COMPUTATIONAL STATISTICS II

Professor: Alessia Pini

PhD program in Economics and Statistics (ECOSTAT)

PRACTICAL INFORMATION

PROGRAM OF THE COURSE

1. **Validation of a model**
    - **Validation set approach**
    - **K-fold cross-validation**
    - **Leave-one-out cross validation**
2. **Bootstrap**
    - **Introduction to Bootstrap**
    - **Bootstrap confidence intervals**
    - **Bootstrap tests**
3. **Introduction to EM**

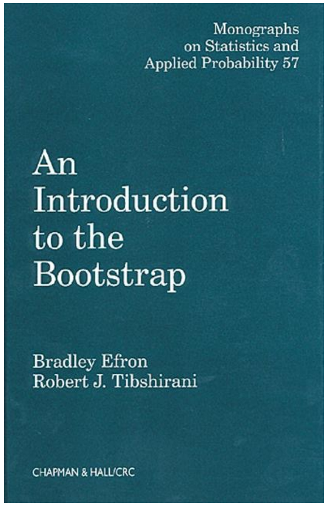
COURSE TEACHER

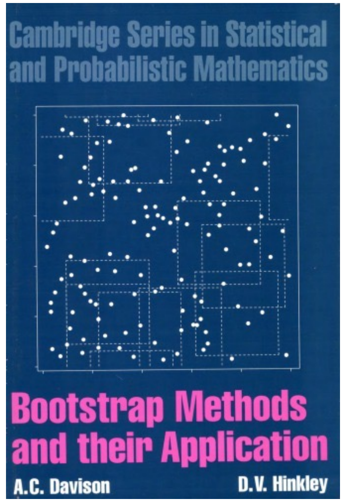
Alessia Pini (Università Cattolica del Sacro Cuore)

Email: alessia.pini@unicatt.it

Web page: http://docenti.unicatt.it/ita/alessia_pini/

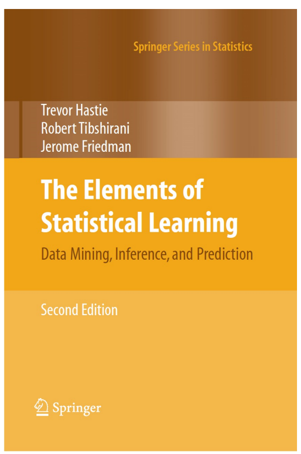
MAIN TEXTBOOKS



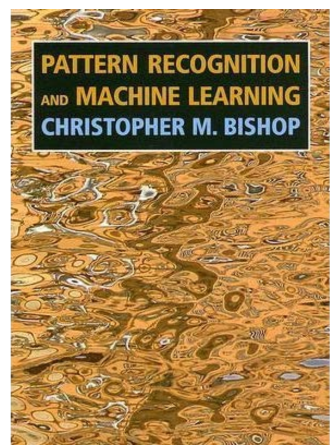


Bootstrap:
An Introduction to the Bootstrap
By Efron, Tibshirani

Bootstrap Methods and their Applications
By Davison, Hinkley



Model validation:
The Elements of Statistical Learning
By Hastie, Tibshirani, Friedman



EM:
Pattern Recognition and Machine Learning
By Bishop
Download at this link

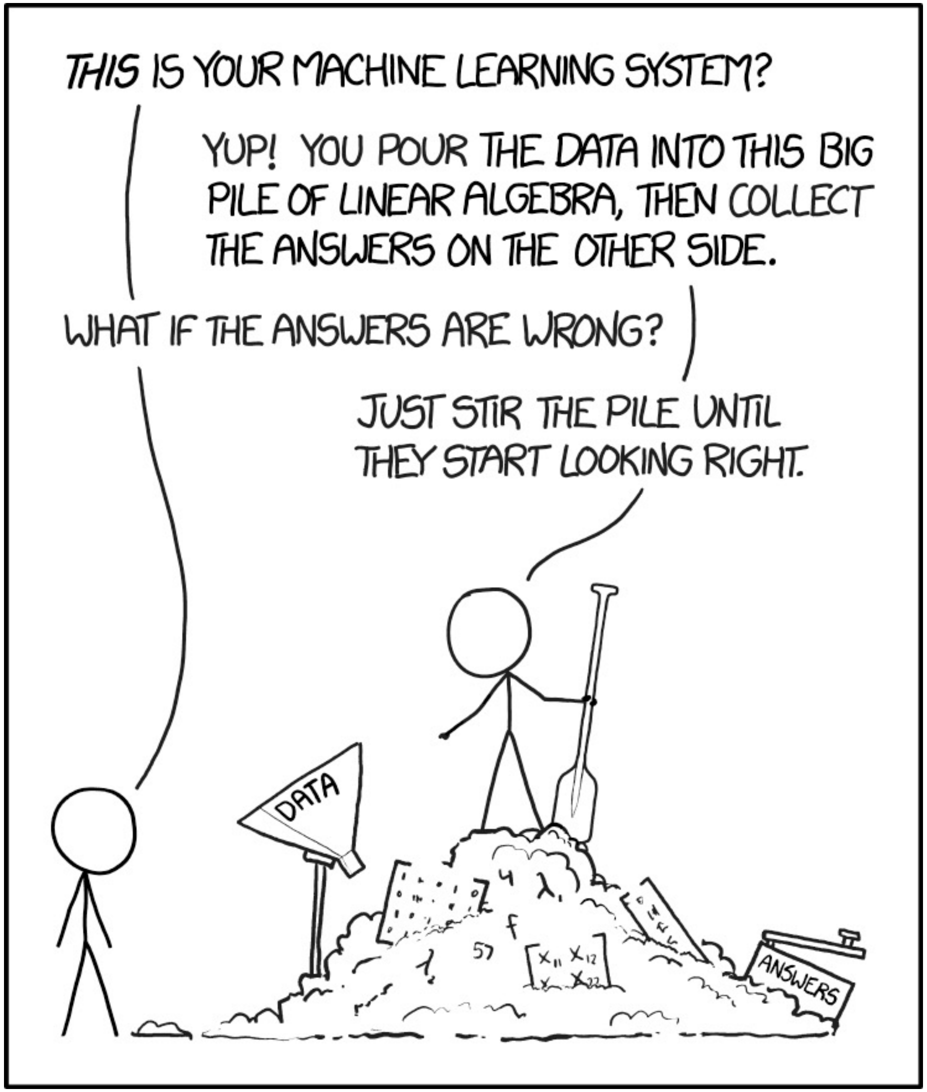
METHODS FOR MODEL VALIDATION

WHAT IS STATISTICS?

