Classification and Regression Trees

1. Introduction

There are numerous algorithms for predicting a variable from a set of continuous or categorical predictors. The mostly used one is regression model.

Depending on the type of outcome variable two types of problems are distinguished:

- Regression-type problems: for continuous dependent variable
- Classification-type problems: for categorical dependent variable

The purpose of the analysis via Classification and Regression Trees (C&RT) is to determine a set of if-then logical conditions (splits) that permit accurate prediction or classification of cases.

C&RT is a nonparametric statistical recursive procedure that identifies mutually exclusive subgroups of a population whose members share common characteristics that influence the dependent variable of interest. C&RT produces visual output that is a multilevel structure that resembles branches of a tree.

2. In the beginning…

In 1984 Berkeley (Breiman) and Stanford (Friedman) statisticians announce a new classification tool which:

- Could separate relevant from irrelevant predictors
- Did not require any kind of variable transformation
- Impervious to outliers and missing values
- Could yield relatively simple and easy to comprehend models
- Require little to no supervision by the analyst

Took long time to become known and popular: method was not covered in many textbooks; was taught only in advanced grad courses; was not part of standard statistical packages

3. These days…

Gained popularity in the data mining

- Availability of huge data sets requiring analysis
- Dealing with non-linear dependency structures
- Increase in data storage capacity
- Exponential growth in computational power of CPUs

4. Example of Classification/Regression Task

- Predicting tumor as benign or malignant
- Classifying credit card transaction as legitimate or fraudulent
- Predicting consumer preferences towards different kinds of vehicles

- Predicting efficacy of drags based on demographic factors
- Predicting medium house value based on crime rate, pollution, age, industrialization level, etc

