TEACHING AND LEARNING ABOUT TREE-BASED METHODS FOR EXPLORATORY DATA ANALYSIS

Joachim Engel[1], Tim Erickson[2] and Laura Martignon[1]
[1]Ludwigsburg University of Education, Ludwigsburg, Germany
[2]Epistemological Engineering, Oakland, CA, USA
engel@ph-ludwigsburg.de

Quantitative information about important societal topics is increasingly accessible to the general public and to individual citizens. Making sense of these data requires the ability to explore, understand, and reason about complex multivariate data. For such data, we need flexible and robust analytical methods that can deal with nonlinear relationships, mixed type variables, high-order interactions and missing values. Classification and regression trees are well suited for the representation and analysis of such complex data. We introduce the digital tool Arbor, which is designed to help understanding about the construction and interpretation of decision and regression trees, and supports critical appreciation of the strengths and pitfalls of tree methods.

INTRODUCTION

Quantitative information about burning societal topics (like demographic change, global warming, crime, civil rights issues, social (in)equality, health hazards and many others) are increasingly accessible to the general public and to individual citizens. Making sense of these data requires the ability to explore, understand, and reason about complex *multivariate* data, because social phenomena do not happen in a vacuum, and their understanding requires awareness of how variables co-vary, or affect each other, or are situated in a network of causal factors and may change over time in complex ways. For such data, we need flexible and robust analytical methods that can deal with nonlinear relationships, mixed type variables, high-order interactions and missing values. Despite such challenges, the methods should be simple to understand and give easily interpretable results. Classification and regression trees are well suited for the representation and analysis of such complex data. Trees explain variation of a single response variable by repeatedly splitting the data into more homogeneous subgroups, using combinations of explanatory variables that may be categorical or numeric to obtain good splits. Each subgroup is characterized by a typical value of the response variable, the number of observations in the group, and the values of the explanatory variable that define it. The tree is represented graphically, and this aids exploration and understanding.

After discussing tree-based methods for understanding complex social data, we introduce the digital tool Arbor, which is freely accessible and can be integrated into educational data analysis platforms such as CODAP (http://codap.concord.org/). Arbor is designed to help understanding about construction and interpretation of decision trees and supports critical evaluation of the strength and pitfalls of tree methods.

TREES AS ROBUST DECISION AND ESTIMATION TOOLS

Tree-based methods for exploratory data analysis are a child of the computer age. Since the inception of the CART algorithm in 1984 by Breiman, Friedman, Olshen and Stone (1984), trees have become a popular tool among biostatisticians and social scientists. Trees are intuitive, conceptually easy to understand, and result in nice representations of highly complex datasets. Trees form the foundation for deeper statistical learning methods in data science under such sonorous names as bagging, boosting, and random forests (Hastie et al., 2006).

CART is based on a generic binary decision tree which proceeds by recursively subdividing the sample space into two subsets. Trees are associated with recursive partitioning schemes (RPS). They proceed by successively partitioning the sample space into finer sets where the data are in some respect more homogeneous. The partitioning scheme is associated with a binary tree. The nodes of the tree correspond to partition sets. If the set divided is into two parts in the process of recursive partitioning, then the corresponding “branching” node has two “child” nodes. The terminal nodes of a tree correspond to final partition sets and are called *leaves*. Tree construction proceeds by recursively partitioning the data space in increasingly homogeneous

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

subsets and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree.

Figure 1 shows an example, based on the 2006 German Income Structure data provided by the German Statistics Office which covers microdata from a random selection of 59,504 adult employees and 16 variables (such as monthly income, gender, type of job, region, etc.). For purposes of illustrating the equivalence of recursive partitioning and trees we restrict ourselves to the two covariates age (`AGE`) and weekly working hours (`WorkHrs`). Feeding the data into the `R` package `rpart` results in the tree shown on the right of Figure 1. While the overall average of the variable hourly pay for 59012 employees is `HourlyPay`=16.235 €, subdividing the whole sample by age (depending if `AGE` < 26.5 years) results in two subgroups. While the younger employees earn only 9.45 Euro per hour on average, the older people have a mean hourly pay of 17.55 Euro. Terminal nodes are marked as rectangles. The left panel illustrates the scheme by displaying the final splits obtained after recursively dividing up the sample space. While earnings apparently increase with age, observe that the relationship with hourly pay and the number of weekly working hours seems more intricate. It appears that more weekly working hours results in a higher hourly pay, but this relationship seems to be reversed for those employees above the age of 56 who work at least 15 hours. Notice that the tree representation uncovers an interaction. The resulting tree can be seen as a multivariate function which is locally constant on multivariate rectangular sets of the sample space; in the example `HourlyPay = ` *f* (`AGE`, `WorkHrs`). Thus, the tree is a *nonparametric* function estimator, i.e., the dependency between predictor and response is described without assuming any specific parametric functional model as is the case with linear regression.

