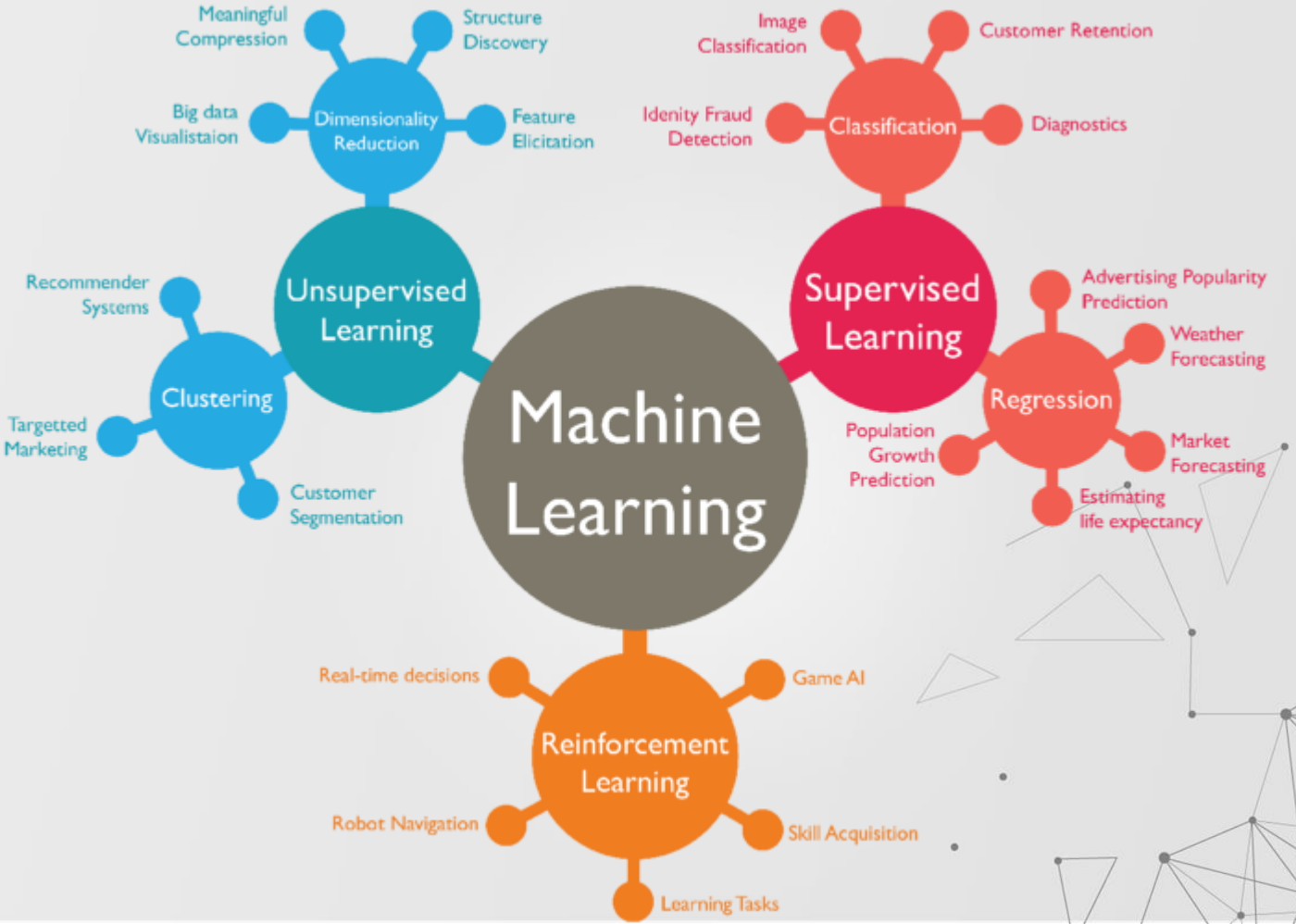
Machine Learning Pipelines

Fei GAO

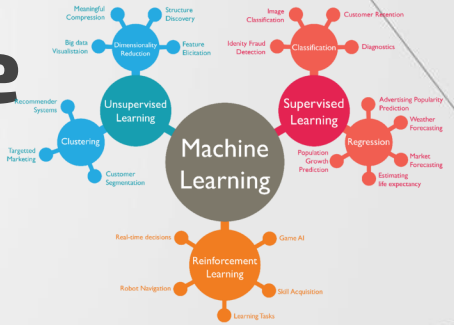
Jan 2025

01

Machine Learning predictive data analysis


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating life expectancy
Population Growth Prediction
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

