data.frame(Sepal.Length=rep(0,5), Sepal.Width=rep(0,5),
Petal.Length=c(1,4,5,4,5), Petal.Width=c(1,2,1,1,2))
predict(iris.ct, newdata = new.iris)
```

```
##          1          2          3          4          5
##     setosa  virginica versicolor versicolor  virginica
## Levels: setosa versicolor virginica
```

```
predict(iris.ct, newdata = new.iris, type="node")
```

```
## 1 2 3 4 5
## 2 7 6 5 7
```

Air Quality data

Load Air quality dataset:

```
airq <- subset(airquality, !is.na(Ozone))
airct <- ctree(Ozone ~ ., data = airq)
print(airct)
```

```
##
## Model formula:
## Ozone ~ Solar.R + Wind + Temp + Month + Day
##
## Fitted party:
## [1] root
## |   [2] Temp <= 82
## |   |   [3] Wind <= 6.9: 55.600 (n = 10, err = 21946.4)
## |   |   [4] Wind > 6.9
## |   |   |   [5] Temp <= 77: 18.479 (n = 48, err = 3956.0)
## |   |   |   [6] Temp > 77: 31.143 (n = 21, err = 4620.6)
## |   [7] Temp > 82
## |   |   [8] Wind <= 10.3: 81.633 (n = 30, err = 15119.0)
## |   |   [9] Wind > 10.3: 48.714 (n = 7, err = 1183.4)
##
## Number of inner nodes:    4
## Number of terminal nodes: 5
```

Summarize the data (TODO : use dplyr::summarise)

```
tapply(airq$Ozone, predict(airct, type = "node"), function(y)
  c("n" = length(y), "Avg." = mean(y),
  "Variance" = var(y), "SSE" = sum((y - mean(y))^2))
  )
```

```
## $`3`
##         n      Avg.  Variance       SSE
##    10.000    55.600  2438.489 21946.400
##
## $`5`
##          n       Avg.   Variance        SSE
##   48.00000   18.47917   84.16977 3955.97917
##
## $`6`
##          n       Avg.   Variance        SSE
##   21.00000   31.14286  231.02857 4620.57143
##
## $`8`
##           n        Avg.    Variance         SSE
```

```
##    30.00000    81.63333   521.34368 15118.96667
##
## $`9`
##          n       Avg.   Variance        SSE
##    7.00000   48.71429  197.23810 1183.42857
```

Test the significance of changes:

```
library("strucchange")
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
sctest(airct)
```

```
## $`1`
##               Solar.R         Wind         Temp     Month         Day
## statistic 13.34761286 4.161370e+01 5.608632e+01 3.1126596 0.02011554
## p.value    0.00129309 5.560572e-10 3.467894e-13 0.3325881 0.99998175
##
## $`2`
##             Solar.R         Wind         Temp     Month      Day
## statistic 5.4095322 12.968549828 11.298951405 0.2148961 2.970294
## p.value   0.0962041  0.001582833  0.003871534 0.9941976 0.357956
##
## $`3`
## NULL
##
## $`4`
##               Solar.R     Wind         Temp      Month       Day
## statistic 9.547191843 2.307676 11.598966936 0.06604893 0.2513143
## p.value   0.009972755 0.497949  0.003295072 0.99965679 0.9916670
##
## $`5`
##              Solar.R      Wind      Temp     Month       Day
## statistic 6.14094026 1.3865355 1.9986304 0.8268341 1.3580462
## p.value   0.06432172 0.7447599 0.5753799 0.8952749 0.7528481
##
## $`6`
##             Solar.R       Wind      Temp    Month       Day
```

```
## statistic 5.1824354 0.02060939 0.9270013 0.165171 4.6220522
## p.value   0.1089932 0.99998062 0.8705785 0.996871 0.1481643
##
## $`7`
##             Solar.R         Wind       Temp     Month        Day
## statistic 0.8083249 11.711564549 6.77148538 0.1307643 0.03992875
## p.value   0.8996614  0.003101788 0.04546281 0.9982052 0.99990034
##
## $`8`
##             Solar.R      Wind      Temp       Month         Day
## statistic 0.9056479 3.1585094 2.9285252 0.008106707 0.008686293
## p.value   0.8759687 0.3247585 0.3657072 0.999998099 0.999997742
##
## $`9`
## NULL
```

