

Introduction à R

Arbre de décision

Ricco Rakotomalala

http://eric.univ-lyon2.fr/~ricco/cours/cours_programmation_R.html

(1) R est un langage de programmation. L'objet de base est un vecteur de données.

C'est un « vrai » langage c.-à-d. types de données, branchements conditionnels, boucles, organisation du code en procédures et fonctions, découpage en modules.

Mode de d'exécution : transmettre à R le fichier script « .r »

(2) R est un logiciel de statistique et de data mining, pilotée en ligne de commande. Il est extensible (quasiment) à l'infini via le système des packages.

Les instructions servent à manipuler les objets R c.-à-d. les ensembles de données, les vecteurs, les modèles, etc.

Mode de d'exécution : introduire commandes dans le terminal, manipulation interactive

→ C'est le mode que nous exploiterons dans ce tutoriel.

<http://www.r-project.org/>

The screenshot shows the homepage of the R Project for Statistical Computing. The browser address bar displays <http://www.r-project.org/>. The page features the R logo, a navigation menu on the left, and a central area with various statistical plots including a PCA plot, a bar chart, a clustering dendrogram, and two normal distribution curves. A red arrow points to the 'download R' link in the 'Getting Started' section.

Navigation Links:

- About R
 - [What is R?](#)
 - [Contributors](#)
 - [Screenshots](#)
 - [What's new?](#)
- Download, Packages
 - [CRAN](#)
- R Project
 - [Foundation](#)
 - [Members & Donors](#)
 - [Mailing Lists](#)
 - [Bug Tracking](#)
 - [Developer Page](#)
 - [Conferences](#)
 - [Search](#)
- Documentation
 - [Manuals](#)
 - [FAQs](#)
 - [The R Journal](#)
 - [Wiki](#)
 - [Books](#)
 - [Certification](#)
 - [Other](#)
- Misc
 - [Bioconductor](#)
 - [Related Projects](#)
 - [User Groups](#)
 - [Links](#)

Getting Started:

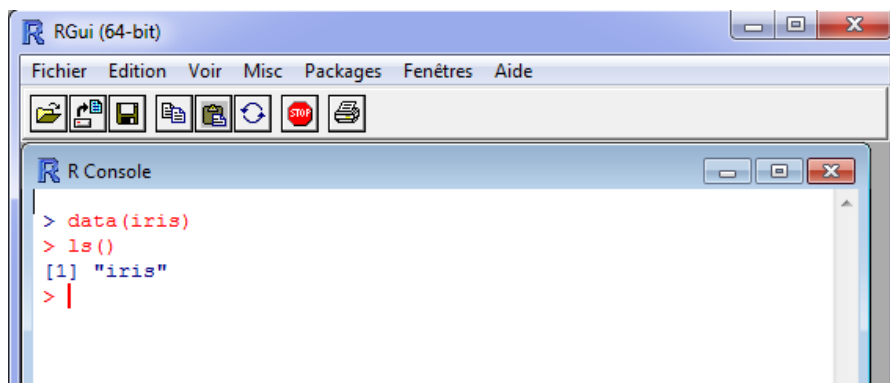
- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please visit your preferred [CRAN mirror](#).
- If you have questions about R like how to download and install software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News :

- R version 2.14.1** (December Snowflakes) has been released on 2011-12-22.
- [The R Journal Vol.3/2](#) is available.
- [useR! 2012](#), will take place at Vanderbilt University, Nashville Tennessee, USA, June 12-15, 2012.
- R version 2.13.2** has been released on 2011-09-30.

This server is hosted by the [Institute for Statistics and Mathematics](#) of the [WU Wien](#).

