Machine Learning applied to Planetary Sciences

PTYS 595B/495B

Leon Palafox

https://leonpalafox.github.io/MLClass/

Validation Methods

This is were we know who is worthy of using ML

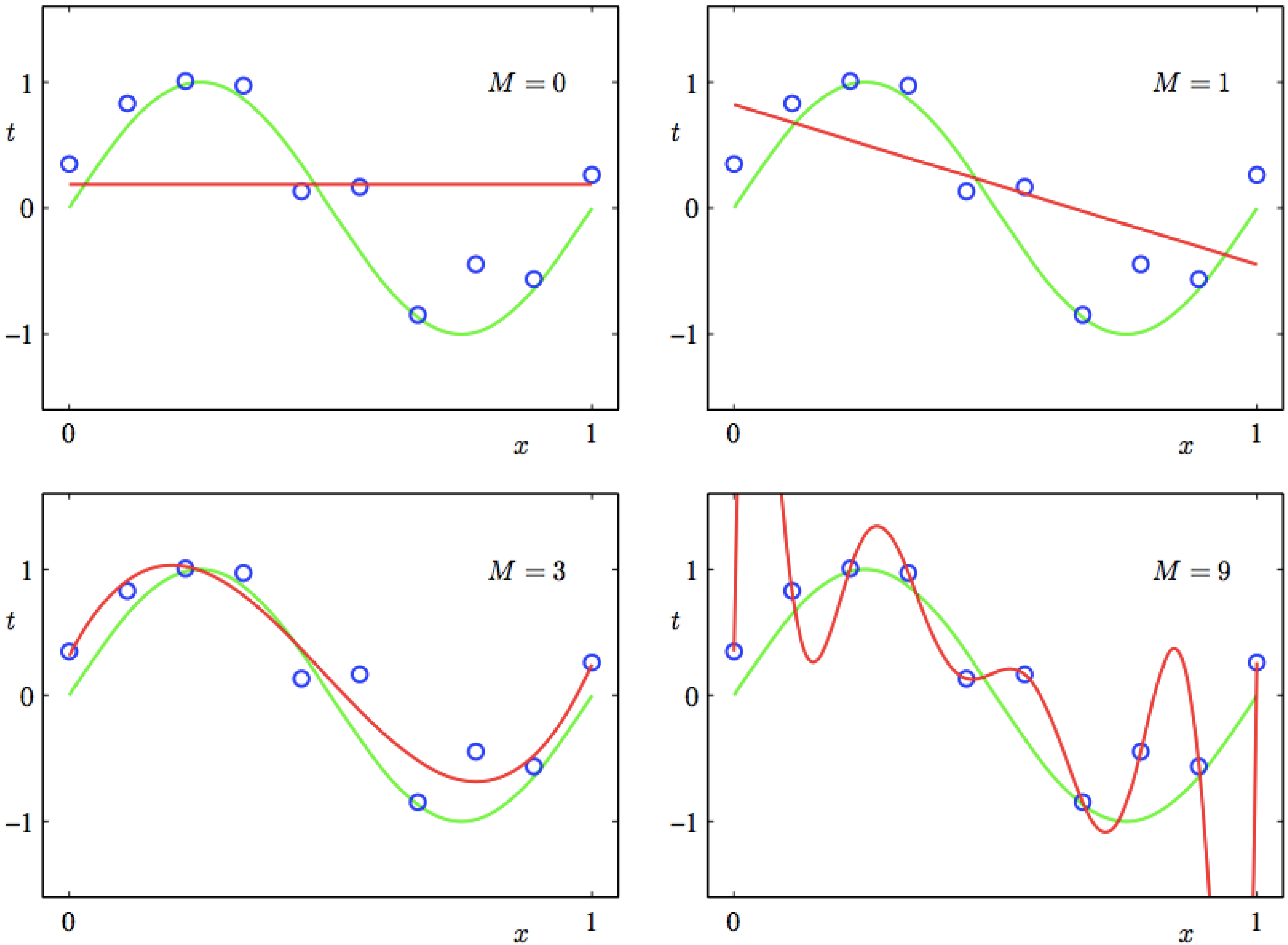
Validation Methods

- Cross validation
  - Test different models
  - Obtain reliable statistics
- Bias -- Variance Analysis
  - Regularization
  - Overfitting
- Area under the curve (AUC)

Cross Validation

- The hypothesis with the smallest training error, won’t be the best.
  - Why?
  - We need test sets and training sets
- Our first tool is called hold-out cross validation.

Smallest training error



Read team review

- What is a red team?
  - Independent (non-biased set of reviewers)
- Why do we need a red team.
  - Avoid journal overfitting.
  - Our public is not us, but a wider audience.
- Ideally there should be red teams for everything.
  - Public talks, presentations, etc.

General elections

- What is the difference between a parliamentary and presidential democracy.
  - Presidential democracy prevents overfitting.
    - Primaries
    - Opinion surveys
  - Parliamentary overfits
    - Prime minister is selected by politicians
    - Brexit!