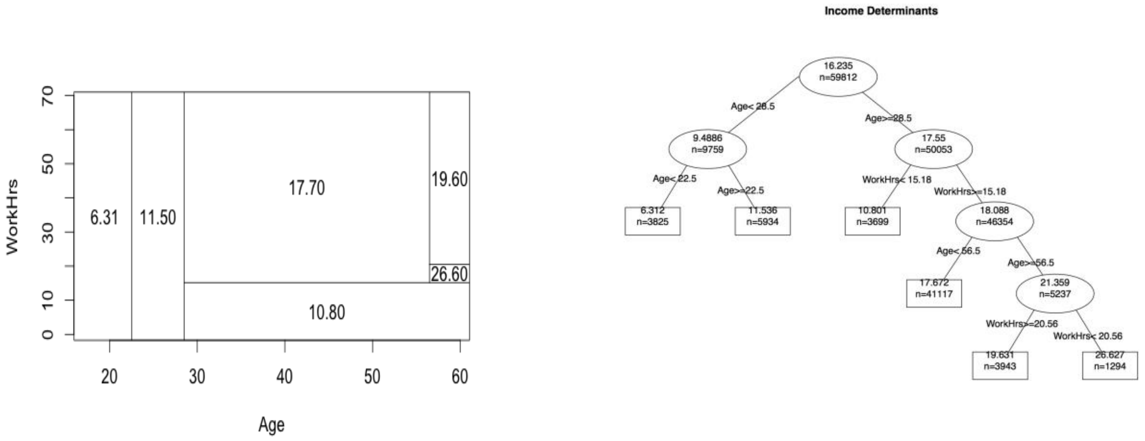


Figure 1: Partition of the German Income Structure Data set (restricted to three continuous variables) and the associated tree

Three important questions that apply equally to classification and regression trees, i.e. to categorical and metric response variables, define the tree methodology:

1. How and where to split a partition set?
2. When to stop the tree growing procedure?
3. How to assign an estimate to every terminal node?

1. For the purpose of choosing a good split, a “purity measure” needs to be defined. Then the goodness of a split is evaluated by how much the overall purity in the daughter nodes is improved over the parent node. One possible (but by far not the only reasonable) choice is the misclassification rate in case of classification and the sum of within-nodes variance for regression trees.
2. We need some type of stopping rule. If we were to allow the tree to become too big, the resulting tree will not perform well on new data. This is similar to choosing too high a degree in polynomial regression. Therefore, we need a stopping criterion to halt the recursive partitioning process. Just like possible purity measures, stopping criteria come in many different forms, including:

  - A maximum number of nodes in the tree: Once this maximum is reached, the process is halted.
  - A minimum number of observations for each node: partitioning of nodes stops if the number of observations goes below a certain number.
  - A threshold for the purity reduction: if the purity improvement with further splitting would be smaller than the threshold, then stop.
  - Powerful algorithms such as `CART` or the `rpart` package in `R` follow a more sophisticated approach: After growing an oversize tree, they prune the tree based on cross-validation.

3. The third question has a straightforward and obvious solution: Calculate a local average or any other location parameter estimate (local median, trimmed local average etc.) over those responses whose predictor belongs to the terminal node under consideration (regression) or classify an observation belonging to the terminal node by a local majority vote (classification).

To summarize,

- Trees can handle a broad range of response types (numeric, rating, categorical) and missing values.
- Trees can be displayed graphically, and are easily interpreted even by a non-expert.
- Trees mirror human decision-making more closely than do sophisticated traditional statistical methods like multivariate regression (Martignon et al., 2003)
- The tree method is very robust, i.e., consistency of the estimation procedure can be shown under very weak assumptions, in contrast to many classical methods that assume linearity and normally-distributed data (Gordon & Olshen, 1980).
- Unfortunately, trees generally do not have the same level of predictive accuracy as some more traditional methods. But for small or moderate training sets simple trees as those presented here or also fast and frugal trees (Martignon et al., 2003) outperform traditional approaches (Leskey & Martignon, 2014)

Thus, trees represent an alternative to many traditional statistical techniques, including multiple regression, analysis of variance, logistic regression, log-linear models, and linear discriminant analysis.