Statistician

Information
Knowledge

THE IMPORTANCE OF MODEL VALIDATION



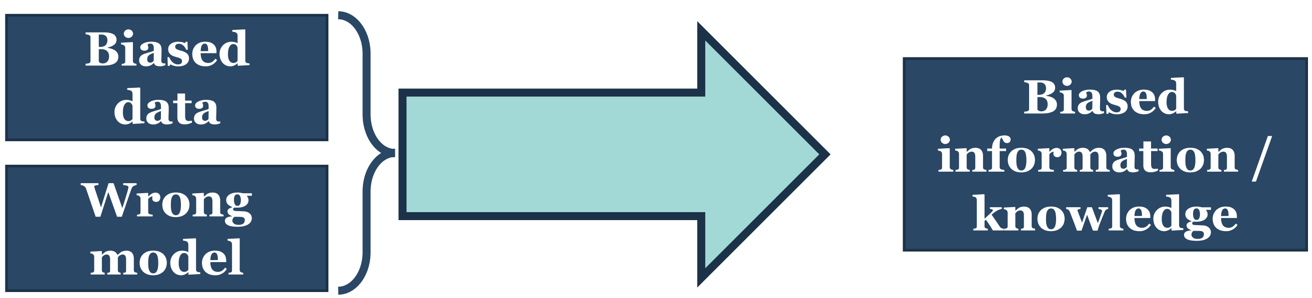
THE IMPORTANCE OF MODEL VALIDATION

REUTERS

Business Markets World Politics TV More

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women



MODEL ACCURACY

How can we assess if a model is working correctly? How to choose between different models?

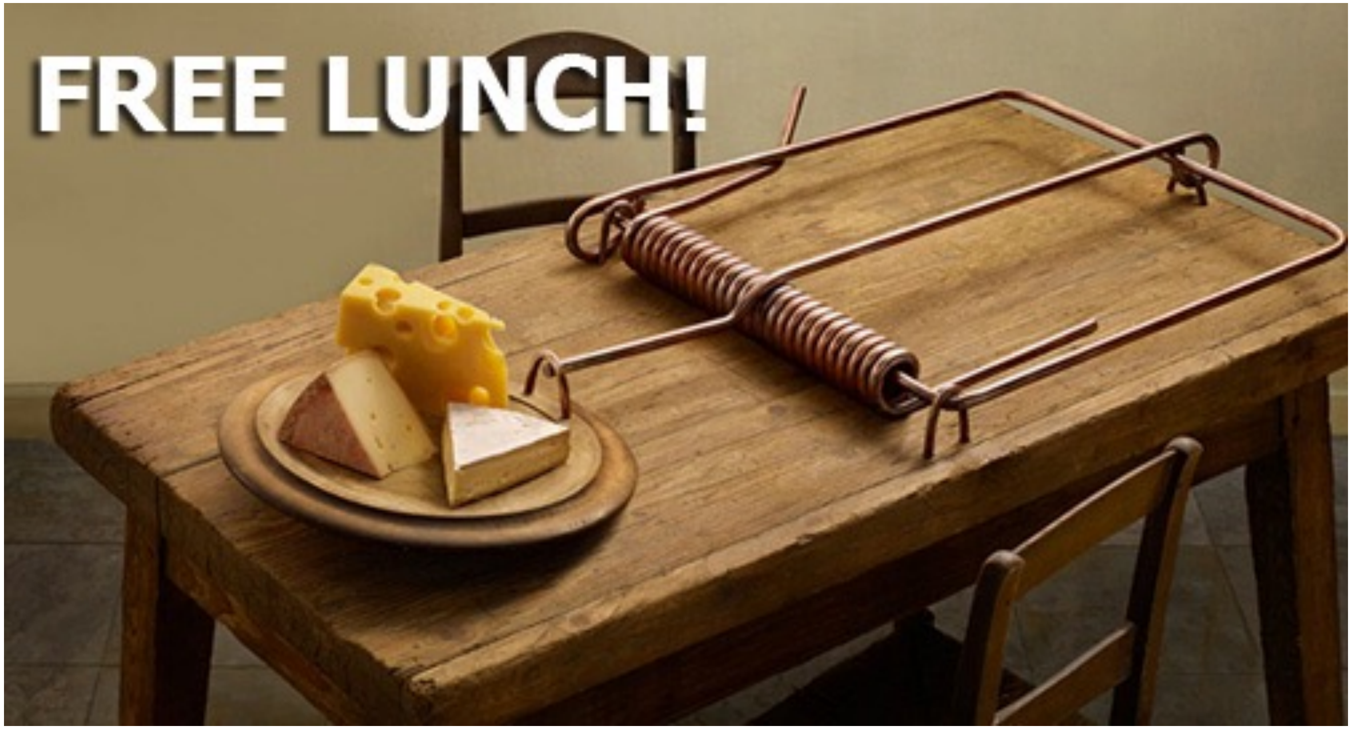
Is there a method that dominates all other methods over all possible data sets?

