4. Machine Learning Overview

4.2 Core Machine Learning Concepts - Part II

Cristiano Bacelar de Oliveira

Agosto, 2020

Universidade Federal do Ceará

Index

- Machine Learning Process
- Data Processing
- Model Building
- Performance Evaluation

Disclaimer

The following content is heavily based on HCIA-AI Course material by Huawei Technologies Co., Ltd., authored by Zhang Luxiao. Distribution is not allowed.

Machine Learning Process

Machine Learning Process

Data collection

Data processing

Model building

Data Collection

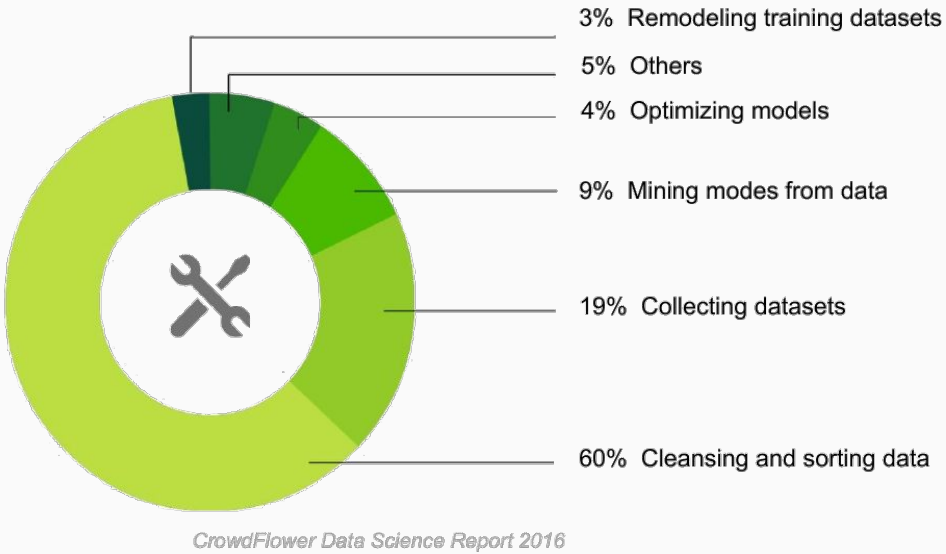
Machine Learning relies on data, so it is essential to collect good information for building the model.

Some concerns related to data collection include:

- Proper number of samples for the task
  - Appropriate volume and balancing
- Data Annotations
- Quality Assurance

Data processing

- Data Cleaning
  - Fill in missing values and ‘fix’ bad or wrong samples
- Data Dimension Reduction
  - Simplify data attributes
  - Feature selection
- Data Normalization
  - Normalize data to reduce noise



CrowdFlower Data Science Report 2016

Data processing (cont.)

