Bayesian classification and regression trees

Hedibert F. Lopes & Paulo Marques
INSPER Institute of Education and Research
São Paulo, Brazil

A not-so-random forest[1]

[1] https://awebic.com/ambiente/arvores-assustadoras

A hard forest[2]

[2] http://www.journeysbychance.com/en/story/view/Madagascar-Private-Plane-Expedition.html

CART - classification and regression trees

Seminal work

Breiman, Friedman, Olshen and Stone (1984) *Classification and Regression Trees* Belmont, Wadsworth.

Concise description and S-plus implementation

Clark and Pregibon (1992) Tree-based models. In *Statistical models in S*.

Bayesian CART

- Chipman, George and McCulloch (1998) Bayesian CART model search
- Denison, Mallick and Smith (1998) A Bayesian CART algorithm

Improvements/extesions

- Wu, Tjelmeland and West (2007) Bayesian CART: Prior specification and posterior simulation. *JCGS*, 16(1), 44-66.
- Chipman, George and McCulloch (2000) Hierarchical priors for Bayesian CART shrinkage. *Statistics and Computing*, 10, 17-24.
- Chipman, George and McCulloch (2002) Bayesian treed models. *Machine Learning*, 48, 299?320.

CART algorithm

Greedy algorithm to grow a tree, then prune it back to avoid overfitting.

Such greedy algorithms typically grow a tree by sequentially choosing splitting rules for nodes on the basis of maximizing some fitting criterion.

This generates a sequence of trees, each of which is an extension of the previous tree.

A single tree is then selected by pruning the largest tree according to a model choice criterion, such as cost-complexity pruning, cross-validation, or even multiple tests of whether two adjoining nodes should be collapsed into a single node.

A tree with 5 terminal nodes

$y \sim N(\theta, 4)$
x_1 is a quantitative predictor taking values in $[0, 10]$
x_2 is a qualitative predictor with categoris $\{A, B, C, D\}$

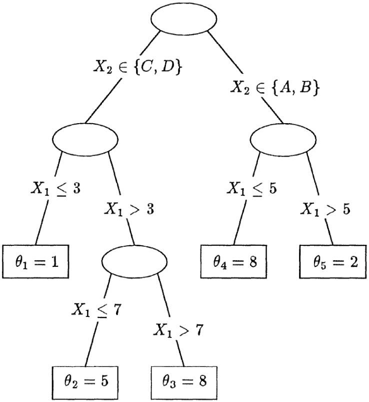


Figure 2. A Regression Tree Where $y \sim N(\theta, 2^2)$ and $\mathbf{x} = (x_1, x_2)$.

houseprice dataset (revisited)

```
house = read.table("houseprice.txt",header=TRUE)

size  = house[,4]*0.092903
price = house[,8]*3.2/1000
Nbhd  = factor(house[,2], labels = c("N1","N2","N3"))

House = data.frame(cbind(price,size,house[,c(3,5,6,7)],Nbhd))

House[1:5,]
   price     size Offers Brick Bedrooms Bathrooms Nbhd
1 365.76 166.2964      2    No        2         2   N2
2 365.44 188.5931      3    No        4         2   N2
3 367.36 161.6512      1    No        3         2   N2
4 303.04 183.9479      3    No        3         2   N2
5 383.36 197.8834      3    No        3         3   N2
```

R package tree

```
install.packages("tree")
library (tree)

tree.fit = tree(price~.,data=House)

summary(tree.fit)

Regression tree:
tree(formula = price ~ ., data = House)
Number of terminal nodes:  10
Residual mean deviance:  1383 = 163200 / 118
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-97.030 -24.820   1.442   0.000  19.690 100.700
```

