Recursive Partitioning and Decision Trees

Malte Schierholz

University of Mannheim, MZES
Institute for Employment Research (IAB)

Malte.Schierholz@iab.de

March 21 and 22, 2018

Overview

Introduction

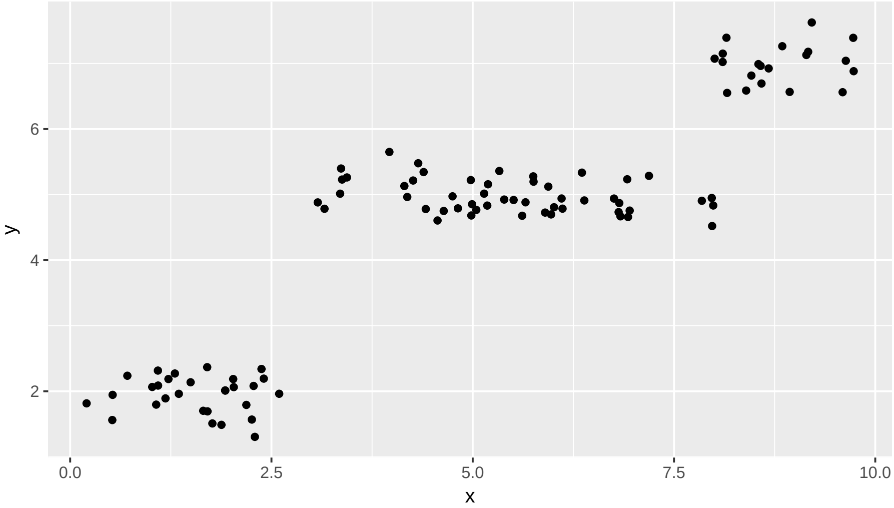
What are decision trees?

- very popular data mining algorithm
- describe the relationship between predictors X and outcome Y
- building block for more complex machine learning algorithms

History:

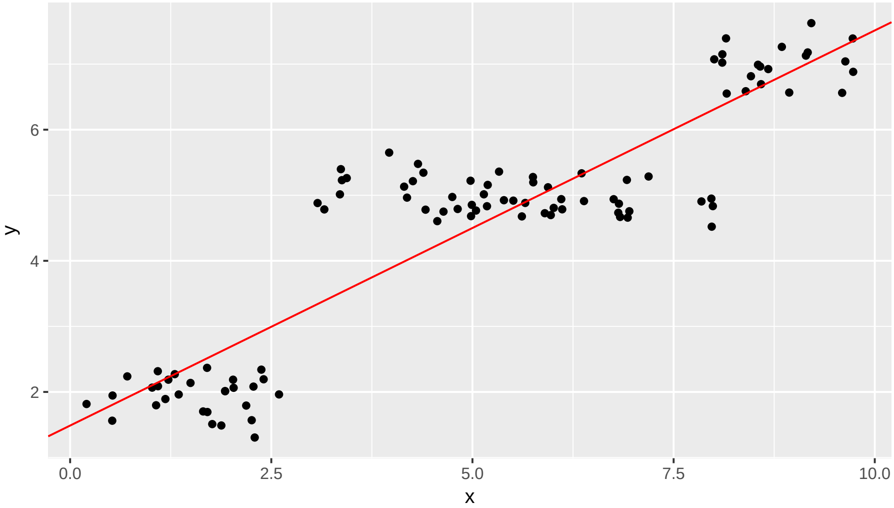
- first suggested by Morgan and Sonquist (1963) to detect interactions
- popularized by Breiman et al. (1984) with a focus on prediction
- today several dozen algorithms fall under this label for a wide variety of problems (Zhang & Singer, 2010; Loh, 2014)