MODEL ACCURACY

How can we assess if a model is working correctly? How to choose between different models?

Is there a method that dominates all other methods over all possible data sets?

There is no such a thing as free lunch.



MODEL ACCURACY

How can we assess if a model is working correctly? How to choose between different models?

Is there a method that dominates all other methods over all possible data sets?

There is no such a thing as free lunch.

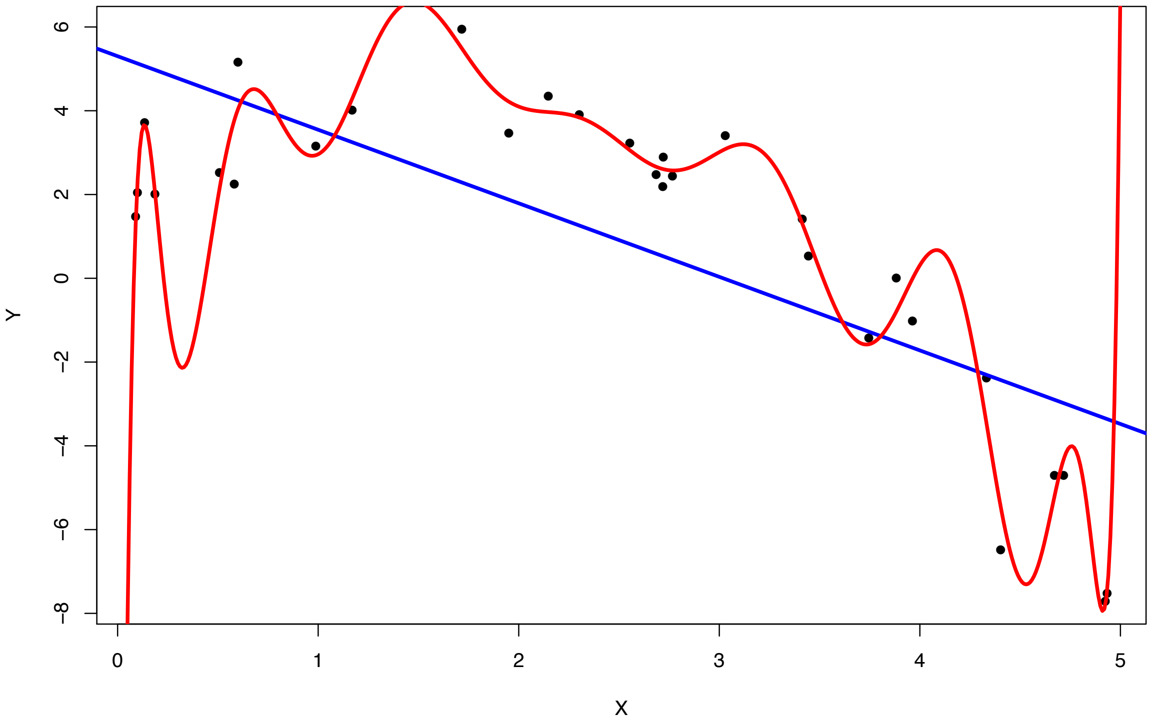
⇓

No one method dominates all other methods over all possible data sets.

We need methods to assess if how well the estimated model matches the data.

RISKS OF A WRONG MODEL

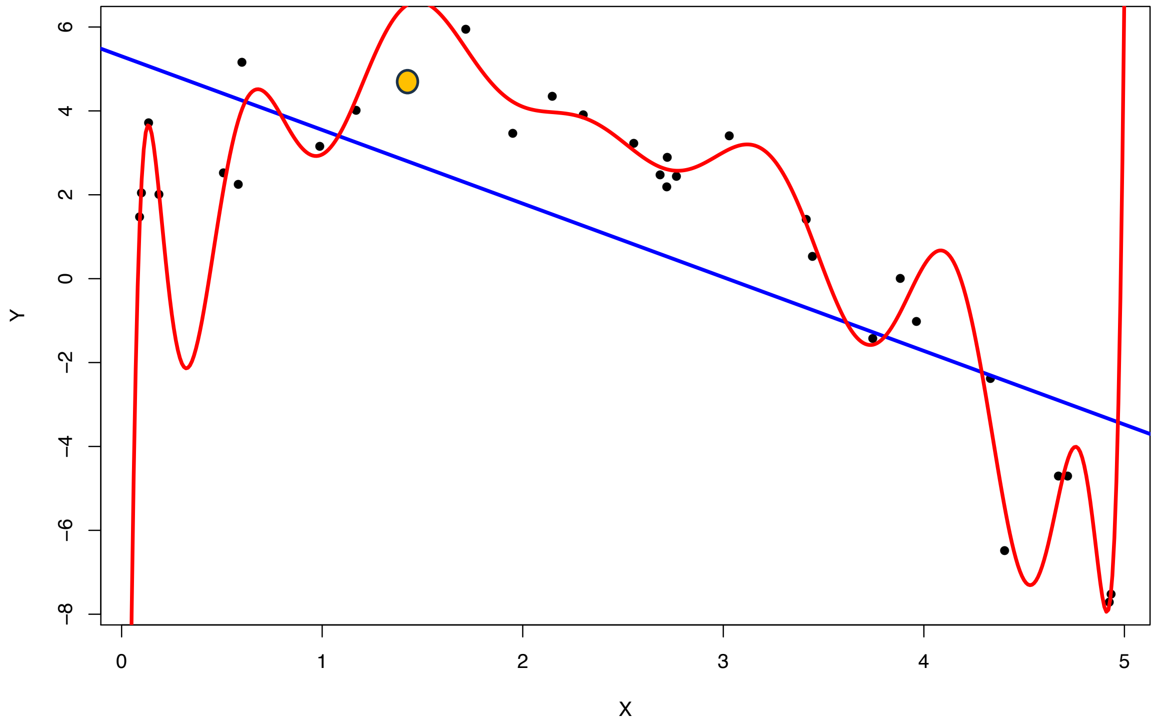
Underfitting / Overfitting



Underfitting: model is too simple to follow data
Overfitting: model is too complex, and follows too closely data (affected by error)

