Machine Learning 101

Jonas Gonzalez

jonas.gonzalez@iit.it

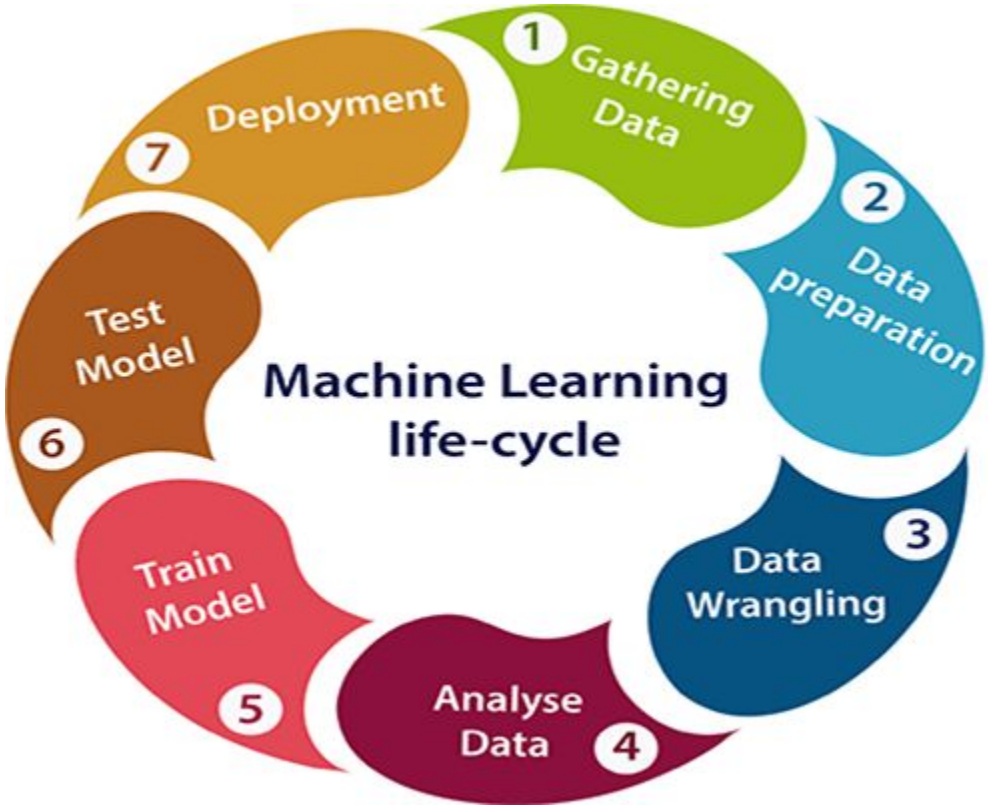
What is Machine Learning ?

*"Teach computers how to learn and act without being **explicitly programmed**"*

*"Machine learning focuses on the development of computer programs that can **access data** and use it to learn for themselves "*

*"The primary aim is to allow the computers to learn **automatically without human intervention** or assistance and adjust actions accordingly"*

Source : expertsystem.com & deepai.org


Machine Learning life-cycle
1 Gathering Data
2 Data preparation
3 Data Wrangling
4 Analyse Data
5 Train Model
6 Test Model
7 Deployment

Data acquisition

- Images
- Audio
- Numeric data

	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Volume	Adj Close
2	02/01/2003	50.65	51.61	50.52	51.6	7545500	44.99
3	03/01/2003	51.61	51.61	49.85	50	8389300	43.59
4	06/01/2003	50.2	50.55	49.67	50.19	7438400	43.76
5	07/01/2003	50.32	50.76	50.1	50.46	6669000	43.99
6	08/01/2003	50.4	51.36	49.86	49.99	7796900	43.58
7	09/01/2003	50.75	52	50.75	51.92	9884800	45.27
8	10/01/2003	51.92	52	51.21	51.62	7426600	45
9	13/01/2003	51.62	52.18	51	51.28	6920800	44.71
10	14/01/2003	51	51.54	50.7	51.41	6759600	44.82
11	15/01/2003	51.45	51.68	50.53	50.59	6503500	44.11
12	16/01/2003	51.1	51.23	49.98	50.3	8086900	43.85
13	17/01/2003	50.3	50.43	49.7	49.97	8661200	43.57
14	21/01/2003	50.07	50.29	48.98	49.01	7827400	42.73
15	22/01/2003	49.02	49.59	47.75	48.07	11097600	41.91
16	23/01/2003	48.07	48.76	47.34	48.57	10896500	42.34
17	24/01/2003	48.4	48.69	47.19	47.3	8425500	41.24

