Hes·so
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences Western Switzerland

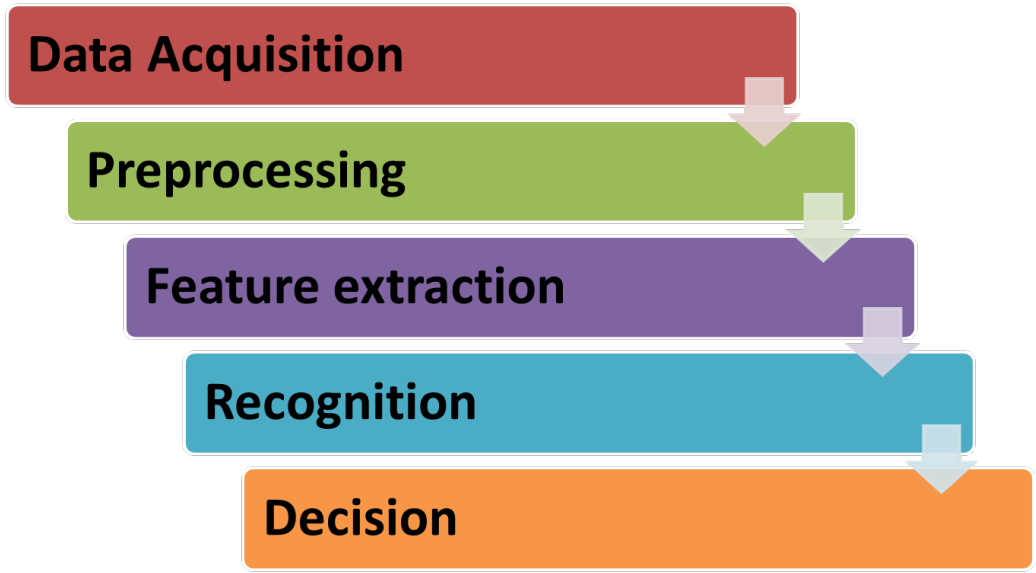
MSE | MASTER OF SCIENCE IN ENGINEERING

Evaluation of classifiers and (general) advice on machine learning

Francesco Carrino, Elena Mugellini, Omar Abou Khaled, Stefano Carrino

Outline

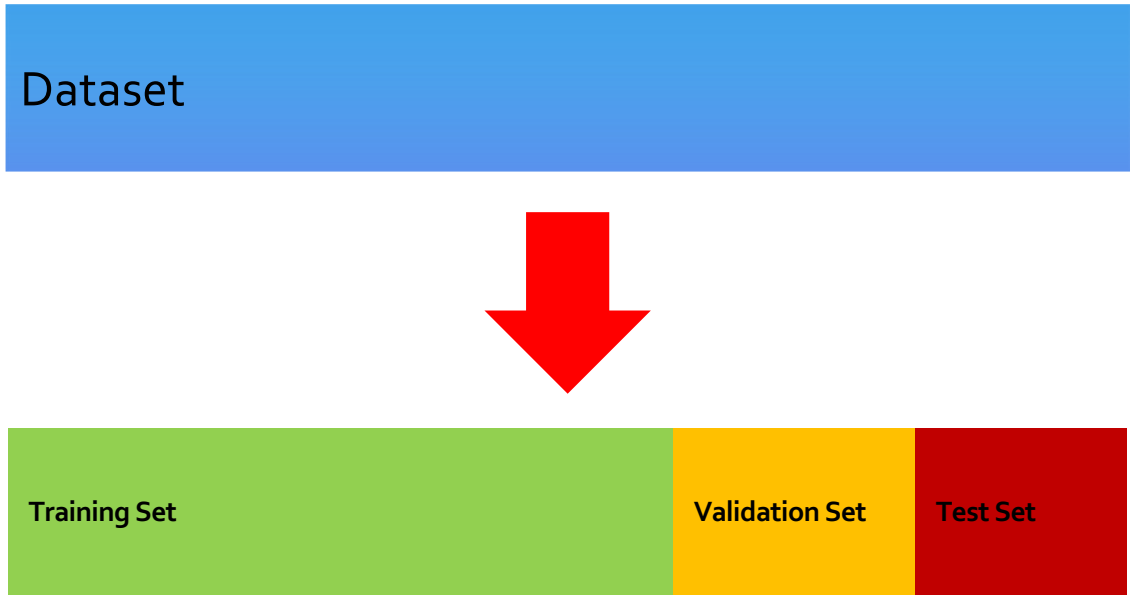
- Introduction
- Unbalanced training set
- Diagnosis bias vs. variance
- Cross-Validation
  - Definition
  - Motivations and goals
  - Procedures and applications
- Performance indicators
  - Confusion matrix
  - Accuracy, Precision, Recall and Specificity



Introduction

- Advices on (supervised) machine learning
  - Random Forest, HMM, SVM, NN, etc.…
- How to **properly** manage data?
  - Feature selection
  - Normalization
  - Dataset splitting
- How to **properly** evaluate the classification results?

