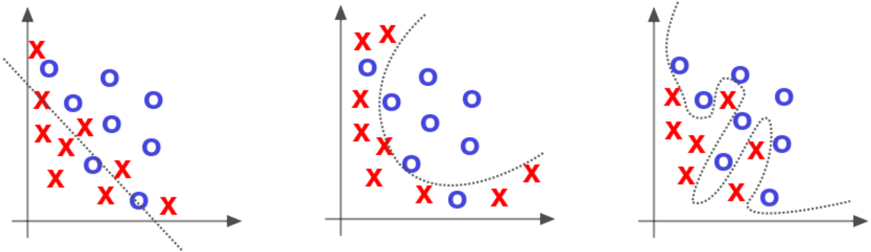
AutoML: Evaluation

Overview and Motivation

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

Training Machine Learning Models

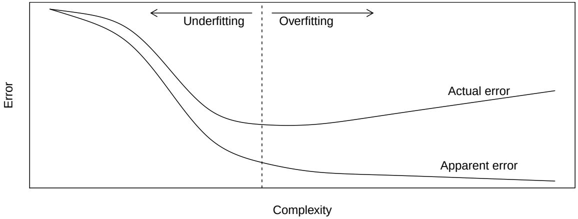
- fundamentally an optimization problem
- determine model parameters such that loss on data is minimized
- quality of fit depends on model class (i.e. degrees of freedom)



Which model is best?

Generalization

- we want models that *generalize* – make "reasonable" predictions on new data
  - ignore outliers
  - smooth
  - captures general trend

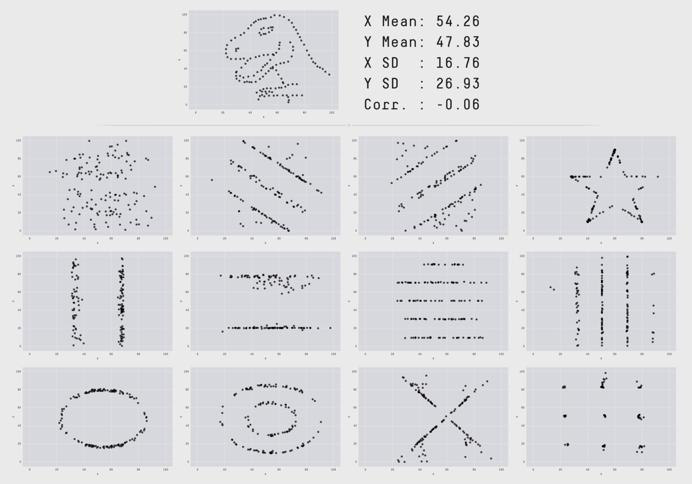


Usually model performance gets better with more data/higher model complexity and then worse, but see [Nakkiran et al. 2019]

Overview

- evaluating machine learning models and quantifying generalization performance
- learning curves
- comparing multiple models/learners on multiple data sets
- statistical tests
- higher levels of optimization, higher levels of evaluation
  - automated machine learning (meta-optimization) can lead to meta-overfitting
  - simple training/testing split(s) no longer sufficient $\rightarrow$ nested evaluation

Evaluate Early, Evaluate Often



https://www.autodeskresearch.com/publications/samestats