University of Zurich[UZH]

EBPI Epidemiology, Biostatistics and Prevention Institute

Transformation Forests

Torsten Hothorn

Joint work with Achim Zeileis, Lisa Schlosser and Heidi Seibold

useR! 2017, 2017-07-05

Machine Learning

Machine Learning methods give

> *computers the ability to learn without being explicitly programmed.*

(Arthur Samuel, 1959)

Actually: Fit statistical models to data by clever optimisation of appropriate target functions.

"Learning": Make statistical model underlying some "learning machine" explicit.

Statistical Learning

An oxymoron, like “Statistical Science”.

Either you learn, or you estimate.

Statistical Modelling

Too dull a term to attract any grant money.

However: Explicitly acknowledges the underlying probabilistic theory.

Today: Understand the statistical model behind a (special) random forest.

Random Forest

What is a random forest?

A model for

$$\mathbb{P}(Y \le y \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}_{Y \mid \mathbf{X} = \mathbf{x}}(y), \quad \forall \mathbf{x} \in \mathcal{X}$$

Parametric (!) Setup

Unconditional model for response

$$\mathbb{P}_{Y,\Theta} = \{\mathbb{P}_{Y,\boldsymbol{\vartheta}} \mid \boldsymbol{\vartheta} \in \Theta\}$$

Conditional model belongs to this family:

$$\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}} = \mathbb{P}_{Y,\boldsymbol{\vartheta}(\mathbf{x})}$$

Task: Estimate $\boldsymbol{\vartheta}$ function

Likelihood Contributions

"Learning" Data: $(y_i, \mathbf{x}_i), i = 1, \dots, N$

$\ell_i : \Theta \to \mathbb{R}$

$\ell_i(\boldsymbol{\vartheta}(\mathbf{x}_i))$ gives the *unconditional* likelihood for observation i with candidate parameters $\boldsymbol{\vartheta}(\mathbf{x}_i)$

Handle censoring and truncation appropriately here

Adaptive Local Likelihood Estimators

$$\hat{\boldsymbol{\vartheta}}^N(\mathbf{x}) := \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w_i^N(\mathbf{x}) \ell_i(\boldsymbol{\vartheta})$$

Conditioning works via weight functions $w_i^N(\mathbf{x})$ only.

Unconditional Maximum Likelihood

$$\hat{\boldsymbol{\vartheta}}_{\text{ML}}^{N} := \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} \ell_i(\boldsymbol{\vartheta})$$

Trees

$$\mathcal{X} = \dot{\bigcup_{b=1,\dots,B}} \mathcal{B}_b$$

$$\begin{aligned} w^N_{\text{Tree},i}(\mathbf{x}) &:= \sum_{b=1}^{B} I(\mathbf{x} \in \mathcal{B}_b \wedge \mathbf{x}_i \in \mathcal{B}_b) \\ \hat{\boldsymbol{\vartheta}}^N_{\text{Tree}}(\mathbf{x}) &:= \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w^N_{\text{Tree},i}(\mathbf{x}) \ell_i(\boldsymbol{\vartheta}) \end{aligned}$$

Forests

$$\mathcal{X} = \dot{\bigcup_{b=1,\dots,B_t}} \mathcal{B}_{tb} \text{ for } t = 1, \dots, T \text{ trees}$$

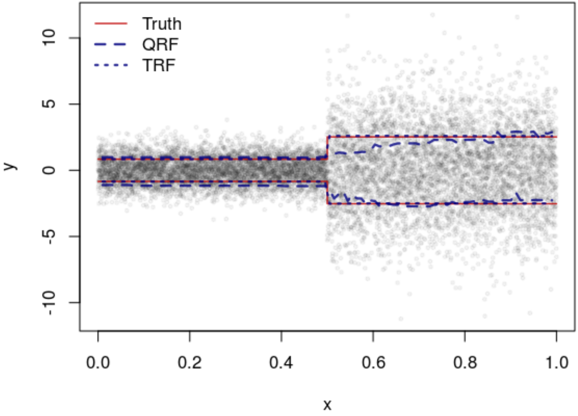
$$\begin{aligned} w^N_{\text{Forest},i}(\mathbf{x}) &:= \sum_{t=1}^{T} \sum_{b=1}^{B_t} I(\mathbf{x} \in \mathcal{B}_{tb} \wedge \mathbf{x}_i \in \mathcal{B}_{tb}) \\ \hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\mathbf{x}) &:= \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w^N_{\text{Forest},i}(\mathbf{x}) \ell_i(\boldsymbol{\vartheta}) \end{aligned}$$

OK, Done! Really?