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing value

Invalid value

Value that should be in another column

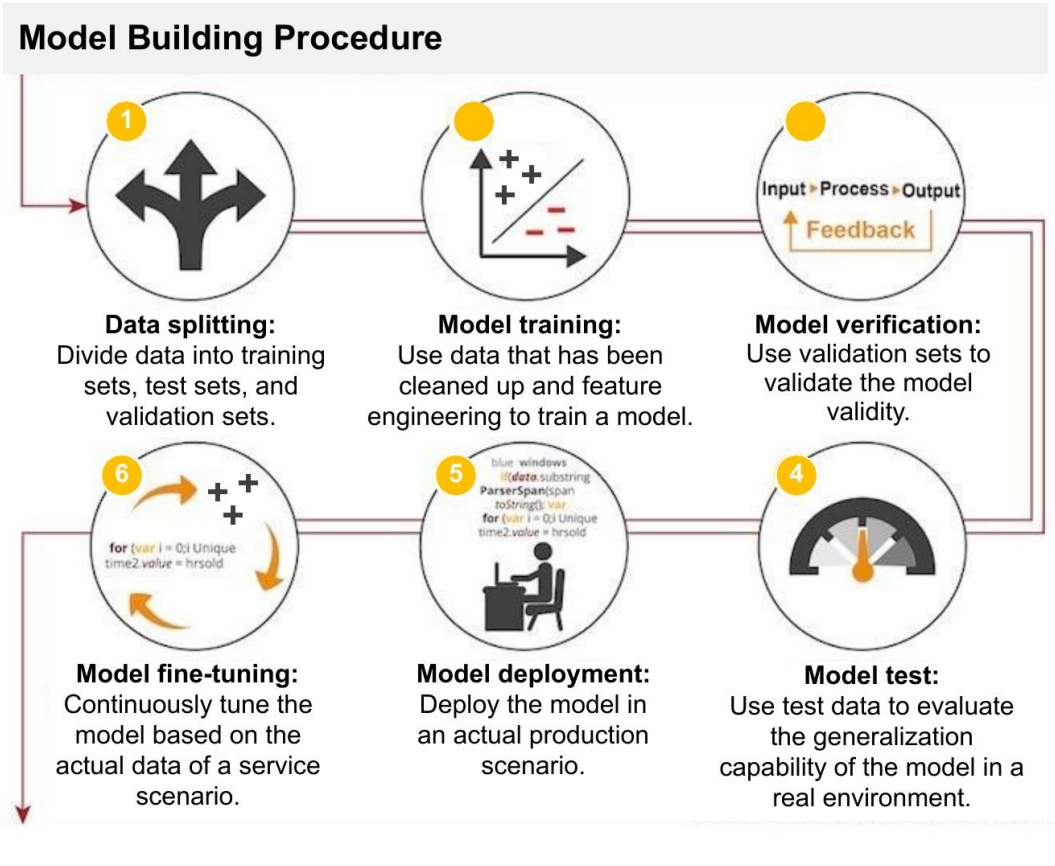
Misspelling

Invalid duplicate item

Incorrect format

Attribute dependency

Model Building Procedure

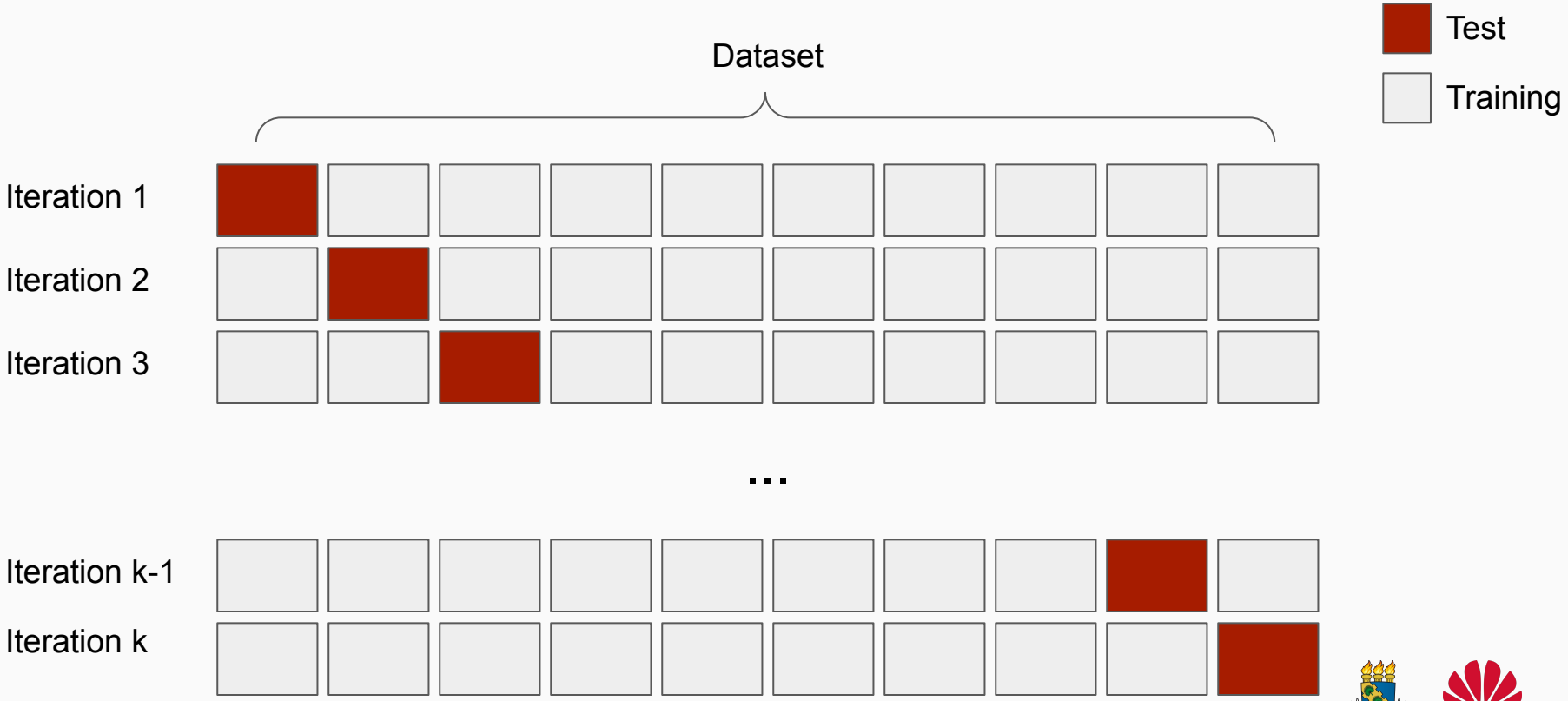


Cross Validation

