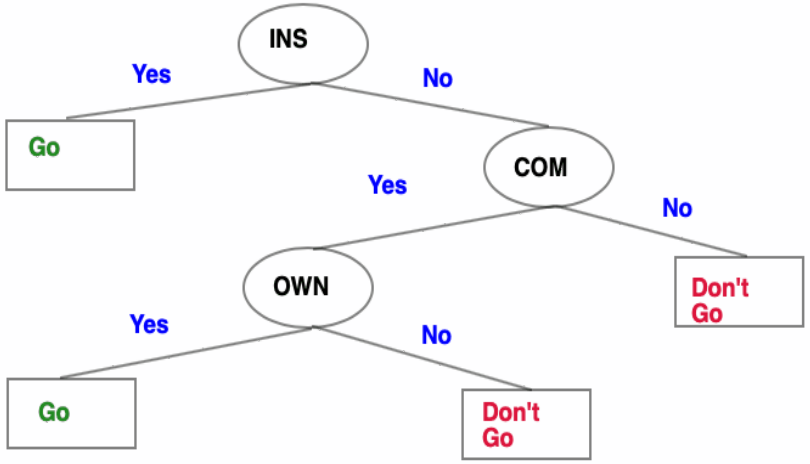
Teaching about decision trees for classificaton problems

Joachim Engel, Ludwigsburg, Germany
Tim Erickson, Oakland, USA
Laura Martignon, Ludwigsburg, Germany

IASE satellite Decision Making Based on Data
Kuala Lumpur, August 15, 2019

How do you decide to attend a conference?

- Inspiration (INS): Will the conference give me new inspirations for my work?
- Community (COM): Will I connect with colleagues?
- Own results (OWN): Do I have new ideas and results to present?



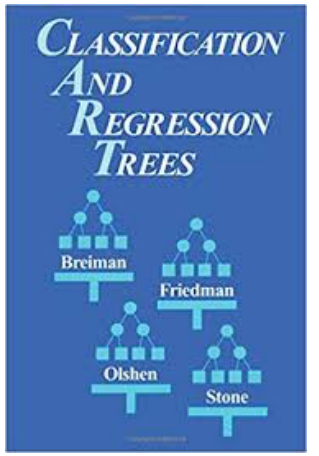
Decision algorithms in complex, „risky“ situations

- Medical diagnosis and intervention
- Financial investment
- Risk analysis
- Pattern recognition

Decision Trees

To make sense of these data, discover structure, etc. we need tools that are

- flexible and robust
- Simple enough to understand
- Provide results that are easy to interpret

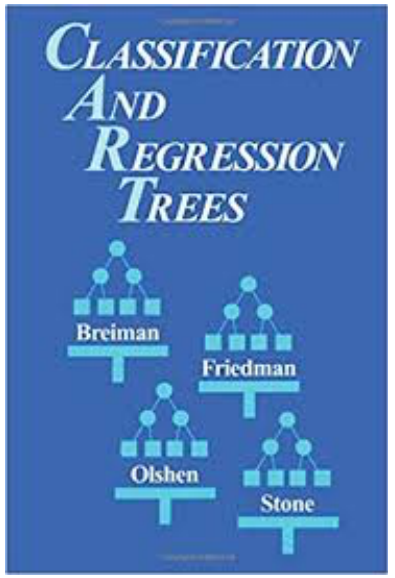


$\Rightarrow$ Classification And Regression Trees (CART)

(Implementation: Powerful Algorithms)

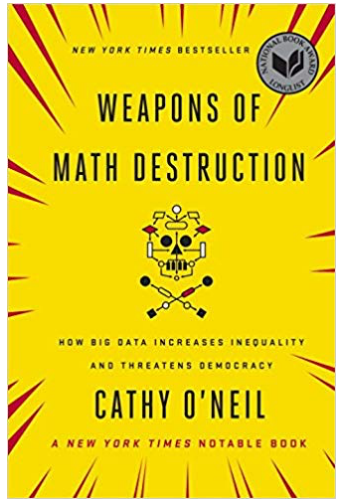
"Our philosophy in data analysis is to look at the data from a number of different viewpoints. Tree structured methods offer an interesting alternative for looking at classification and problems. It has sometimes given clues to data structure not apparent from a linear regression analysis. Like any tool, its greatest benefit lies in its intelligent and sensible application."

