Model selection and resampling methods

MSc Data Science
2nd sememster

Audience

Minimal background in mathematics and statistic

- analysis and calculus (integral, derivatives, study of functions, …)
- basic statistical concepts (expectation, median, covariance, distributions, …)

Minimal knowledge on statistical modeling

- regression
- expectation, variance/covariance, statistical descriptors, …

Basic expertise with Python and Jupyter Notebook

- installing new packages
- writing basic code and running pipelines
- knowledge of standard libraries (numpy, pandas, scikit-learn)

The course

Based on lessons and notebooks

Additional reading material and references are provided at each lesson

All the material available at the course website

https://marcolorenzi.github.io/teaching.html

10 lessons

What is expected from you:

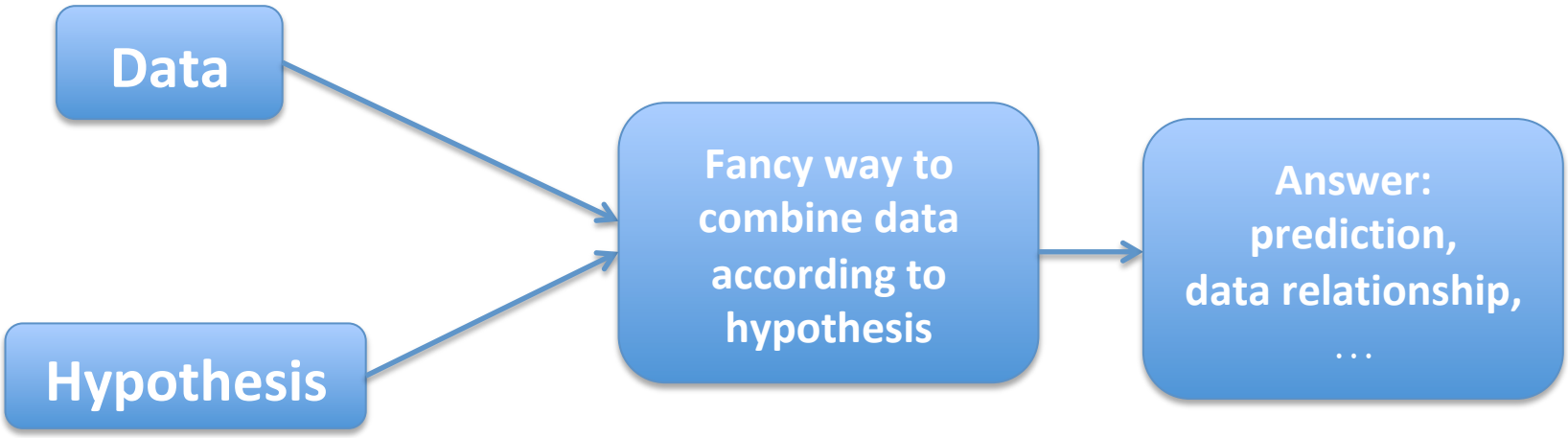
- Bi-weekly assignments:
  4 in total, 10 points

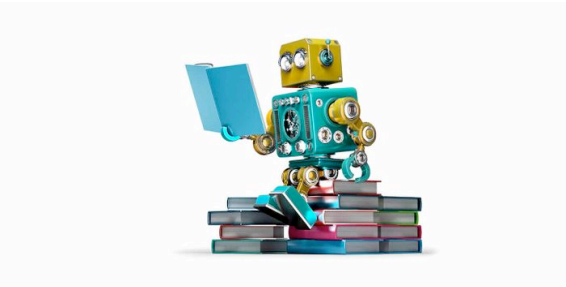
- Final exam:
  10 points

Why model selection?

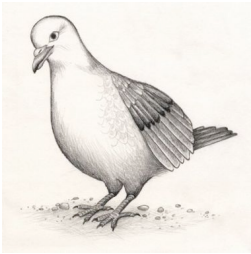
source: https://machinelearningmastery.com

The common denominator





Machine Learning and Pigeons



"… a clock is arranged to present the food hopper at regular intervals with no reference whatsoever to the bird's behavior".