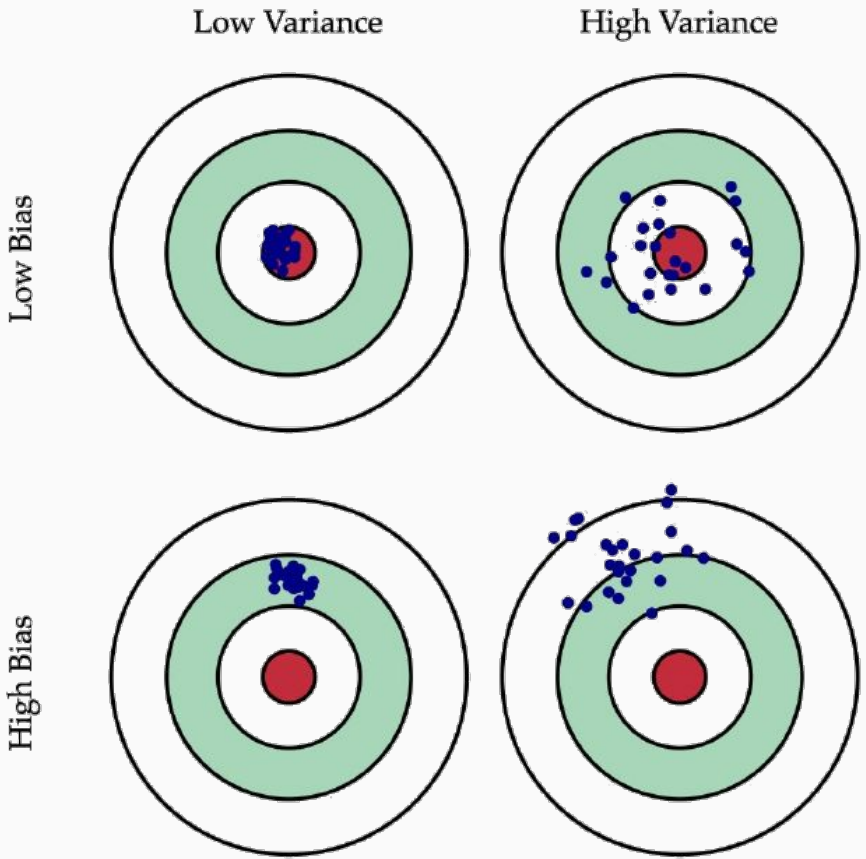
Cross Validation is a method used to validate the performance of a classifier. The basic idea is to divide the original dataset into two parts: training set and validation set. The training set is used to train the classifier and the validation set is used to check the classifier performance.

