Machine Learning for econometrics

Statistical learning and penalized regression

Matthieu Doutreligne

January 10, 2025

Today's program

- Last session: importance of causal variable status
- Today: predictive inference in high dimensions
  - Statistical learning basics
  - Penalized linear regression for predictive inference
  - Hands-on with scikit-learn

Today's program

- Last session: importance of causal variable status
- Today: predictive inference in high dimensions
  - Statistical learning basics
  - Penalized linear regression for predictive inference
  - Hands-on with scikit-learn
- Next session:
  - Flexible models: Trees, Random Forests, Gradient Boosting
  - Practical scikit-learn

Table of contents

Settings: statistical learning

Statistical learning, ie. predictive inference

Goal

- Predict the value of an outcome based on one or more input variables.

Setting

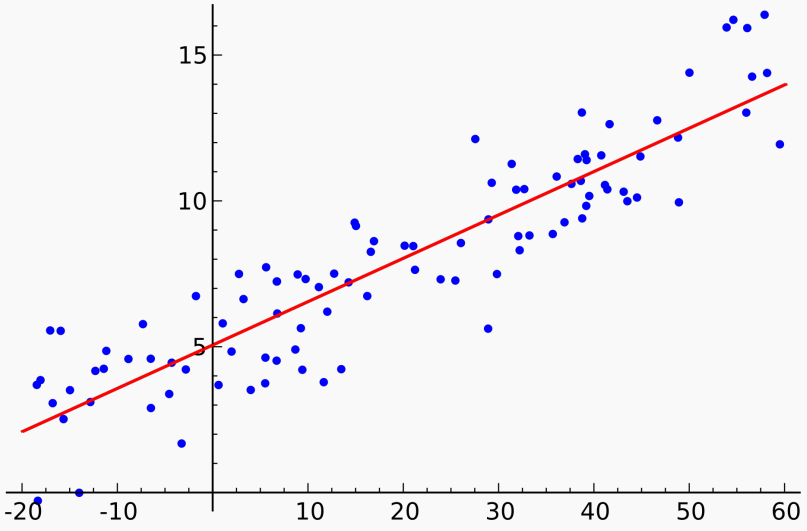
- Data: n pairs of (features, outcome), $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ identically and independently distributed (i.i.d.) from an unknown distribution P.
- Goal: find a function $\hat{f} : \mathcal{X} \to \mathcal{Y}$ that approximates the true value of y ie. for a new pair (x, y), we should have:

$$\hat{y} = \hat{f}(x) \approx y$$

Vocabulary

Finding the appropriate model $\hat{f}$ is called learning, training or fitting the model.

Statistical learning, two types of problems

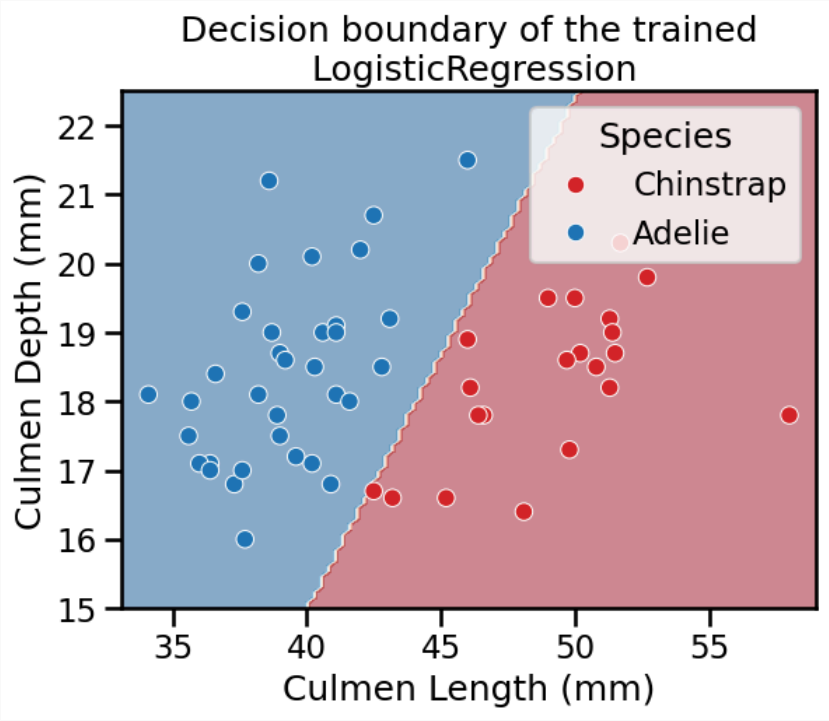


Regression

- The outcome is continuous: eg. wage prediction
- The error is often measured by the mean squared error (MSE):