Data preparation

- Incomplete data
  - NaN, null
- Noisy
  - Age = -20 , Height = 120
- Inconsistent
  - Rating = 18/ 30 or Rating = A-F

Data wrangling

- Choose how to deal with your data
  - NaN, null values
    - Take the mean
    - Remove instance
  - Normalize your data
  - Remove outliers

Analyze your data

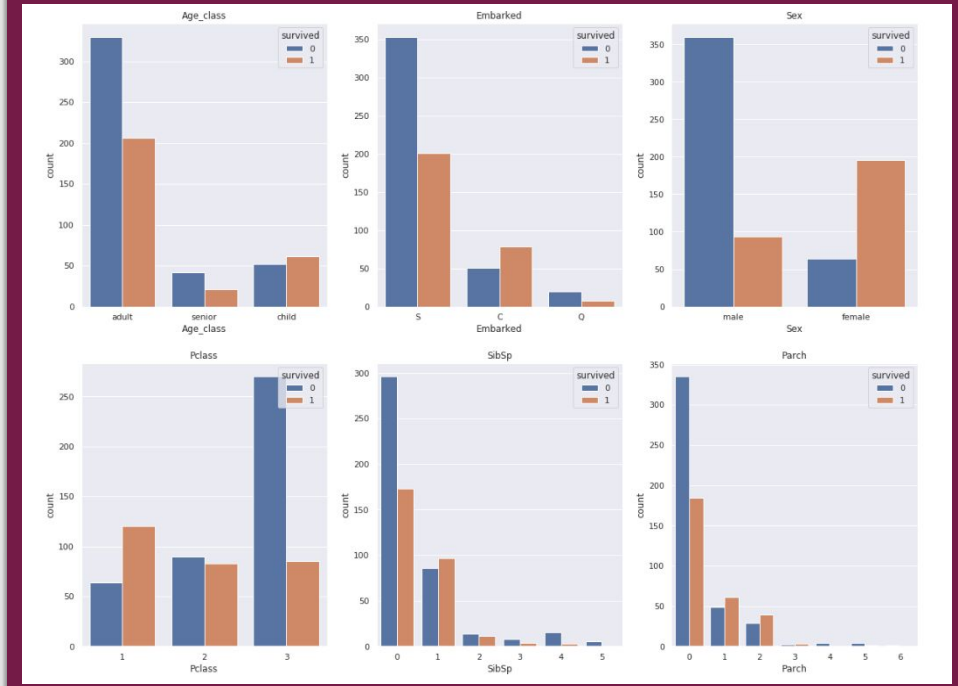
“Data analysis is the process of evaluating data using analytical and statistical tools to discover useful information”

