Bayesian classification and regression trees

Hedibert F. Lopes & Paulo Marques
INSPER Institute of Education and Research
São Paulo, Brazil
A not-so-random forest

1https://awebic.com/ambiente/arvores-assustadoras
A hard forest\textsuperscript{2}

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

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

Improvements/extensions
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 categories \( \{A, B, C, D\} \)

---

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

```r
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      2  N2
  2 365.44 188.5931  3   No      4     2      2  N2
  3 367.36 161.6512  1   No      3     2      2  N2
  4 303.04 183.9479  3   No      3     2      2  N2
  5 383.36 197.8834  3   No      3     3      3  N2
```
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.
Residuals: -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 *
  # 3) Nbhd: N3 39 152300 509.7
    # 6) Brick: No 23  41200 474.3
      # 12) size < 186.735 9   3074 438.3 *
    # 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

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

\[
\begin{align*}
> & \text{fit} = \text{predict(tree.fit,House)} \\
> & \text{sqrt(mean((fit-price)^2))} \\
& [1] 35.70442
\end{align*}
\]

\[
\begin{align*}
> & \text{olsfit} = \text{lm(price~.,data=House)} \\
> & \text{sqrt(mean(olsfit$\text{res}^2))} \\
& [1] 31.04256
\end{align*}
\]
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:

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

\(d_\eta\) is the depth of the node \(\eta\) (i.e., the number of splits above \(\eta\)) and \(\beta \geq 0\).
Prior distribution of the number of terminal nodes

Figure 3. Prior Distribution on Number of Terminal Nodes. (a) $\alpha = .5, \beta = .5$, with prior mean 2.1; (b) $\alpha = .95, \beta = .5$, with prior mean 7; (c) $\alpha = .95, \beta = 1$, with prior mean 3.7; (d) $\alpha = .95, \beta = 1.5$, with prior mean 2.9.
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.
Breast cancer study

Data from UC Irvine repository of machine learning databases\(^4\)

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, \ldots, 10. 683 complete observations were used.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clump thickness</td>
<td>clump</td>
</tr>
<tr>
<td>Uniformity of cell size</td>
<td>size</td>
</tr>
<tr>
<td>Uniformity of cell shape</td>
<td>shape</td>
</tr>
<tr>
<td>Marginal adhesion</td>
<td>adhes</td>
</tr>
<tr>
<td>Single epithelial cell size</td>
<td>secs</td>
</tr>
<tr>
<td>Bare nuclei</td>
<td>bare</td>
</tr>
<tr>
<td>Bland chromatin</td>
<td>bland</td>
</tr>
<tr>
<td>Normal nucleoli</td>
<td>normal</td>
</tr>
<tr>
<td>Mitoses</td>
<td>mitoses</td>
</tr>
<tr>
<td>Class</td>
<td>class (2 = benign, 4 = malignant)</td>
</tr>
</tbody>
</table>

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

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

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