$$\text{MSE} = \mathbb{E}\left[\left(Y - \hat{f}(X)\right)^2\right]$$

Statistical learning, two types of problems



Classification

- Outcome is categorical: eg. diagnosis, loan default, ...
- Error is often measured with accuracy:

$$\text{Misclassification rate} = \mathbb{E}\left[\mathbb{1}\left(Y \neq \hat{f}(X)\right)\right]$$

with $\hat{f} \in \{0, 1\}$ for binary classification

Motivation: why prediction?

Why do we need prediction for ?

Statistical inference

- Goal: infer some intervention effect with a causal interpretation
- Require to regress "away" the relationship between the treatment or the outcome and the confounders -> more on this in sessions on Double machine learning.

Predictive inference

- Some problems in economics requires accurate prediction (Kleinberg et al., 2015) without a causal interpretation
- Eg. Stratisfying on a risk score (loan, preventive care, ...)

Do we need more than linear models?

Let:

- p is the number of features
- n is the number of observations

Maybe no

- Low-dimensional data: $n \gg p$
- High predictive performances

Maybe yes

- High-dimensional data: ie. $p \gg n$
- Poor predictive performances

Do we need more than linear models?

When do we have "high-dimension"?

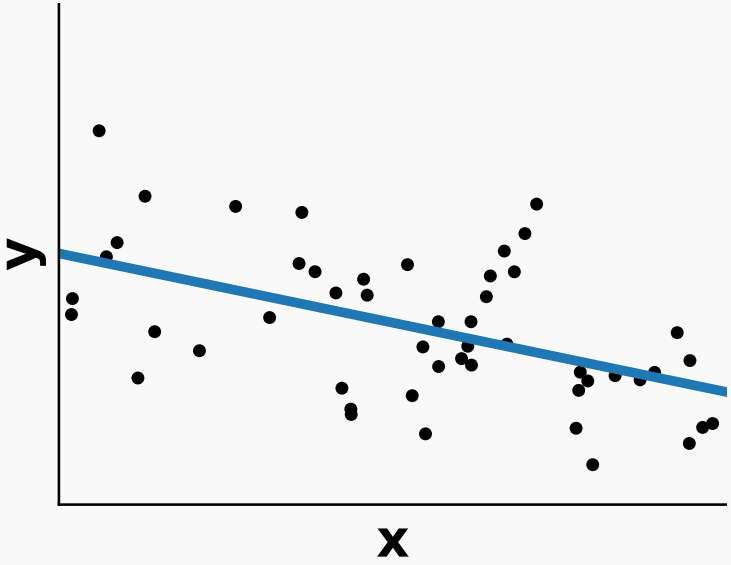
- Is $p \gg n$ a common setting in economics?
- Consider the wage dataset:
  - $n = 5150$ individuals
  - $d = 18$ variables

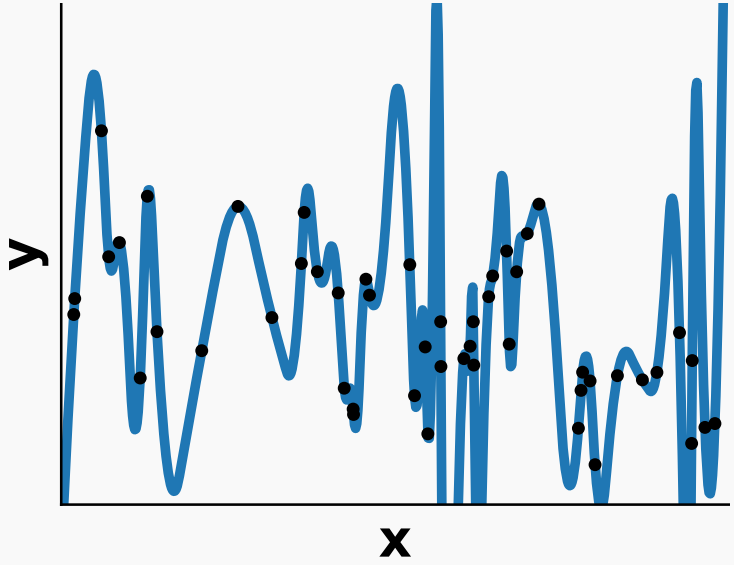
When do we have "high-dimension"?