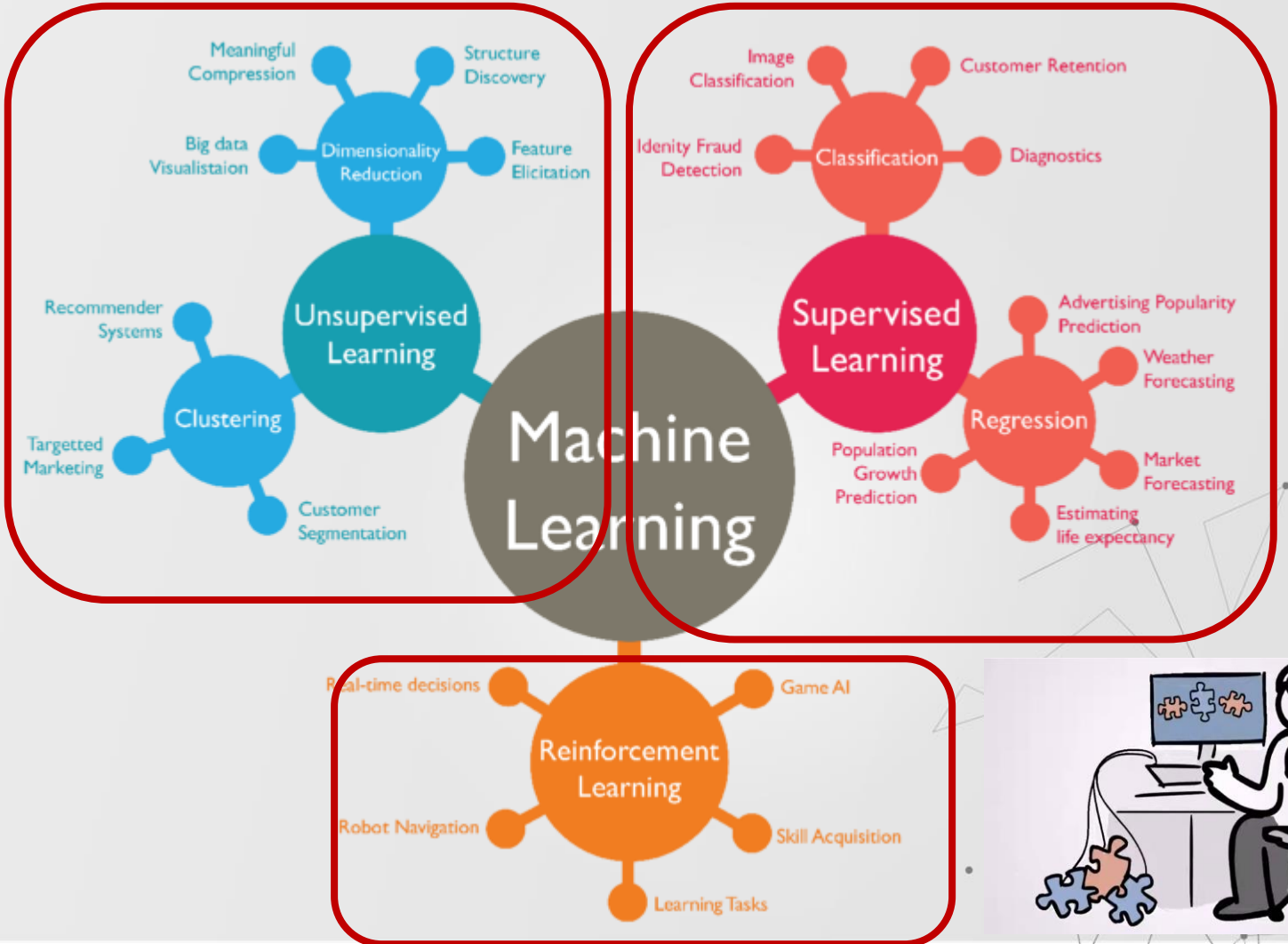
Données historique



Nouvelles données

Mémorisation / apprentissage

Généralisation / Prédiction


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

APPRENTISSAGE SUPERVISÉ

Features + Label

(X1, X2, X3, X4, ….) (Y)

APPRENTISSAGE NON SUPERVISÉ

Features + ~~Label~~

(X_1, X_2, X_3, X_4, ….) (PAS d'Y)

Problématique de l'analyse prédictive supervisée

- Y : label
- X1, X2, … : features
- Nous cherchons f() : une fonction qui essaie d'établir la relation Y = f(X1, X2, …) **pour faire de la prédiction**
- f() doit être aussi précise que possible pour pouvoir prédire sur les nouvelles données

Les principales familles

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

- Logistic Regression
- Decision Tree
- Random Forest
- SVM
- KNN
- XGBoost
- RN
- Naive Bayes Classifier

Predict a number

REGRESSION

«Divide the ties by length»

- Linear Regression
- Decision Tree
- Random Forest
- SVM
- KNN
- XGBoost
- RN

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

Find hidden dependencies

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Identify sequences

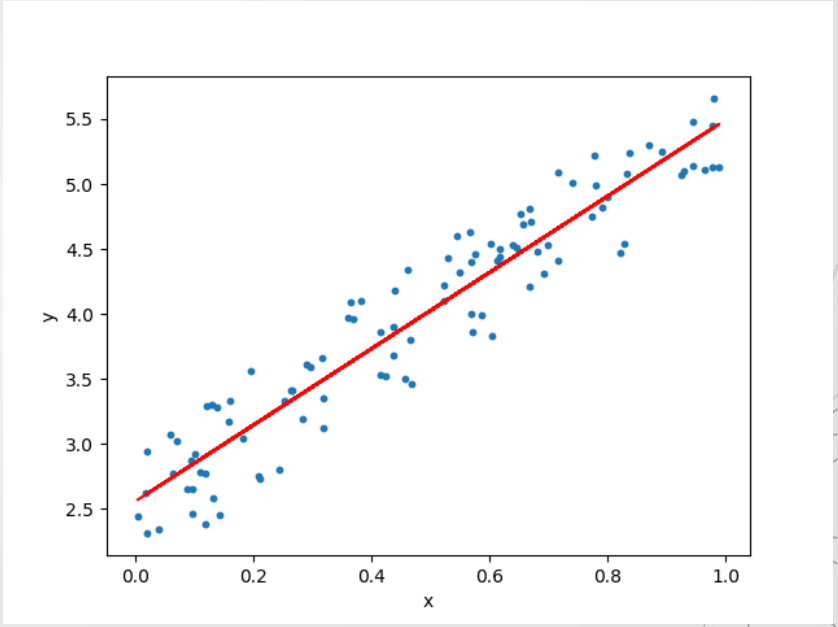
ASSOCIATION

«Find what clothes I often wear together»

Usuel MACHINE LEARNING ALGORITHMS

Linear Regression

Linear Regression tends to establish a relationship between a dependent variable(Y) and one or more independent variable(X) by finding the best fit of the straight line.

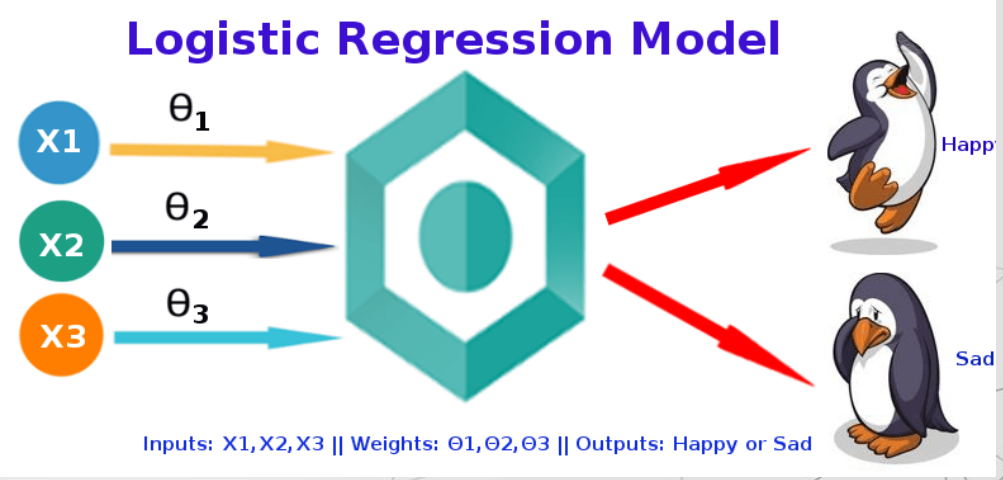


Regressor problems

Usuel MACHINE LEARNING ALGORITHMS

Logistic Regression

The logistic regression technique involves the dependent variable, which can be represented in the binary (0 or 1, true or false, yes or no) values or the probability of a successful or fail event.