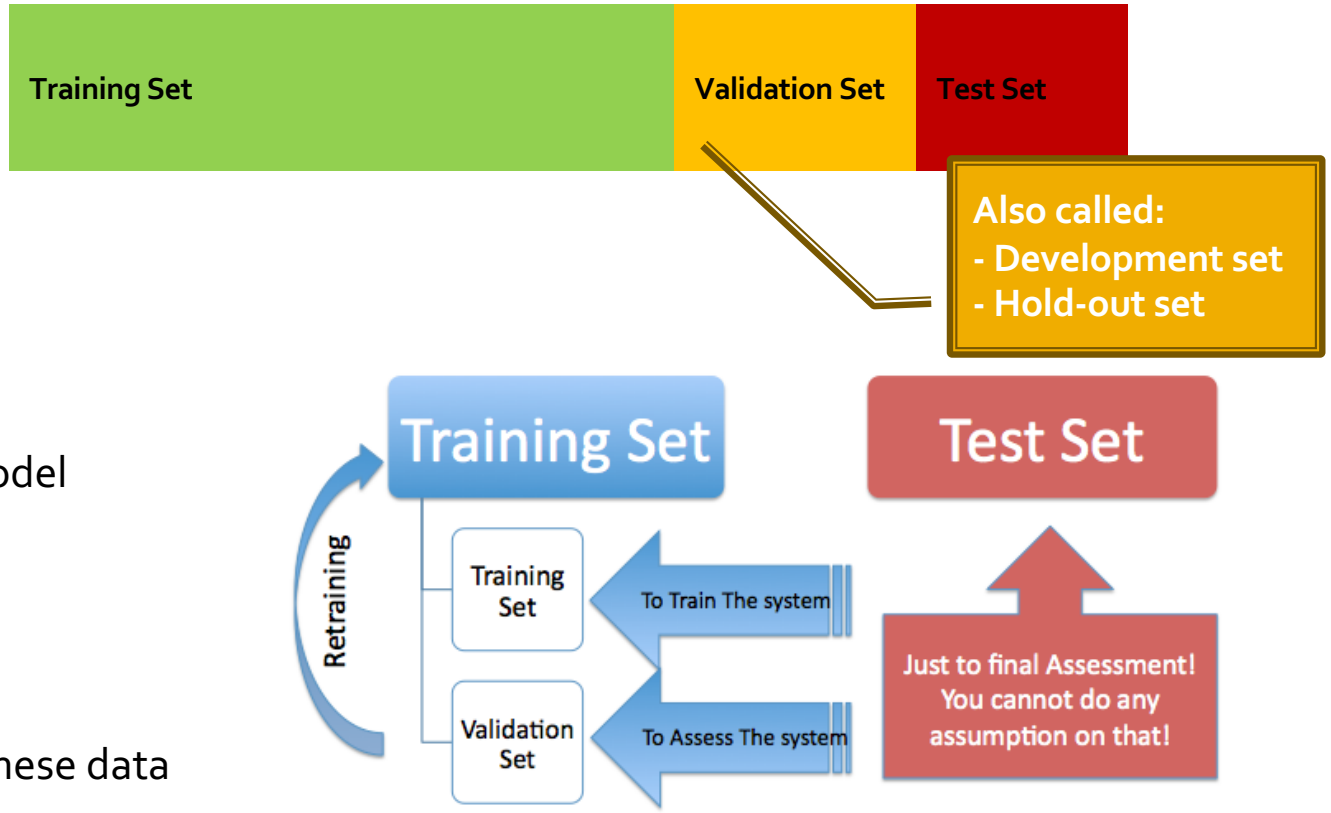
Learning Process– General Schema



Learning Process– General Schema

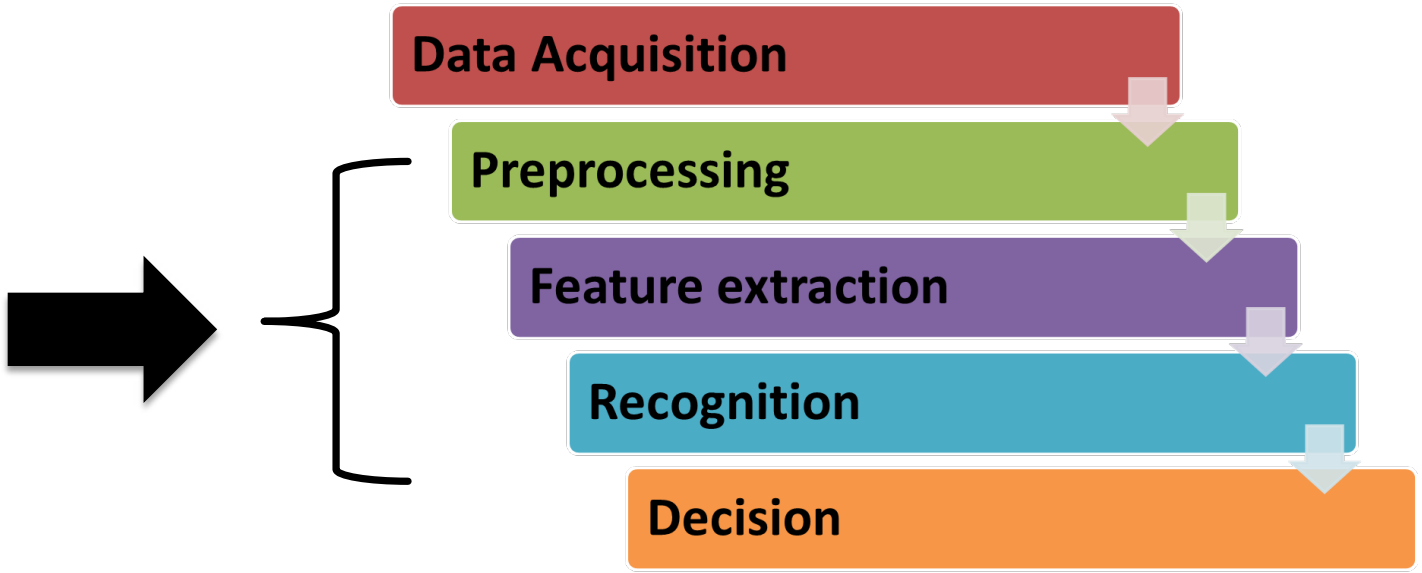
- Steps:

1. Training Set:
   - Feature extraction
   - Data Model
2. Validation Set
   - Optimization of the model
3. Iterate 1 and 2
4. Test Set:
   - Final assessment!
   - No assumption using these data

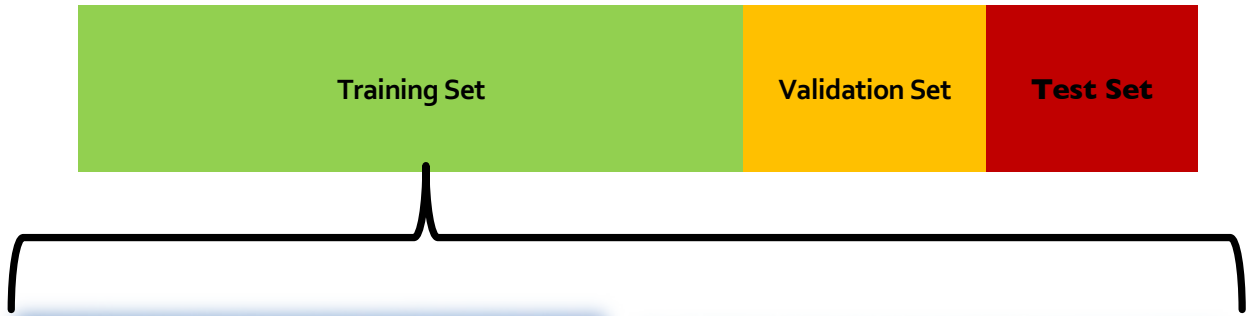


http://textanddatamining.blogspot.ch/2011/09/how-classifier-accuracy-is-conditioned.html

Balancing Training set



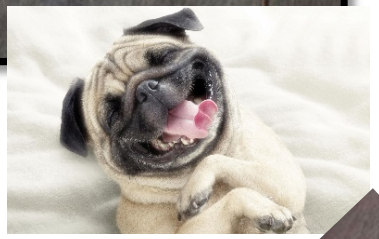
Unbalanced training set?

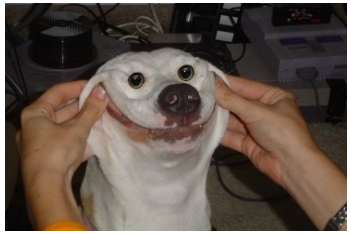


- Fish or duck?
  - Training set "skewed" (or unbalanced)

Unbalanced training set?