I RISCHI DI UN MODELLO ERRATO

Underfitting / Overfitting



Underfitting: model is too simple to follow data

Overfitting: model is too complex, and follows too closely data (affected by error)

In both cases, we make an error in estimating a new observation

MODEL ACCURACY - REGRESSION

Model accuracy in regression can be evaluated using the mean square error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_{i1}, \ldots, x_{ip}))^2$$

Problem

- The model is fitted using the training set, and MSE is computed on the same data.
- The MSE is generally low when the model is flexible.
- It is **always** possible to find a model with zero MSE (e.g., polynomial regression with n-1 coefficients).

MODEL ACCURACY - REGRESSION

Idea:

Compute the MSE on a different data set.

Test MSE: mean square error for test observations (new observations that were not used to train the model).

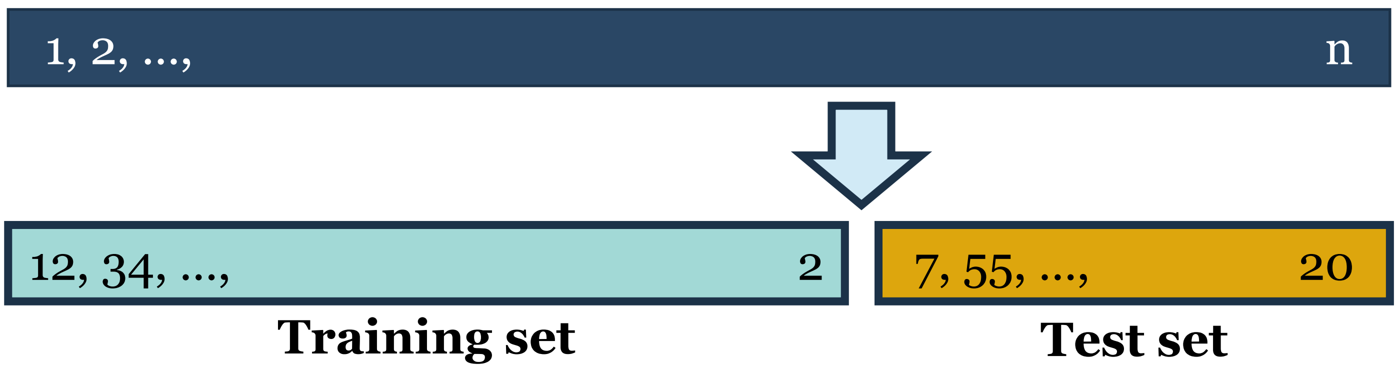
$$\text{MSE}_{\text{TEST}} = \mathbb{E}[(y_{new,i} - \widehat{f}(x_{new,i1}, \ldots, x_{new,ip})^2]$$

Such quantity depends on the data distribution, which is generally unknown. We need a way to estimate it.

We would like to compute the error that a model is committing in estimating a new observation.

MEASURING MODEL ACCURACY: VALIDATION SET

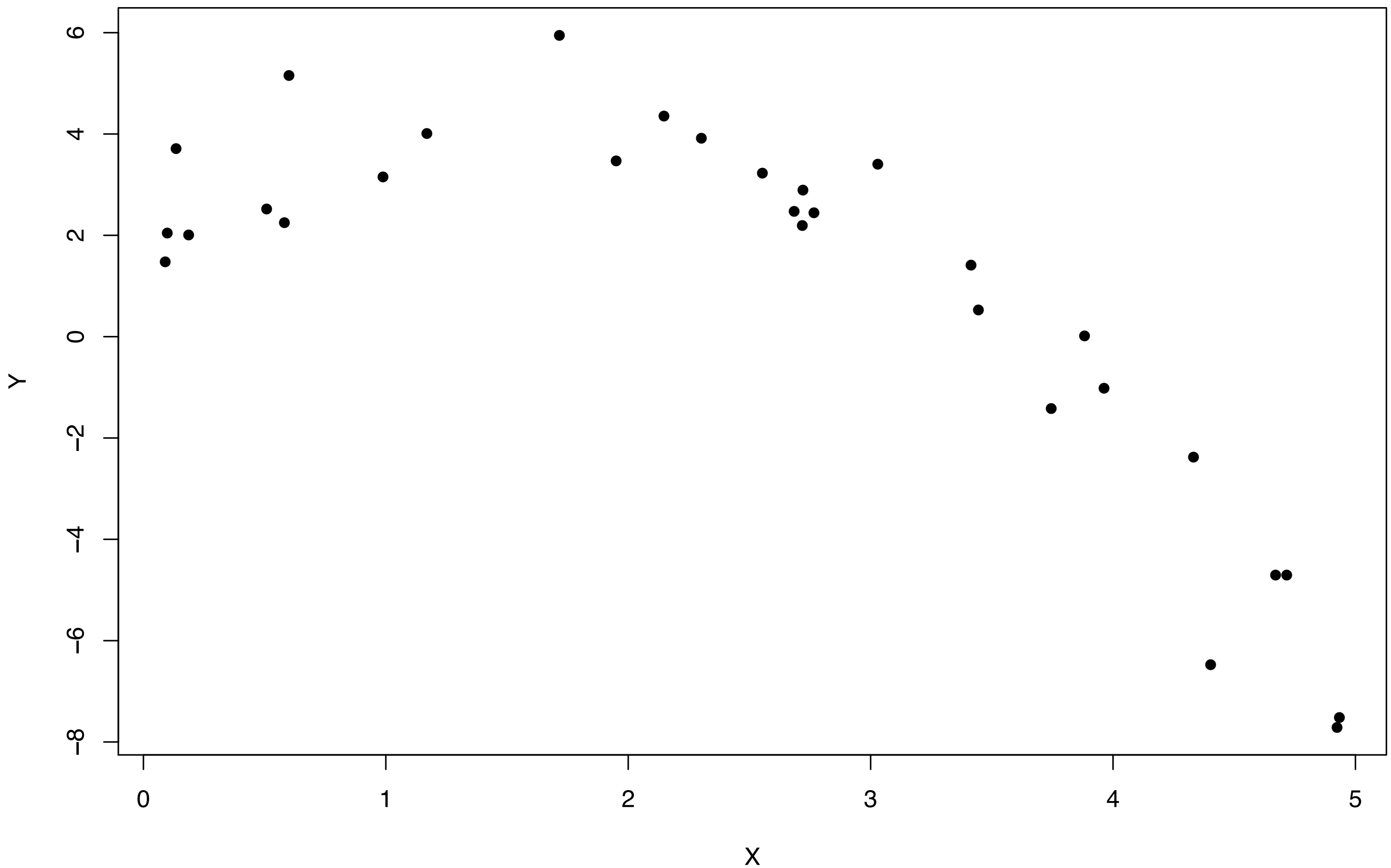
The validation set approach consists in splitting the original dataset into a training set (used for fitting the model) and a test set (used for estimating the MSE).