- One bird was conditioned to turn counter-clockwise about the cage
- Another repeatedly thrust its head into one of the upper corners of the cage
- A third developed a 'tossing' response as if placing its head beneath an invisible bar and lifting it repeatedly
- Two birds developed a pendulum motion of the head and body
- Another bird was conditioned to make incomplete pecking or brushing movements directed toward but not touching the floor

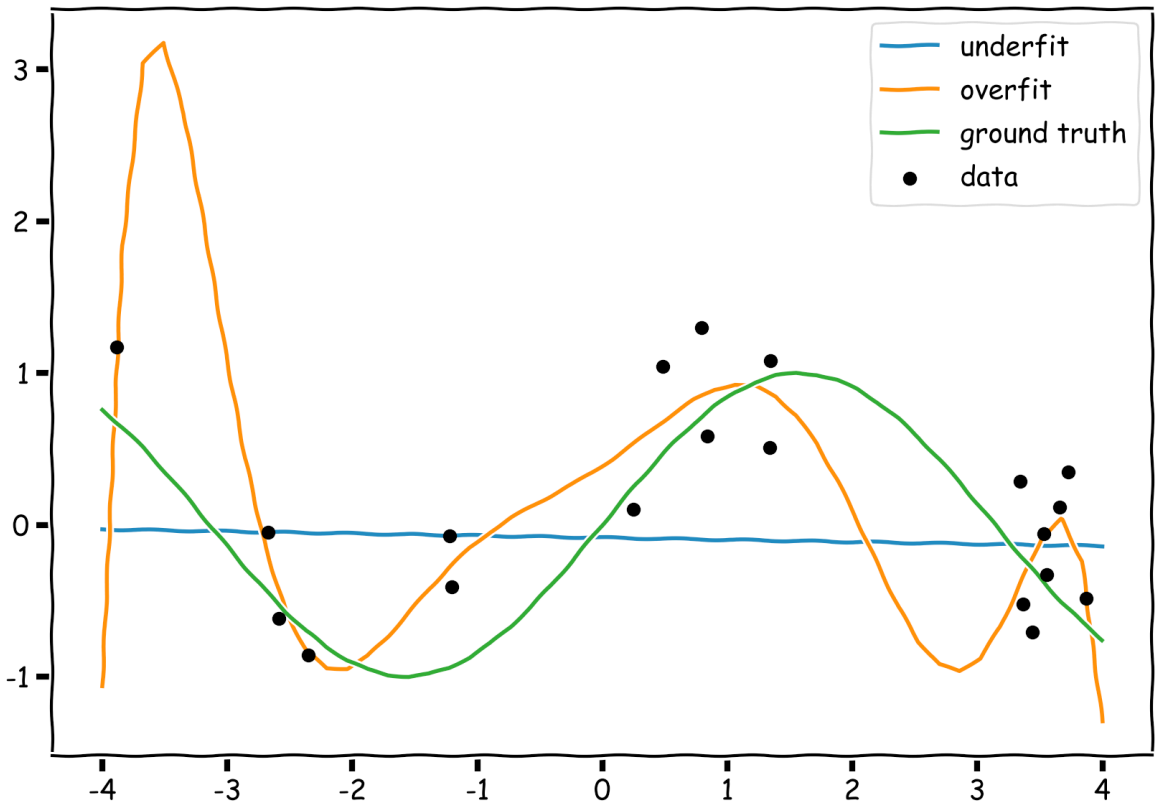
The bird happens to be executing some response as the hopper appears; as a result it tends to repeat this response

'Superstition' in the Pigeon
B.F. Skinner
Journal of Experimental Psychology #38, 1947

Models can be superstitious (and behave like pigeons)

They tend to enhance a behavior (prediction) when a positive response (data) is met.

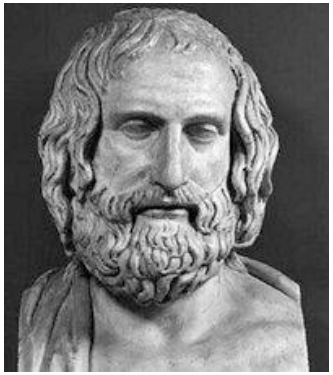
The “enhancing” ability depends on their assumptions:



A model is not **true**

A model provides an **opinion:**

- based on some sense of reality under the form of data
- based on its own assumptions



Relativism

"… there is no absolute evaluation of the nature … because the evaluation will be relative to who is perceiving it. Therefore, to Person X, the weather is cold, whereas to Person Y, the weather is hot. This philosophy implies that there are no absolute "truths". The truth, according to Protagoras, is relative, and differs according to each individual"

Protagoras, wikipedia.org

The job of a data scientist is to determine whether an opinion can be trusted

When different opinions are available, this is called **model selection**

Why model selection (2)

Tackling a whatever machine learning problem (regression, classifications, …)