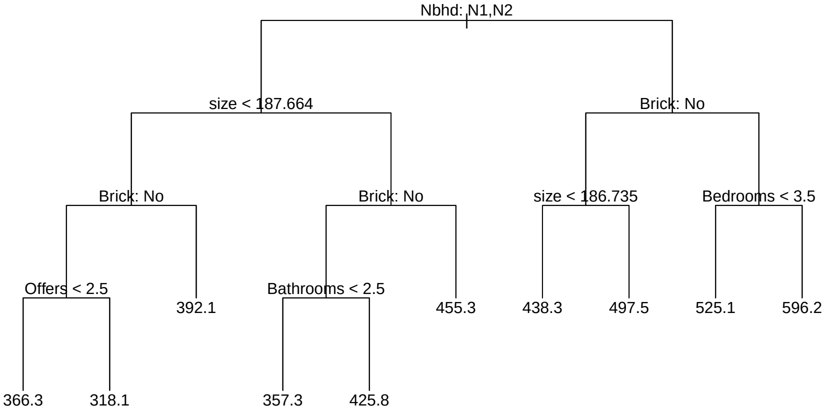
Fitted tree

```
tree.fit
# node), split, n, deviance, yval
      # * denotes terminal node

 # 1) root 128 938900 417.4
   # 2) Nbhd: N1,N2 89 308000 376.9
     # 4) size < 187.664 55 128700 353.2
       # 8) Brick: No 40  68160 338.6
        # 16) Offers < 2.5 17  11780 366.3 *
        # 17) Offers > 2.5 23  33740 318.1 *
       # 9) Brick: Yes 15  29260 392.1 *
     # 5) size > 187.664 34  98480 415.2
      # 10) Brick: No 23  64670 396.0
        # 20) Bathrooms < 2.5 10  13830 357.3 *
        # 21) Bathrooms > 2.5 13  24300 425.8 *
      # 11) Brick: Yes 11   7707 455.3 *
   # 3) Nbhd: N3 39 152300 509.7
     # 6) Brick: No 23  41200 474.3
      # 12) size < 186.735 9   3074 438.3 *
      # 13) size > 186.735 14  18880 497.5 *
     # 7) Brick: Yes 16  40790 560.6
      # 14) Bedrooms < 3.5 8   4420 525.1 *
      # 15) Bedrooms > 3.5 8  16180 596.2 *
```

Fitted tree

```
plot(tree.fit,type="uniform")
text(tree.fit,pretty=0,cex=0.5)
```



Root MSE

```
> fit = predict(tree.fit,House)
> sqrt(mean((fit-price)^2))
[1] 35.70442
```

```
> olsfit = lm(price~.,data=House)
> sqrt(mean(olsfit$res^2))
[1] 31.04256
```

Bayesian CART algorithm

The two basic components of this approach consist of prior specification and stochastic search.

The basic idea is to have the prior induce a posterior distribution that will guide the stochastic search toward more promising CART models.

As the search proceeds, such models can then be selected with a variety of criteria, such as posterior probability, marginal likelihood, residual sum of squares, and misclassification rates.

Stochastic search: GROW, PRUNE, CHANGE and SWAP.

In a sense our procedure is a sophisticated heuristic for finding good models rather than a fully Bayesian analysis, which presently seems to be computationally infeasible.

Tree structure

Joint distribution of the tree and its parameters

$$p(\Theta, T) = p(\Theta|T)p(T)$$

The growing process is determined by the specification of two functions:

$$p_{\mathrm{SPLIT}}(\eta, T) = \alpha(1 + d_\eta)^{-\beta} \quad \text{and} \quad p_{\mathrm{RULE}}(\rho|\eta, T)$$

d_η is the depth of the node η (i.e., the number of splits above η) and $\beta \geq 0$.

Prior distribution of the number of terminal nodes

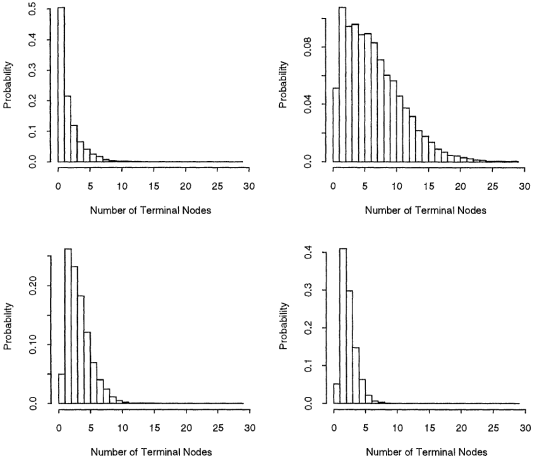


Figure 3. Prior Distribution on Number of Terminal Nodes. (a) α = .5, β = .5, with prior mean 2.1; (b) α = .95, β = .5, with prior mean 7; (c) α = .95, β = 1, with prior mean 3.7; (d) α = .95, β = 1.5, with prior mean 2.9.

