Data Modeling Mindsets

Christoph Molnar

Data Modeling Mindsets

Bayesian and frequentist statistics, machine learning and causal inference – these approaches share common methods and models. They differ in assumptions about the data-generating process and when a model is a good generalization of the real world.

Machine Learning

$$\arg\min_{f} L(X, Y, f(X))$$

Machine learning minimizes a loss function L by finding the best function f that to predict target Y from features X. A good machine learning model has a low loss on the test data.

Statistical Inference

$$\arg\max_{\theta} P(\theta, X)$$

Statistical inference fits the best parameters of a chosen probability distribution for variables X. A good statistical model has a high goodness-of-fit: the data fitthe distribution.

Bayesian Inference

$$P(\theta|X) = \frac{P(X|\theta) \cdot P(\theta)}{P(X)}$$

Bayesian inference assumes that the distribution parameters θ are random variables with an a-priori distribution. A good Bayesian model has a high posterior probability (Bayes factor).

Causal Inference

$$P(Y|do(X))$$

Causal inference operates on the principles of causality, intervention and counterfactuals..
A good causal model has high goodness-of-fit and solid causal assumptions.

Which One is the Best?

The smart way is to be pragmatic about the modeling choices. Need a causal interpretation? Think causal inference. Only predictive performance is important? Pick machine learning. Want to include prior information about model parameters? -> Bayesian stats.