- Fish c
  - Trai
    "ske
    (or u
- With
  classi





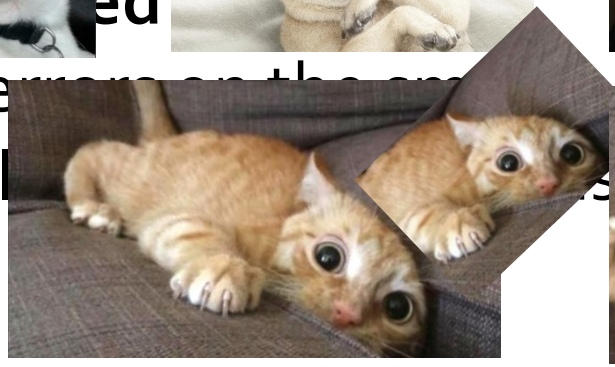
ion

e t set ed

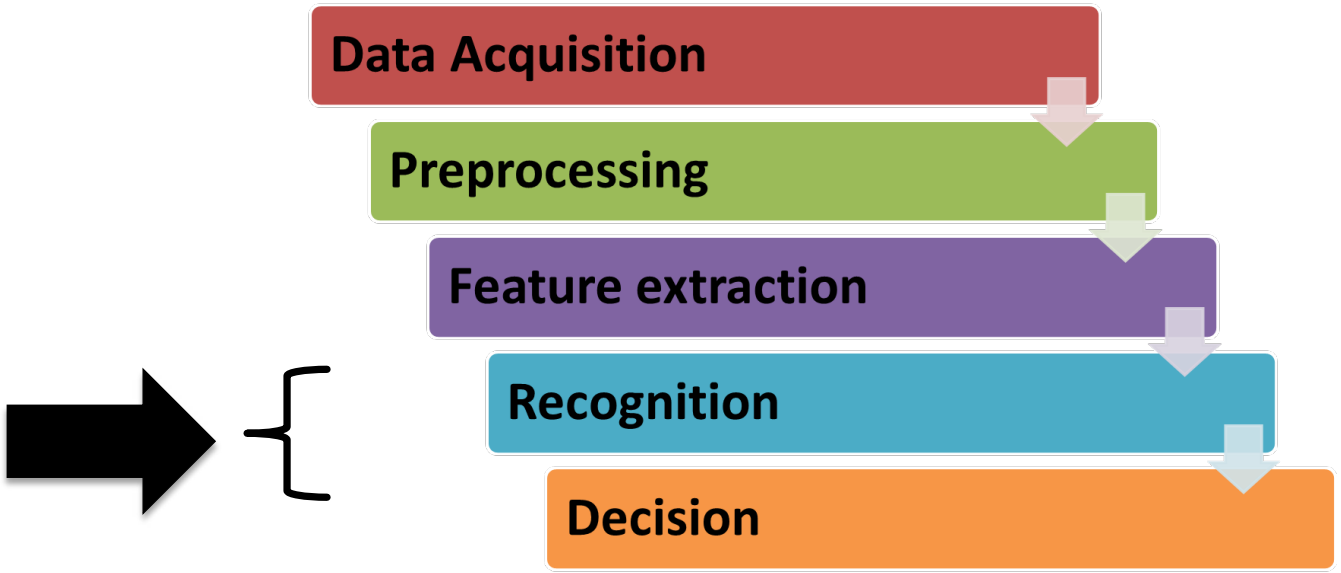
ortance to e

ging to the

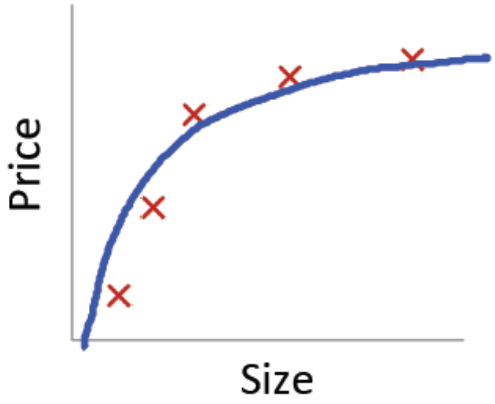
epeated



Diagnosis Bias Vs. Variance Or Underfitting Vs. Overfitting [7]

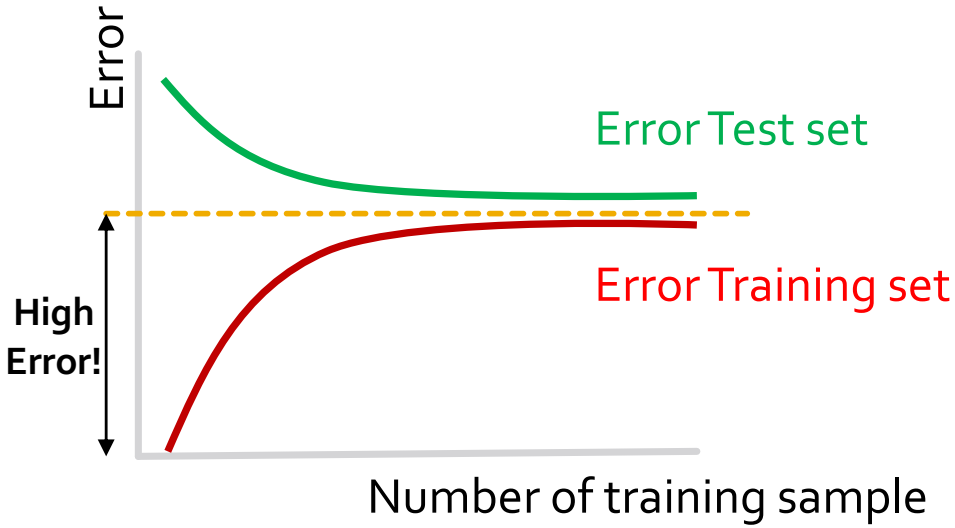


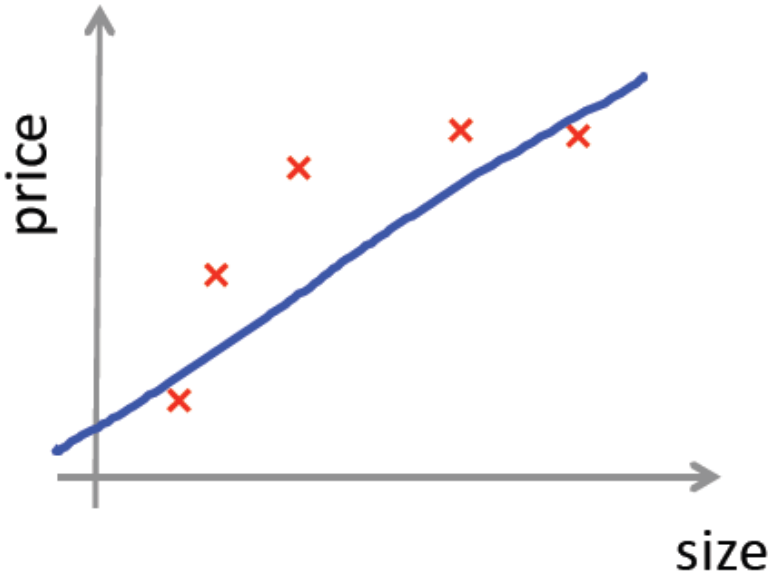
Bias Vs Overfitting



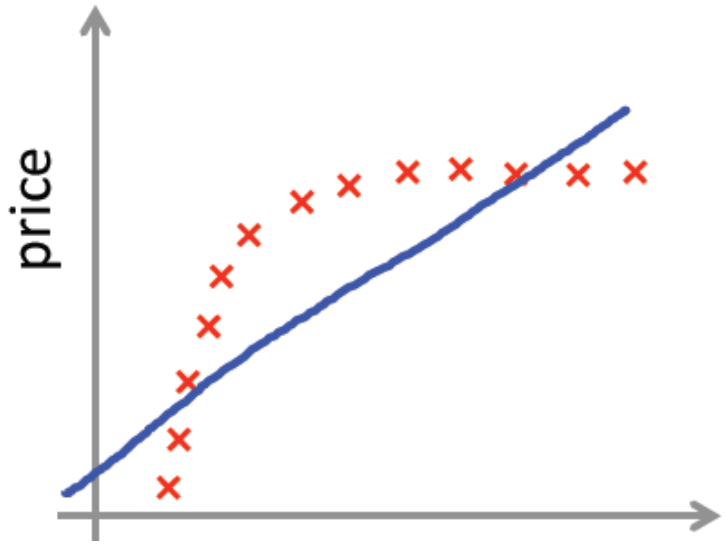
$$\theta_0 + \theta_1 x + \theta_2 x^2$$