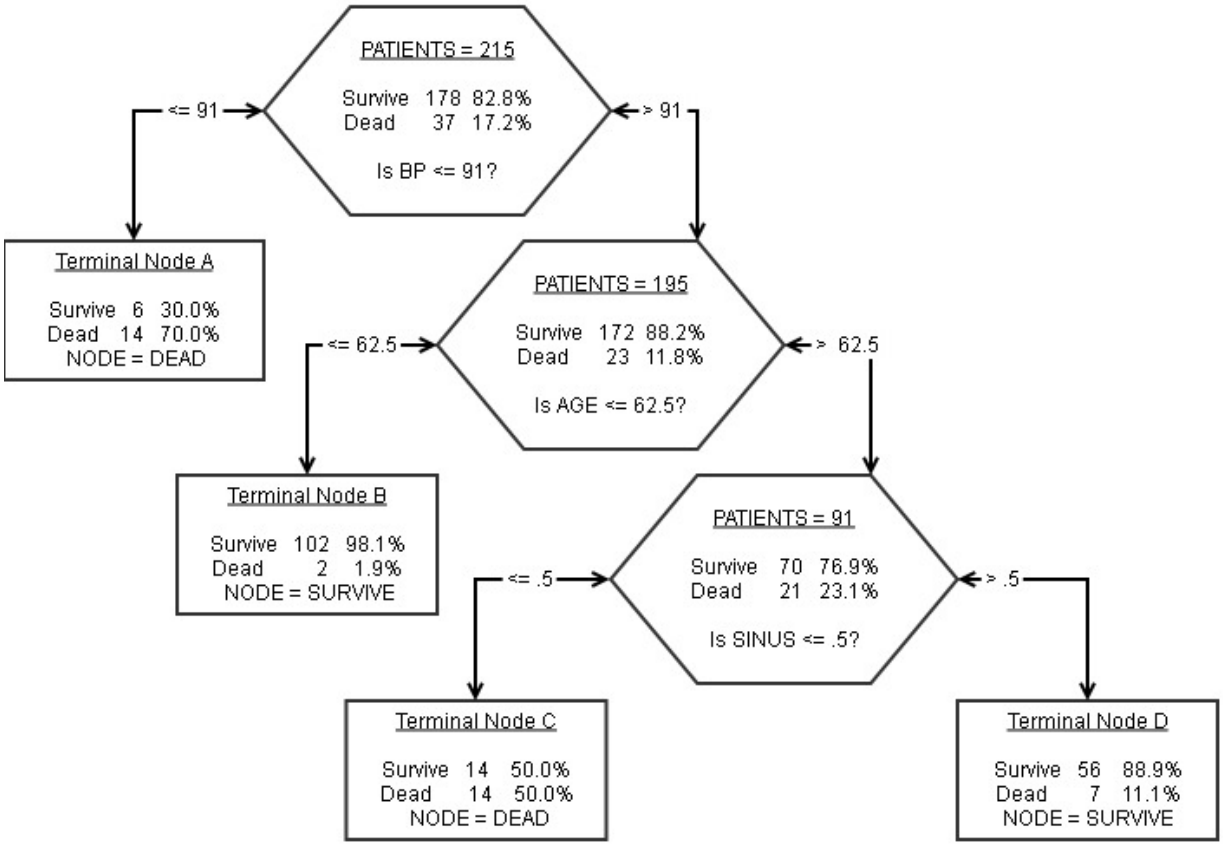
5. C&RT Example in Public Health

C&RT methodology has been applied to health sciences and clinical research and to a lesser degree in public health

- Epidemiologic studies of morbidity and mortality from specific diseases
- Comparison of cost-effectiveness of colorectal cancer screening technologies
- Influenza treatment strategies
- Development of screening and diagnostic tools

6. Typical C&RT analysis

UCSD Heart Disease Study
Given the diagnostic of a heart attack predict who is at risk of 2nd heart attack and early death within 30 days. Prediction will determine treatment program (intensive care or not) . For each patient about 100 variables were available



Comments: only two answers are possible: binary partitioning
YES always goes to the left
Tree is a classifier

Ref: Introduction to CART by Matthew Magistrado, Salford Systems 2007

7. How to Read It

- Entire tree represents a complete analysis or model
- Has the form of a decision tree
- Root of inverted tree contains all data
- Root gives rise to child (son, daughter) nodes
- Child nodes can in turn give rise to their own children
- At some point a given path ends in a terminal node
- Terminal node classifies object
- Path through the tree governed by the answers to QUESTIONS or RULES

8. Classification – A two-step process

1. Model construction
    a. Model presented as classification tree, decision tree, of formulae
2. Model usage
    a. Real world decision making
    b. Selection of the most important prognostic variables. Use results to find variables to include in logistic regression
    c. Suggests interaction: AGE is not relevant if BP is not high
    d. To classify future or unknown objects
    e. Accuracy rate is the percentage of test set samples that are correctly classified by the model
    f. Test set should be independent of training set, otherwise over-fitting will occur

9. Tree Growing Procedure

1) How do we choose the Boolean conditions for splitting at each node?

2) Which criterion should we use to split a parent node into its two daughter nodes?

3) How do we decide when a node becomes a terminal node (i.e., stop splitting)?

4) How do we assign a class to a terminal node?

10. Splitting Strategies

At each node, the tree-growing algorithm has to decide on which variable it is “best” to split. We need to consider every possible split over all variables present at that node, then enumerate all possible splits, evaluate each one, and decide which is best in some sense.

The best split will generate the greatest improvement in predictive accuracy.