R peut fonctionner sous **Windows**, **Mac OS X**, **Linux**



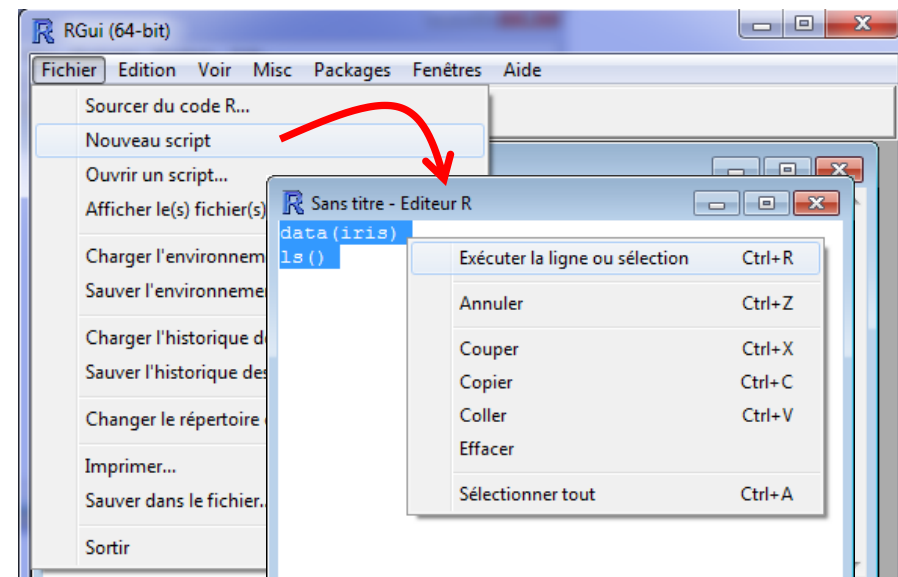
```
> data(iris)
> ls()
[1] "iris"
> |
```

Mode « terminal »

+ interactivité, visualisation immédiate des résultats

+ avec « ↑ », on retrouve les anciennes commandes

- pas de sauvegarde des commandes si fermeture de R (si en fait, avec *FICHIER / SAUVER L'HISTORIQUE DES COMMANDES*)



```
data(iris)
ls()
```

Mode « script »

+ interactivité, visualisation immédiate des résultats (CTRL + R)

+ maintien d'une liste « propre » des commandes utiles uniquement

+ possibilité d'E/S (chargement ou sauvegarde d'un fichier script « .r »)

→ mode conseillé pour nous

Si on veut programmer (mode programmation), mieux vaut passer par un éditeur externe (ex. TINN-R, R-STUDIO, ECLISPE + StatET,...)



demo_reglog.txt - Bloc-notes

age	taux	angine	coeur
50	126.0	1	presence
49	126.0	0	presence
46	144.0	0	presence
49	139.0	0	presence
62	154.0	1	presence
35	156.0	1	presence
67	160.0	0	absence
65	140.0	0	absence
47	143.0	0	absence
58	165.0	0	absence
57	115.0	1	absence
59	145.0	0	absence
44	175.0	0	absence
41	153.0	0	absence
54	152.0	0	absence
52	169.0	0	absence
57	168.0	1	absence
50	158.0	0	absence
44	170.0	0	absence
49	171.0	0	absence

Nom du data.frame

Séparateur de colonnes

1^{ère} ligne = nom des variables

Nom du fichier

Point décimal

```
R Console
> heart <- read.table(file="demo_reglog.txt", sep="\t", dec=".", header=T)
> class(heart)
[1] "data.frame"
> summary(heart)
      age      taux      angine      coeur
Min.   :35.00  Min.   :115.0  Min.   :0.00  absence :14
1st Qu.:46.75  1st Qu.:142.2  1st Qu.:0.00  presence: 6
Median :50.00  Median :153.5  Median :0.00
Mean   :51.75  Mean   :151.4  Mean   :0.25
3rd Qu.:57.25  3rd Qu.:165.8  3rd Qu.:0.25
Max.   :67.00  Max.   :175.0  Max.   :1.00
> |
```

« âge », « taux » et « angine » sont considérées comme quantitatives.

« cœur » est une variable qualitative.

Fichier texte, séparateur tabulation.

data.frame = matrice de données = liste de vecteurs de même longueur.

Vecteur = variable.

Les variables sont typées. Les plus utilisées sont « numeric / integer » (variables quantitatives) et « factor » (variables qualitatives)

Remarque : on peut accéder aux variables d'un data.frame avec l'opérateur \$

```
R Console
> class(heart$age)
[1] "integer"
> class(heart$taux)
[1] "numeric"
> class(heart$angine)
[1] "integer"
> class(heart$coeur)
[1] "factor"
> mean(heart$age)
[1] 51.75
> |
```

Package ?