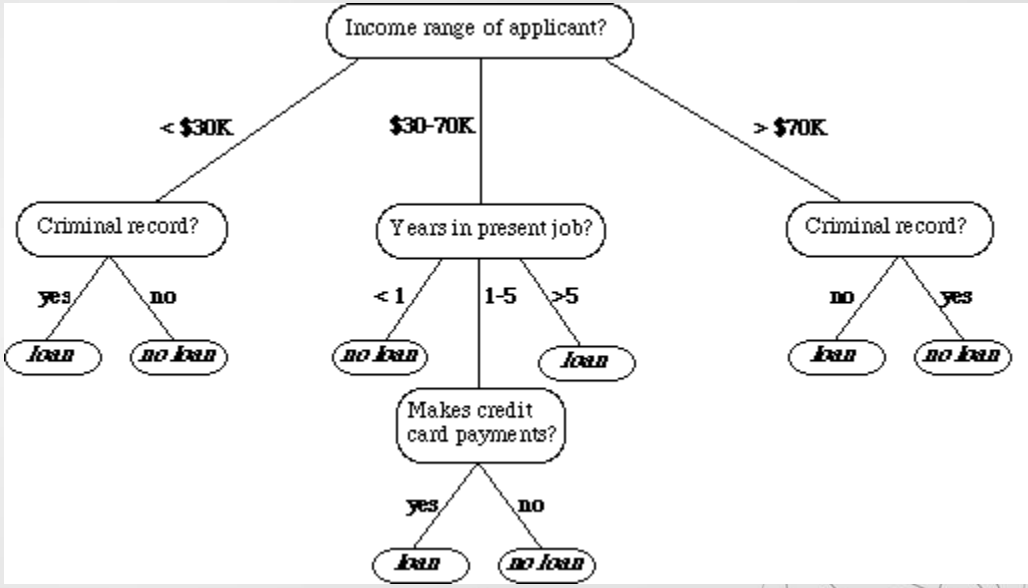


Classification problems

Usuel MACHINE LEARNING ALGORITHMS

Decision Tree

The decision tree works on an if-then statement. Decision tree tries to solve a problem by using tree representation

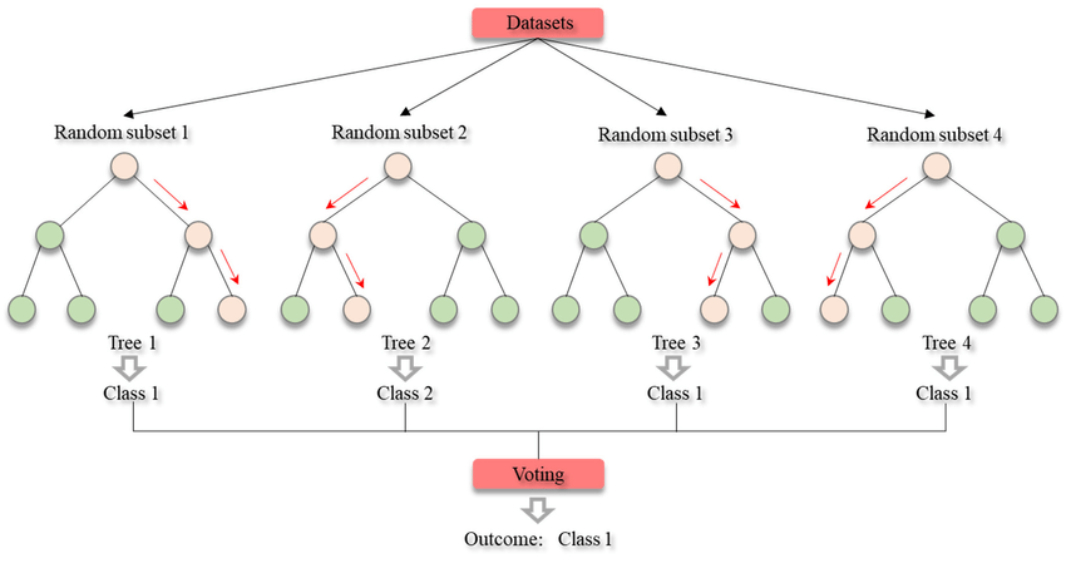


Classification as well as Regressor problems

Usuel MACHINE LEARNING ALGORITHMS

Random Forest

Random Forest is an ensemble machine learning algorithm that follows the bootstrapping technique. Random forest randomly selects a set of features that are used to decide the best split at each node of the decision tree.