Stochastic search MCMC[3]

Chipman, George and McCulloch (1998) use a Metropolis-Hastings step with a transition kernel choosing randomly among four steps:

- Grow: Pick a terminal node and split into two children nodes,
- Prune: Pick a parent of two terminal nodes and collapse,
- Change: Pick an internal node and reassign the splitting rule,
- Swap: Pick a parent-child pair and swap splitting rules, unless the other child of the parent has the same pair, in which case give both children the splitting rule of the parent.

All steps are reversible, so the Markov chain is reversible.

[3] Slide from `https://classes.soe.ucsc.edu/ams225/Winter10/ams225-jacobson-presentation.pdf`

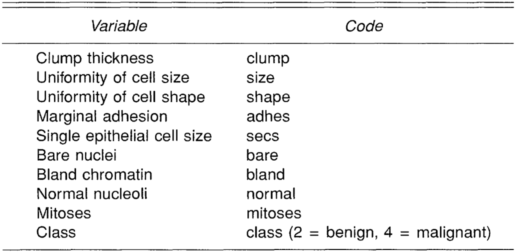
Breast cancer study

Breiman (1996) and Tibshirani and Knight (1995)

Data from UC Irvine repository of machine learning databases[4]
`ftp://ftp.ics.uci.edu/pub/machine-learning-databases`

The dataset comprises nine cellular characteristics that might be useful in predicting whether a tumor is benign or malignant (the binary variable class)
All cellular characteristics are ordered numeric variables, each with levels 1,2,. . . ,10.
683 complete observations were used.

Table 1. Variables for the Breast Cancer Data



Variable	*Code*
Clump thickness	clump
Uniformity of cell size	size
Uniformity of cell shape	shape
Marginal adhesion	adhes
Single epithelial cell size	secs
Bare nuclei	bare
Bland chromatin	bland
Normal nucleoli	normal
Mitoses	mitoses
Class	class (2 = benign, 4 = malignant)

[4] Wolberg and Mangasarian (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *PNAS*, 87, 9193-9196.

A 5-node tree found by stochastic search

Bayesian CART, stochastic search (18%) versus CART greedy search (30%)

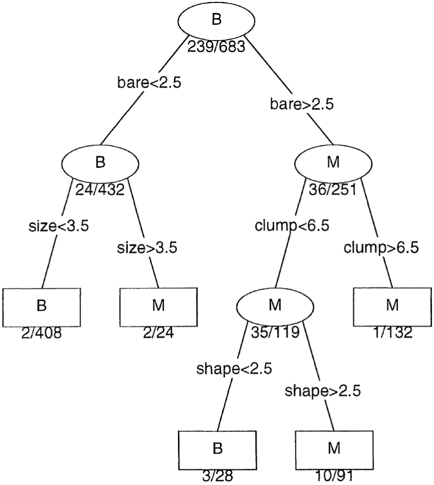


Figure 1. A Five-Node Tree Found by by Stochastic Search. The overall misclassification rate is 18. The letters B and M, which refer to benign and malignant tumors, indicate the response which is in the majority in each node. The misclassification rates and number of observations are given below each node.

Variables in the Boston housing dataset

Boston Housing data, Harrison and Rubinfeld (1978)[5]
The data consist of 14 characteristics of 506 census tracts in the Boston area.
The response is the logged median value of owner occupied homes in each tract.