The Basic Idea



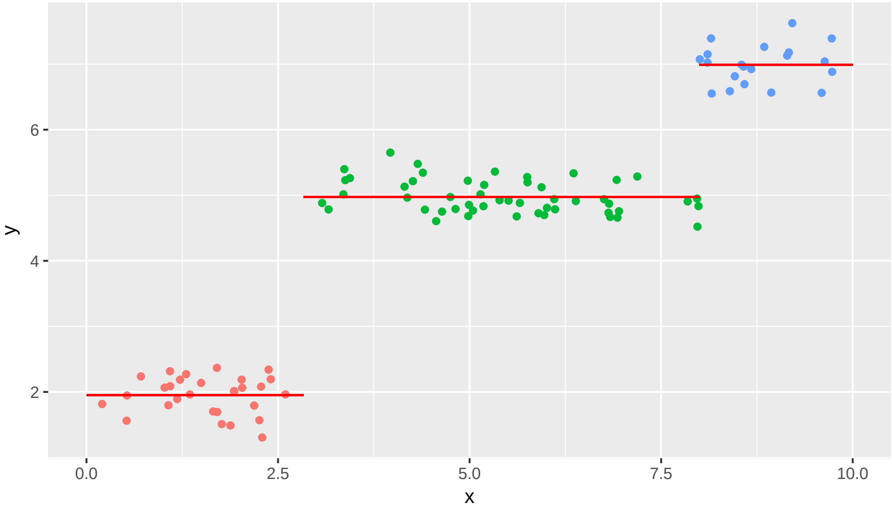
Simulated data: $y = 2 \cdot I(x < 3) + 5 \cdot I(x > 3 \& x < 8) + 7 \cdot I(x > 8) + \epsilon$

The Basic Idea



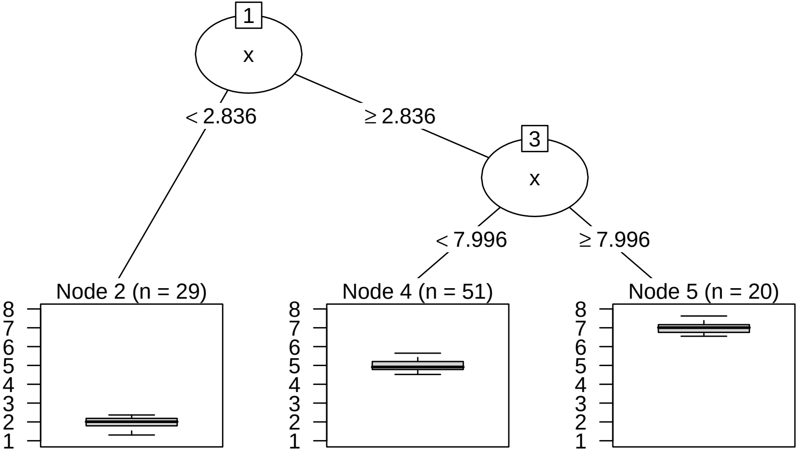
Is least squares regression really appropriate?

The Basic Idea



Split data into groups! This requires a search for optimal splitting points.

A Regression Tree



Tree representation is intuitive and easy to interpret.

Classification Trees

Can decision trees classify different Iris species (= Arten von Schwertlilien) based on the length and width of their flowers?

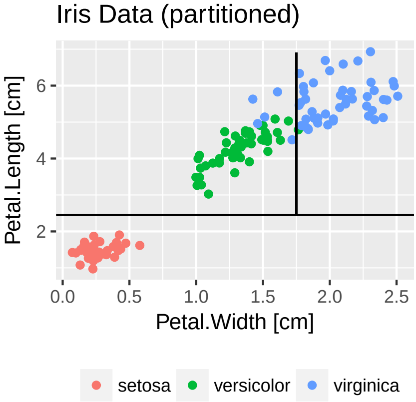
Iris setosa

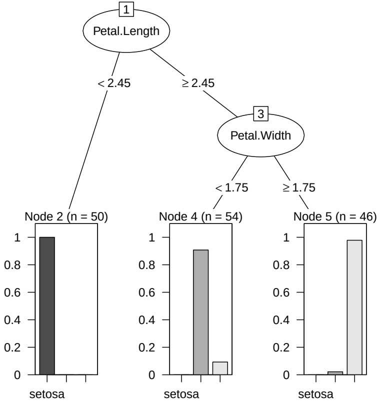
Iris versicolor

Iris virginica

Source: Wikipedia
Iris setosa CC BY-SA 3.0, [1]
Iris versicolor by Danielle Langlois, CC BY-SA 3.0, [2]
Iris virginica by Frank Mayfield - originally posted to Flickr as Iris virginica shrevei BLUE FLAG, CC BY-SA 2.0, [3]