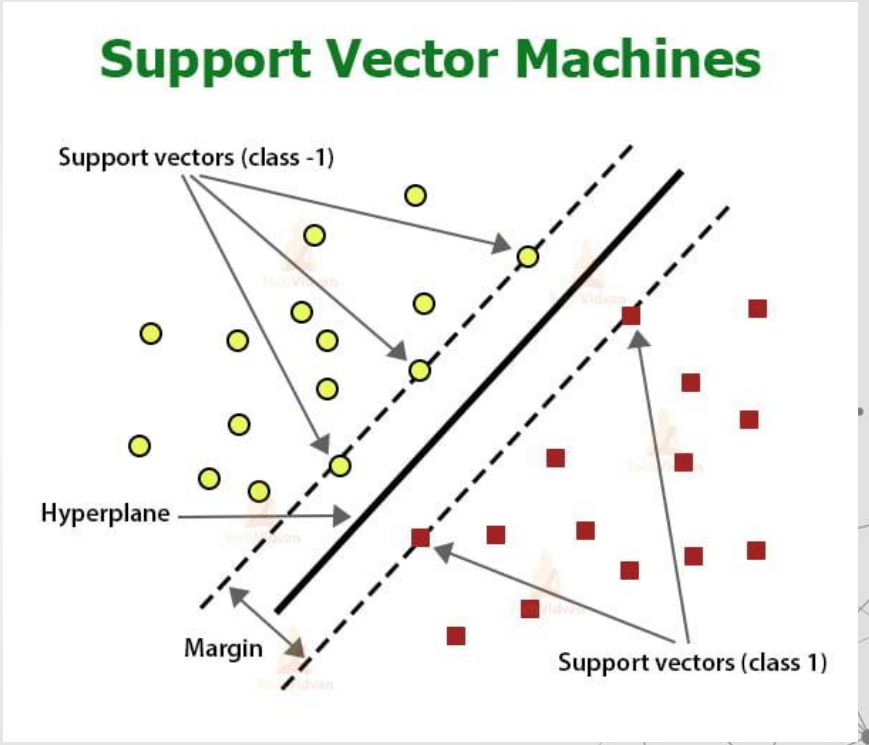


Classification as well as Regressor problems

Usuel MACHINE LEARNING ALGORITHMS

Support Vector Machine

SVM tries to find a line/hyperplane (in multidimensional space) that separates these two classes. Then it classifies the new point depending on whether it lies on the positive or negative side of the hyperplane depending on the classes to predict.



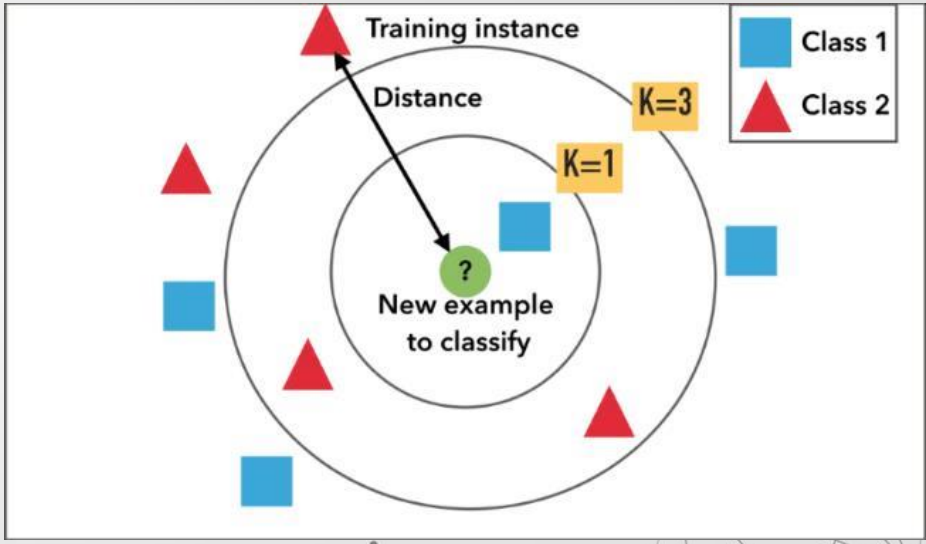
Classification as well as Regressor problems

Usuel MACHINE LEARNING ALGORITHMS

K Nearest Neighbor

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

It does not create a generalized model during the time of training. Testing is very costly.

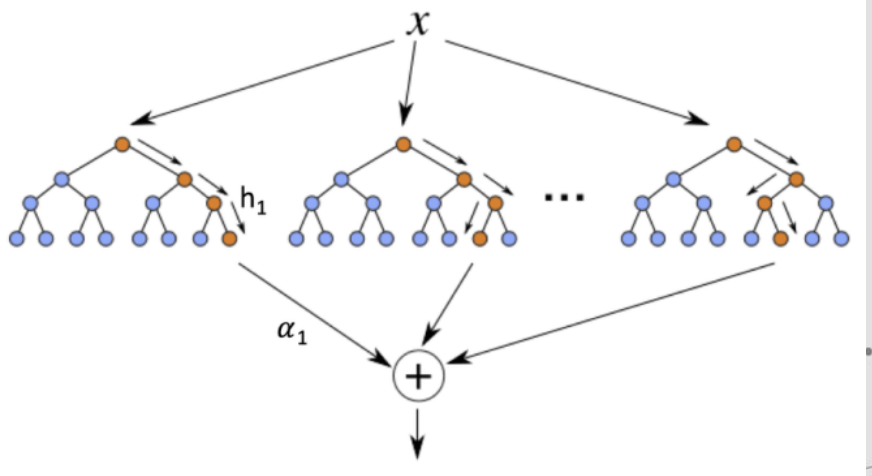


Classification as well as Regressor problems

Usuel MACHINE LEARNING ALGORITHMS

Gradient Boosting

XGBoost is a decision-tree-based ensemble Machine
Learning algorithm that uses
a gradient boosting framework.
The approach consists in using a gradient descent to minimize the empirical risk

