Machine Learning with Python: A Hands-On Introduction

https://github.com/cbrownley/2023MLWEEK_MLWITHPYTHON

Clinton Brownley, PhD
https://cbrownley.github.io/
https://www.linkedin.com/in/clintonbrownley/

Agenda (too much…we’ll take our time :)

8:30-8:45 – Setup and Overview

8:45-9:30 – **Data preprocessing**

9:30-10:00 – Hands-on: Data preprocessing

10:00-10:30 – **Cross-validation**

10:30-11:00 – Hands-on: Cross-validation

11:00-11:30 – Hands-on: K-fold cross-validation

11:30-12:00 – **Classification** (breast tumor diagnosis)

12:00-12:30 – Hands-on: Classification

12:30-1:00 – Hands-on: Decision Trees

1:00-1:30 – **Regression** (california housing)

1:30-2:00 – Hands-on: Regression

2:00-2:30 – Hands-on: **Shrinkage methods**

2:30-3:00 – Classification (credit card fraud)

3:00-3:30 – Regression (cycling counts)

3:30-4:00 – Hands-on: Classification (hotel bookings)

4:00-4:30 – Hands-on: **Explainable ML**

Prediction vs Causal Inference

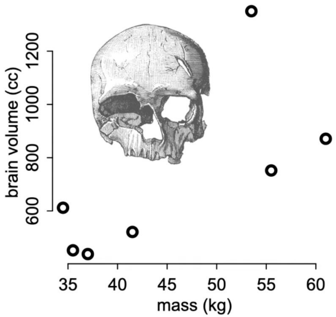
Problems of Prediction

What function describes these points? (fitting, compression)

What function explains these points? (causal inference)

What would happen if we changed a point's mass? (intervention)

What is the next observation from the same process? (prediction)



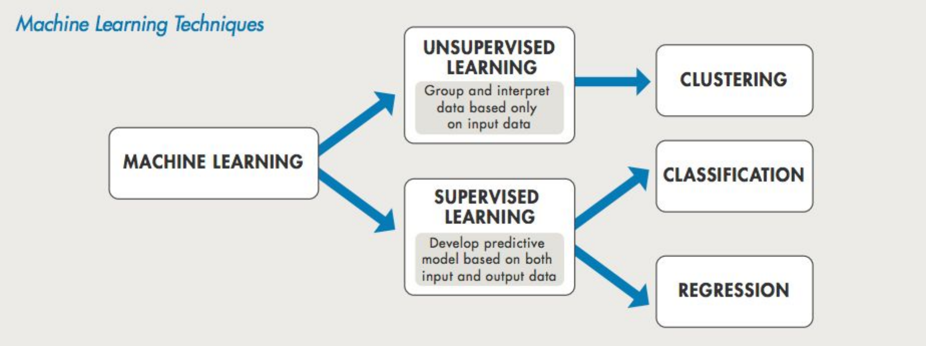
Good & Bad Controls

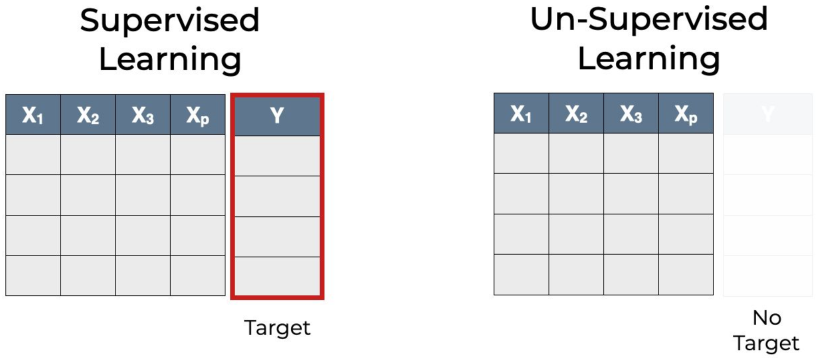
"Control" variable: Variable introduced to an analysis so that a causal estimate is possible

Common **wrong** heuristics for choosing control variables

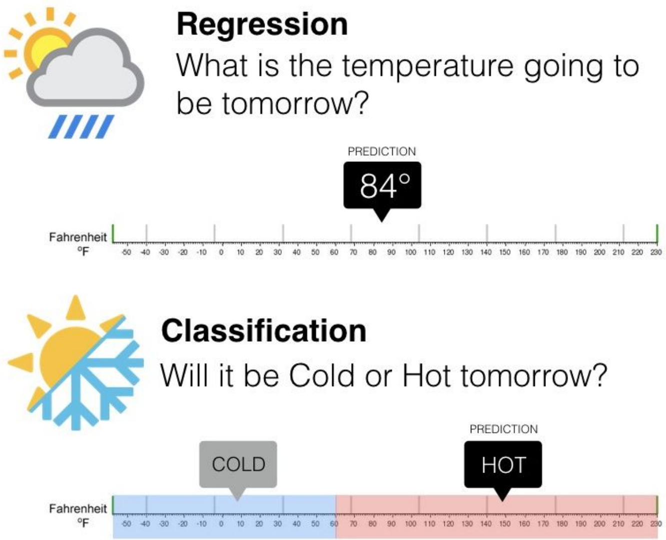
- Anything in the spreadsheet **YOLO**!
- Any variables not highly **collinear**
- Any **pre-treatment** measurement (baseline)

