Thoughts on Modeling

Data Science Studio

The “Two Cultures”

Statist. Sci.
Volume 16, Issue 3 (2001), 199-231.

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

Leo Breiman

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)

$$\hat{\beta} \quad \text{vs} \quad \hat{\gamma}$$

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

- Inference (β)
- Prediction (y)

Modeling

Modeling

$$y = f(X)$$

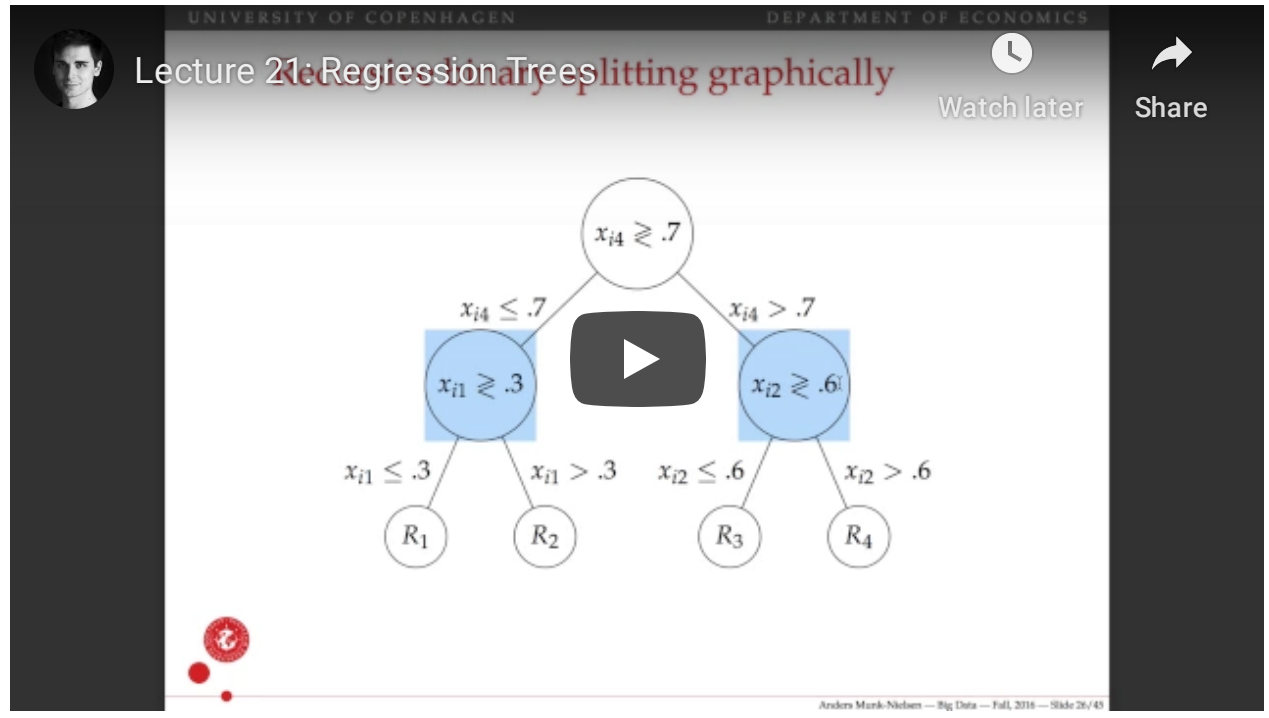
Regression $\rightarrow f(X) = \alpha + \beta X + \epsilon$

CART $\rightarrow f(X) =$ Tree

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

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

Regression Trees



Random Forest

- Several regression trees
- Random subsets of N and k
- Averaged for predictions

Today

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