$$\widehat{MSE}_{TEST} = \frac{1}{n_{test}} \sum_{i \in test} (y_i - \hat{y}_i)^2$$

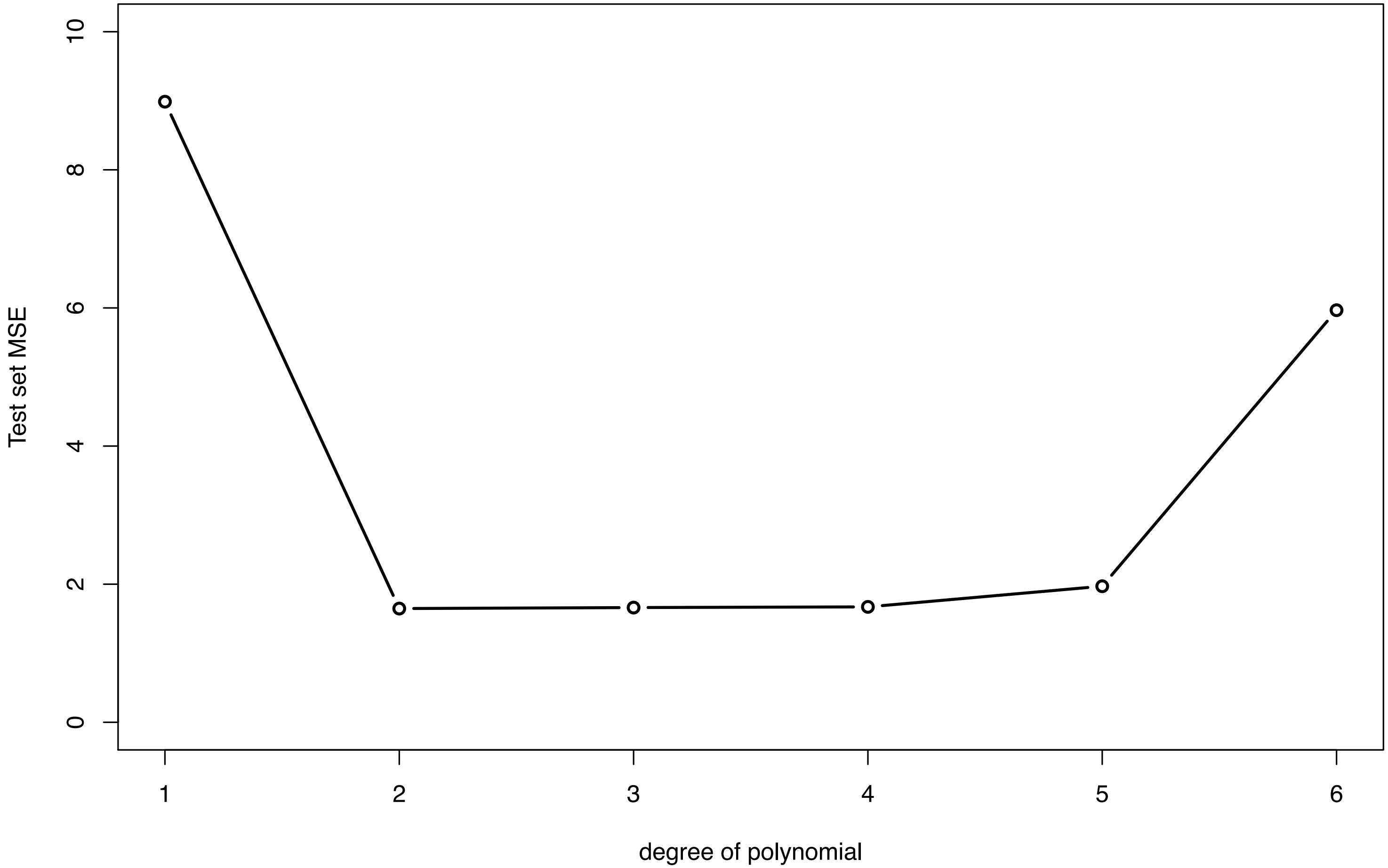
MEASURING MODEL ACCURACY: VALIDATION SET