- Un package est une bibliothèque externe
- Sous Windows → fichiers binaires pré-compilés
- Extension .zip
- Il est toujours documenté : fichier HTML (aide sous R) et PDF

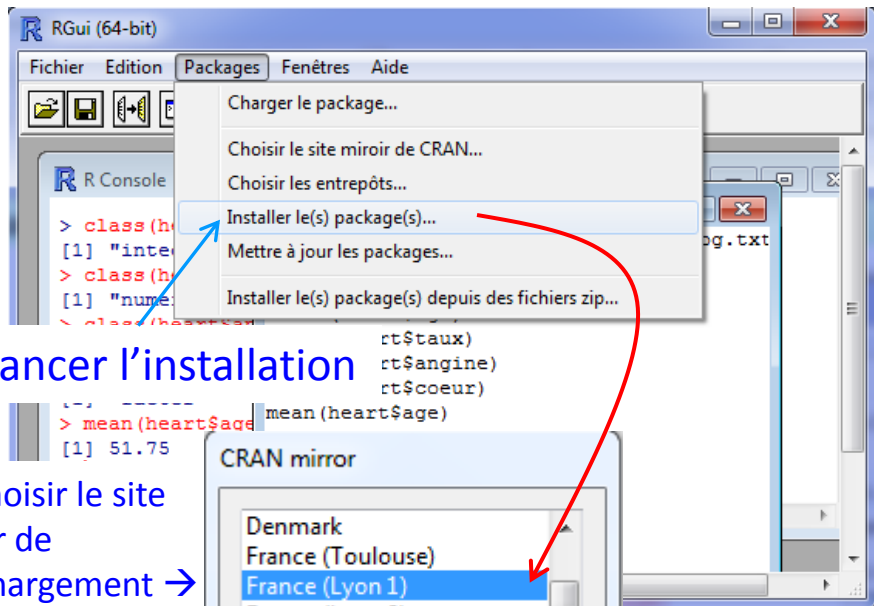
Quel intérêt ?

- Un package contient des collections de fonctions utilisables sous R
- Souvent centrés sur un sujet particulier (ex. *rpart* pour les arbres de décision, etc.)
- Gestion affinée des packages : nous pouvons les installer, désinstaller, charger, décharger et mettre à jour à notre guise

Ce système permet d'augmenter considérablement la puissance de R !!!

Ex. installer et charger le package « xlsx » permettant de lire directement les fichiers Excel (*.xls et *.xlsx)

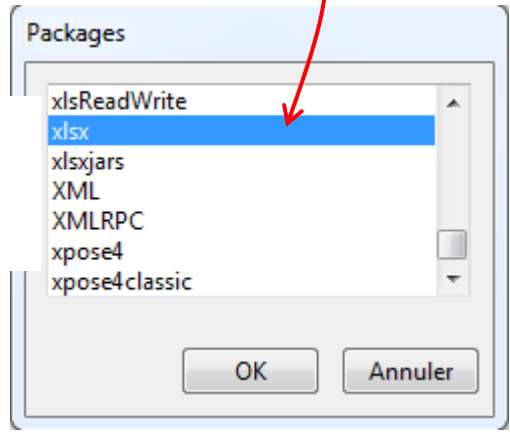
Installation (une fois) et chargement d'un package (à chaque utilisation)



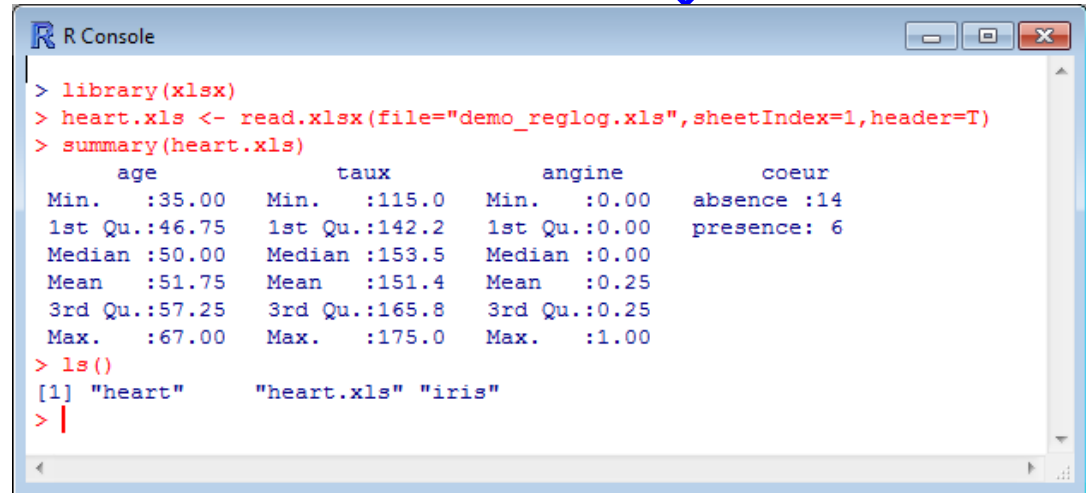
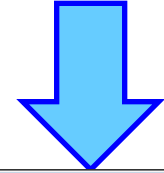
(1) Lancer l'installation