High Bias

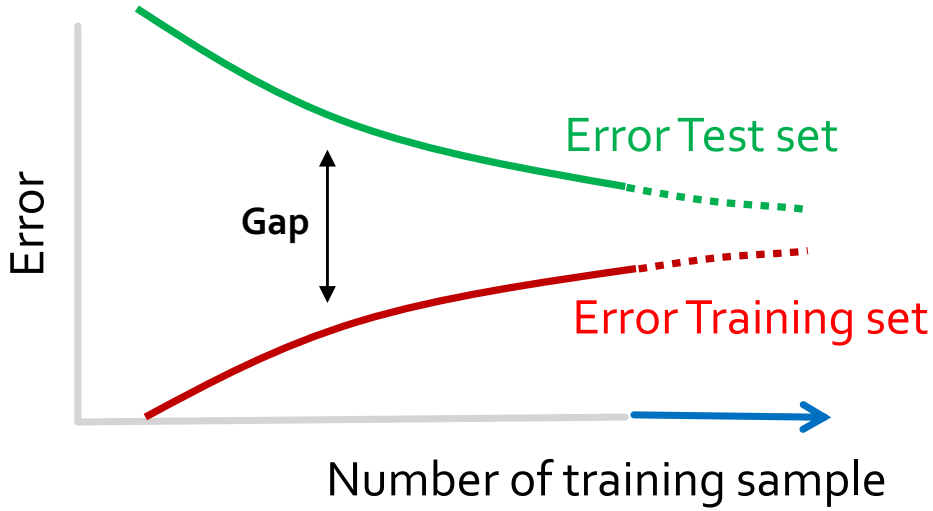


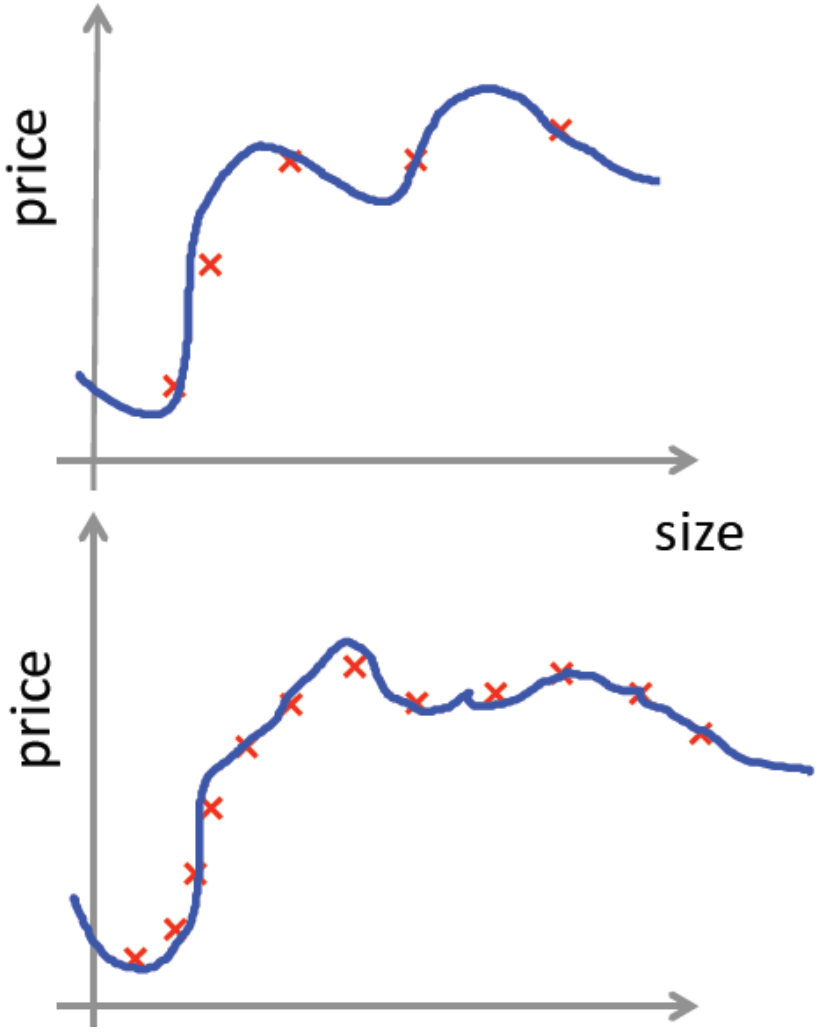


- **High error in the beginning**
- **Getting more training data will NOT help!**



High Variance

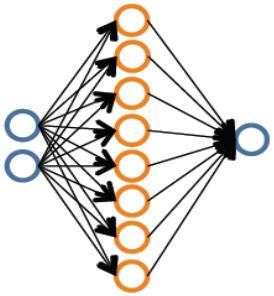


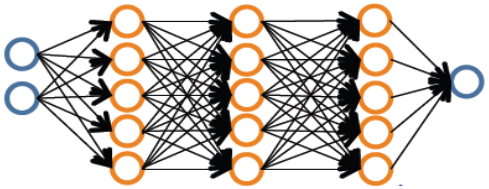


- **Larger gap between the two errors**
- **Getting more training data is likely to help!**

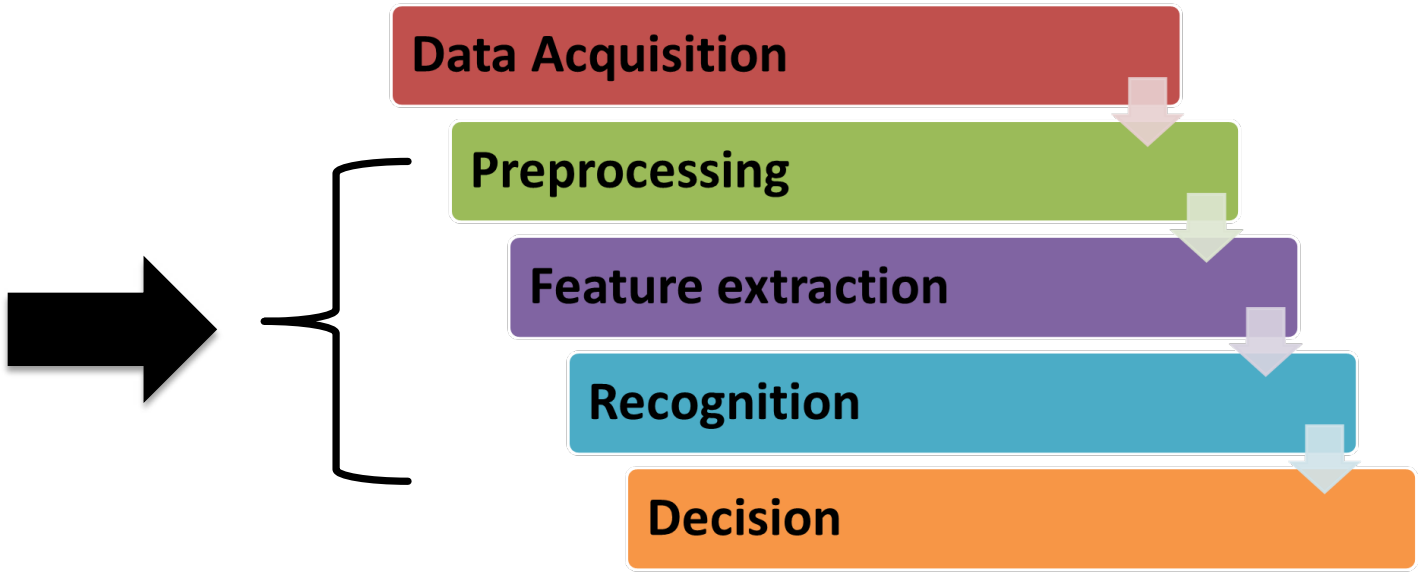
What to try next

- **High Bias problem (underfitting)**
  - Try getting additional features
- **High Variance problem (overfitting)**
  - Get more training example
  - Try smaller sets of features
- **Example: Neural networks**
  - *Small* neural networks
    - Fewer parameters
    - more prone to underfitting
  - *Large* neural networks
    - more parameters
    - more prone to overfitting