- Is $p \gg n$ a common setting in economics?
- Consider the wage dataset:
  - $n = 5150$ individuals
  - $d = 18$ variables
  - But, categorical variables, non-linearities and interactions increase the real number of features:
    - non-linearities: add polynomials of degree 2: $p = 2 \times 18 = 36$
    - interactions:
      - Of degree 2: $\binom{d}{2} = \binom{18}{2} = 153$
      - All interactions: $2^d = 2^{18} - k - 1 = 262125$

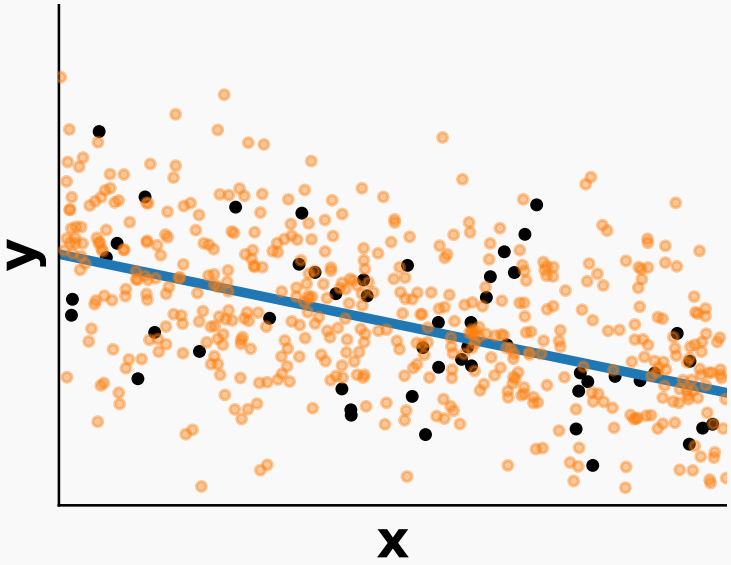
Statistical learning theory

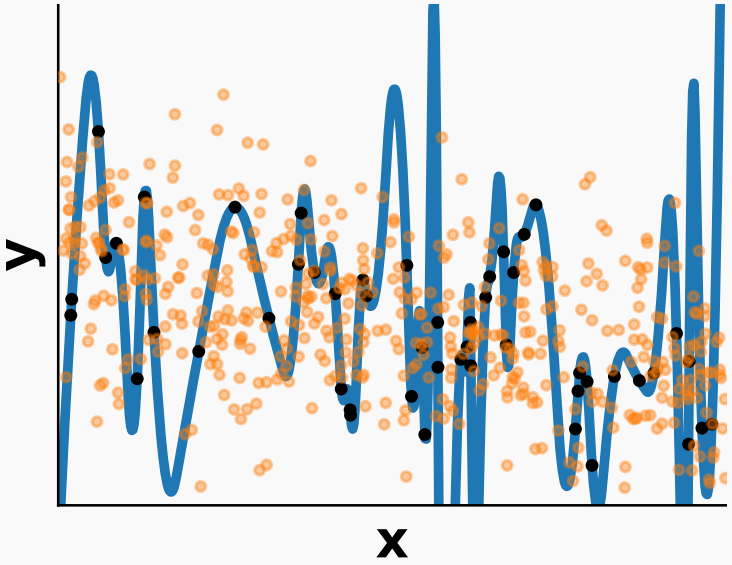
Which data fit do you prefer?





Which data fit do you prefer? (new data incoming)