Impurity measures – provides an indication of the relative homogeneity of cases in the node:

- Gini index
- Entropy
- Chi-square
- Least-squares deviation criteria

11. Pruning

Splitting can continue until all cases are perfectly classified – complete tree. It is large, complex, and not useful or accurate for predicting new observation.
Stopping rules:
Minimum n or Fraction of objects
Complexity parameter: measures the ‘cost’ of adding another variable to the tree. Found through cross-validation.
Strategies:

- Grow tree to the right size which is determined by previous research
- Use well-documented, structured procedure developed by Breiman (1984)

12. Random Forest

A random forest is a collection of single trees grown in a special way.
The overall prediction is determined by voting (in classification) or averaging (in regression)
The key to accuracy is low correlation and bias. To keep bias low, trees are grown to maximum depth

Each tree is grown on a bootstrap sample from the learning set
A number R is specified (square root by default) such that it is noticeably smaller than the total number of available predictors. During tree growing phase, at each node only R predictors are randomly selected and tried

All major advantages of a single tree are automatically preserved

Since each tree is grown on a bootstrap sample, one can

◦ Use out of bag samples to compute an unbiased estimate of the accuracy

◦ Use out of bag samples to determine variable importances

There is no overfitting as the number of trees increases - generalizes well to new data

High levels of predictive accuracy delivered automatically

Only a few control parameters to experiment with

Strong for both regression and classification

13. Conditional Classification Trees

14. Software:

SPSS via add-on module "SPSS classification trees"

R: tree, rpart, party

Breiman, Leo, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth Int Group, 1984.

Chen, Xiang, Minghui Wang, and Heping Zhang. "The Use of Classification Trees for Bioinformatics." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, no. 1 (January 2011): 55–63. doi:10.1002/widm.14.

Christoph Molnar. "Conditional Trees." December 19, 2012. http://www.slideshare.net/christophmolnar/conditional-trees.

"Classification and Regression Trees - Springer." Accessed October 17, 2013. http://link.springer.com/article/10.1007%2Fs00038-011-0315-z/fulltext.html.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15, no. 3 (2006). http://amstat.tandfonline.com/doi/full/10.1198/106186006X133933.

Izenman, Alan Julian. "Recursive Partitioning and Tree-Based Methods." In *Modern Multivariate Statistical Techniques*, by Alan J. Izenman, 281–314. New York, NY: Springer New York, 2013. http://link.springer.com/10.1007/978-0-387-78189-1_9.

Lemon, Stephenie C., Jason Roy, Melissa A. Clark, Peter D. Friedmann, and William Rakowski. "Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison With Logistic Regression." *Annals of Behavioral Medicine* 26, no. 3 (November 2003): 172–181.

Loh, Wei-Yin. "Classification and Regression Trees." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, no. 1 (2011): 14–23. doi:10.1002/widm.8.

Marshall, Roger J. “The Use of Classification and Regression Trees in Clinical Epidemiology.” *Journal of Clinical Epidemiology* 54, no. 6 (June 2001): 603–609. doi:10.1016/S0895-4356(00)00344-9.

Nunn, Martha E., Juanjuan Fan, Xiaogang Su, Richard A. Levine, Hyo-Jung Lee, and Michael K. McGuire. “Development of Prognostic Indicators Using Classification and Regression Trees for Survival.” *Periodontology 2000* 58, no. 1 (2012): 134–142. doi:10.1111/j.1600-0757.2011.00421.x.

“Random Forests - Classification Description.” Accessed October 1, 2013. http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm#intro.

Siciliano, Roberta, and Francesco Mola. “Multivariate Data Analysis and Modeling through Classification and Regression Trees.” *Computational Statistics & Data Analysis* 32, no. 3–4 (January 28, 2000): 285–301. doi:10.1016/S0167-9473(99)00082-1.

Strobl, Carolin, James Malley, and Gerhard Tutz. “An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests.” *Psychological Methods* 14, no. 4 (December 2009): 323–348. doi:http://dx.doi.org/10.1037/a0016973.

Tseng, George C. “Classification and Regression Tree.” In *Encyclopedia of Measurement and Statistics*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc. Accessed October 17, 2013. http://knowledge.sagepub.com/view/statistics/n83.xml.

“Classification and Regression Tree.” In *Encyclopedia of Measurement and Statistics*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc. Accessed October 17, 2013. http://knowledge.sagepub.com/view/statistics/n83.xml.