(2) Choisir le site miroir de téléchargement →

(3) Choisir le package à installer →



- > Chargement du package [library]
- > Lecture du fichier Excel (1^{ère} feuille, nom de variable sur 1^{ère} ligne) [read.xlsx]
- > Statistiques descriptives [summary]
- > ls() liste le contenu de la mémoire



Fichier utilisé pour la présentation des arbres de décision

	A	B	C	D	E	F	G	H	I	J
1	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	classe
2		2	1	1	1	2	1	1	1	1 begin
3		1	2	2	1	2	1	1	1	1 begin
4		1	1	1	1	1	1	2	1	1 begin
5		2	1	1	1	2	1	1	1	1 begin
6		1	1	1	1	2	1	2	1	1 begin
7		3	1	1	1	2	1	2	1	1 begin
8		5	10	6	1	10	4	4	10	10 malignant
9		1	1	1	1	1	1	3	1	1 begin
10		1	1	1	1	2	1	1	1	1 begin
11		10	10	10	10	6	10	8	1	5 malignant

Echantillon d'apprentissage, 399 observations

```
breast.app <- read.xlsx(file="breast.xls",sheetIndex=1,header=T)
```

	A	B	C	D	E	F	G	H	I	J
1	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	classe
2		4	1	1	1	2	3	1	1	1 begin
3		1	5	8	6	5	8	7	10	1 malignant
4		3	2	2	3	2	1	1	1	1 begin
5		3	1	1	3	8	1	5	8	1 begin
6		10	4	6	1	2	10	5	3	1 malignant
7		10	5	7	3	3	7	3	3	8 malignant
8		6	3	3	3	3	2	6	1	1 begin
9		9	10	10	1	10	8	3	3	1 malignant
10		7	5	6	3	3	8	7	4	1 malignant
11		5	3	2	8	5	10	8	1	2 malignant

Echantillon de test, 300 observations

```
breast.test <- read.xlsx(file="breast.xls",sheetIndex=2,header=T)
```


Arbres de décision avec le package « rpart »

Chargement du package « rpart », spécialisé dans les arbres
Toujours installé avec R

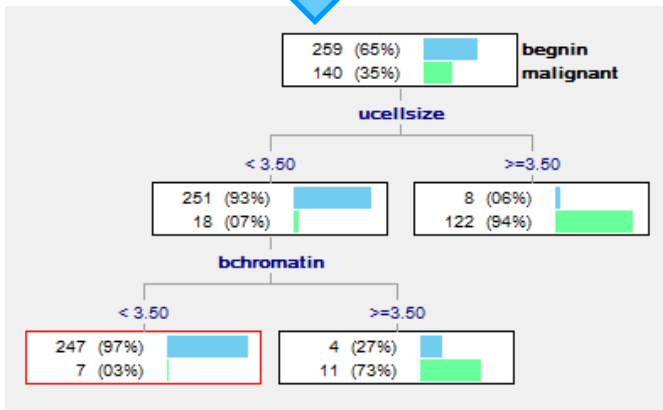
```
R Console
> library(rpart)
> arbre.1 <- rpart(classe ~ ., data=breast.app, method="class")
> print(arbre.1)
n= 399

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

1) root 399 140 beginn (0.64912281 0.35087719)
 2) ucellsize< 3.5 269 18 beginn (0.93308550 0.06691450)
   4) bchromatin< 3.5 254 7 beginn (0.97244094 0.02755906) *
   5) bchromatin>=3.5 15 4 malignant (0.26666667 0.73333333) *
 3) ucellsize>=3.5 130 8 malignant (0.06153846 0.93846154) *
```