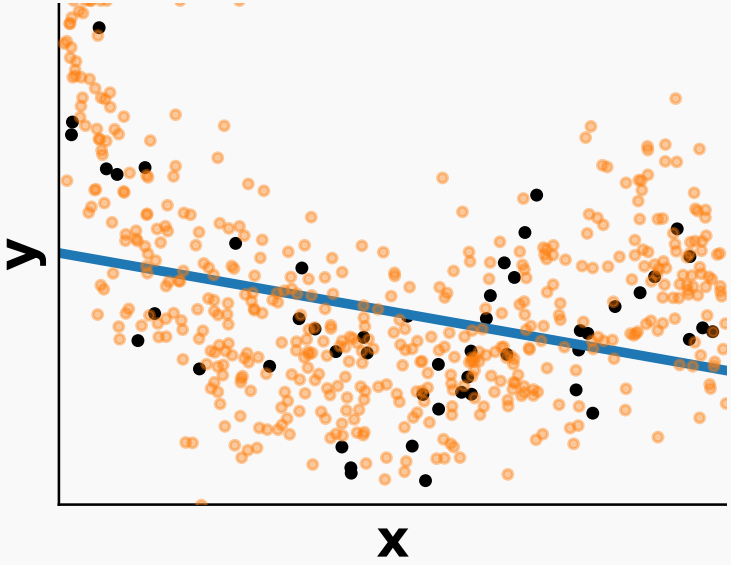
- Answering this question might be hard.
- Goal: create models that generalize.
- The good way of framing the question is: how will the model perform on new data?

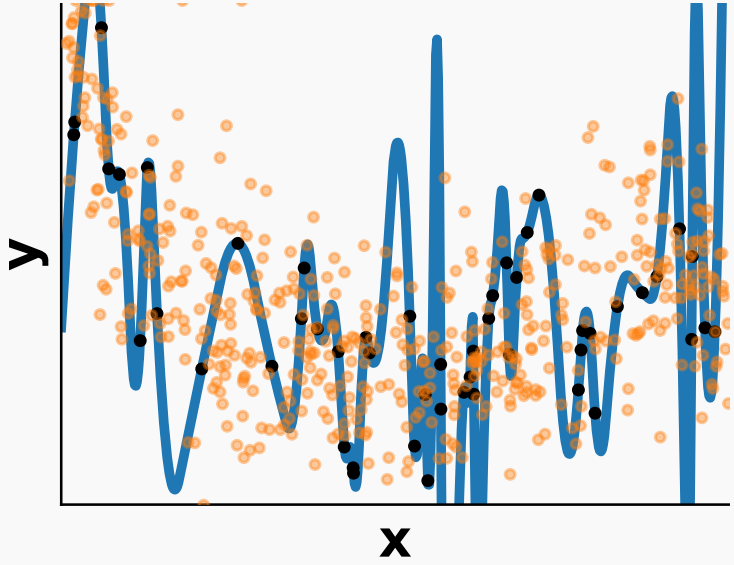
Train vs test error

Measure the performances on test data = generalization

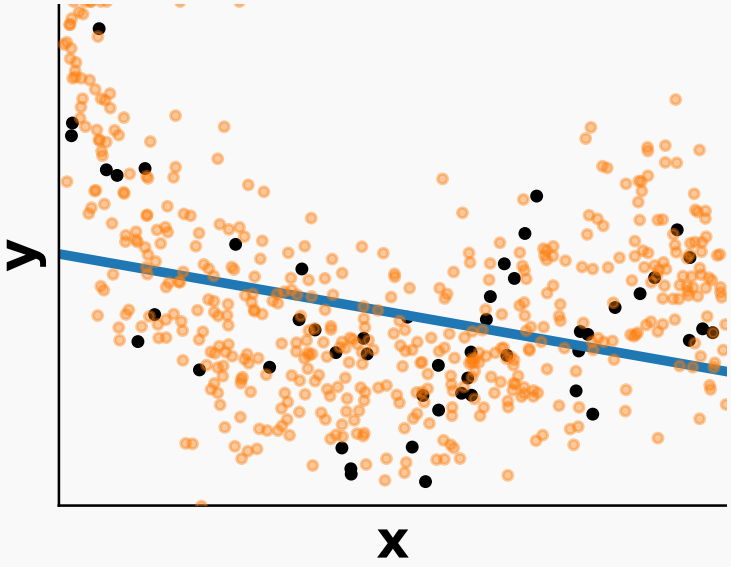
Measure the errors on the training data = fitting