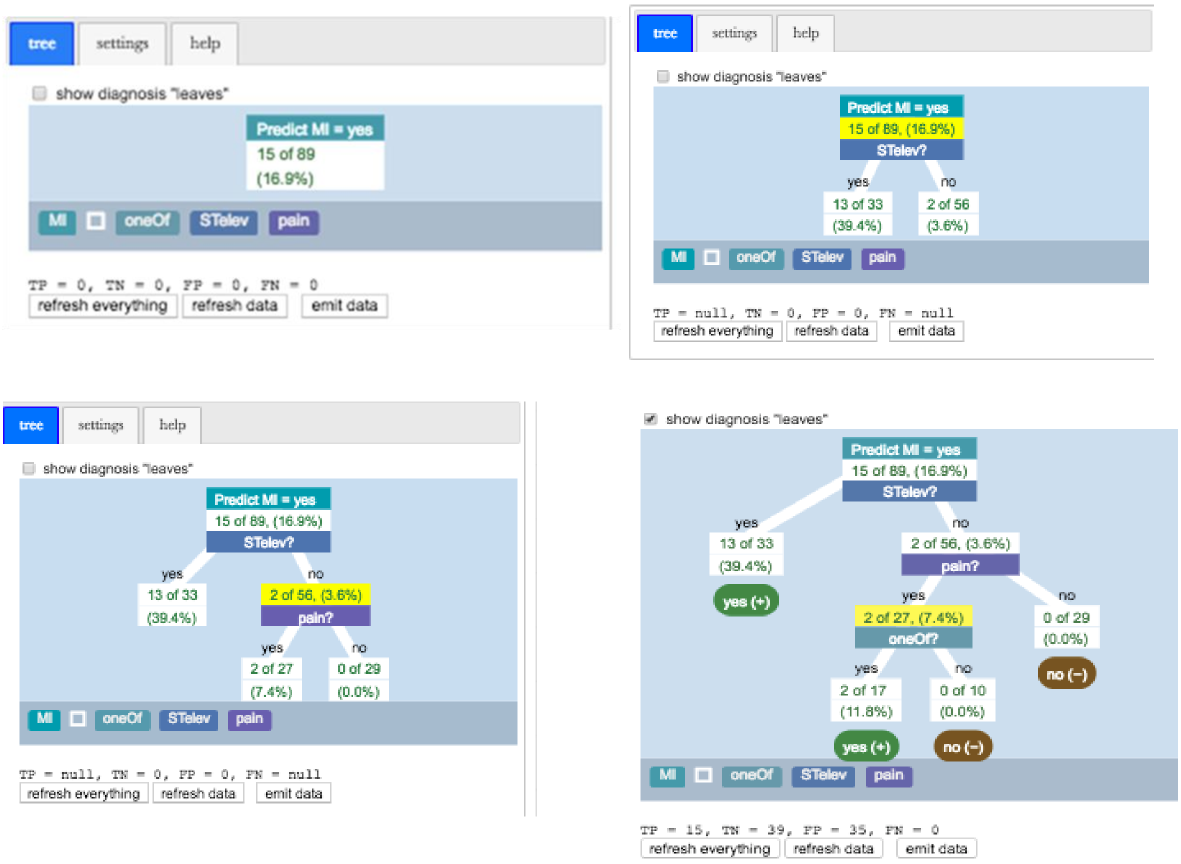
LEARNING ABOUT TREES

Understanding multivariate data is substantially supported by modern statistical methods and modes of presentation. Arbor, developed as a digital learning tool by the second author during his visit to Ludwigsburg in fall of 2017, is a plug-in to the freely available data science education platform CODAP (Finzer, 2017). It is designed to familiarize learners with the powerful and versatile tree method for data exploration and prediction. Unlike the implementation of trees in software for professional data analysis, Arbor has no automatized algorithms that compute optimal splits and right-sized trees. The user, through drag and drop moves, decides step by step which variables to use for consecutive splits, how to specify the split, and when to terminate tree growth. Statistics (e.g., misclassification rates for classification trees or sums of squares in the case of regression trees) evaluating the goodness of the chosen split are immediately reported, which allows comparison with alternative splits. The purpose of Arbor is not the derivation of an optimal tree, but to let the user explore the flexibility of the tree method and the consequences of various splits, thus to gain appreciation for trees as a way to represent complex data. Arbor supports the understanding of how classification and regression trees are constructed, thereby facilitating the interpretation of tree structures. While trees for classification and for regression have much in common, there are also distinct differences in the implementation. Therefore, in the following we give a rough account of these two applications along two examples. For more in-depth exploration we encourage the reader to follow the link given at the end of this paper and try out the following and other examples.