Classification Trees



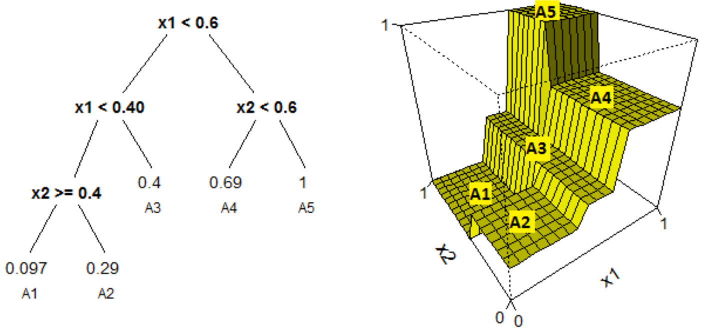


Predictions in each subgroup can be

- the majority class
- relative frequencies of each class

Tree Representations

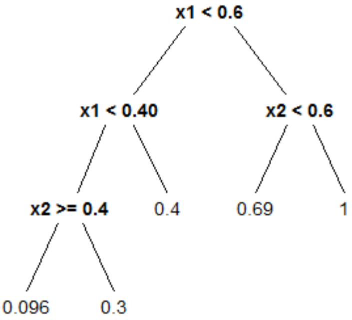
Different Representations for Trees



Left, a hypothetical regression tree.
Right, the corresponding regression surface. If subgroups were known, we could estimate this by least squares regression,

$$y = \beta_1 \cdot I(X_1 < 0.4 \& X_2 \geq 0.4) + \beta_2 \cdot I(X_1 < 0.4 \& X_2 < 0.4) + ... + \epsilon$$

Different Representations for Trees



IF $X_1 < 0.4$ AND $X_2 \geq 0.4$
THEN $Y \leftarrow 0.096$
IF $X_1 < 0.4$ AND $X_2 < 0.4$
THEN $Y \leftarrow 0.3$
IF $X_1 < 0.6$ AND $X_1 \geq 0.4$
THEN $Y \leftarrow 0.4$
IF $X_1 \geq 0.6$ AND $X_2 < 0.6$
THEN $Y \leftarrow 0.69$
IF $X_1 \geq 0.6$ AND $X_2 \geq 0.6$
THEN $Y \leftarrow 1$

Left, a hypothetical regression tree.
Right, the corresponding logical rules.

Decision Trees vs Rule Sets

Rules derived from decision trees

- are non-overlapping and
- cover the complete predictor space.

Rule learning can also proceed without these constraints.

Example:

IF $X_1 < 0.2$ THEN *Approved* $\leftarrow$ yes
IF $X_2 < 0.8$ THEN *Approved* $\leftarrow$ no
DEFAULT THEN *Approved* $\leftarrow$ yes

Rule set induction is not covered in this course. See Fürnkranz et al. (2012) for an overview.

Algorithms

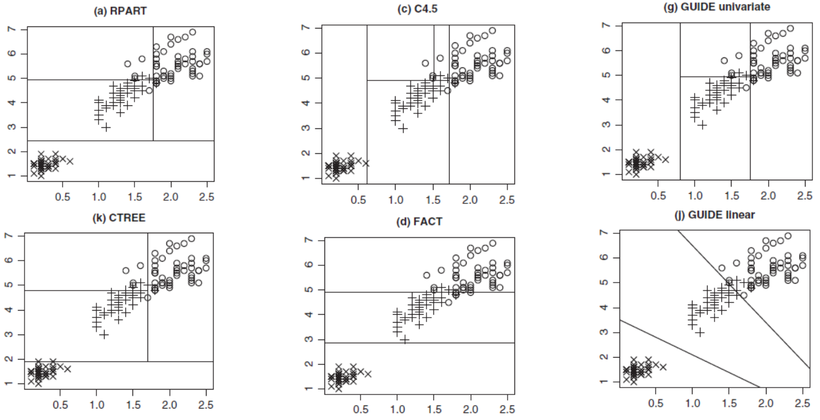
General Algorithm

Algorithm 1: Tree growing process

```
Define stopping criteria;
Assign training data to root node;
if stopping criterion is reached then
    end splitting;
else
    find the optimal split point (variable and threshold);
    split node into two subnodes at this split point;
    for each node of the current tree do
        continue tree growing process;
    end
end
(optional) Prune tree
```