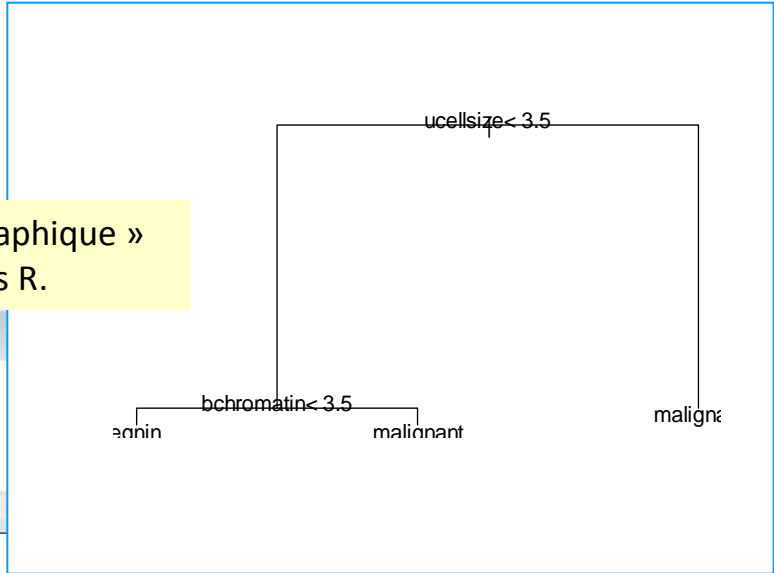
[data = breast.app] - Construction de l'arbre sur l'échantillon « breast.app ».
[classe ~ .] - Prédire « classe » à partir des autres variables de la base.
[method = « class »] - On construit un arbre de décision c.-à-d. apprentissage supervisé.

Affichage « format texte » de l'arbre.



Affichage « graphique » de l'arbre dans R.

```
R Console
> plot(arbre.1)
> text(arbre.1)
> |
```



Evaluation de l'arbre sur un échantillon test

Construction de la prédiction avec `predict(...)`

Modèle utilisé pour la prédiction

`newdata` - Appliquer la prédiction sur l'échantillon test (data.frame « breast.test »)

`type` - Prédiction est une variable qualitative (classement)

```
R Console  
> pred.classe <- predict(arbre.1,newdata=breast.test,type="class")  
> print(summary(pred.classe))  
  begin malignant  
    191      109
```

« `pred.classe` » est un vecteur de taille 300 ; 191 individus ont été classés « begin », 109 « malignant ».

Matrice de confusion

```
R Console  
> mc <- table(breast.test$classe,pred.classe)  
> print(mc)  
      pred.classe  
begin 184      15  
malignant 7      94  
> erreur <- (mc[2,1]+mc[1,2])/sum(mc)  
> print(erreur)  
[1] 0.07333333  
>
```

`table(...)` construit un tableau croisé entre la cible observée (classe) et la prédiction du modèle (`pred.classe`)

La table `mc` se comporte comme une matrice à 2 dimensions, on en déduit le taux d'erreur

`erreur` = Somme des éléments hors diagonale principale / Nombre total des observations.

Modifier les paramètres de construction de l'arbre

Modifier les paramètres de construction de l'arbre

Minsplit = taille minimale pour segmenter

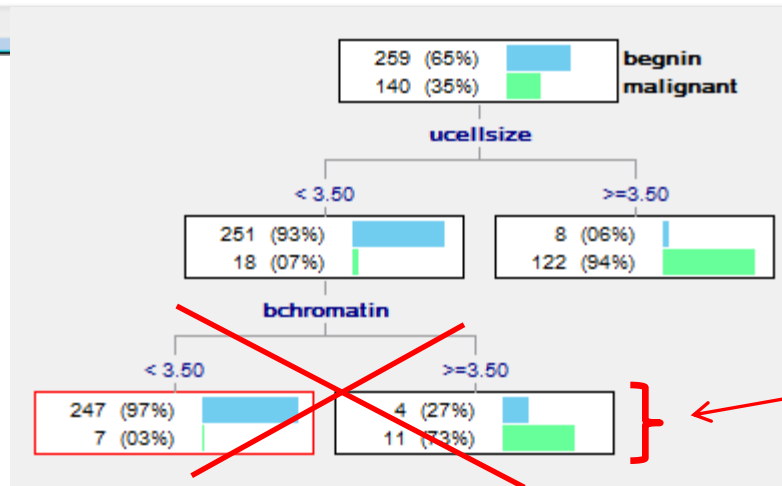
Minbucket = effectif d'admissibilité

```
R Console
> parametres <- rpart.control(minsplit=50, minbucket=20)
> arbre.2 <- rpart(classe ~ ., data=breast.app, method="class", control=parametres)
> print(arbre.2)
n= 399

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

1) root 399 140 beginn (0.64912281 0.35087719)
2) ucellsize< 3.5 269 18 beginn (0.93308550 0.06691450) *
3) ucellsize>=3.5 130 8 malignant (0.06153846 0.93846154) *
>
> |
```

Construction de l'arbre avec les paramètres modifiés.