Example on simulated data



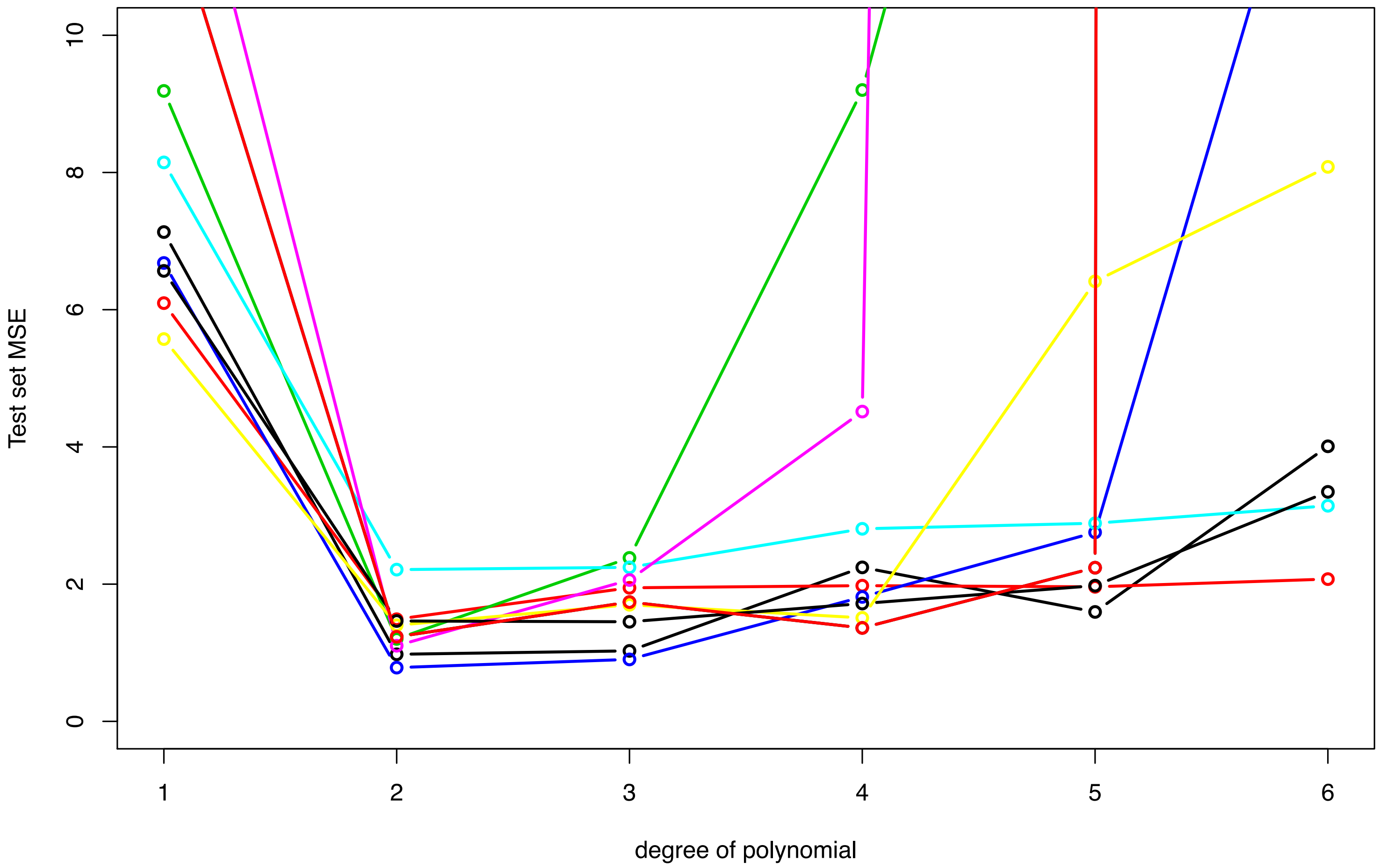
MEASURING MODEL ACCURACY: VALIDATION SET

Example on simulated data



MEASURING MODEL ACCURACY: VALIDATION SET

Example on simulated data



MEASURING MODEL ACCURACY: VALIDATION SET

Pros / Cons:

+ Easy to implement, very fast to run.

− The error estimate depends on the initial choice of training/test set.

− Only a subsample of the original data set is used to train the model. Hence, the fitting error on the entire dataset is overestimated.

MEASURING MODEL ACCURACY: VALIDATION SET

Example: estimation of MSE of a linear regression.

MEASURING MODEL ACCURACY: K-FOLD CROSS-VALIDATION

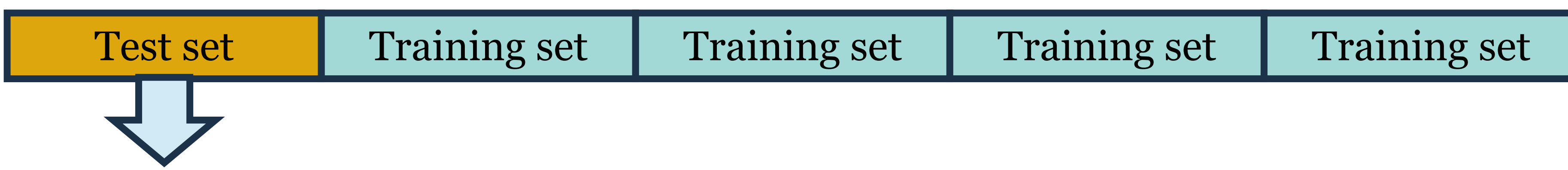
- The dataset is randomly split into K parts (folds) of approximately equal dimension.

Example: K=5

MEASURING MODEL ACCURACY: K-FOLD CROSS-VALIDATION

- The dataset is randomly split into K parts (folds) of approximately equal dimension.
- Repeat for each fold k=1,2,…,K:
  - The fold k is used as test set and all other are together the training set.
  - Compute the average squared prediction error for each fold.