Algorithmic Variations

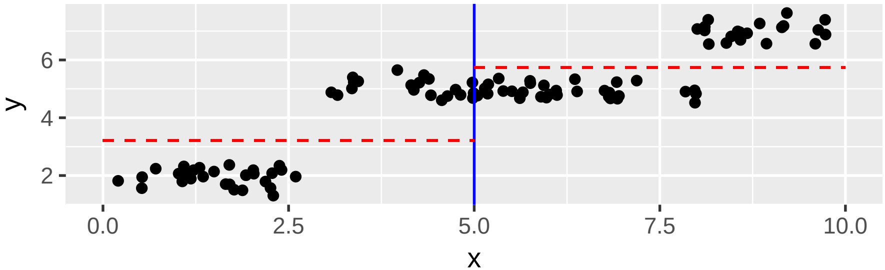
Many different algorithms exist ⇒ Many different partitions of Iris data



(Source: Loh 2014, p. 336)

Algorithms

Split point discovery in CARTs (rpart)



For each group (k = left / right) calculate "Impurity" (= residual sum of squares in regression problems),

$$m_k = \frac{1}{N_k} \sum_{i \in \text{group}_k} y_i \qquad s_k^2 = \sum_{i \in \text{group}_k} (y_i - m_k)^2 \tag{1}$$

Choose splitting variable X_i and threshold t which minimize $s^2_{\text{left}} + s^2_{\text{right}}$

Split point discovery in CARTs (rpart)

Same idea for classification problems (2 classes here): *Define impurity* I_k based on proportion of positives $\hat{p}_k$, ($k =$ left / right)

Split point discovery in CARTs (rpart)

Same idea for classification problems (2 classes here): *Define impurity* I_k based on proportion of positives $\hat{p}_k$, (k = left / right)

Possible impurity functions:

Missclassification error:

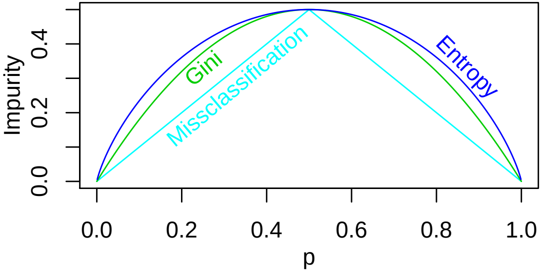
$$I_k = 1 - \max(\hat{p}_k, 1 - \hat{p}_k)$$

Gini index:

$$I_k = 2\hat{p}_k(1 - \hat{p}_k)$$

Entropy:

$$I_k = -\hat{p}_k \log \hat{p}_k - \ldots$$



Split point discovery in CARTs (rpart)

Same idea for classification problems (2 classes here): *Define impurity* I_k based on proportion of positives $\hat{p}_k$, (k = left / right)

Possible impurity functions:

Missclassification error:

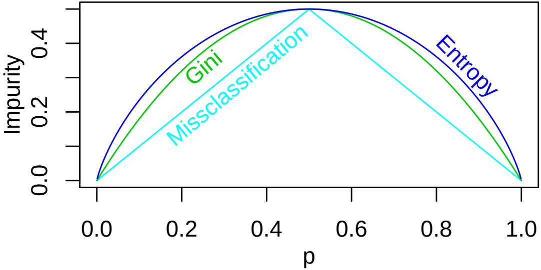
$$I_k = 1 - \max(\hat{p}_k, 1 - \hat{p}_k)$$