Parce que moins de 20 individus sur cette feuille

Autres packages pour les arbres – Le package « tree »

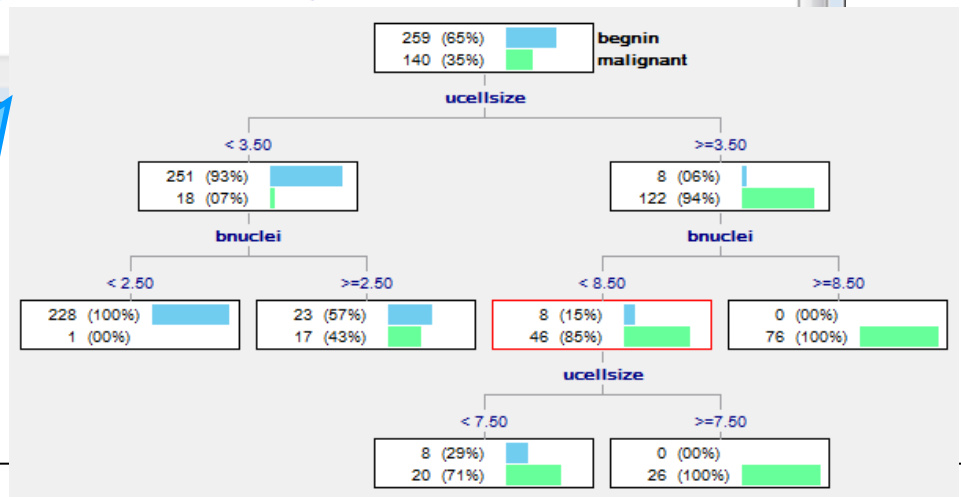
Chargement du package « tree », il faut l'installer au préalable

Paramètres d'apprentissage :
nobs = nombre d'obs. dans l'éch. d'apprentissage ; **mincut** = effectif d'admissibilité, **minsize** = taille min. pour segmenter