Quick exploration to see patterns

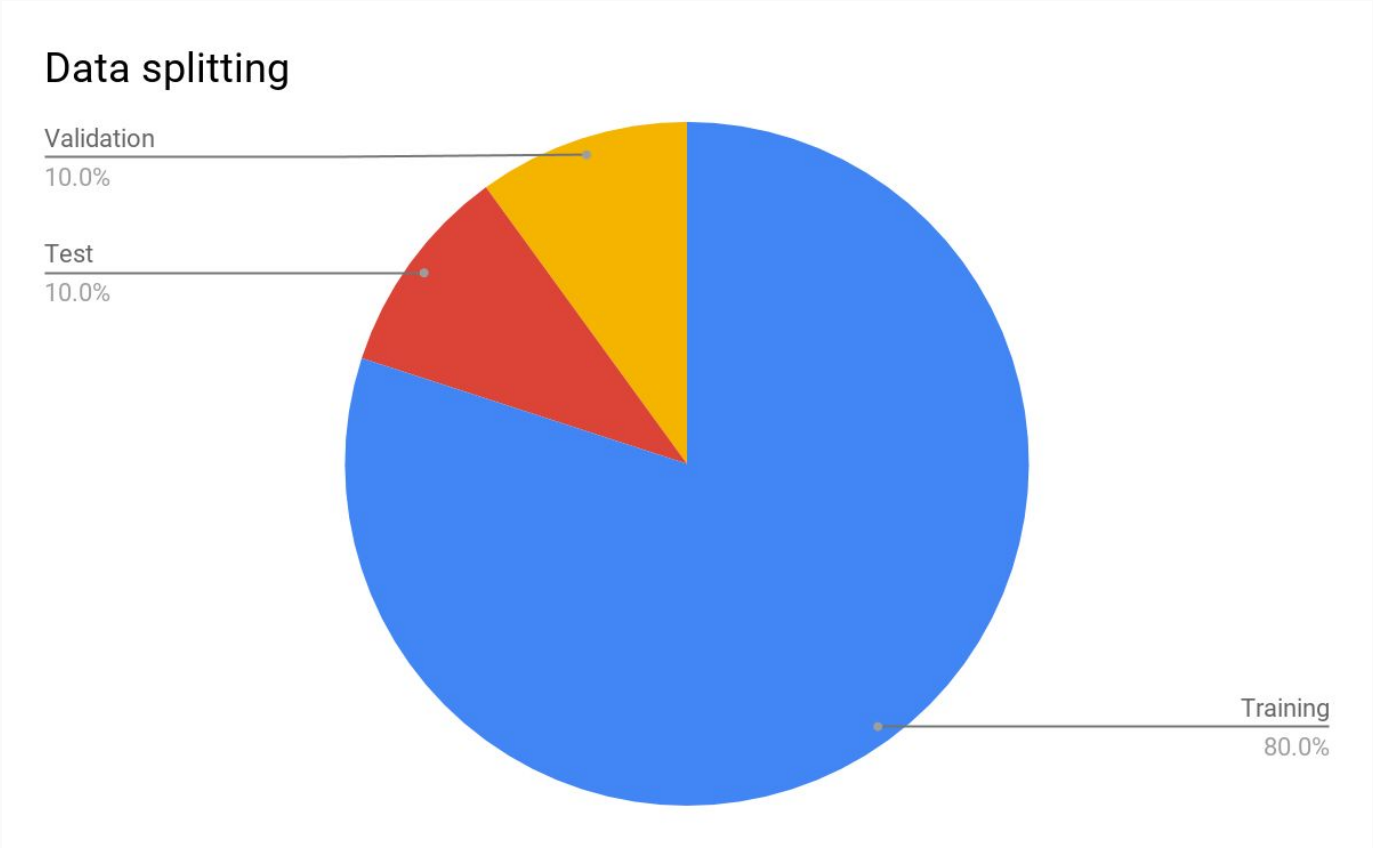
- Statistical Distributions
- Plotting (Histogram, scatter plot, box plot)
- Correlation

Analyze your data

“Data analysis is the process of evaluating data using analytical and statistical tools to discover useful information”

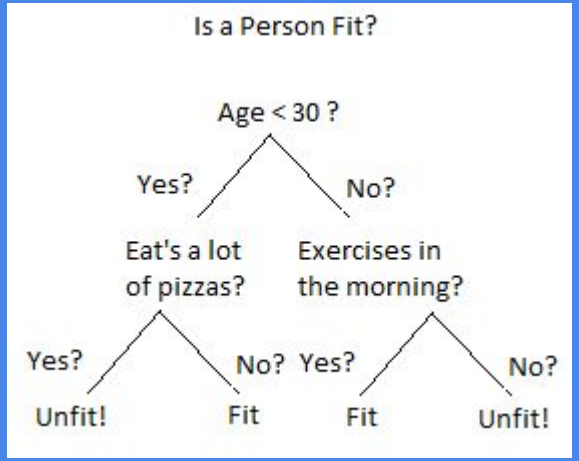


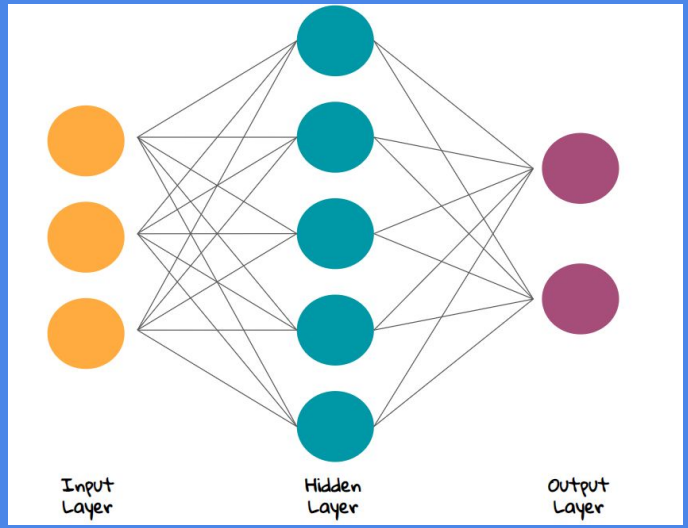
Preparing your data for learning



Ready to make learning :

- Data is clean
- Exploratory analysis give us the features that might be useful for our initial question
- Dataset





Now lets start doing ML !