```
nodeapply(airct, ids = nodeids(airct), FUN = function(n) info_node(n))
```

```
## $`1`
## $`1`$criterion
##                Solar.R          Wind          Temp      Month          Day
## statistic 13.347612859  4.161370e+01  5.608632e+01  3.1126596   0.02011554
## p.value    0.001293090  5.560572e-10  3.467894e-13  0.3325881   0.99998175
## criterion -0.001293926 -5.560572e-10 -3.467894e-13 -0.4043478 -10.91135399
##
## $`1`$p.value
##         Temp
## 3.467894e-13
##
## $`1`$unweighted
## [1] TRUE
##
## $`1`$nobs
## [1] 116
##
##
## $`2`
## $`2`$criterion
##               Solar.R         Wind         Temp      Month        Day
## statistic  5.4095322 12.968549828 11.298951405  0.2148961  2.9702941
## p.value    0.0962041  0.001582833  0.003871534  0.9941976  0.3579560
## criterion -0.1011517 -0.001584087 -0.003879048 -5.1494901 -0.4430985
##
## $`2`$p.value
##        Wind
```

```
## 0.001582833
##
## $`2`$unweighted
## [1] TRUE
##
## $`2`$nobs
## [1] 79
##
##
## $`3`
## NULL
##
## $`4`
## $`4`$criterion
##                Solar.R       Wind         Temp       Month         Day
## statistic  9.547191843  2.3076758 11.598966936  0.06604893  0.2513143
## p.value    0.009972755  0.4979490  0.003295072  0.99965679  0.9916670
## criterion -0.010022817 -0.6890536 -0.003300512 -7.97715435 -4.7875322
##
## $`4`$p.value
##        Temp
## 0.003295072
##
## $`4`$unweighted
## [1] TRUE
##
## $`4`$nobs
## [1] 69
##
##
## $`5`
## $`5`$criterion
##                Solar.R       Wind       Temp      Month         Day
## statistic  6.14094026  1.3865355  1.9986304  0.8268341  1.3580462
## p.value    0.06432172  0.7447599  0.5753799  0.8952749  0.7528481
## criterion -0.06648358 -1.3655507 -0.8565604 -2.2564164 -1.3977520
##
## $`5`$p.value
##    Solar.R
## 0.06432172
##
## $`5`$unweighted
## [1] TRUE
##
## $`5`$nobs
```

```
## [1] 48
##
##
## $`6`
## $`6`$criterion
##              Solar.R         Wind       Temp     Month        Day
## statistic  5.1824354   0.02060939  0.9270013  0.165171  4.6220522
## p.value    0.1089932   0.99998062  0.8705785  0.996871  0.1481643
## criterion -0.1154032 -10.85112902 -2.0446807 -5.767027 -0.1603616
##
## $`6`$p.value
##   Solar.R
## 0.1089932
##
## $`6`$unweighted
## [1] TRUE
##
## $`6`$nobs
## [1] 21
##
##
## $`7`
## $`7`$criterion
##              Solar.R         Wind        Temp      Month          Day
## statistic  0.8083249 11.711564549  6.77148538  0.1307643  0.03992875
## p.value    0.8996614  0.003101788  0.04546281  0.9982052  0.99990034
## criterion -2.2992049 -0.003106609 -0.04652868 -6.3228783 -9.21378864
##
## $`7`$p.value
##        Wind
## 0.003101788
##
## $`7`$unweighted
## [1] TRUE
##
## $`7`$nobs
## [1] 37
##
##
## $`8`
## $`8`$criterion
##              Solar.R       Wind       Temp          Month            Day
## statistic  0.9056479  3.1585094  2.9285252   0.008106707   0.008686293
## p.value    0.8759687  0.3247585  0.3657072   0.999998099   0.999997742
## criterion -2.0872214 -0.3926849 -0.4552446 -13.173367586 -13.001213570
```

```
##
## $`8`$p.value
##      Wind
## 0.3247585
##
## $`8`$unweighted
## [1] TRUE
##
## $`8`$nobs
## [1] 30
##
##
## $`9`
## NULL
```

```
plot(airct)
```