CROSS-VALIDATION [1]



Motivation and Goals

- Reminder: the goal of machine learning is automatically extracting relevant information from data and applying it to analyze new data
  - Regression
  - Classification
- Problem
  - Good prediction capability on the training data
  - **But** might fail to predict future *unseen* data
- **We need a procedure for estimating the generalization performance!**

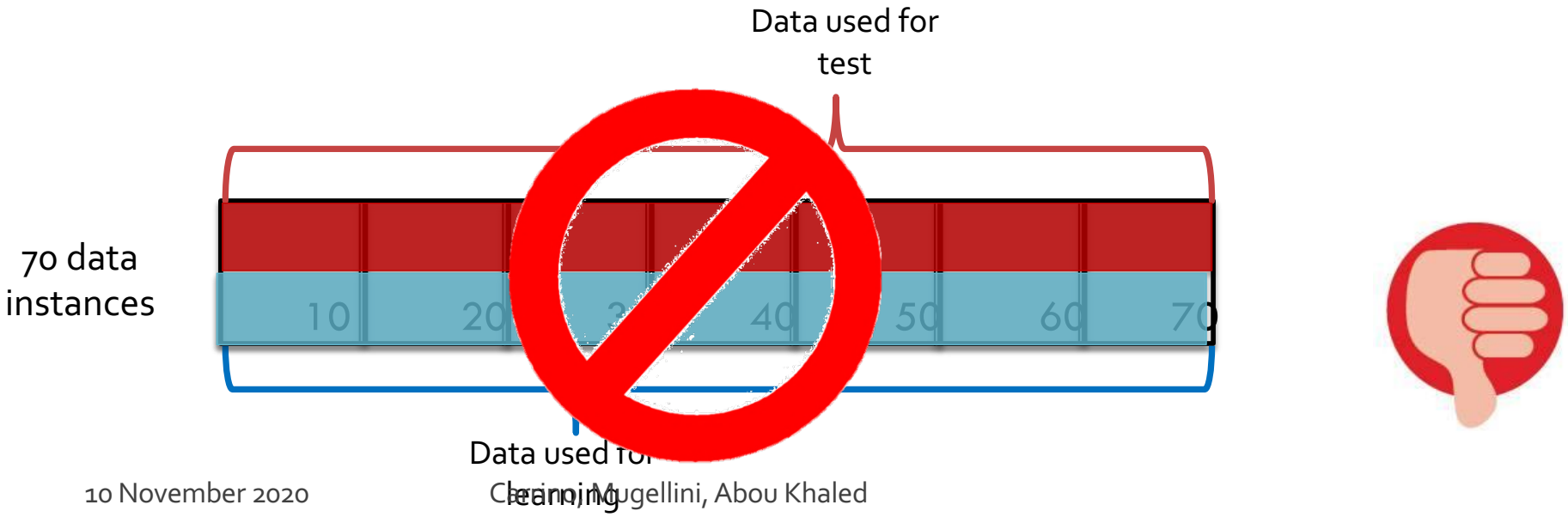
Definition

« Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn (or train) a model and the other used to validate the model. » [1]

- **K-fold cross-validation** is a popular form of cross-validation

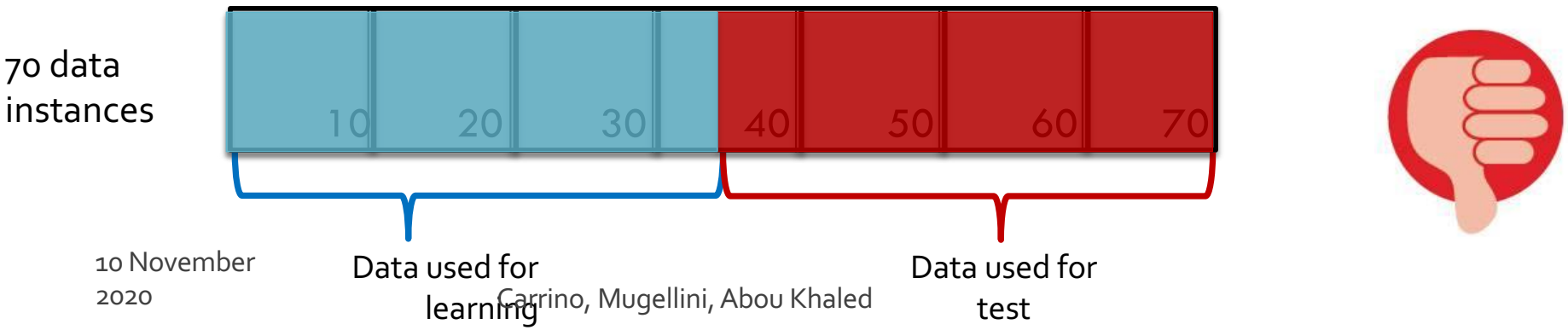
Procedures

- **Resubstitution Validation**
  - Learning from all the available data
  - Test on all the available data
    - Pros: it uses all the available data
    - Cons: it suffers **seriously** from over-fitting



Procedures

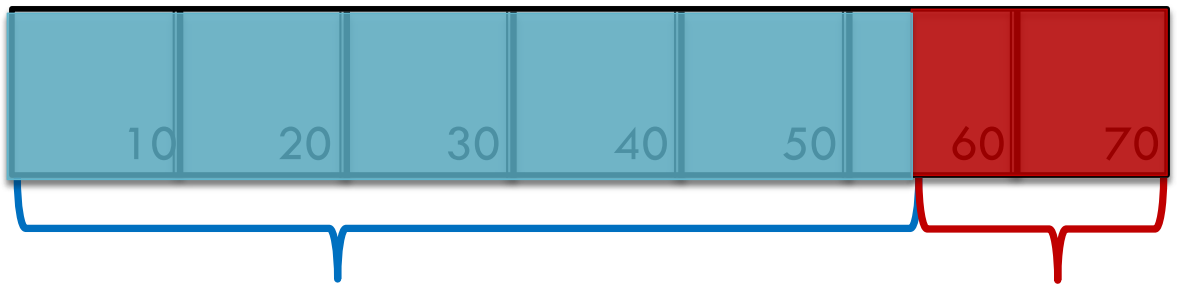
- **Hold-Out Validation**
  - Learning from half of the available data
  - Test on the other half of data. The test data is held out and not looked at during training.
    - Pros: it avoids the overlap between training data and test data
    - Cons:
      - Do not use all the available data for the training
      - Results highly dependent on the choice for the training/test split



Procedures

- **Hold-Out Validation**
  - Learning from half of the available data
  - Test on the other half of data. The test data is held out and not looked at during training.
    - Pros: it avoids the overlap between training data and test data
    - Cons:
      - Do not use all the available data for the training
      - Results highly dependent on the choice for the training/test split

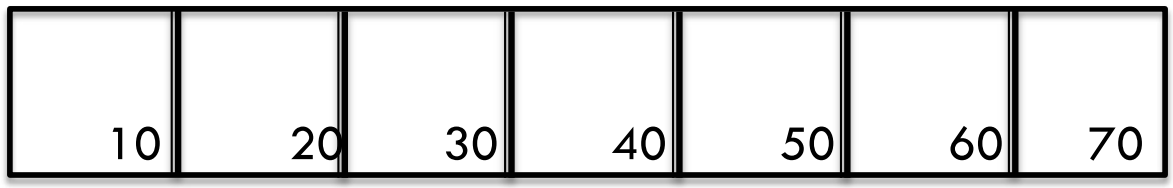
70 data instances



Procedures

- **K-fold Cross-validation**
  - The data is first partitioned into *k* equally sized segments (or folds)
  - K iterations of training and validation, where:
    - Learning on k-1 folds
    - Test on the held out fold