*“A computer program is said to learn from experience **E** with respect to someclass of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**”*

Source : http://www.deeplearningbook.org/

The Task *T*

“Machine learning aims at generalizing a task that will be to difficult to solve with fixed program”

- Classification:

From a vector $x \in R^n$ find the class its belong

That is finding $f : R^n \rightarrow \{1, \ldots, k\}$

- Regression:

From a vector $x \in R^n$ predict a numerical value

That is finding $f : R^n \rightarrow \{1, \ldots, k\}$

The Performance

P

“How evaluate the performance of the model on the task T”

	Predicted Positives	Predicted Negatives
Actual: Positives	**TP**	**FN**
Actual: Negatives	**FP**	**TN**

ACCURACY = **TP + TN / (TP + TN + FP + FN)**

The Performance

P

“How evaluate the performance of the model on the task T”

N = 100	**Predicted:** Positives	**Predicted:** Negatives
Actual: Positives	**45**	**5**
Actual: Negatives	**5**	**45**

ACCURACY = **45 + 45 / 100 = 0.9**

The Performance

P

“How evaluate the performance of the model on the task T”

N = 100	**Predicted:** Positives	**Predicted:** Negatives
Actual: Positives	**0**	**0**
Actual: Negatives	**2**	**98**

ACCURACY = 0 + 98 / 100 = 0.98% !!

The Performance

P

“How evaluate the performance of the model on the task T”

N = 100	**Predicted:** Positives	**Predicted:** Negatives
Actual: Positives	**0**	**1**
Actual: Negatives	**1**	**98**

PRECISION = TP / TP + FP -> 0 / 2 = 0%

The Performance

P

“How evaluate the performance of the model on the task T”

N = 100	**Predicted:** Positives	**Predicted:** Negatives
Actual: Positives	**0**	**1**
Actual: Negatives	**1**	**98**

RECALL = TP / TP + FN -> 0 / 1 = 0%

The Performance

P

“How evaluate the performance of the model on the task T”

N = 165	**Predicted:** Positives	**Predicted:** Negatives
Actual: Positives	**50**	**10**
Actual: Negatives	**5**	**100**

ACCURACY = 0.9 PRECISION = 0.91 RECALL = 0.95

F-measure = 2*PRECISION*RECALL / (PRECISION +RECALL)
F-measure = 0.92

The Experience *E*

Usually ML models experience a dataset, a collection of data points.

Supervised Learning

“ The model is given a dataset of pair (x,y) and it has to learn a function that successfully predict the label Y from new input X.“

Unsupervised Learning

“ The dataset is only data points with different features f. The goal generally is to learn some useful properties (the probability distribution)

It can also be applied to find clusters”