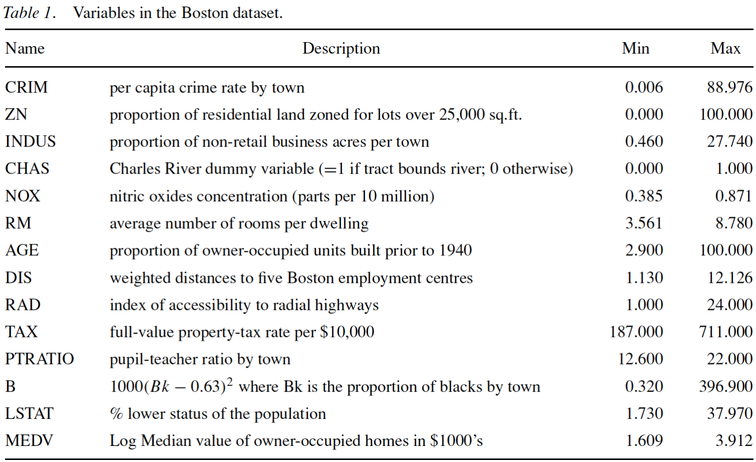


Table 1. Variables in the Boston dataset.

Name	Description	Min	Max
CRIM	per capita crime rate by town	0.006	88.976
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.	0.000	100.000
INDUS	proportion of non-retail business acres per town	0.460	27.740
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)	0.000	1.000
NOX	nitric oxides concentration (parts per 10 million)	0.385	0.871
RM	average number of rooms per dwelling	3.561	8.780
AGE	proportion of owner-occupied units built prior to 1940	2.900	100.000
DIS	weighted distances to five Boston employment centres	1.130	12.126
RAD	index of accessibility to radial highways	1.000	24.000
TAX	full-value property-tax rate per $10,000	187.000	711.000
PTRATIO	pupil-teacher ratio by town	12.600	22.000
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town	0.320	396.900
LSTAT	% lower status of the population	1.730	37.970
MEDV	Log Median value of owner-occupied homes in $1000's	1.609	3.912

[5] Harrison and Rubinfeld (1978) Hedonic prices and the demand for clean air. *Journal of Environmental Economic and Management*, 5, 81-102.

RMSE for training data

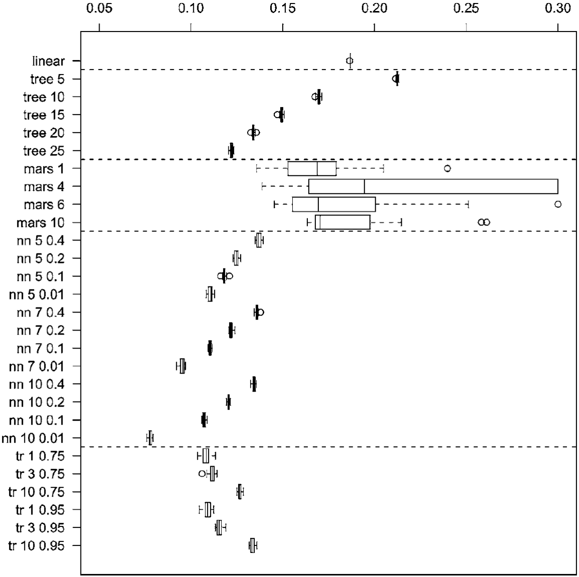


Figure 8. RMSE for training data, Boston housing example.

RMSE for test data

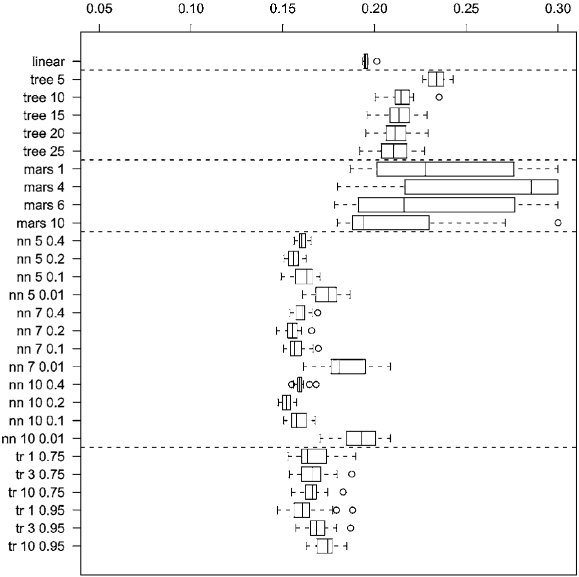


Figure 9. RMSE for test data, Boston housing example.