Example: K=5



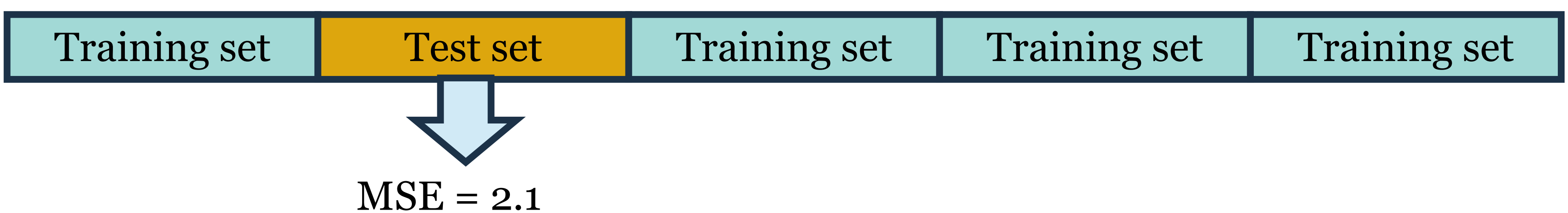
MSE = 1.4

UNIVERSITÀ CATTOLICA del Sacro Cuore

MEASURING MODEL ACCURACY: K-FOLD CROSS-VALIDATION

- The dataset is randomly split into K parts (folds) of approximately equal dimension.
- Repeat for each fold k=1,2,…,K:
  - The fold k is used as test set and all other are together the training set.
  - Compute the average squared prediction error for each fold.

Example: K=5



MEASURING MODEL ACCURACY: K-FOLD CROSS-VALIDATION

- The dataset is randomly split into K parts (folds) of approximately equal dimension.
- Repeat for each fold k=1,2,…,K:
  - The fold k is used as test set and all other are together the training set.
  - Compute the average squared prediction error for each fold.
- Average the obtained results.

Example: K=5

1.4	2.1	1.6	1.3	1.8

1.64

MEASURING MODEL ACCURACY: K-FOLD CROSS-VALIDATION

Pros / Cons:

+ Largely used.
- The error estimate still depends on the initial partition into folds, even though the dependence is weaker than in the case of validation set.
- Only a subsample of the original data set is used to train the model. Hence, the fitting error on the entire dataset is overestimated.
+ However, the test set is usually of a smaller size wrt the validation set, so the bias is lower.
+ Computationally more expensive than validation set approach, but generally affordable.

MEASURING MODEL ACCURACY: LOOCV

Special case: if K=n we obtain a method called leave-one out cross validation (LOOCV). At each iteration, the test set only contains one observation.

$$\widehat{\text{MSE}}_{\text{TEST}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f_{(-i)}}(x_{i1},\ldots,x_{ip}))^2$$

Prediction error on the *i*th observation

Model estimated using as training set all observations except the *i*th one

MEASURING MODEL ACCURACY: LOOCV

Special case: if K=n we obtain a method called leave-one out cross validation (LOOCV). At each iteration, the test set only contains one observation.

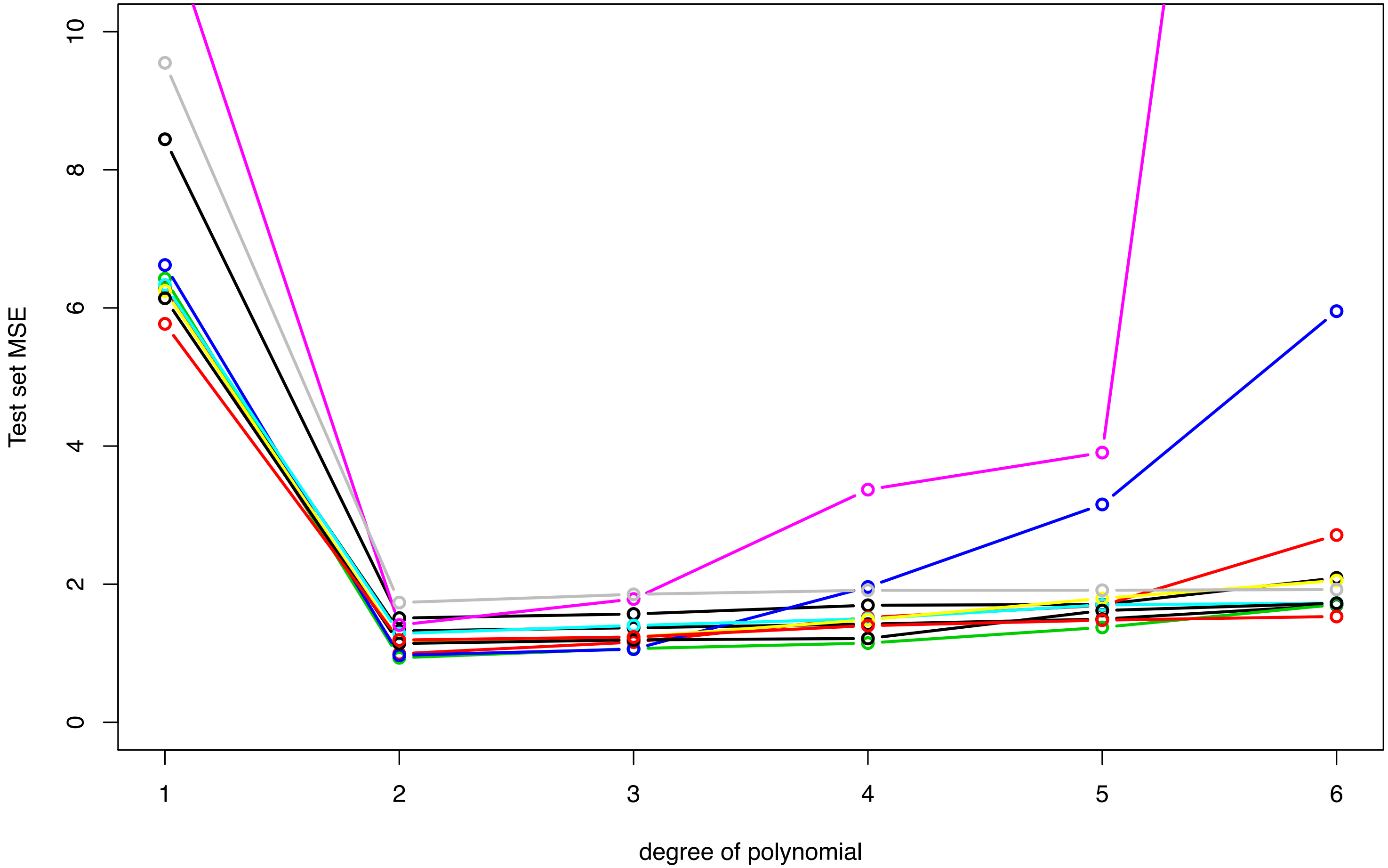
Pros / Cons:

+ The error estimate does not depend on the initial partition into folds, since in this case it is not random.
+ Almost all data are used for fitting the model, so the error is not overestimated.
- Different iterations gives correlated error estimates, since the training sets are very similar between each other. Therefore, the final estimate is affected by high variance.
- If n is large, LOOCV is computationally very expensive.

⇨ A k-fold cross-validation with 5-10 folds is typically a good compromise.

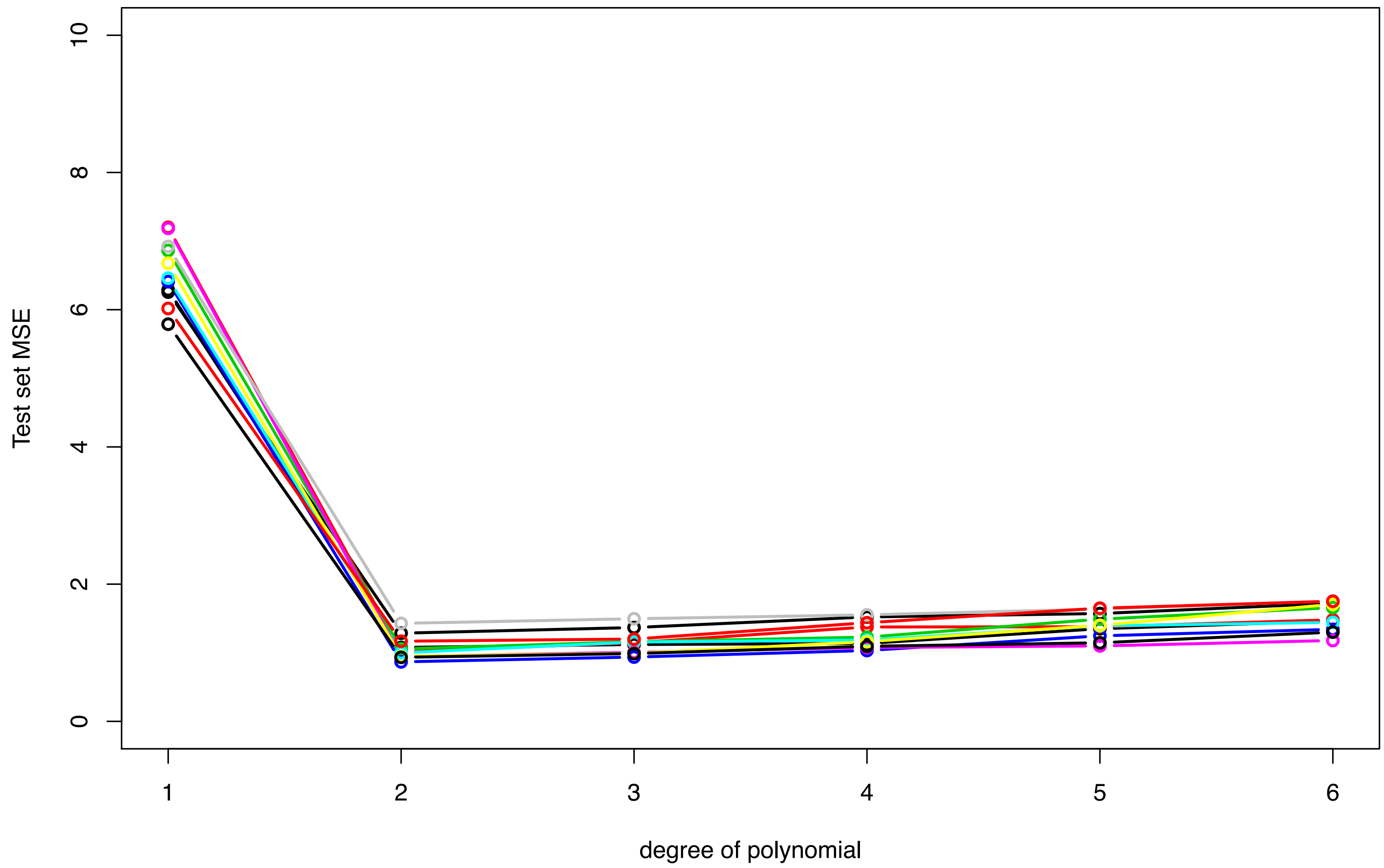
MEASURING MODEL ACCURACY: LOOCV

Example: 3-folds



MEASURING MODEL ACCURACY: LOOCV

Example: 5 folds



MEASURING MODEL ACCURACY: LOOCV

Example: LOOCV