1- Writing some code implementing a model idea

2- Getting the data from some repository

3- Training the model:

- 3a. Bug fixes, parameters hacking
- 3b. Use on all or on some part of the data, cross-validating, collecting results, …

4- Repeat point 3 many (many) times. Trying, trying, trying, …

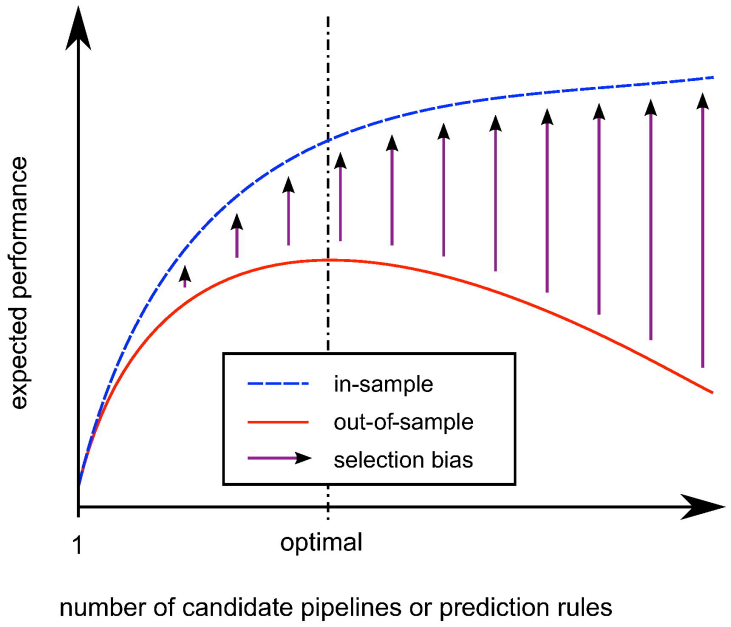
5- Once everything works reasonably, fixing all the parameters and processing steps

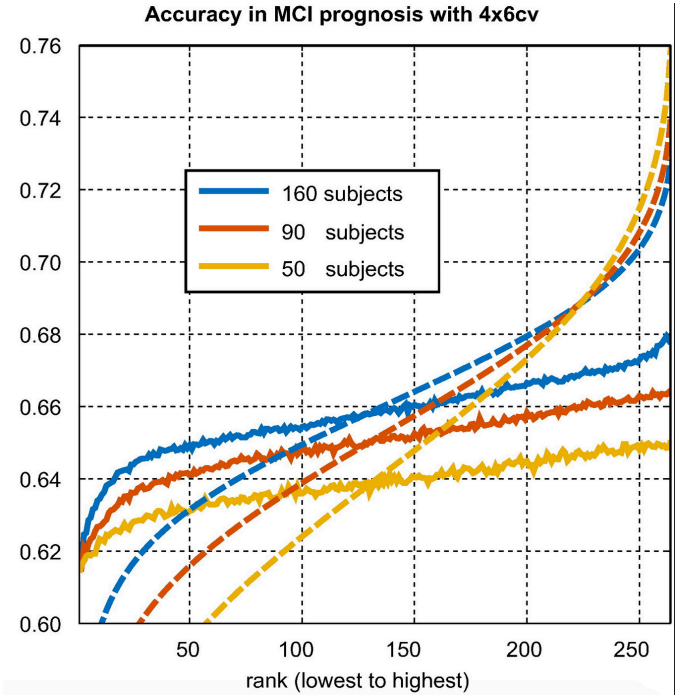
6- Publishing/reporting the results

Why model selection (2)

A case study:

automated clinical diagnosis of Alzheimer's disease





Mendelson et al. NeuroImage: Clinical, 2017

Standard approaches to model selection

- Empirical (part 1)

Jackknife
Bootstrap
Cross-validation
Step-wise comparison

- Theoretical (part 2)

Information Criteria
Bayesian model comparison

Part 1. A focus on resampling methods.

*"Pull yourself up by your **bootstraps**"*

… widely thought to be based on one of the eighteenth century Adventures of Baron Munchausen, by Rudolf Erich Raspe.

Efron & Tibshirani
An Introduction to Bootstrap, 1993

Key concept of **data resampling.** Doing our best out of the available resources.

At the end of the course you will be able to

- Critically assess the performance of the model on a specified task
- Identify and prevent the sources of assessment bias
- Create your own benchmark for a variety of modeling problem
- Identify modeling alternatives and evaluation strategies
- Visualize and present performances across models
- Understand the basis of theoretical approaches to model selection

Questions?