Apprentissage : classe vs. toutes les autres variables du data.frame

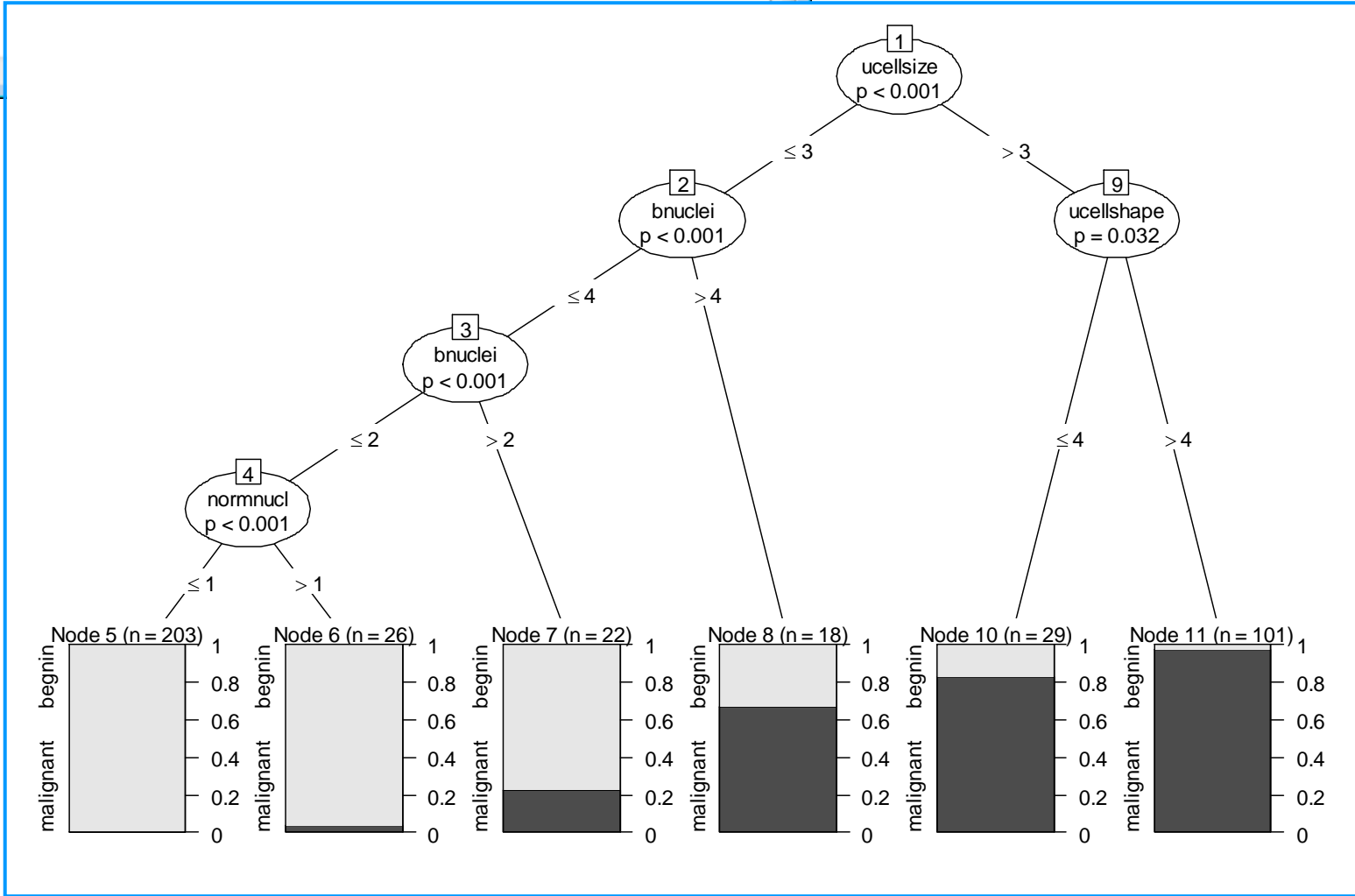
```
R Console
> library(tree)
> param.3 <- tree.control(nobs=nrow(breast.app), mincut=20, minsize=50)
> arbre.3 <- tree(classe ~ ., data=breast.app, control=param.3)
> print(arbre.3)
node), split, n, deviance, yval, (yprob)
* denotes terminal node
1) root 399 517.10 benign ( 0.649123 0.350877 )
2) ucellsize < 3.5 269 132.10 benign ( 0.933086 0.066914 )
4) bnuclei < 2.5 229 12.86 benign ( 0.995633 0.004367 ) *
5) bnuclei > 2.5 40 54.55 benign ( 0.575000 0.425000 ) *
3) ucellsize > 3.5 130 60.11 malignant ( 0.061538 0.938462 )
6) bnuclei < 8.5 54 45.30 malignant ( 0.148148 0.851852 )
12) ucellsize < 7.5 28 33.50 malignant ( 0.285714 0.714286 ) *
13) ucellsize > 7.5 26 0.00 malignant ( 0.000000 1.000000 ) *
7) bnuclei > 8.5 76 0.00 malignant ( 0.000000 1.000000 ) *
```

$$\text{DEVIANCE} = -2 * n * [0.649 * \text{LN}(0.649) + 0.351 * \text{LN}(0.351)]$$



```
R Console  
> library(party)  
> param.4 <- ctree_control(minsplit=20,minbucket=10)  
> arbre.4 <- ctree(classe ~ ., data = breast.app, controls=param.4)  
> plot(arbre.4)  
>  
> |
```

Un des avantages, affichage « sympathique » de l'arbre.



De la documentation à profusion (n'achetez jamais des livres sur R)

Site du cours

http://eric.univ-lyon2.fr/~ricco/cours/cours_programmation_R.html

RPART

T. M. Therneau, E.J. Atkinson, « An introduction to Recursive Partitioning using RPART Routines »

<http://www.mayo.edu/hsr/techrpt/61.pdf>

Programmation R

<http://www.duclert.org/>

Quick-R

<http://www.statmethods.net/>

POLLS (Kdnuggets)

Data Mining / Analytics Tools Used - <http://www.kdnuggets.com/polls/2011/tools-analytics-data-mining.html>

(Mai 2011, R en 2nde position)

What languages you used for data mining / data analysis?

<http://www.kdnuggets.com/polls/2011/languages-for-data-mining-analytics.html>

(Août 2011, langage R en 1^{ère} position)