Gini index:

$$I_k = 2\hat{p}_k(1 - \hat{p}_k)$$

Entropy:

$$I_k = -\hat{p}_k \log \hat{p}_k - \ldots$$

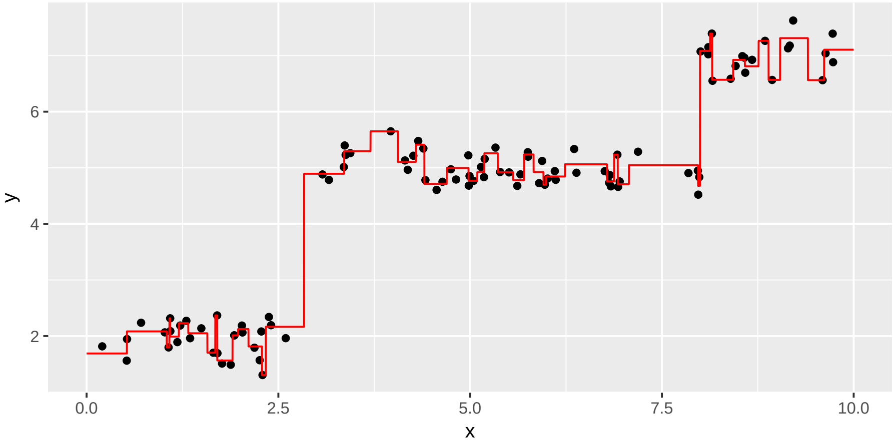


Choose split point that minimizes the (weighted) sum of impurities

$$N_{\text{left}} \cdot I_{\text{left}} + N_{\text{right}} \cdot I_{\text{right}}$$

with weights N_k, the number of cases that will be send to the kth node

Overfitting



- Overfitting = Poor generalization to new data
- Function approximates training data well, but the number of terminal nodes is high ($|\mathcal{T}| = 47$) (Trade-off: Model fit $\leftrightarrow$ model complexity)

Tree pruning in CARTs (rpart)

Stopping rules

- Minimum number of cases in terminal nodes
- Improvement below some threshold

$\rightarrow$ However, worthless splits can be followed by good splits

Tree pruning in CARTs (rpart)

Stopping rules

- Minimum number of cases in terminal nodes
- Improvement below some threshold

$\rightarrow$ However, worthless splits can be followed by good splits

Cost complexity pruning

Balance tree quality $SSE(\mathcal{T}) = \sum(y_i - \hat{y}_i(\mathcal{T}))^2$ and tree size $|\mathcal{T}|$.
Remove internal nodes until we find a subtree $\mathcal{T}_\alpha$ which minimizes

$$C_\alpha(\mathcal{T}) = SSE(\mathcal{T}) + \alpha|\mathcal{T}|$$

- α controls the penalty on the number of terminal nodes
- α can be chosen through cross-validation

Coding Session

Coding session with

- Tree building with `rpart`
- Cost-complexity pruning
- Prediction
- Using misclassification costs and priors
- Missing data
- Binary outcome

Not shown:

- Many outcomes other than binary are possible: ordinal, continuous, multivariate, censored, longitudinal (see Zhang (2010) and Loh (2014) for details)

Algorithms

1. Introduction

2. Tree Representations

3. Algorithms

- Classification and Regression Trees (Breiman et al. 1984)
- Model-based Recursive Partitioning (Zeileis et al. 2008)

4. Cautionary Notes and Suggested Applications

Model-based Recursive Partitioning

Idea:

- Parametric model is specified in advance, e.g.:
  - Constant leaves, normally distributed: $y \sim N(\theta, \sigma^2)$
  - Least Square Regression: $y|x \sim N(x^T\theta, \sigma^2)$
  - logit, glm, survival, ...