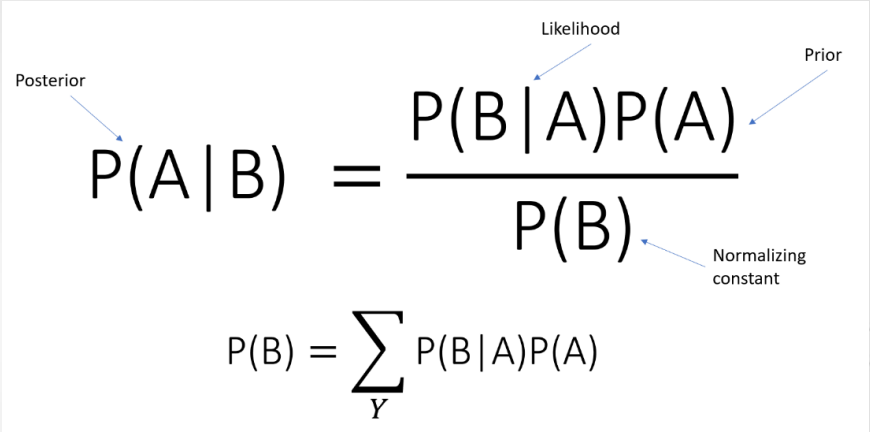


Classification as well as Regressor problems

Usuel MACHINE LEARNING ALGORITHMS

Naive Bayes Classifier

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

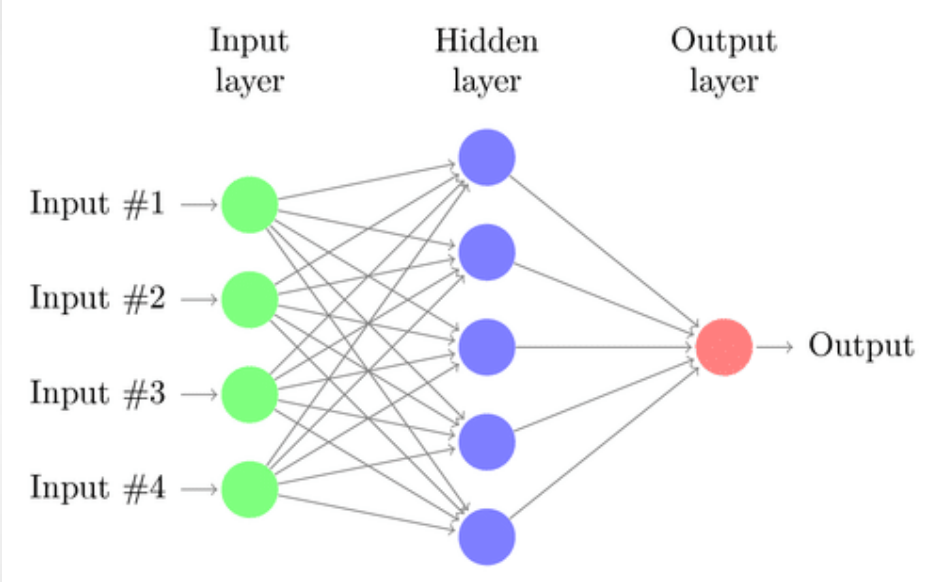


Classification problems

Usuel MACHINE LEARNING ALGORITHMS

Neural Network

As neural suggests, they are brainspired systems which are intended to replicate the way that we humans learn. NNs consist of input and output layers, as well as a hidden layer consisting of units that transform the input.



Classification as well as Regressor problems

Pourquoi existe-t-il autant d'algorithmes ?

Le « théorème » du « No Free Lunch »

Il n'existe pas d'algorithme qui soit le meilleur quelque soit le problème d'apprentissage…

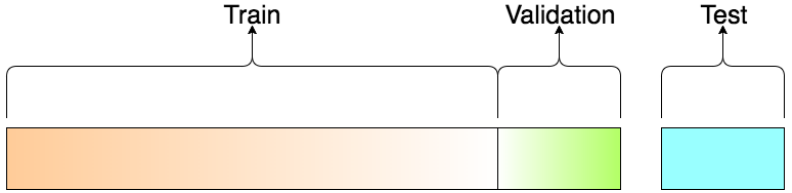
Schéma typique (standard) de l'analyse prédictive supervisée

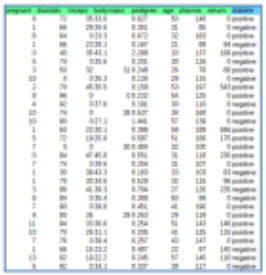
Y : variable cible

X1, X2, … : variables explicatives

f(.) une fonction qui essaie d'établir la relation Y = f(X1, X2, …)

f(.) doit être « aussi précise que possible »…





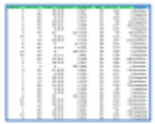
Construction de la fonction f(.) à partir des données d'apprentissage

Training set

$Y = f(X1,X2,\ldots) + \varepsilon$

Application du modèle (prédiction) sur l'ensemble de test

Validation set



$(Y, \hat{Y})$

Y : valeurs observées

Y^ : valeurs prédites par f(.)

Mesures de performances par confrontation entre Y et Y^ : matrice de confusion + mesures

Metrics Regression

1. **MAE : Mean Average Error**

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

2. **RMSE : Root Mean Squared Error**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

3. **Root Mean Squared Logarithmic Error**

$$RMSLE = \sqrt{\frac{1}{n}\sum(\log(1 + \text{prédiction}) - \log(1 + \text{ci}$$

4. **Autres mesures**

Mean Squared Error, Weigthed Mean Average Error, ...

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\text{prédiction}_i - \text{cible}_i)^2$$

Metrics Classification

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

1. Accuracy, Precision, Recall…

Accuracy = (TP+TN)/(TP+FP+FN+TN)

Precision = TP/(TP+**FP**) : is the proportion of all positive cases that were correctly classified => **FP intolerable**

Sensitivity/Recall = (TP)/(TP+**FN**) => **FN intolerable**

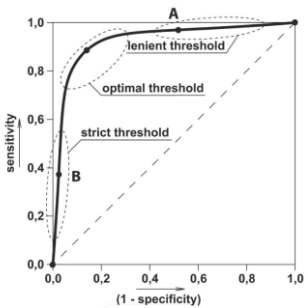
Specificity = TN/(TN+FP) : is the proportion of all negative cases that were correctly classified

False positive rate (FPR) =1−specificity = 1-(FP/(FP+TN)) : is the proportion of false positives amor

2

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Plus d'information : https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226