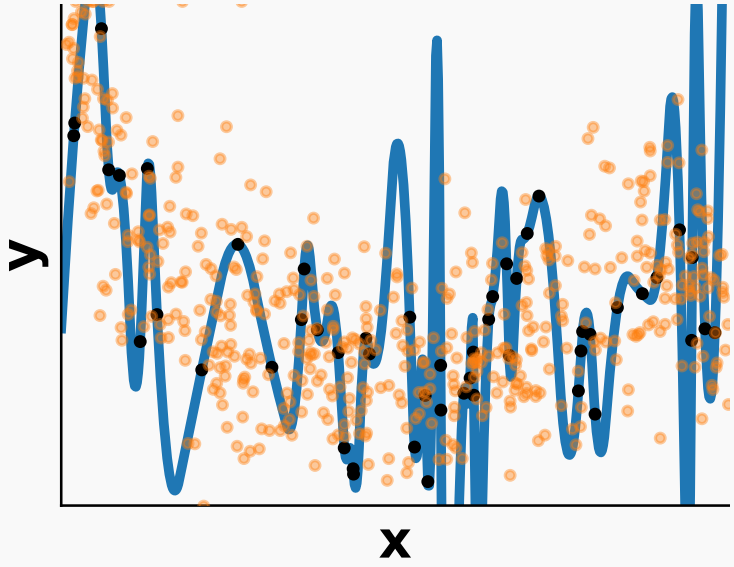
How to choose the complexity of the model?





How to choose the complexity of the model?





This trade-off is is called Bias variance trade-off.

- Let's recover this trade-off in the context of statistical learning theory.

Empirical Risk Minimization

- Define a loss function ℓ that defines proximity between the predicted value $\hat{y} = f(x)$ and the true value y: $\ell(f(x), y)$
- Usually, for continuous outcomes, the squared loss is used: $\ell(f(x), y) = (f(x) - y)^2$
- We choose among a (finite) family of functions $f \in \mathcal{F}$, the best possible function $f^\star$ minimizes the risk or expected loss $\mathcal{E}(f) = \mathbb{E}\left[(f(x) - y)^2\right]$:

$$f^\star = \text{argmin}_{f \in \mathcal{F}} \mathbb{E}\left[(f(x) - y)^2\right]$$

Empirical risk minimization: estimation error

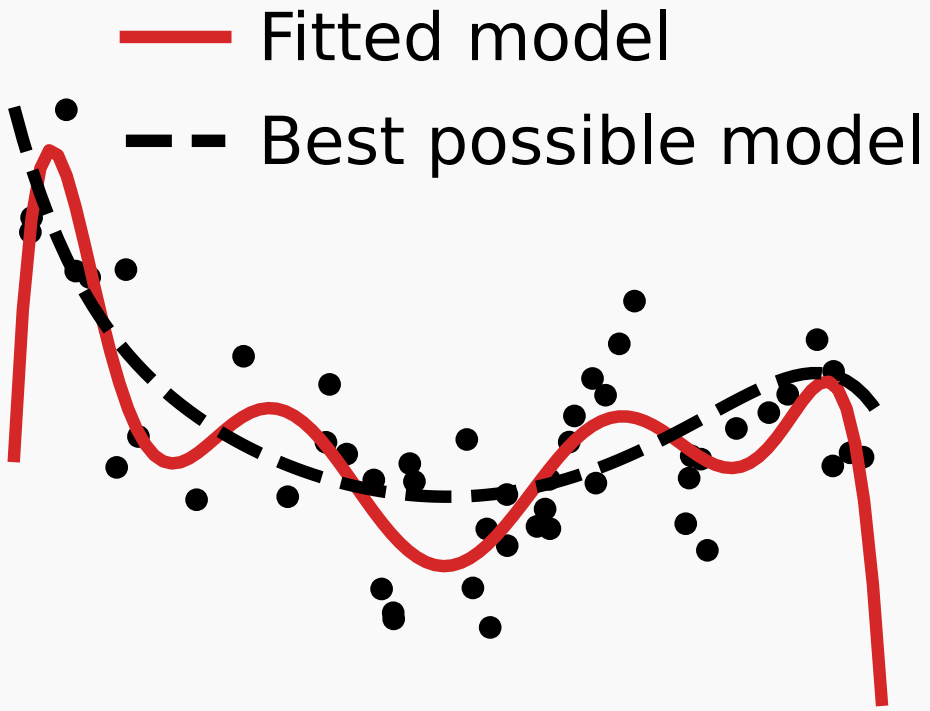
- In finite sample regimes, the expectation is not accessible since we only have access to a finite number of data pairs
- In practice, we minimize the empirical risk or average loss $R_{\text{emp}} = \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2$:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2$$