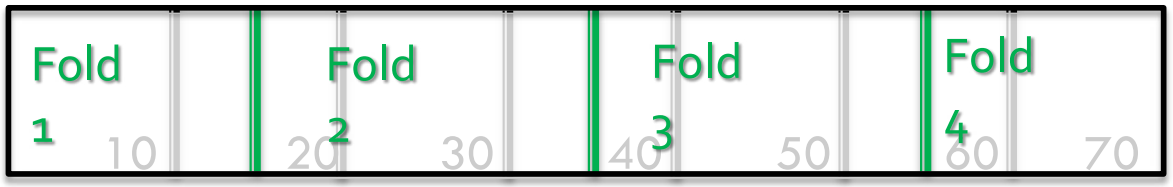
70 data instances



Procedures

- **K-fold Cross-validation**
  - Example: 4-folds

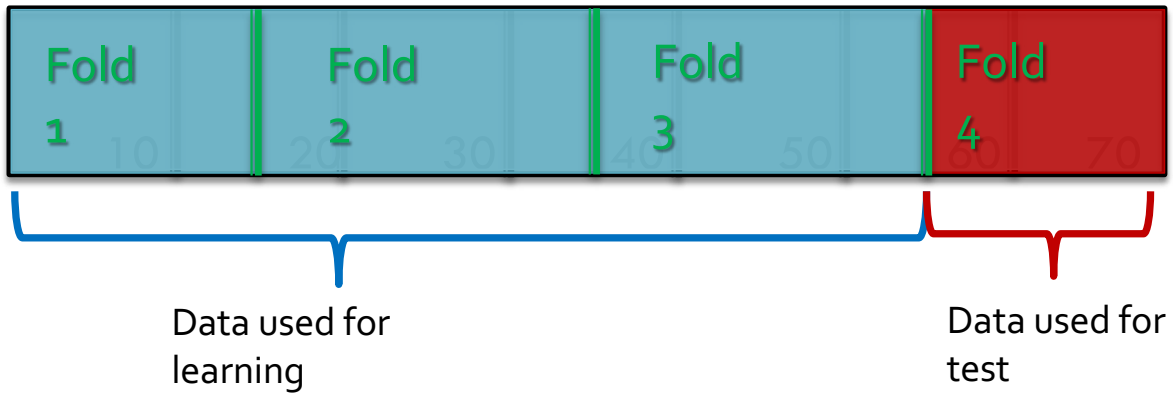
70 data instances, K=4



Procedures

- **K-fold Cross-validation**
  - Example k = 4

70 data instances, K=4
1st iteration



Procedures

- **K-fold Cross-validation**
  - Example k = 4

70 data instances, K=4

1st iteration

2nd iteration

3rd iteration

4th iteration

Procedures

- **K-fold Cross-validation**

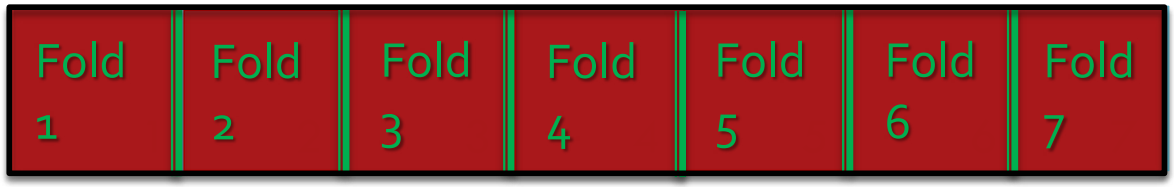
70 data instances,
K=4

- Note: data are commonly **stratified**
  - Rearranging the data ensuring that each fold is a good representative of the whole.
- Pros
  - It uses all the available data
  - It avoids the overlap between training data and test data
  - Accurate performance estimation **also if few samples are available**
- Cons*[3]*
  - Limited samples for performance estimation;

Procedures

- **Leave-One-Out Cross-validation**
  - Special case of *k-fold cross* validation where k is equal to the data instance number
  - Learning on k-1 folds
  - Test on the held out fold
    - Pros: it is almost **unbiased**
    - Cons: high variance → unreliable accuracy estimations [2]

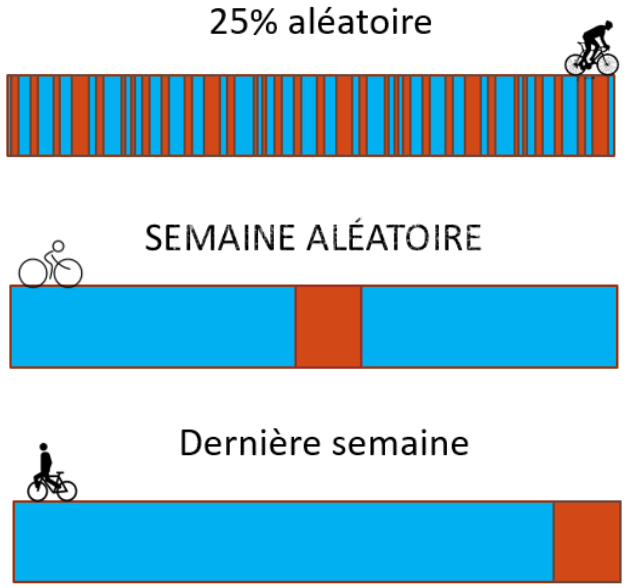
7 data instances,
K=**7**



- Note: widely used when the available data are rare

Error Measures - Partitioning

- How to split data into train/test sets ?
  - Randomly select samples ?
  - Select a "*week*" for test and the rest for train ?
  - Last "*week*" of the set ?
  - Weekly cross-validation ?

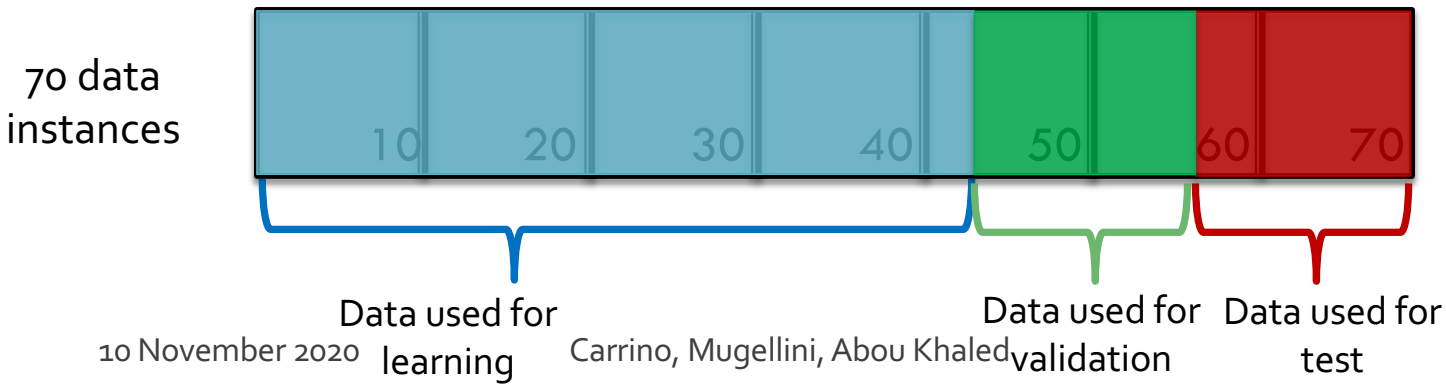


→ **In our case**, do you think it has an influence ?

→ What are the advantage(s) of the "Weekly Cross-validation approach ?