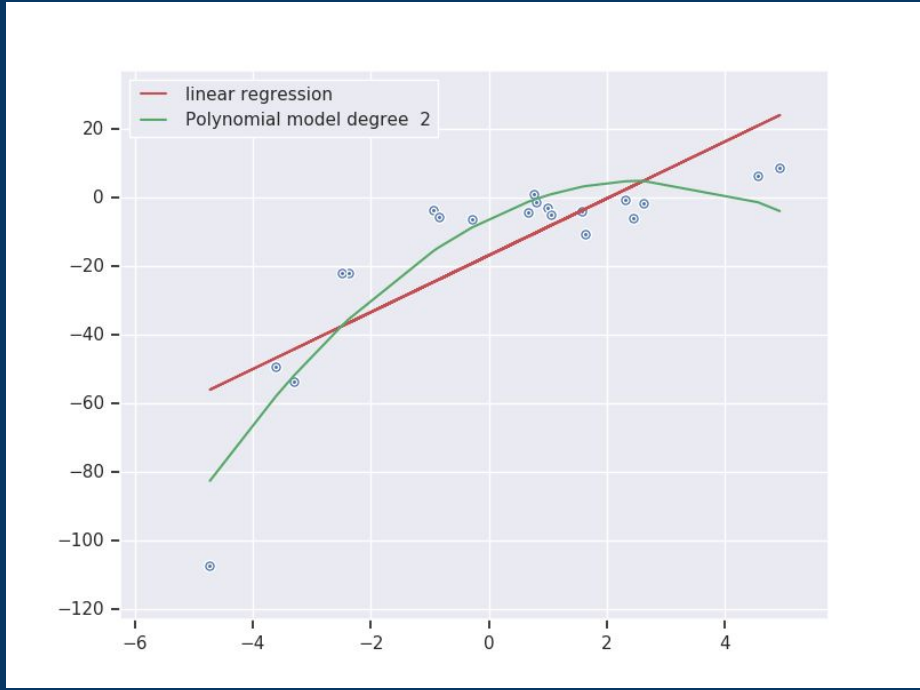
A Simple Example

Linear Regression

$$Y = aX + b$$

A Simple Example

Linear Regression

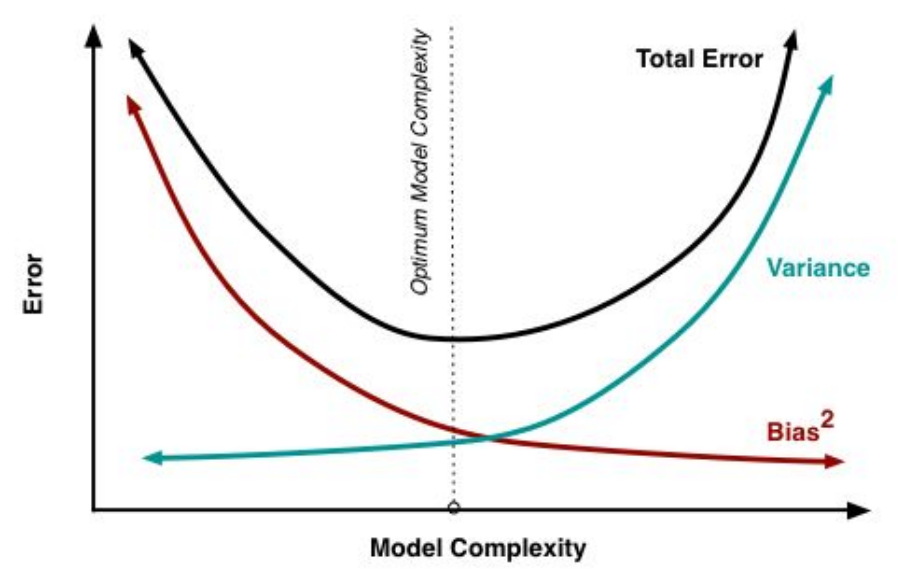


Underfitting vs Overfitting

Bias vs Variance tradeoff

High bias -> underfitting

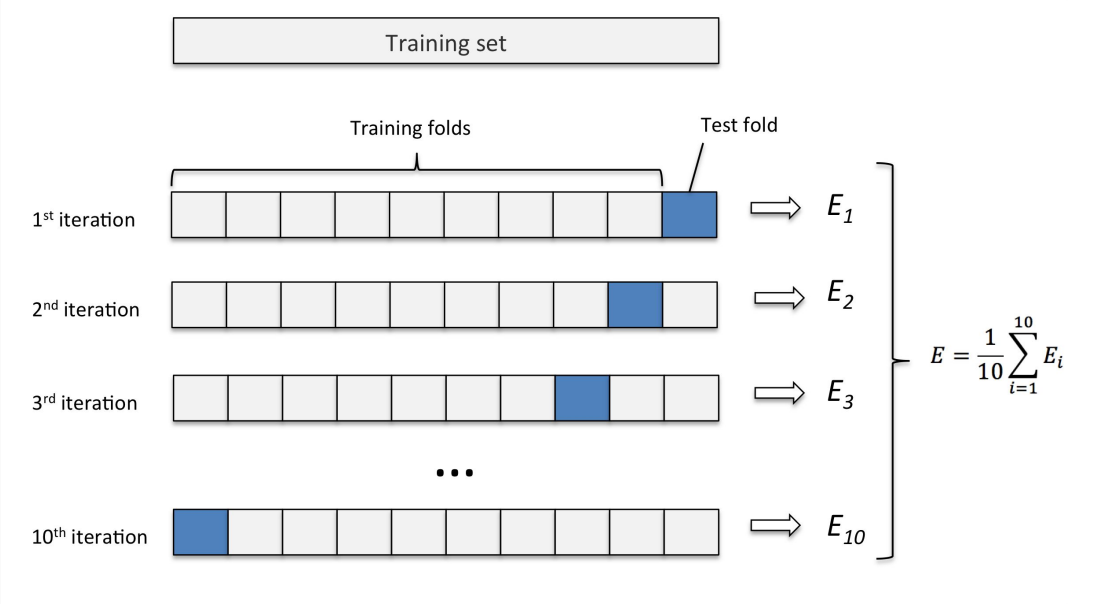
High variance -> overfitting



Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Cross-Validation : k-fold

Evaluate machine learning models on a limited data sample



Next time :
In depth Machine learning
(Math and hand on session)


"That's all Folks!"