- This creates the estimation error, related to sampling noise:

$$\mathcal{E}\left(\hat{f}\right) - \mathcal{E}(f^\star) = \mathbb{E}\left[\left(\hat{f}(x) - y\right)^2\right] - \mathbb{E}\left[(f^\star(x) - y)^2\right]$$

High estimation error means overfit



Model is too complex

- The model is able to recover the true generative process
- But its flexibility captures noise

Too much noise

Not enough data

Bayes error rate: Randomness of the problem

- For interesting problems, there is some randomness: ie. $y = g(x) + e$ with $E(e|x) = 0$ and $\mathrm{Var}(e|x) = \sigma^2$
- The best possible estimator is g, yielding the Bayes error, the unavoidable error:

$$\mathcal{E}(g) = \mathbb{E}\Big[(g(x) + e - g(x))^2\Big] = \mathbb{E}[e^2]$$

Bias variance trade-off

- Decomposition of the empirical risk of a fitted model $\hat{f}$:

$$\mathcal{E}\left(\hat{f}\right) = \mathcal{E}(g) + \mathcal{E}(f^\star) - \mathcal{E}(g) + \mathcal{E}\left(\hat{f}\right) - \mathcal{E}(f^\star)$$

Lasso for predictive inference

Bias-variance trade-off, take home messages

High bias == underfitting

- systematic prediction errors
- the model prefers to ignore some aspects of the data
- mispecified models

High variance == overfitting:

- prediction errors without obvious structure
- small change in the training set, large change in model
- unstable models

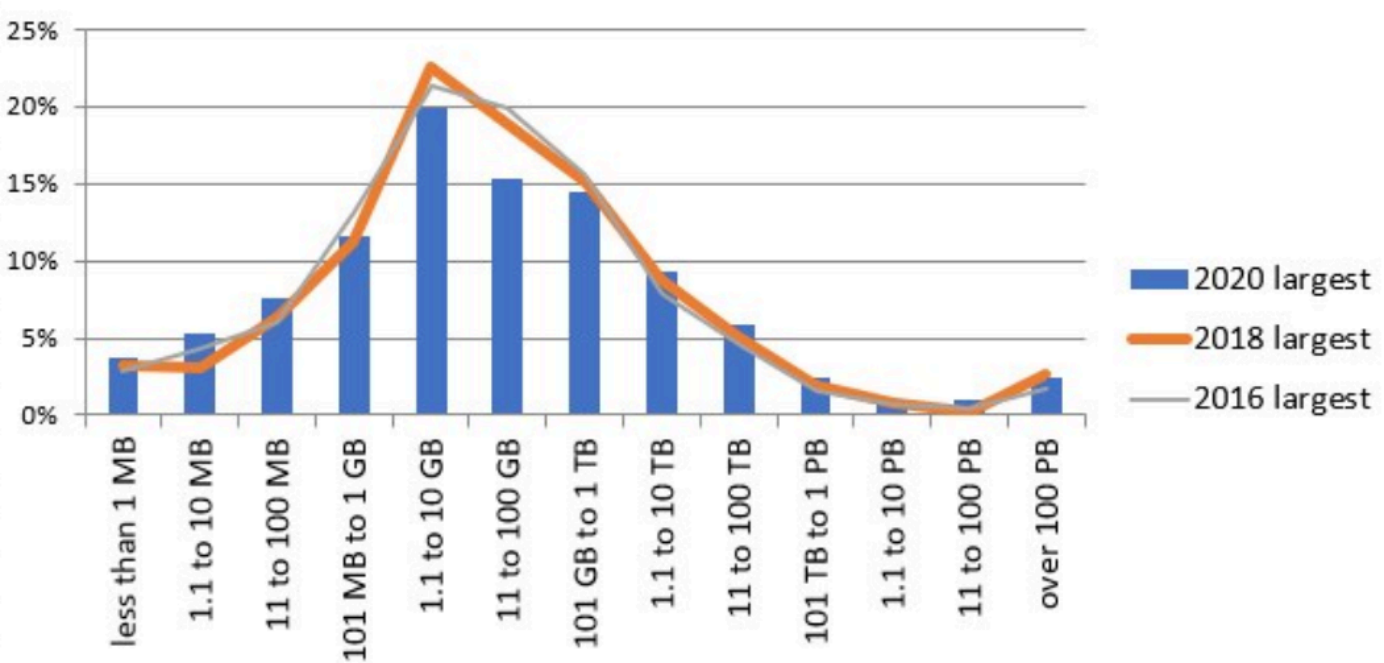
A word on deep learning

Why not use deep learning everywhere?