These "nearest neighbor weights" have been used before, for example in conditional inference forests (**party**, **partykit**) or quantile regression forests (**quantregForest**) with *STANDARD* trees.

Unfortunately, there is a catch.

The Problem



The Solution

We need splits sensitive to *distributional* and not just *mean* changes.

Transformation model (google "MLT useR! 2016"):

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = \Phi(\mathbf{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\mathbf{x}))$$

- $\mathbf{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\mathbf{x})$ is a smooth, monotone Bernstein of degree d
- $d = 1$ means $\mathbb{P}_{Y \mid \mathbf{X} = \mathbf{x}} = \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$
- $d = 5$ is surprisingly flexible

All "classical" distributions: Distribution forests (Lisa, in 20min)

Transformation Trees (TRT)

- Start with $\hat{\boldsymbol{\vartheta}}_{\text{ML}}^N$
- Search for parameter instabilities in $\hat{\boldsymbol{\vartheta}}_{\text{ML}}^N$ as a function of $\mathbf{x}$ using model-based recursive partitioning (a beefed-up version)
- Potentially find changes in the mean AND higher moments
- Forests: Aggregate these trees via adaptive local likelihood estimation

Transformation Forests (TRF)

$$\hat{\mathbb{P}}(Y \le y \mid \mathbf{X} = \mathbf{x}) = \Phi(\mathbf{a}_{\text{Bs},d}(y)^\top \hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\mathbf{x}))$$

makes the forest "parametric" with

- Forest likelihood
- Prediction intervals
- Likelihood-based variable importance
- Parametric bootstrap
- ...

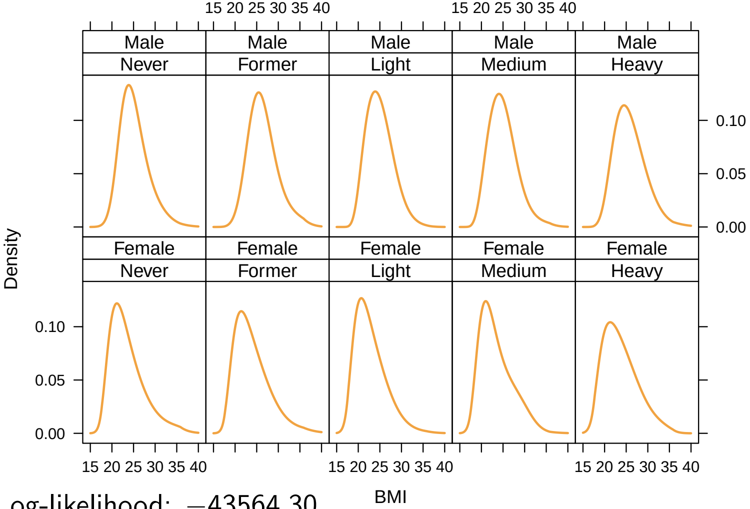
and applicable to censored and truncated data.