--Breiman, Friedman, Olshen, Stone



Purpose of this talk

- Help to understand how decision trees built
  - How created?
  - What kind of choices?
  - Which measures of quality?
- Intro to data science: „supervised learning”
- Critically reflect on algorithms



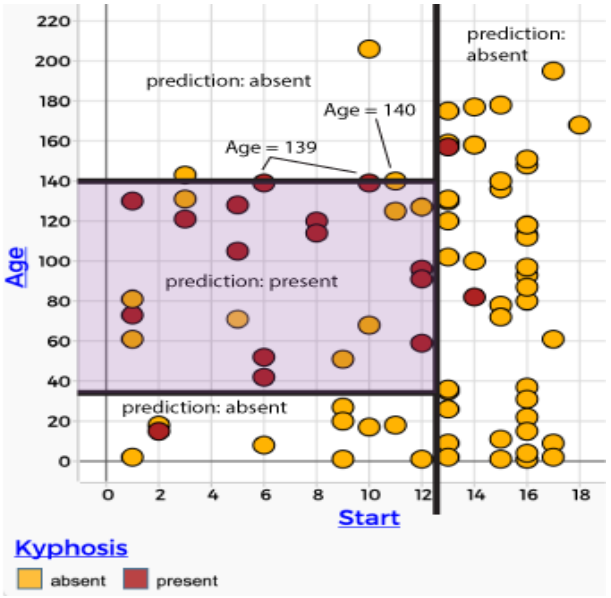
Decision algorithms

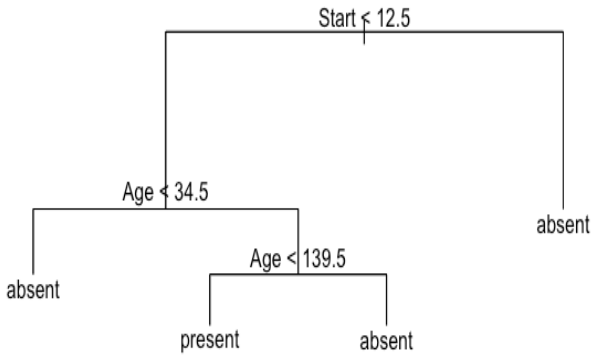
- Training sample: used for learning a decision rule
- Test sample: used for evaluating the quality of the decision rule

Example: Kyphosis after spinal surgery (Chambers & Hastie 1992)

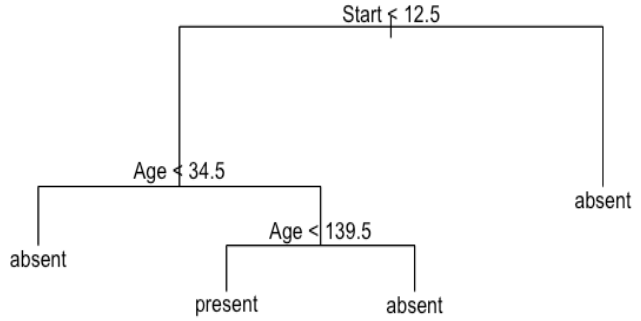
Sample 81 „cases“= children who after spinal surgery had or didn‘t have postoperative kyphosis (=a type of spinal deformation)

- `Age:` age of the child in months
- `Start:` the number of the first (topmost) vertebra operated on
- `Kyphosis:` indicates presence or absence of kyphosis





What makes a tree a good tree



- How and where to split?
  - Create regions where the data are more homogeneous;
    - Need some purity measure (open to creativity; e.g. Gini diver cross entropy, misclassification rate, ..)
- How to assign value to terminal node?
  - Majority vote (possibly weighted)
- When to stop tree growth? (Similar to variable selection in linear regression -> tendency for overfitting)
  - CART prunes an oversize tree based on cross-validation
  - Threshold for purity reduction
  - Minimum number of observations per node
  - Maximum number of nodes

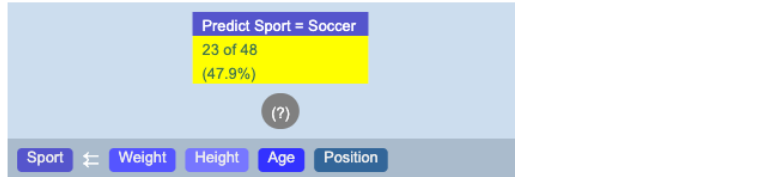
Teaching about classification trees: Crucial issues

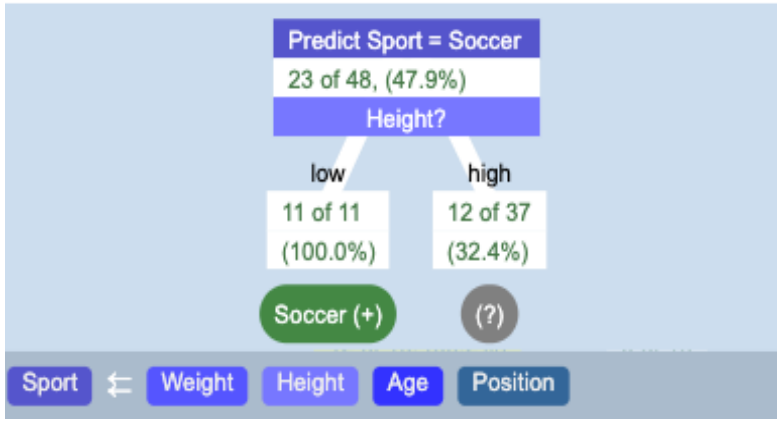
Based on experience at high school (data science course) and college (math teacher students):