The result of this approach depends on how the division into training and test sets is done. In order to mitigate such dependency, a better approach is to use a k-fold cross validation method.

k-fold Cross Validation



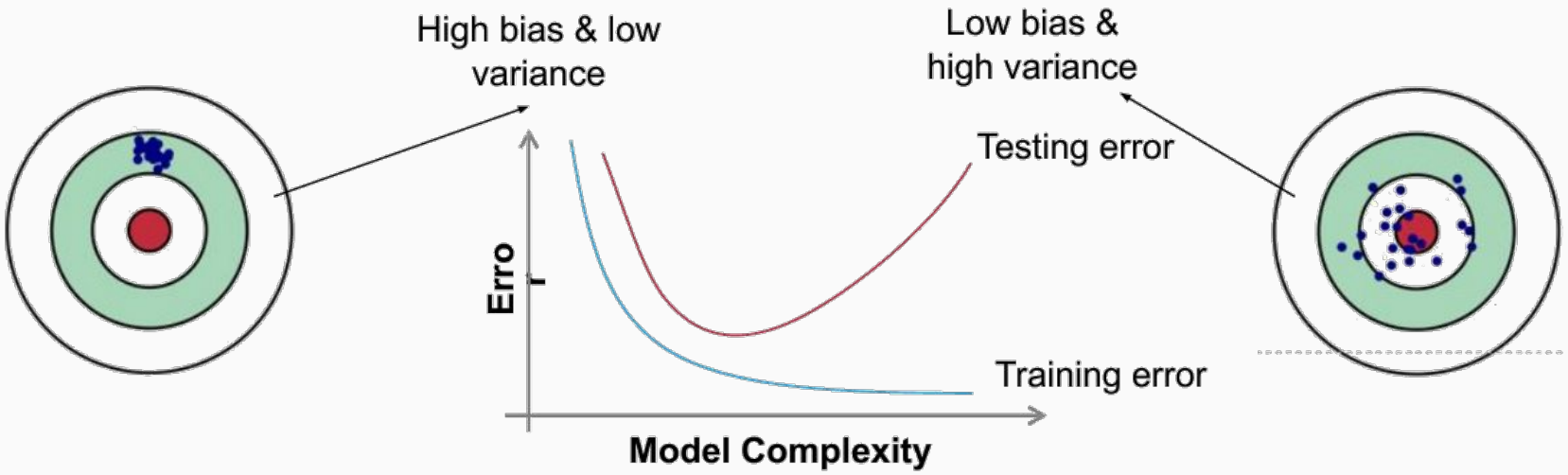
Bias and Variance Error



Generally, the prediction error can be divided into two subforms:

- Error caused by "bias"
- Error caused by "variance"

Bias and Variance Error



Performance Evaluation - Classification

Confusion Matrix

- Observed
  - P: Positive
  - N: Negative
- Predicted
  - TP: True Positive
  - FP: False Positive
  - TN: True Negative
  - FN: False Negative

Predicted / Observed	yes	no	Total
yes	TP	FN	P
no	FP	TN	N
Total	P'	N'	$P + N$

Confusion matrix

Performance Evaluation - Classification

Measure	Formula
Accuracy and correct classification rate	$\frac{TP + TN}{P + N}$
Error rate and false classification rate	$\frac{FP + FN}{P + N}$
Sensitivity, true positive rate, and $recall$	$\frac{TP}{P}$
Specificity and true negative rate	$\frac{TN}{N}$
Precision	$\frac{TP}{TP + FP}$
F score: harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β (β is a non-negative real number)	$\frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Performance Evaluation - Regression

$$MAE = \frac{1}{\mathrm{m}} \sum_{i=1}^{m} |y_i - \hat{y}_i|$$

$$MSE = \frac{1}{\mathrm{m}} \sum_{i=1}^{\mathrm{m}} (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y}_i)^2}$$

- Mean Absolute Error (MAE)
- Mean Square Error (MSE)
- R^2

Performance Evaluation - General

What is a good model?

- Generalization capability
  - Can it accurately predict the actual service data?
- Interpretability
  - Is the prediction result easy to interpret?
- Prediction speed
  - How long does it take to predict each piece of data?
- Plasticity / Scalability
  - Is the prediction rate still acceptable when the service volume increases with a huge data volume?

Thank You!

Next: 4.3 - Common Machine Learning Algorithms