EXAMPLE 1: MEDICAL DIAGNOSIS

In a medical emergency room incoming patients with certain symptoms have to be classified as high risk patients for myocardial infarct (YES or NO)) or not in order to be assigned either to the coronary care unit or to a regular nursing bed. Available clues (with the possible response: YES or NO) include (1) ST segment elevation `STelev` in the electrocardiogram (ECG), (2) severe chest pain `pain` (3) any one or more from a list of 5 possible symptoms `oneOf`. Notice that the response variable (and all the predictors) are categorical, hence we construct a classification tree. Based on a training sample of 89 patients (whose physical conditions on above clues is known as well as if they in fact had a heart attack or not) a classification and decision tree is then constructed. Obviously, there are two types of misclassifications : a patient who will have a myocardial infarction may wrongfully be placed in an ordinary hospital ward, or a patient without infarct will be sent to the coronary care unit—obviously with different consequences. While the capacity of the CCU may be limited, sending an infarct patient to the ordinary ward may have the patient’s death as its consequence.

Figure 2 displays consecutive steps for tree construction with Arbor. We begin by seeing only the root node (upper left). In the upper right, we drop the variable `STelev` onto the root node, splitting it into two child nodes. Then we drop `pain` onto the “no STelev” node (Lower left). In the last illustration, we have included the third predictor variable and assigned diagnoses to the leaves Also notice the misclassification of the training sample data (TP=True positive, TN=True Negative, FP=false positive, and FN=false negative). For further exploration, we recommend the reader visit https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=33781



```
TP = 0, TN = 0, FP = 0, FN = 0
```

```
TP = null, TN = 0, FP = 0, FN = null
```

```
TP = null, TN = 0, FP = 0, FN = null
```

```
TP = 15, TN = 39, FP = 35, FN = 0
```

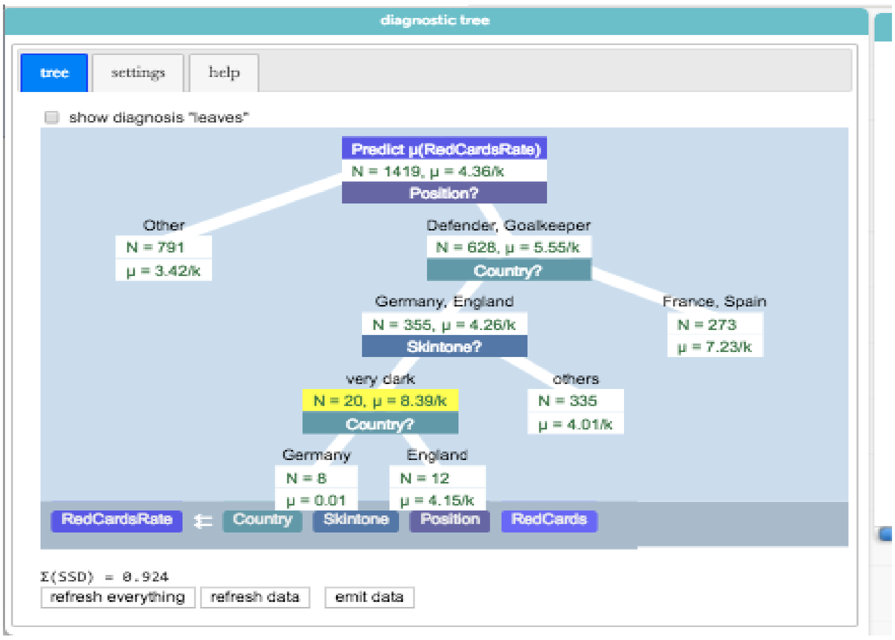
Figure 2: Successive steps in construction of a classification tree with Arbor, to be read from upper left (root node , indicating that 15 out of 89 patents in the training sample have myocardial infarct) to the final tree with terminal node (lower right; notice the true/false & positive/ negative counts in the final panel)

EXAMPLE 2: ARE REFEREES IN EUROPEAN SOCCER RACIALLY BIASED?