3. AUC

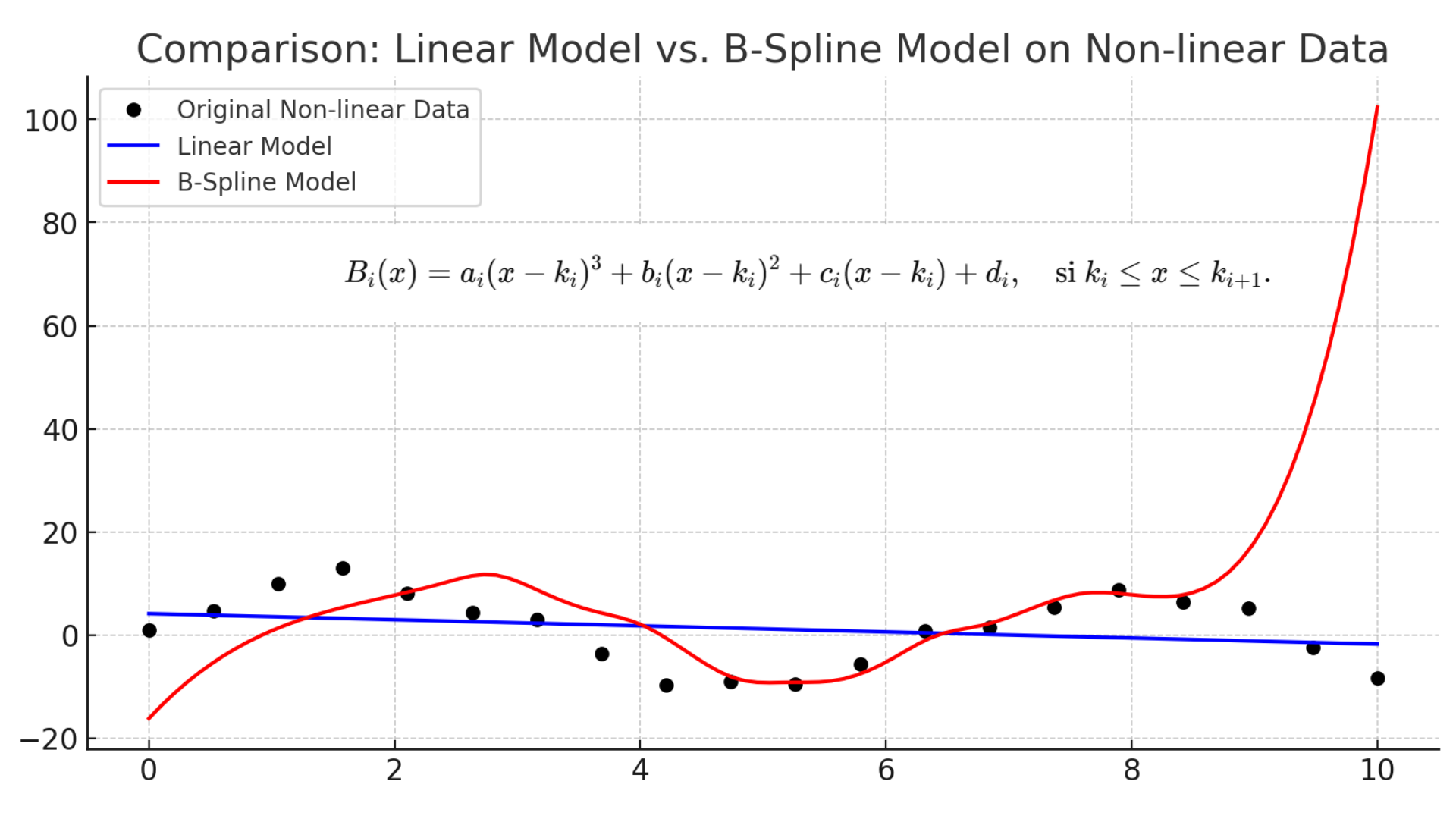
AUC ROC indicates how well the probabilities from the positive classes are separated from the negative classes => Allowing threshold selection

Interaction

- Une promotion (variable 1) augmente les ventes de 10 unités.
- La publicité (variable 2) augmente les ventes de 15 unités.
- Cependant, si vous combinez une promotion **et** une publicité, les ventes augmentent de 40 unités.
- Ici, l'effet combiné (40) est supérieur à la somme des effets individuels (10 + 15 = 25).

Modèle : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$

Si $\beta_3 > 0$, cela signifie qu'il existe une interaction amplificatrice entre x_1 et x_2.


Comparison: Linear Model vs. B-Spline Model on Non-linear Data
Original Non-linear Data
Linear Model
B-Spline Model
$B_i(x) = a_i(x - k_i)^3 + b_i(x - k_i)^2 + c_i(x - k_i) + d_i, \quad \text{si } k_i \leq x \leq k_{i+1}.$
100
80
60
40
20
0
−20
0
2
4
6
8
10

Méthode	Quand l'utiliser
Interactions	Quand il existe une relation forte entre deux variables (effets combinés) ou à tester avec une hypothèse.
Polynomial Features	Quand la relation non linéaire est globale et simple (par exemple, courbe quadratique ou cubique).
Splines	Quand la relation non linéaire est complexe et locale (varie dans différentes parties des données).

Processus général d'un problème de data science

1

Définition de la cible et des variables explicatives

- Identification ou construction de la cible
- Identification des variables explicatives

2

Construction du jeu de données

- Filtres de la population concernée
- Exploration et visualisation

3

Préparation des données

- Feature Engineering : création de variables explicatives **metier et DS**
- Traitement des valeurs manquantes
- Encodage numérique des variables catégoriques **et La mise à l'échelle**
- Séparation des données en échantillons