Supervised vs Unsupervised

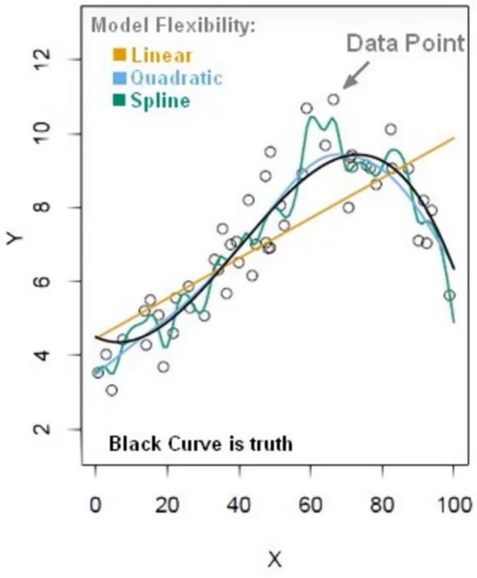


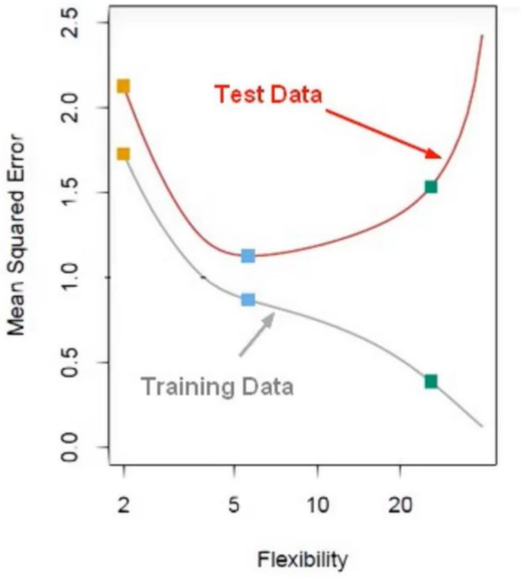


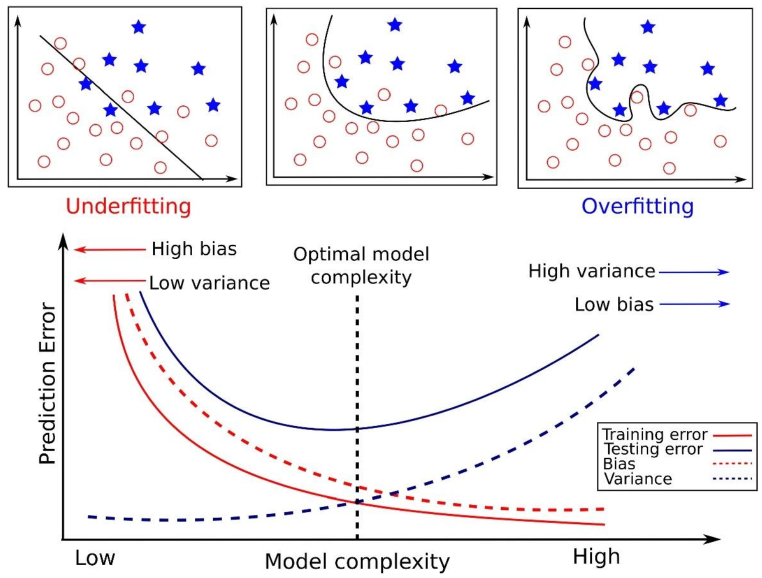
Regression vs Classification



The Bias-Variance Trade-off







$$\mathrm{E}\left[\left(y-\hat{f}(x)\right)^2\right]=\mathrm{Bias}\left[\hat{f}(x)\right]^2+\mathrm{Var}\left[\hat{f}(x)\right]+\sigma^2$$

Where:

$$\mathrm{Bias}\left[\hat{f}(x)\right]=\mathrm{E}\left[\hat{f}(x)-f(x)\right]$$

and

$$\mathrm{Var}\left[\hat{f}(x)\right]=\mathrm{E}[\hat{f}(x)^2]-\mathrm{E}[\hat{f}(x)]^2$$

Flexibility vs Interpretability

FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*