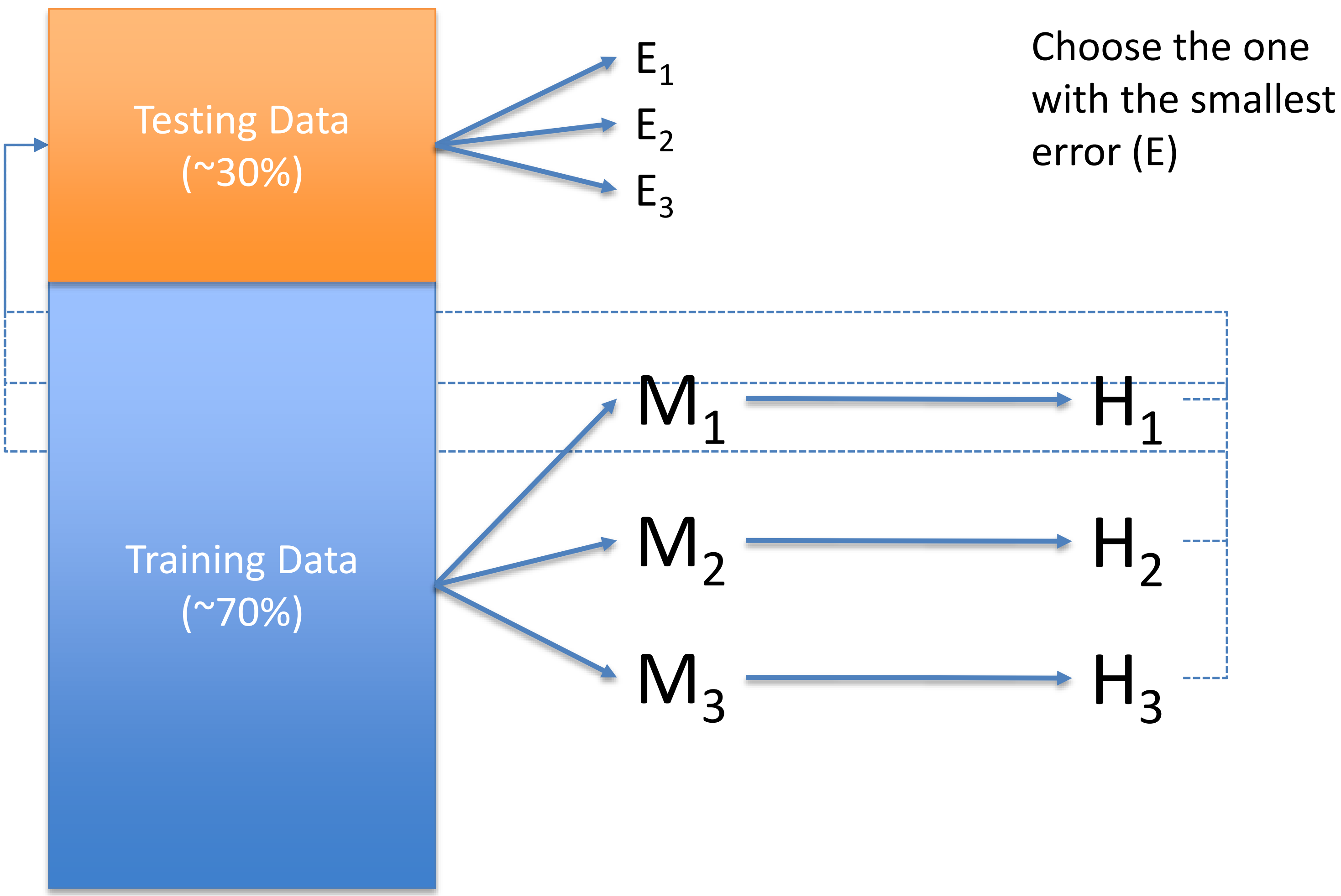
Movies

- Blockbusters try to generalize as much as possible.
  - Test audiences.
  - Big budget actors (not necessarily good).
  - “Wide appeal”
- Oscar winners (generally) overfit to a select group of movie critics.

Cross-Validation

- Is the most basic tool to prevent overfitting.
  - Machine Learning 101
- Is a systematic approach to find the best set of parameters in our algorithms.
  - SVM parameters
  - Regularization weights.
  - Size of the Neural Network

Hold-out cross validation



What is M

- Everything that we have assigned arbitrarily is fair game.
- Linear Regression
  - Order of the polynomial, regularization parameter
- SVM
  - Kernel, variables associated with kernel
- NN
  - Number of layers, activation functions, number of units.

Problems with Hold-out CV

- We are “wasting” ~70% of our data.

- For problems with few data points, this is just not desirable

- Be wary of papers that used CV, but have only few data points.
  – Be even more skeptic of papers that don’t mention CV at all.

An even better CV

- K-fold CV
  - Split the data into k subsets (disjoint)
  - For each j = 1..k
    - Train model (M_i) in every subset, except j
    - Get an error (E_{ij}) for Model i in iteration j
  - Total error for M_i is going to be the average of all the errors (E_{ij})

K-Fold Cross validation

K-Fold Cross validation

M_1

M_4

M_2

M_3

Advantages of CV

- We can run each fold in an independent CPU.
  - Sklearn is optimized to do the folds in parallel.

- Unlike other validation approaches is straightforward in its implementation.
  - Bayesian techniques are particularly convoluted when it comes to validation methods.

Disadvantages

- If we don't have much data, the folds are going to be very correlated.
  - This still results in an overfitting to the data.
- It obviously takes more time to run the algorithm (K x #Models) times.
  - Neural Networks take remarkably long times to be run a single time.
- The big one:

How many folds are good?

Say hi to your first hyperparameter, which is a parameter to set parameters.

How many folds are good

- K = 3 does a decent job
- If is a simple algorithm, like SVM or Logistic regression, you can always use 10.
- Leave-one out (K=N-1) is a travesty and should not be used.
  - We have as many folds as data points -1.
  - Glorified bootstrapping.