Model-based Recursive Partitioning

Idea:

- Parametric model is specified in advance, e.g.:
  - Constant leaves, normally distributed: $y \sim N(\theta, \sigma^2)$
  - Least Square Regression: $y|x \sim N(x^T\theta, \sigma^2)$
  - logit, glm, survival, ...
- Three types of variables:
  - Response variable y
  - Predictor variables $x_1, ..., x_p$ (relationship with Y known from theory)
  - Partitioning variables $z_1, ..., z_j$ (relationship with Y unclear)

Model-based Recursive Partitioning

Idea:

- Parametric model is specified in advance, e.g.:
  - Constant leaves, normally distributed: $y \sim N(\theta, \sigma^2)$
  - Least Square Regression: $y|x \sim N(x^T\theta, \sigma^2)$
  - logit, glm, survival, ...
- Three types of variables:
  - Response variable y
  - Predictor variables $x_1, ..., x_p$ (relationship with Y known from theory)
  - Partitioning variables $z_1, ..., z_j$ (relationship with Y unclear)
- Should one partition the n observations with respect to partitioning variables $z_1, ..., z_j$ and estimate separate models $y|x$ within each partition?

Model-based Recursive Partitioning

Idea:

- Parametric model is specified in advance, e.g.:
  - Constant leaves, normally distributed: $y \sim N(\theta, \sigma^2)$
  - Least Square Regression: $y|x \sim N(x^T\theta, \sigma^2)$
  - logit, glm, survival, ...
- Three types of variables:
  - Response variable y
  - Predictor variables $x_1, ..., x_p$ (relationship with Y known from theory)
  - Partitioning variables $z_1, ..., z_j$ (relationship with Y unclear)
- Should one partition the n observations with respect to partitioning variables $z_1, ..., z_j$ and estimate separate models $y|x$ within each partition?
- Many possible applications:
  - Model exploration and checking,
  - Determine if separate models for subsets are needed, ...

Algorithm

Algorithm 2: Recursive partitioning with GLMs

Parameter : *p*-value threshold
Initialization: Fit initial model using all observations

Test for parameter instability (M-fluctation test) on each partitioning variable;
if *minimum p-value exceeds threshold* **then**
    end partitioning;
else
    choose partitioning variable associated with the smallest *p*-value;
    find the optimal split point;
    split node into two subnodes at this split point;
    for *each node of the current tree* **do**
        continue partitioning process;
    end
end

Coding Session

Parameter instability tests are essential:

- Tests if θ is constant over all partitioning variables Z,
- or if θ varies among Z, $\theta = f(Z)$

Coding Session

Parameter instability tests are essential:

- Tests if θ is constant over all partitioning variables Z,
- or if θ varies among Z, $\theta = f(Z)$

Coding session with

- Model-based Recursive Partitioning with `partykit`
- Explanation of the algorithm

Tree Representations

1. Introduction
2. Tree Representations
3. Algorithms
4. Cautionary Notes and Suggested Applications

Interpretability

Interpretation of trees is easy! Really???

Interpretability

Interpretation of trees is easy! Really???

- Instability of tree structure: With small changes in the training data, different splitting points might get selected and changes propagate through the complete tree

Interpretability

Interpretation of trees is easy! Really???

- Instability of tree structure: With small changes in the training data, different splitting points might get selected and changes propagate through the complete tree
- Inference unclear: How to calculate standard errors and confidence intervals?

Interpretability

Interpretation of trees is easy! Really???

- Instability of tree structure: With small changes in the training data, different splitting points might get selected and changes propagate through the complete tree
- Inference unclear: How to calculate standard errors and confidence intervals?
- Missing predictors: Correlated predictors compete for the same split and one significant predictor may keep away another

Interpretability

Interpretation of trees is easy! Really???

- Instability of tree structure: With small changes in the training data, different splitting points might get selected and changes propagate through the complete tree
- Inference unclear: How to calculate standard errors and confidence intervals?
- Missing predictors: Correlated predictors compete for the same split and one significant predictor may keep away another
- Biased variable selection: variables with many categories, numeric variables, or variables with many missings are preferred for splitting (solved in the `partykit`-package)

(Strobl et al. 2009, Loh and discussants 2014)

Prediction

Most of the literature on decision trees emphasizes medical diagnosis and prediction. Examples from Breiman (1984):

- Diagnosis: Based on patients' reports of chest pain and medical measurements, which ones are having a heart attack?
- Prediction: After having a heart attack, which groups of patients are under high risk of dying within the next 30 days?