Swiss Body Mass Index Distributions

2012 survey ($N = 16427$) in Switzerland
Explain conditional distribution of BMI given

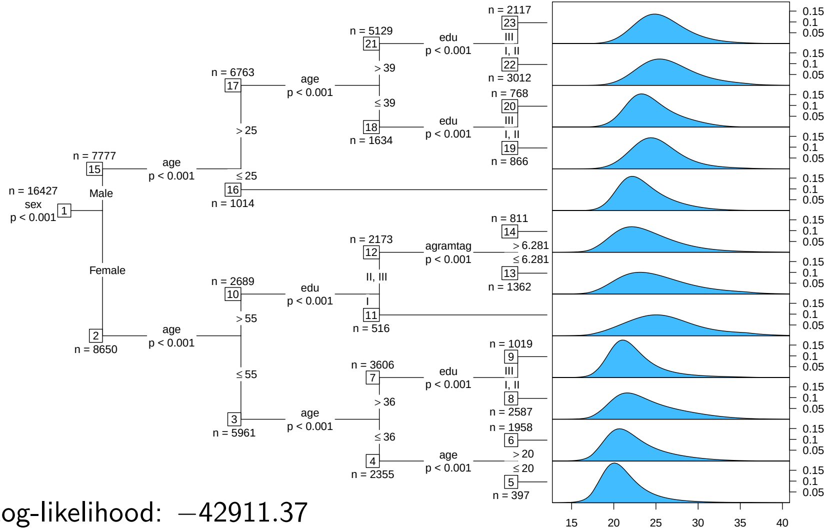
- Sex,
- Smoking status,
- Age,
- Education,
- Physical activity,
- Alcohol intake,
- Fruit and vegetable consumption,
- Region, and
- Nationality.

Sex- and Smoking



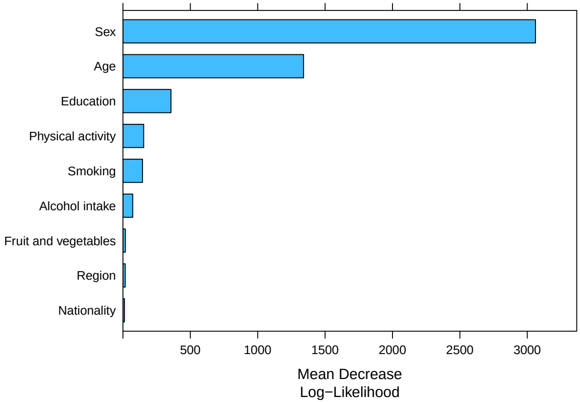
Log-likelihood: -43564.30

Transformation Tree



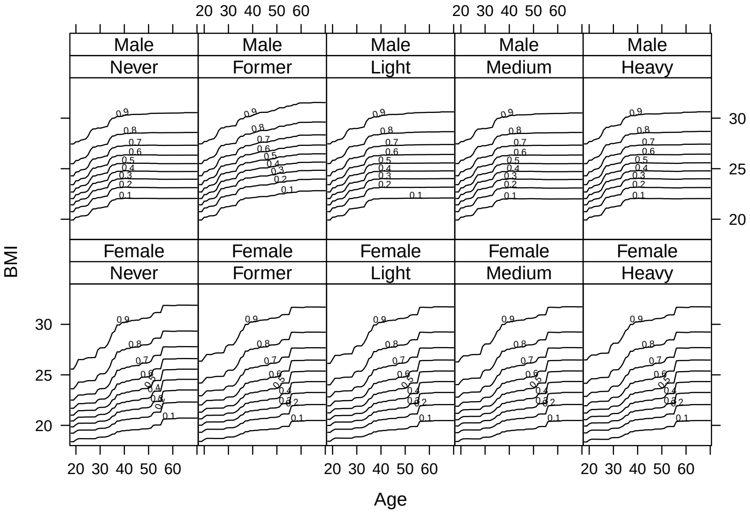
Log-likelihood: -42911.37

Transformation Forest: Variable Importance



In-bag log-likelihood: -42629.63; out-of-bag: -42856.93

Transformation Forest: Partial Deciles



Summary

- Transformation Trees and Forests are adaptive local likelihood estimators of a conditional distribution function
- Inherit the nonparametric freedom and the parametric simplicity
- Great as supermodels to compare simpler ones to
- Can predict distributions, not just means
- Make model evaluation (parametric bootstrap) and inference (variable importance) easier and more generally applicable
- Applicable to censored and truncated responses

https://arxiv.org/abs/1701.02110
http://arxiv.org/abs/1706.08269
https://r-forge.r-project.org/projects/ctm/