An introduction to conditional inference trees in R

Assoc. Prof. Dr. Martin Schweinberger

University of Queensland | Arctic University of Norway, Tromsø

About me

Quantitative corpus-linguist

- Lecturer in Applied linguistics at the University of Queensland (UQ)
- Associate Professor II at the Arctic University of Norway in Tromsø (UiT)
- Co-Director of the Language Technology and Data Analysis Laboratory (LADAL) at UQ
- Principal Data Science Advisor to the AcqVA Aurora Lab at UiT
- Projects: Australian Text Analytics Platform (ATAP) | Language Data Commons of Australia (LDaCA)
- Studied Philosophy, English Philology and Psychology in Kassel and Galway
- PhD in Hamburg, Post docs (or similar positions) at the Linguistic Diversity in Urban Areas Excellence Cluster (LiMA), the FU Berlin, the Universities of Greifswald, Luneburg, and Kassel

AcqVA Aurora Lab

atap australian text analytics platform atap.edu.au

LADAL

LDaCA

About the workshop

Timeline | Table of Contents

- Introduction
- When to use trees
- What are pros and cons?
- Case study | Practice
- Outro

During practice, we will use a Jupyter notebook and you can

1. sit back and follow
2. you can practice using the data provided by me
3. you can try and use your own data (but I cannot help you in modifying the code)

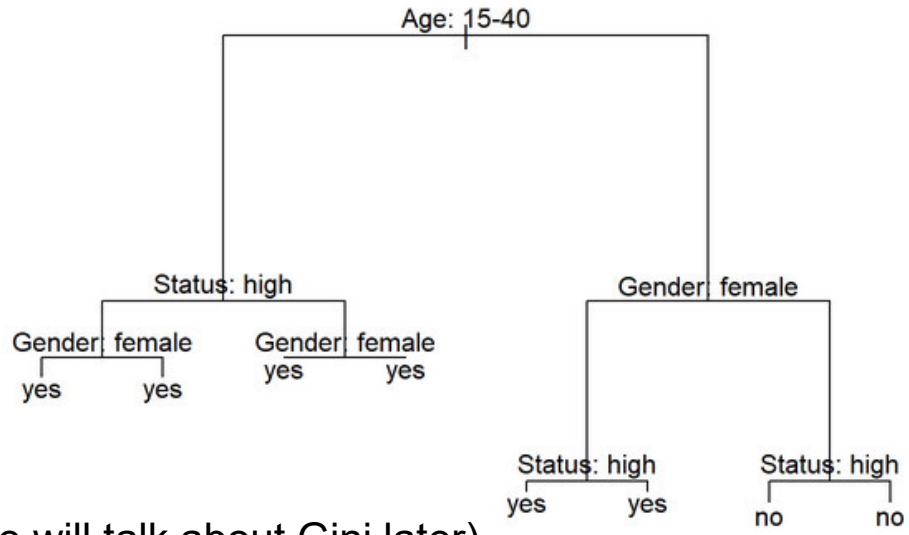
Introduction: what are tree-based models?

Tree-based models are a multivariate statistical method used in machine learning and it is now becoming more common in the language sciences. It is non-parametric and thus relies on few distributional requirements, it is easy to use, and it takes any type of dependent and predictor variables. It works by recurrent partitioning (splitting) of the data.

Better than chi-square tests because

- it is multivariate
- is robust
- takes all types of variables.

Example



Classification and Regression Trees

- **Decision tree**
  all possible nodes (outcomes) are shown
- **Classification and Regression Tree**
  only splits if justified by information measure (e.g., Gini; we will talk about Gini later)
- **Conditional Inference Tree**
  only splits based on statistical test (splits represent statistical difference)

Basic principle

Take a data set and see if there is a maximally distinct distribution of the dependent variable if we split the data set on a level of any of the independent variables (repeat until the resulting distribution is no longer different)

Think break

Let's assume we have 10,000 adults and we want to split them into men and women: what would be the best variable to split them by (assuming we can only go be looks including height)?