Key advantage: Tree-based predictions are easy to comprehend

Prediction

Most of the literature on decision trees emphasizes medical diagnosis and prediction. Examples from Breiman (1984):

- Diagnosis: Based on patients' reports of chest pain and medical measurements, which ones are having a heart attack?
- Prediction: After having a heart attack, which groups of patients are under high risk of dying within the next 30 days?

Key advantage: Tree-based predictions are easy to comprehend

But:

- The prediction surface is not smooth, but a step function. An obstacle for optimal prediction ...
- More complex algorithms (e.g., bagging, boosting) often outperform simple decision trees

Main application: as a building block within more complex algorithms

Decision Trees in Survey Research

Survey researchers showed interest in “Automatic interaction detection” (=decision trees) long before Breiman (1984) popularized them for predictive tasks.

Original intention from Morgan and Sonquist (1963):

- Interaction effects in regression models are bothersome. Existing alternatives:
  - Specify possible interaction terms in advance
  - Run separate regressions (e.g. male/female)
  - ...
- Data-driven solution: Automatic, sequential identification of subgroups

Explicit acknowledgement of possible non-linearities and interactions

Decision Trees in Survey Research

Potential use cases:

- Exploratory data analysis: Researcher may find hidden structures in the data
- Feature Engineering: Find interaction terms to be included in a regression model
- Subgroup identification: Find common characteristics of specific subgroups
- ...

(suggested by Fielding and O'Muircheartaigh, 1977, p. 26)

The following ideas are more speculative and are open to discussion ...

Key idea: Segments/subgroups matter!

How to Characterize the Unemployed?

Using multiple variables selected by hand?

Tabelle 1

Ausgewählte Strukturmerkmale der Kurzzeit- und Langzeitarbeitslosen im Zeitvergleich

Juni 2010 und 2014, Anteile in Prozent

	Kurzzeitarbeitslose (unter 1 Jahr)		Langzeitarbeitslose (1 Jahr und länger)	
	Juni 2010	Juni 2014	Juni 2010	Juni 2014
Berufsausbildung				
Ohne abgeschlossene Berufsausbildung	39,9	42,5	46,9	50,6
Betriebliche/schulische Ausbildung	49,5	46,9	42,2	42,2
Akademische Ausbildung	6,6	8,6	3,8	4,6
Anforderungsniveau der gesuchten Tätigkeit				
Helfer	33,4	40,1	42,3	51,8
Fachkraft	45,4	41,5	40,4	37,1
Spezialist	5,6	5,7	3,7	3,5
Experte	6,1	7,2	3,2	3,3
Alter				
15 – 24 Jahre	13,5	12,0	2,4	2,5
25 – 34 Jahre	26,0	27,5	19,7	18,4
35 – 44 Jahre	22,9	21,0	26,0	22,6
45 – 54 Jahre	23,5	22,7	30,6	29,8
55 – 64 Jahre	14,0	16,7	21,3	26,4

(taken from Bruckmeier et al. (2015), IAB-Kurzbericht)

Using combinations of variables that are most predictive?

TABLE 1. SPENDING UNIT INCOME AND THE NUMBER IN THE UNIT WITHIN VARIOUS SUBGROUPS

Group	Spending unit average (1958) income	Number in unit	Number of cases
Nonwhite, did not finish high school	$ 2489	3.3	191
Nonwhite, did finish high school	5005	3.4	67
White, retired, did not finish high school	2217	1.7	272
White, retired, did finish high school	4520	1.7	72
White, nonretired farmers, did not finish high school	3950	3.6	87
White nonretired farmers, did finish high school	6750	3.6	24
The Remainder			
0–8 grades of school			
18–34 years old	4150	3.8	72
35–54 years old	4670	3.8	240
55 and older—not retired	4846	2.2	208
9–11 grades of school			
18–34 years old	5032	3.7	112
35–54 years old	6223	3.4	202
55 and older—not retired	4720	2.1	63
12 grades of school			
18–34 years old	5458	3.3	193
35–54 years old	7765	3.8	291
55 and older—not retired	6850	2.0	46
Some college			
18–34 years old	5378	3.0	102
35–54 years old	7930	3.8	112

(taken from Morgan and Sonquist (1963))

Feature Engineering

"The secret sauce" for building good predictive models

Difficulty:

- Variables can have many infrequent categories (e.g. > 200 countries in the IEB).
- How to use the country variable in a regression $y = x^T\beta + \gamma_{\text{country}} + \epsilon$?