Procedures

- **How to chose some parameters of the model?**
  - Learning from 60% of the available data
  - Validation from 20% of the available data
    - Here we choose the best parameters
  - Test from 20% of the available data



Procedures

- **Exemple: tuning the RF (i.e. , find the optimal number of trees in the forest, *n_estimators*)**
- Step 1: put aside the test set (*remember* to stratify de data)

Procedures

- **Exemple: tuning the RF (i.e. , find the optimal number of trees in the forest, *n_estimators*)**
- Step 1: put aside the test set (*remember* to stratify de data)

- Step 2: use the ***k*** fold cross-validation method to determine the parameter ***n_estimators*** that optimizes the accuracy

For each iteration

*for **n** = 1 : 20*
Calculate the accuracy

.....

.....

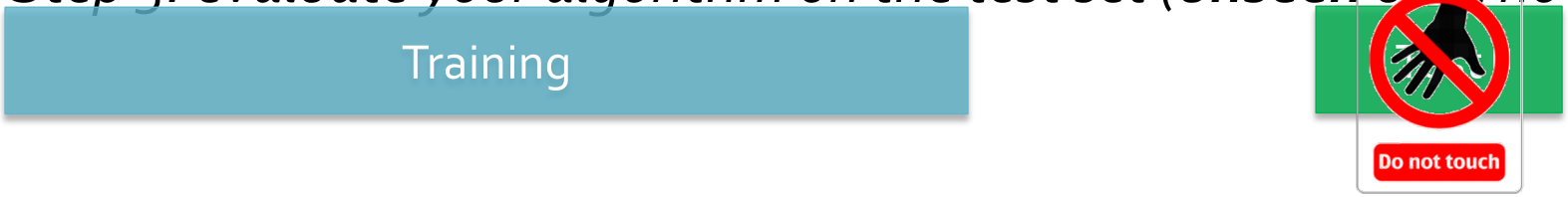
(K times)

Procedures

- **Exemple: tuning the RF (i.e. , find the optimal number of trees in the forest, *n_estimators*)**
- Step 3: calculate the accuracy mean as a function of *n_estimators*. Goal: to find the **best** *n_estimators (n_estimators*)*

- *Step 4: train your algorithm using **n_estimators*** over the **whole** dataset*

- *Step 5: evaluate your algorithm on the test set (**unseen** until now!)*



For a good performance estimation…

- Requirements
  - Independent measurements
  - *Sufficiently large number of measurements*
- Factors for the independence
  - the training data set
  - the test data set
- Conditions *[3]*
  - No overlap between **test data** in more than *one round*
  - No overlap between **training data** and **test data** in the same round
  - Sufficiently large amount of samples for the **training**

For a good performance estimation…

- Solution: **K-fold cross-validation!**
  - Keep the test sets independent among rounds
  - Ensure no overlap between training data and test data in a same round
  - Optimize the amount of data used for training

- Question: what's the appropriate number of **k**?

Kkkkkkkkkkkkkk?

- Larger k…
  - More performance estimations
  - The training set size is closer to the full data size
    - Good generalization
  - The overlap between training sets increases
  - The test set size is very reduced
    - Less precise measurements of the accuracy

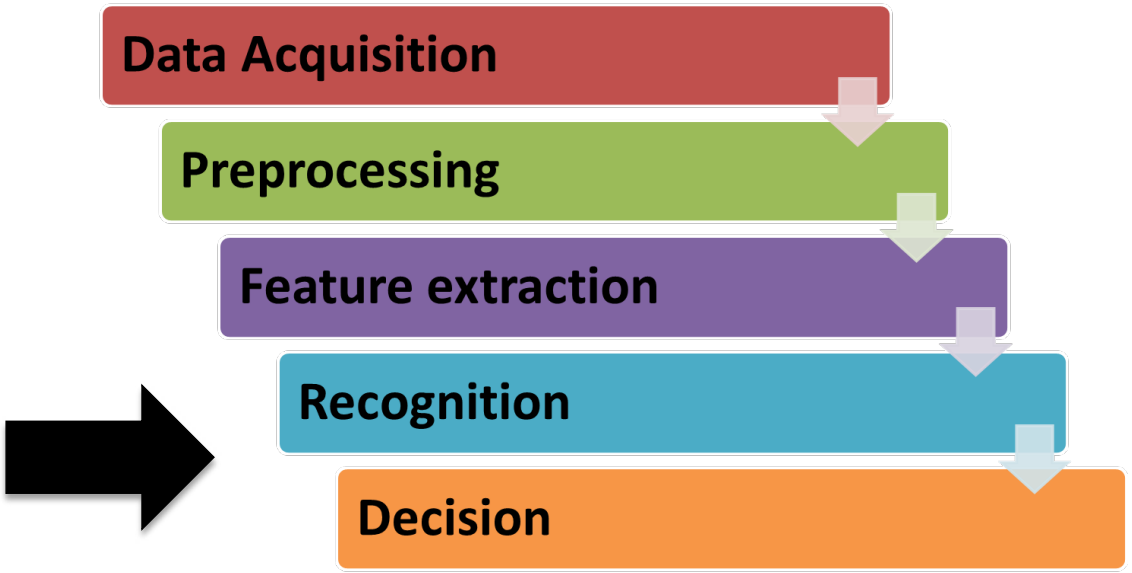



- In practice…
  - Bigger the k means longer computation time
  - **K=10** is a good compromise

Applications

- **Obtain reliable performances estimation**
  - *Accuracy*
  - *Precision*
  - *Recall*
  - *F-score*
  - *…*
- **Algorithm Tuning**
  - Feature selection to maximize classifiers performances on a particular dataset
  - Find the parameters that optimize the classifiers
    - K for the k-NN
    - Parameter C for SVM
    - Number of Gaussians for GMM
    - etc.

Performance Indicators [4]



Confusion Matrix

*A **confusion matrix** is a specific table layout that allows visualization of the **performance of an algorithm***

Confusion Matrix

		Actual Class		
		Tuna	Codfish	Salmon
Predicted Class	Tuna	15	4	7
	Codfish	3	20	4
	Salmon	6	1	15
		24	25	26

		Actual Class	
		True	False
Predicted Class	Positive	True Positive	False Positive *(Type I error)*
	Negative	False Negative *(Type II error)*	True Negative

Confusion Matrix - TP

		Actual Class		
		Tuna	Codfish	Salmon
Predicted Class	Tuna	15	4	7
	Codfish	3	20	4
	Salmon	6	1	15

Salmon		Actual Class	
		True	False
Predicted Class	Positive	True Positive 15	False Positive 7
	Negative	False Negative 11	True Negative 42

Confusion Matrix - FP

		Actual Class		
		Tuna	Codfish	Salmon
Predicted Class	Tuna	15	4	7
	Codfish	3	20	4
	Salmon	6	1	15

Salmon		Actual Class	
		True	False
Predicted Class	Positive	True Positive 15	False Positive 7
	Negative	False Negative 11	True Negative 42

Confusion Matrix - FN

		Actual Class		
		Tuna	Codfish	Salmon
Predicted Class	Tuna	15	4	7
	Codfish	3	20	4
	Salmon	6	1	15

Salmon		Actual Class	
		True	False
Predicted Class	Positive	True Positive 15	False Positive 7
	Negative	False Negative 11	True Negative 42

Confusion Matrix - TN

		Actual Class		
		Tuna	Codfish	Salmon
Predicted Class	Tuna	15	4	7
	Codfish	3	20	4
	Salmon	6	1	15

Salmon		Actual Class	
		True	False
Predicted Class	Positive	True Positive 15	False Positive 7
	Negative	False Negative 11	True Negative 42

Accuracy and Precision

		Actual Class	
		True	False
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy = the proportion of true results in the population

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision = the proportion of the true positives against all the positive results

$$Precision = \frac{TP}{TP + FP}$$

Accuracy and Precision

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

		Actual Class	
		Men	Wom
Predicted Class	Men	10 **TP**	3 **FP**
	Wom	9 **FN**	9 **TN**

		Actual Class	
		Men	Wom
Predicted Class	Men	10 **TN**	3 **FN**
	Wom	9 **FP**	9 **TP**

Accuracy and Precision

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Accuracy = Overall Accuracy (class independent. Attention: this is true **only** in the 2 classes case)

		Actual Class	
		Men	Wom
Predicted Class	Men	10	3
	Wom	9	9

Precision ≠ Overall Precision (class dependent)

- **For each class you get a different precision!!**
- Overall Precision ??