- Success of deep learning in image, speech recognition and text
- Why not so used in economics?

Limited data settings

- Typically in economics everywhere, we have a limited number of observations

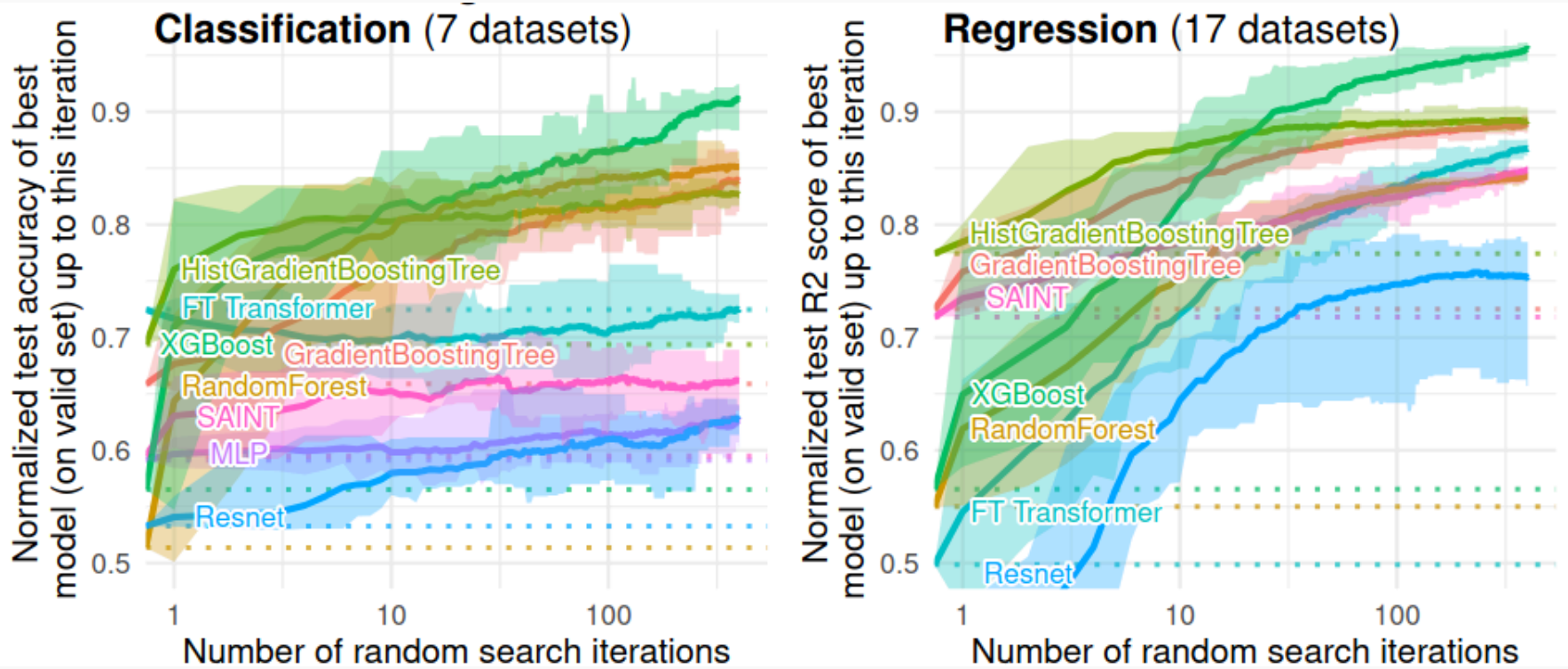


Typical dataset are mid-sized. This does not change with time.[1]

[1] https://www.kdnuggets.com/2020/07/poll-largest-dataset-analyzed-results.html

Deep learning underperforms on data tables

Tree-based methods outperform tailored deep learning architectures (Grinsztajn et al., 2022)



DAG for a RCT: the treatment is independent of the confounders

Resources

- https://web.stanford.edu/~swager/stats361.pdf
- https://www.mixtapesessions.io/
- https://alejandroschuler.github.io/mci/
- https://theeffectbook.net/index.html

Bibliography

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35, 507–520.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. American Economic Review, 105(5), 491–495.