

Thoughts on Modeling

ENVS615 | Dani Arribas-Bel

The “Two Cultures”

Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)

Leo Breiman

Full-text: Open access

[PDF File \(300 KB\)](#)

[Abstract](#)

[Article info and citation](#)

[First page](#)

[References](#)

Abstract

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the **data are generated by a given stochastic data model**. The other uses **algorithmic models** and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Machine Learning: An Applied Econometric Approach

Sendhil Mullainathan

Jann Spiess

JOURNAL OF ECONOMIC PERSPECTIVES

VOL. 31, NO. 2, SPRING 2017

(pp. 87-106)

[Download Full Text PDF](#)
(Complimentary)

$\hat{\beta}$

vs

\hat{y}

$$y = \beta X + \epsilon$$

- Inference (β)
- Prediction (y)

Modeling

$$y = f(X)$$

Regression $\rightarrow f(X) = \alpha + \beta X + \epsilon$

CART $\rightarrow f(X) = \text{Tree}$

Random Forest $\rightarrow f(X) = \text{Tree Ensemble}$

Gradient boost $\rightarrow f(X) = \text{Sequence of predictors (e.g. trees)}$

This block

- Baseline model ($\beta + \gamma$)
- Fitting Random Forests
- Predictive checking ($\beta + \gamma$)
- Uncertainty (β)
- Model performance (γ)
- Overfitting & Cross-Validation (γ)



“Thoughts on modelling” by Dani Arribas-Bel is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.