When to use tree-based models

- You have one dependent variable (outcome | response)
- You have more than one independent variable (predictor)
- You expect there to be interactions between independent variables (effects or not only additive)
- You are interested in how predictors impact a response (outcome)
- Your categorical variables have less than 54 levels
- Effects are non-linear
- The most important effect is not an interaction

Variable types

- Nominal (two levels, e.g. yes, no)
- Categorical (more than 2 levels, e.g., nationalities)
- Ordinal (ranked or ordered categories, e.g., disagree, neutral, agree)
- Numeric
  - Counts (e.g., 2, 15, 212)
  - True numeric (e.g., -5.8, 1.4, 113.0)

Pros

Several advantages have been associated with using tree-based models:

1. Tree-structure models are very useful because they are **extremely flexible** as they can deal with different types of variables and provide a very good understanding of the structure in the data.
2. Tree-structure models have been deemed particularly interesting for linguists because they can handle moderate sample sizes and many high-order interactions better then regression models.
3. Tree-structure models are (supposedly) better at detecting non-linear or non-monotonic relationships between predictors and dependent variables. This also means that they are better at finding and displaying interactions involving many predictors.
4. Tree-structure models are easy to implement in R and do not require the model selection, validation, and diagnostics associated with regression models.
5. Tree-structure models can be used as variable-selection procedure which informs about which variables have any sort of significant relationship with the dependent variable and can thereby inform model fitting.

Cons

Despite these potential advantages, a word of warning is in order. For instance,

1. Simple tree-structure models have been shown to fail in detecting the correct predictors if the variance is solely determined by a **single interaction** (Gries 2021, chap. 7.3). This failure is caused by the fact that the predictor used in the first split of a tree is selected as the one with the strongest main effect (Boulesteix et al. 2015, 344). This issue can, however, be avoided by hard-coding the interactions as predictors plus using ensemble methods such as random forests rather than individual trees (see Gries 2021, chap. 7.3).

2. Another shortcoming is that tree-structure models partition the data (rather than "fitting a line" through the data which can lead to more coarse-grained predictions compared to regression models when dealing with numeric dependent variables (again, see Gries 2021, chap. 7.3).

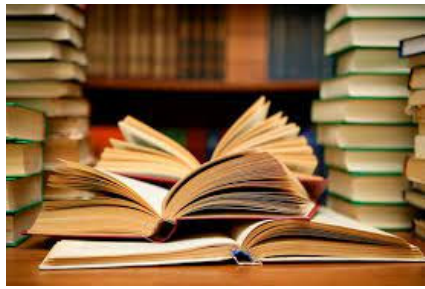
3. Boulesteix et al. (2015, 341) state that high correlations between predictors can hinder the detection of interactions when using small data sets. However, regression do not fare better here as they are even more strongly affected by (multi-)collinearity (see Gries 2021, chap. 7.3).

4. Tree-structure models are bad a detecting interactions when the variables have strong main effects which is, unfortunately, common when dealing with linguistic data (Wright, Ziegler, and König 2016).

5. Tree-structure models cannot handle factorial variables with many levels (more than 53 levels) which is very common in linguistics where individual speakers or items are variables.

Let’s now learn how to implement a conditional inference tree in R
Thank you very much!

Assoc. Prof. Dr. Martin Schweinberger



References

Boulesteix, Anne-Laure, Silke Janitza, Alexander Hapfelmeier, Kristel Van Steen, and Carolin Strobl. **2015**. “Letter to the Editor: On the Term ‘Interaction’ and Related Phrases in the Literature on RandomForests.” *Briefings in Bioinformatics* 16 (2): 338–45. https://academic.oup.com/bib/article/16/2/338/246566.

Gries, Stefan Th. **2021**. *Statistics for Linguistics Using r: A Practical Introduction*. Berlin & New York: Mouton de Gruyter.

Wright, Marvin N., Andreas **Ziegler**, and Inke R. **König**. **2016**. “Do Little Interactions Get Lost in Dark Random Forests?” 17 (145). https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-0995-8.

Let’s now learn how to implement a conditional inference tree in R

An introduction to conditional inference trees in R

Assoc. Prof. Dr. Martin Schweinberger

University of Queensland | Arctic University of Norway, Tromsø