4

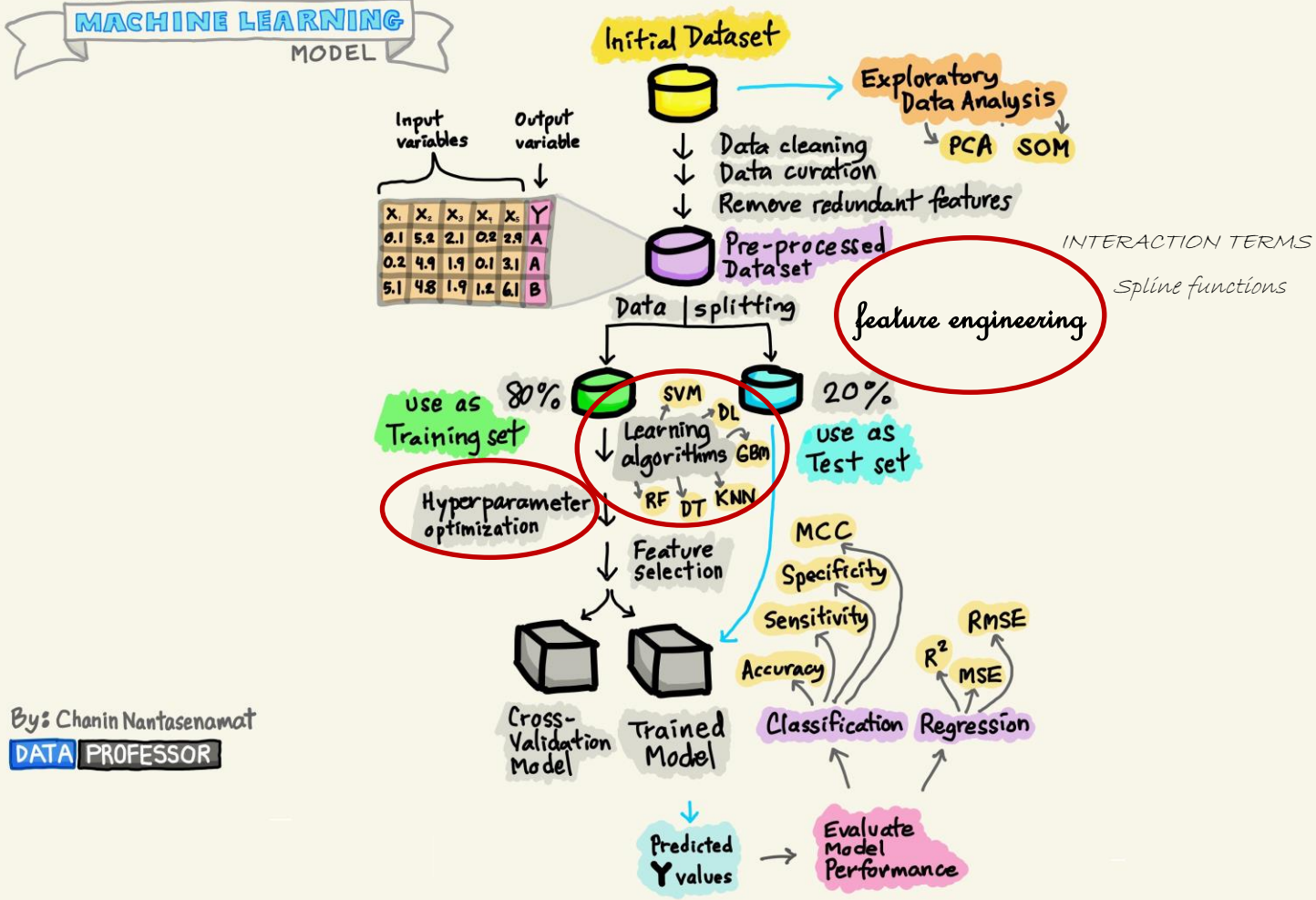
Entraînement du modèle

- Sélection du meilleur jeu de paramètres
- Entraînement du modèle

Résultats et performance du modèle

- Application du modèle sur l'échantillon de test
- Etude des caractéristiques clés du modèle

BUILDING THE MACHINE LEARNING MODEL

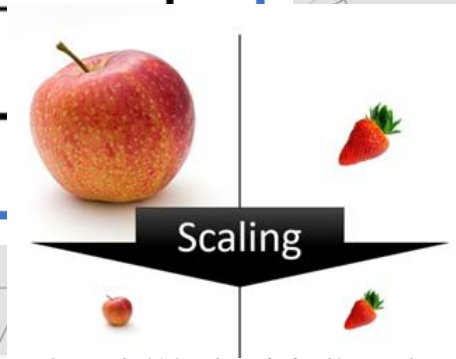


X_1	X_2	X_3	X_4	X_5	Y
0.1	5.2	2.1	0.2	2.9	A
0.2	4.9	1.9	0.1	3.1	A
5.1	4.8	1.9	1.2	6.1	B

By: Chanin Nantasenamat
DATA PROFESSOR

La mise à l'echelle

Name	Weight	Price
Orange	15	1
Apple	18	3
Banana	12	2
Grape	10	5



La mise à l'echelle

Sr No.	Algorithms	Feature Scaling
1.	Linear/Non-Linear Regressions	Yes
2.	Logistic Regression	Yes
3.	KNN	Yes
4.	SVM	Yes
5.	Neural Networks	Yes
6.	K-means clustering	Yes
7.	CART	No
8.	Random Forests	No
9.	Gradient Boosted Decision Trees	No
10.	Naïve Bayes	N0
11.	PCA	Yes
12.	SVD	Yes
13.	Factorization Machines	Yes

Fig: Feature Scaling requires based on Machine Learning

Source : Ref : https://www.kaggle.com/getting-started/159643

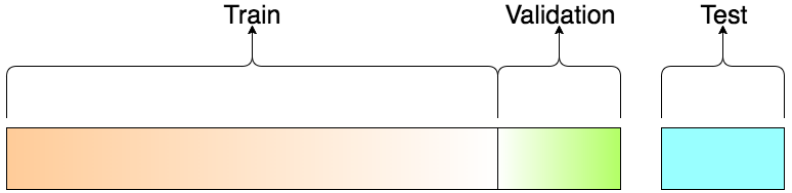
Schéma typique (standard) de l'analyse prédictive supervisée

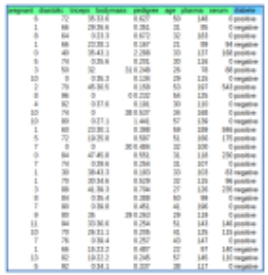
Y : variable cible

X1, X2, … : variables explicatives

f(.) une fonction qui essaie d'établir la relation Y = f(X1, X2, …)

f(.) doit être « aussi précise que possible »…

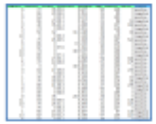




Ensemble de données (dataset)