$OverallPrecision = \frac{1}{N}\sum_{i=1}^{N} P_i$, where **N** is the number of classes

$WeightedPrecision = \frac{(P_{c1} * |c1|) + (P_{c2} * |c2|)}{|c1| + |c2|}$, where $|ci|$ is the number of instances in the class **i**

Sensitivity (or recall) and Specificity

Example:

True positive: Sick people correctly diagnosed as sick

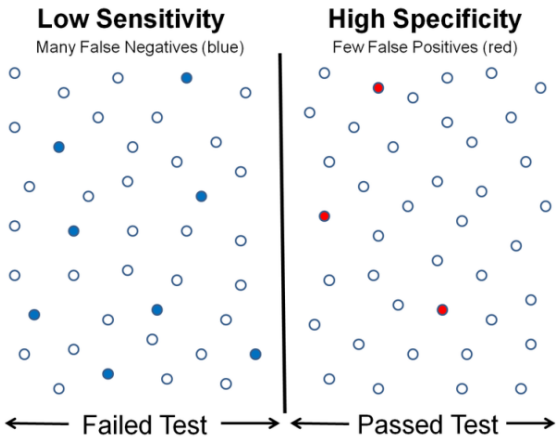
False positive: Healthy people incorrectly identified as sick

True negative: Healthy people correctly identified as healthy

False negative: Sick people incorrectly identified as healthy.

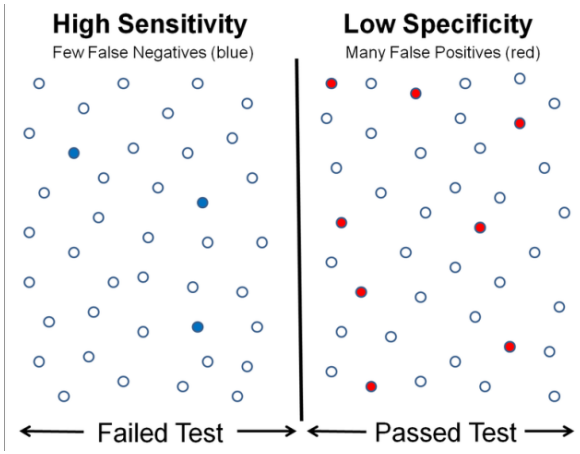
Sensitivity = the probability of a positive test given that the patient is ill

$$Sensitivity = \frac{TP}{TP + FN}$$



Specificity = the probability of a negative test given that the patient is well

$$Specificity = \frac{TN}{TN + FP}$$



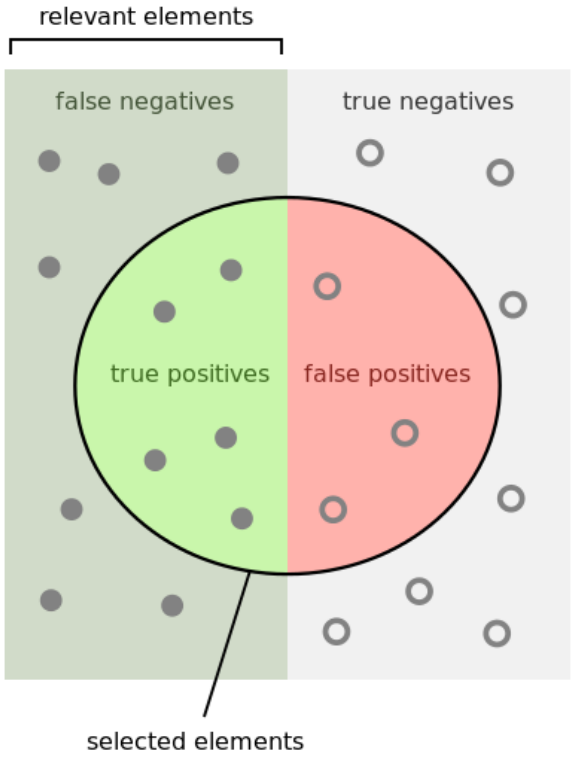
Precision and Recall – Other interpretations

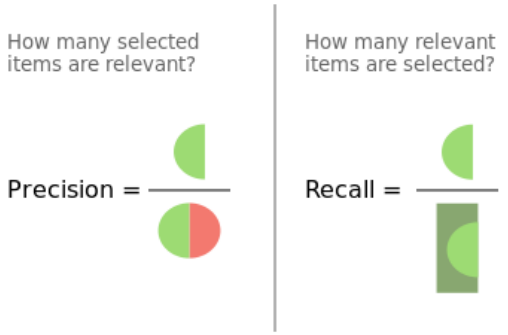
Information retrieval

- **Precision** is the fraction of retrieved instances that are relevant
- **Recall** is the fraction of relevant instances that are retrieved

Probabilistic interpretation

- **Precision** is the probability that a (randomly selected) retrieved document is relevant.
- **Recall** is the probability that a (randomly selected) relevant document is retrieved in a search.





Other performance indicators *[6]*

F-score (or F-measure or F1 score)

Measure of a test's accuracy

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \qquad F = (1 + \beta) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Matthews correlation coefficient

Measure of the quality of binary (two-class) classifications

On the contrary of F-score, it takes the **true negative** rate into account

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \qquad F = (1 + \beta) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Conclusion

Conclusion

- Normalization
- Training set balanced/unbalanced
- Diagnosis bias vs. variance
- Cross-Validation
- Confusion matrix & Performance indicators

- To ANN, or to SVM (or whatever) **that is** the question!
  - **Not really…**

Conclusion

- **The question is:**
  - To have good data (a lot of data…)
  - To know how to manage the data
  - To know how to evaluate the results
  - …and know your context!

What you should know

- When and why is important to use Feature scaling & normalization
- What does it involve to have unbalanced classes in a training set?
- Cross-Validation
  - Motivations, goals, what it is…

What you should know

- Performance indicators
  - What is a confusion matrix
  - How to use the performance indicators
- Overfitting (High Variance) Vs Underfitting (High Bias)

Questions?

Refereces

- [1] **Cross-Validation**. Payam Refaeilzadeh, Lei Tang, Huan Liu, Arizona State University
- [2] **Estimating the error rate of a prediction rule: improvement on cross-validation**. Efron B. J. Am. Stat. Assoc., 78:316–331,1983.
- [3] Approximate statistical tests for comparing supervised classification learning algorithms. Dietterich T.G. Neural Comput., 10(7):1895–1923
- [4] International vocabulary of metrology — Basic and general concepts and associated terms
- [5] An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. John Robert Taylor. University Science Books. pp. 128–129
- [6] Assessing the accuracy of prediction algorithms for classification: an overview. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Bioinformatics 2000 000, 16, 412–424

Refereces

- [7] Coursera, Machine learning, Andrew Ng, Stanford University, https://www.coursera.org/course/ml
- [8] http://en.wikipedia.org/wiki/Confidence_interval