Soccer, the most popular mass spectator sport in the world, is a game where humanity comes alive. Behind the façade of its obvious entertainment aspect, it has proved to reflect cultural nationalism, communal identity, and cultural specificity. But one of the ugly aspects of its vast popularity is the fact that spectators can be racist, homophobic, and discriminatory—especially when it comes down to unsettling the away team. But what about the referees? Are they applying the rules in a fair manner, irrespective of any personal or demographic characteristics of the players?

A rich dataset contains demographic information with 23 variables about soccer players (n=1419) playing in the first male divisions of England, France, Germany, and Spain (Silverzahn et al., 2014). The data include referee calls, player demographics such as team position, nation league, height, weight and more. It also includes a rating of players' skin colour which was coded by two independent raters on a 5-point scale ranging from 0=very light skin to 4=very dark skin. Obviously, there are relevant covariates that may determine a player's likelihood to receive a red card. A serious analysis has to take into account, for example, the variable POSITION. Defenders may receive many more cards than the average forward since their role of preventing the other teams from scoring may cause them to make more fouls. Other possibly relevant variables to consider are a player's physical condition (i.e., weight and height) which may be confounded with SKINTONE as darker players may be stronger or weaker and heavier players may be a bit slower in their motions and hence seem to have a higher chance of getting a red card. Another relevant variable to be controlled for may be the league COUNTRY. SKINTONE is not evenly distributed across league countries, and in some leagues the rules may be applied more strictly than in others, a cause for potential bias due to a different distribution of skin tone and the different distribution of red cards within each league.

Figure 3 shows a regression tree created with Arbor showing the dependency of the number of red cards per game ("RedCardsRate") in European football on the covariates skin color, league (England, France, Germany or Spain) and the position of the player. For further details of tree-structured statistical methods, reference should be made to the literature (Breiman et al 1984, Hastie et al., 2006).



```
Σ(SSD) = 0.924
```

Figure. 3: Regression tree for the number of red cards per game depending on the skin color, the player's position and the country league (created with Arbor, a plug-in tool for CODAP)

While (possibly weighted) misclassification rate is a natural purity measure for classification trees, how shall we measure purity in the regression case? One possible choice is the relative sum of square deviation (SSD) within nodes, a measure well known from analysis of variance. For tree growing, this implies to search for the split that results in maximum decrease of Sum of Squares. In the case of the tree in Figure 3, we see that we have explained only about 7.6% of the variance with our tree model.

SUMMARY

Trees are a modern tool for classification and nonparametric regression, suitable for large mixed-type multivariate data. The digital learning tool Arbor is designed to familiarize learners with this powerful and versatile method for data exploration and prediction. Instead of relying on powerful algorithms, it requires the user to make successive choices about which variable to split, how to split, and when to stop growing the tree. By using Arbor, students can come to understand what the algorithms accomplish, and perhaps even more to the point, come to understand the nature of trees themselves.

Arbor is a free plug-in to CODAP, the freely available data analysis platform (Finzer, 2017). The reader can explore Examples 1 and 2 using CODAP-based electronic documents with additional technical descriptions:

Ex 1: https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=35800
Ex 2: https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=35801

A more detailed description of these two examples can be found under www.procivicstat.org

ACKNOWLEDGMENT

The work reported in this paper was supported in part by the ProCivicStat project, a strategic partnership of the Universities of Durham, Haifa, Ludwigsburg, Paderborn, Porto and Szeged, funded by the ERASMUS+ program of the European Commission. However the views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the funding agency.

REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984). *Classification and regression trees*. Belmont, California: Wadsworth.

Finzer, W. (2017). Common Online Data Analysis Platform. https://codap.concord.org

Gordon, L. and Olshen, R. (1980). Consistent nonparametric regression from recursive partitioning schemes. *J. Mutivariate Analysis*, *10*, 611- 627

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference and Prediction* (2. Ed.). New York: Springer.

Laskey, K. & Martignon, L. (2014). Comparing fast and frugal trees and baysian networks for risk assessment. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute.

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. (2003). Naïve and yet enlightened: from natural frequencies to fast and frugal decision trees. In: D. Hardman, L. Macchi (Eds), *Psychological Perspectives on Reasoning, Judgment and Decision Making*. Wiley

Silberzahn, R. Uhlmann, E.L, Martin, D. & Nosek, B. (2014). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. Online: https://osf.io/47tnc/ (retrieved March 11, 2018)