Possible solutions: Group countries first ...

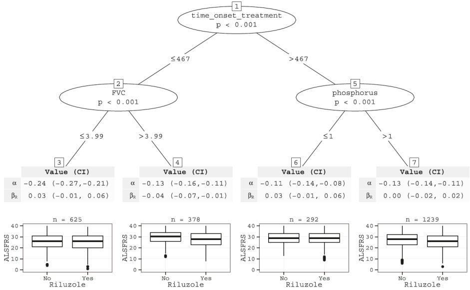
- manually (by continent?)
- automatically with trees, based on similar outcome

Subgroup Identification in Causal Analysis

Setting:

- Randomized trial with treatment (T = 1) and control groups (T = 0)
- $y = \alpha + \beta T + \epsilon$
- β identifies the average treatment effect

What about heterogeneous treatment effects? Subgroups might differ ...



Can we detect parameter instabilities in α or β?

Model-based Recursive Partitioning detects four subgroups:
Functional status of ALS patients six months after treatment with Riluzole started

(taken from Seibold et al. 2016)

Summary

- Divide-and-conquer strategy that splits the data into subgroups
- Surface from decision trees is a step function (as compared to continuous function in OLS regression)
- No need to specify the functional form in advance (unlike regression)
- Non-linearities and interactions are handled automatically
- Easy to interpret and easy to overinterpret
- May be used for ...
    - Data exploration: Discover subgroups, parameter instabilities and interactions
    - Prediction: Self-contained tool or used within more complex algorithms

Software Resources

Resources for R

- General overview: `https://cran.r-project.org/web/views/MachineLearning.html`
- Standard package to build CARTSs: `rpart`
- Conditional Inference Trees and Model-based recursive partitioning: `partykit`
- Unified infrastructure for tree representation: `partykit`

More open-source software

- Weka implements many rule learning algorithms: `https://www.cs.waikato.ac.nz/ml/weka/`

References: Overviews

Fürnkranz, Johannes, Gamberger, Dragan & Lavrač, Nada (2012)
Foundations of Rule Learning. Springer.

Hastie, Trevor, Tibshirani, Robert & Friedman, Jerome (2009)
The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. Chapter 9.

Loh, Wei-Yin (2014)
Fifty Years of Classification and Regression Trees. *International Statistical Review* **82**(3). 329–348

McArdle, John & Ritschard, Gilbert (Ed.) (2014)
Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences. Routledge.

Zhang, Heping & Singer, Burton (2010)
Recursive Partitioning and Applications. Springer.

References: Special Papers

Breiman, Leo, Friedman, Jerome, Olshen, Richard & Stones, Charles (1984)
Classification and Regression Trees. Brooks/Cole Publishing.

Fielding, A. and O'Muircheartaigh, Colm (1977)
Binary Segmentation in Survey Analysis with Particular Reference to AID. *The Statistician* **26**(1). 17–28

Hothorn, Torsten, Hornik, Kurt & Zeileis, Achim (2006)
Unbiased Recursive Partitioning: A conditional inference Framework. *Journal of Computational and Graphical Statistics* **15**(3). 651–674

Morgan, James and Sonquist, John (1963)
Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association* **58**(302). 415–434

References: Special Papers

Seibold, Heidi, Zeileis, Achim & Hothorn, Torsten (2016)
Model-Based Recursive Partitioning for Subgroup Analyses. *International Journal of Biostatistics* **12**(1). 45–63

Strobl, Carolin, Malley, James & Tutz, Gerhard (2009)
An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods* **14**(4). 323–348

Zeileis, Achim, Hothorn, Torsten & Hornik, Kurt (2008)
Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics* **17**(2). 492–514