Maching Learning Basics

Jae Yeon Kim

17 April, 2019

DGP

- Statistics starts think of the data as being generated by a black box.
  - y <- nature <- x
- Data analysis
  - Prediction (algorithms; e.g., regression line)
  - Information (inference; e.g., confidence intervals)

Two cultures (Breiman 2001)

- Statistical inference
  - y <- some probability models (e.g., linear regression, logistic regression, Cox model) <- x
  - $y = X\beta + \epsilon$
  - The goal is to estimate β

- Machine learning
  - y <- unkown <- x
  - y <-> decision trees, neutral nets <-> x
  - “The problem is to find an algorithm f(x) such that for future x in a test set, f(x) will be a good predictor of y.”

Data Modeling vs. Algorithmic Modeling

"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**."

"Algorithmic models, both in theory and practice, has developed rapidly in fields of outside statistics. It can be used on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets."

Statistical model

- The following discussions come from Athey and Imbens's review paper (2019)
- Specifying a target (i.e., an estimand; a joint distribution of data)
- Fitting a model to data using an objective function (e.g., the sum of squared errors)
- Reporting point estimates and standard errors
- Validation by yes-no using goodness-of-fit tests and residual examination

ML

- Developing algorithms (estimating F)
- Prediction power not structural/causal parameters
- Basically, high-dimensional data statistics ($N < P$)
- The major problem is to avoid “the curse of dimensionality”

- How to deal with high-P
  - Parametric apprach (regression)
    - Easy, fast, but could be far from true F
    - Making algorithm more flexible can cause overfitting
  - Non-parametric approach (support vector machines)
    - No assumptions made (agnostic approach)
    - But requires more data and slow
    - Still, overfitting issue

- Validation: out-of-sample comparisons (cross-validation) not in-sample goodness-of-fit measures
- So, it's curve-fitting but the primary focus is unseen not seen data
- Machine learning as p-hacking problem in the 21st century is correct and wrong at the same time

ML - Terminology

- Sample to estimate parameters = Training sample
- Estimating the model = Being trained
- Regressors, covariates, or predictors = Features
- Regression parameters = weights
- Prediction problems = Supervised (some response variables are known) + Unsupervised (response variables are not known)

Underdeveloped in the ML literature

- Two types of errors: reducible and irreducible (upper bound)
- Missing data problem persists (Holland)
- No interventions $\rightarrow$ no causal arguments
- Machines that won Go games against masters and drive cars still have hard time with causality (Judea Pearl's criticism)