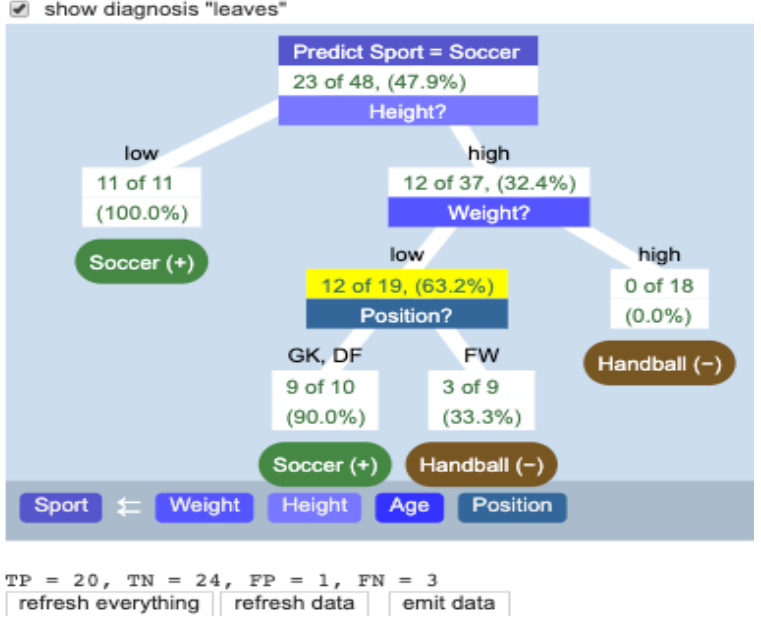
- Role of training sample
- Some choices seem haphazardly: how to justify?
- Meaning of nodes and purity measure
- Problem of oversized trees
- Assumptions and limitations

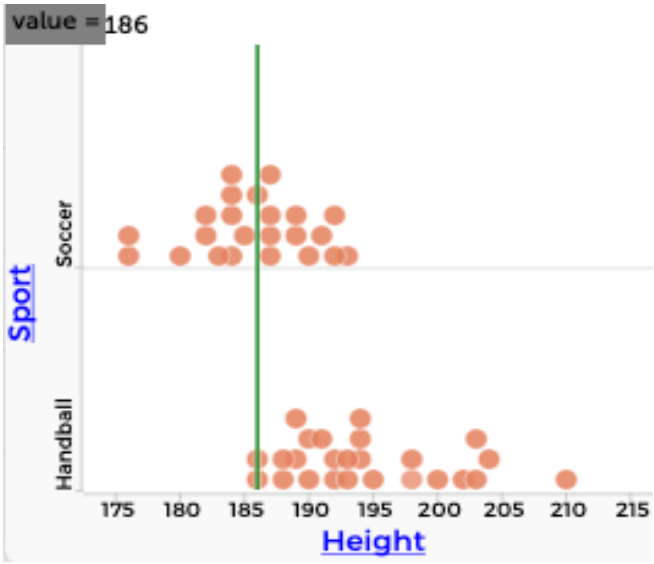
Teaching about classification trees: Solutions

- Step 1: Hands-on, with data cards
  - E.g. given a deck of 48 cards with infos about football and handball players (the „training sample“) with physical characteristics such as height, weight, age (as predictor) and their sport (target variable), develop a classification rule to „predict“ their sport. Then take another 10 cards (the „test sample“) to measure the quality of your decision rule
- Step 2: Technology, ARBOR - a plug-in tool to CODAP (developed by the second author)









- ARBOR slows down the algorithm, let‘s YOU decide next steps
- Lets YOU do the decisons stepwise and records a measure of tree quality
- YOU can define YOUR OWN measure of tree quality
- Very unlikely to result in the optimal tree, but YOU can investigate consequences of your choice of splitting criteria
- ARBOR allows you to use the CODAP capacity to find reasonable or good next steps

https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=100528

Xenobiologist: a data game diagnosing extraterrestrials

https://codap.xyz

https://codap.concord.org/releases/latest/static/dg/en/cert/index.html?di=https://codap.xyz/plugins/xeno/xeno.html

https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=112182

https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=33583

Summary

- Classification and Regression Trees are an exploratory method to find structure is messy data
- Easy to interpret result
- Based on heavy algorithms,
- Entrance to Data Science methods like Random Forests, Bagging, Boosting
- ARBOR is a learning tool to better understand how CART works