Ensemble d'apprentissage

Training set

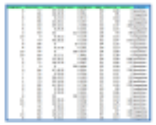


Construction de la fonction f(.) à partir des données d'apprentissage

$Y = f(X1,X2,\ldots) + \varepsilon$

Application du modèle (prédiction) sur l'ensemble de test

Test set



Ensemble de test

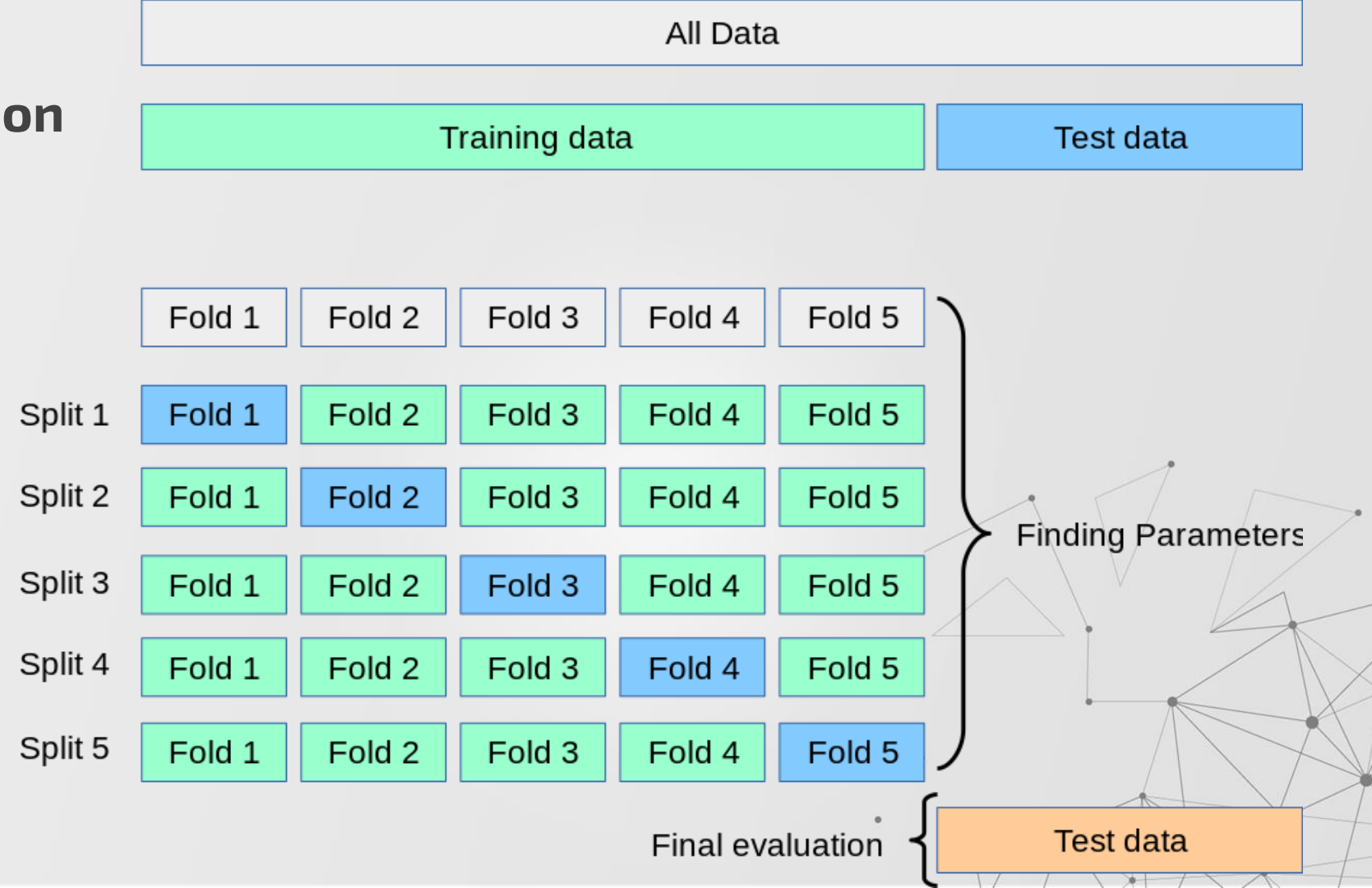
$(Y, \hat{Y})$

Y : valeurs observées

Y^ : valeurs prédites par f(.)

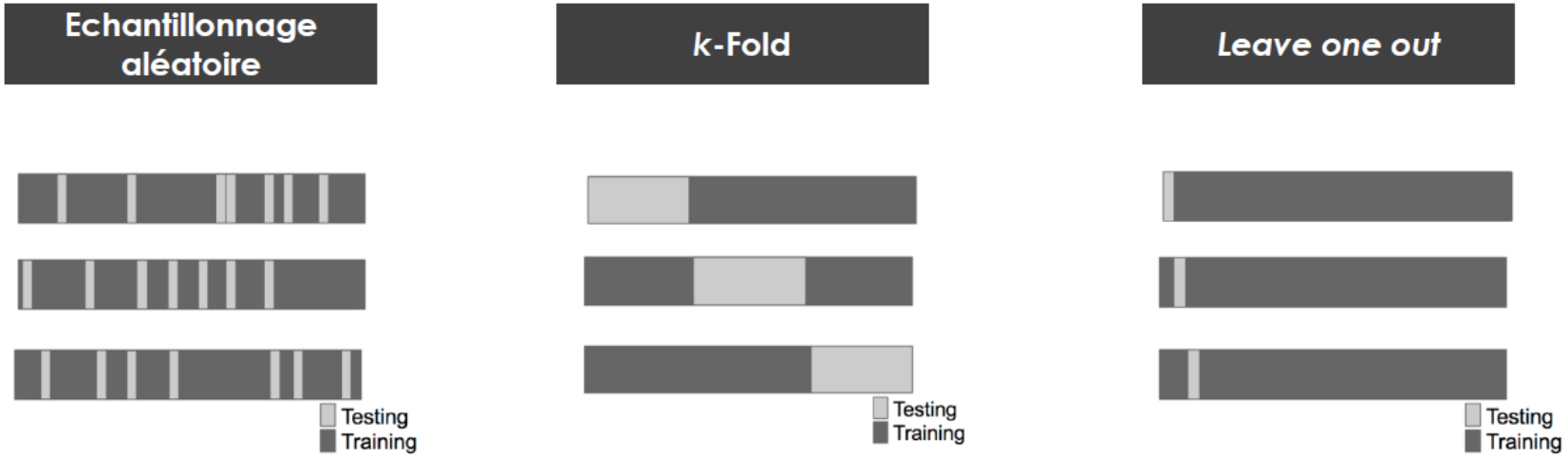
Mesures de performances par confrontation entre Y et Y^ : matrice de confusion + mesures

Cross validation



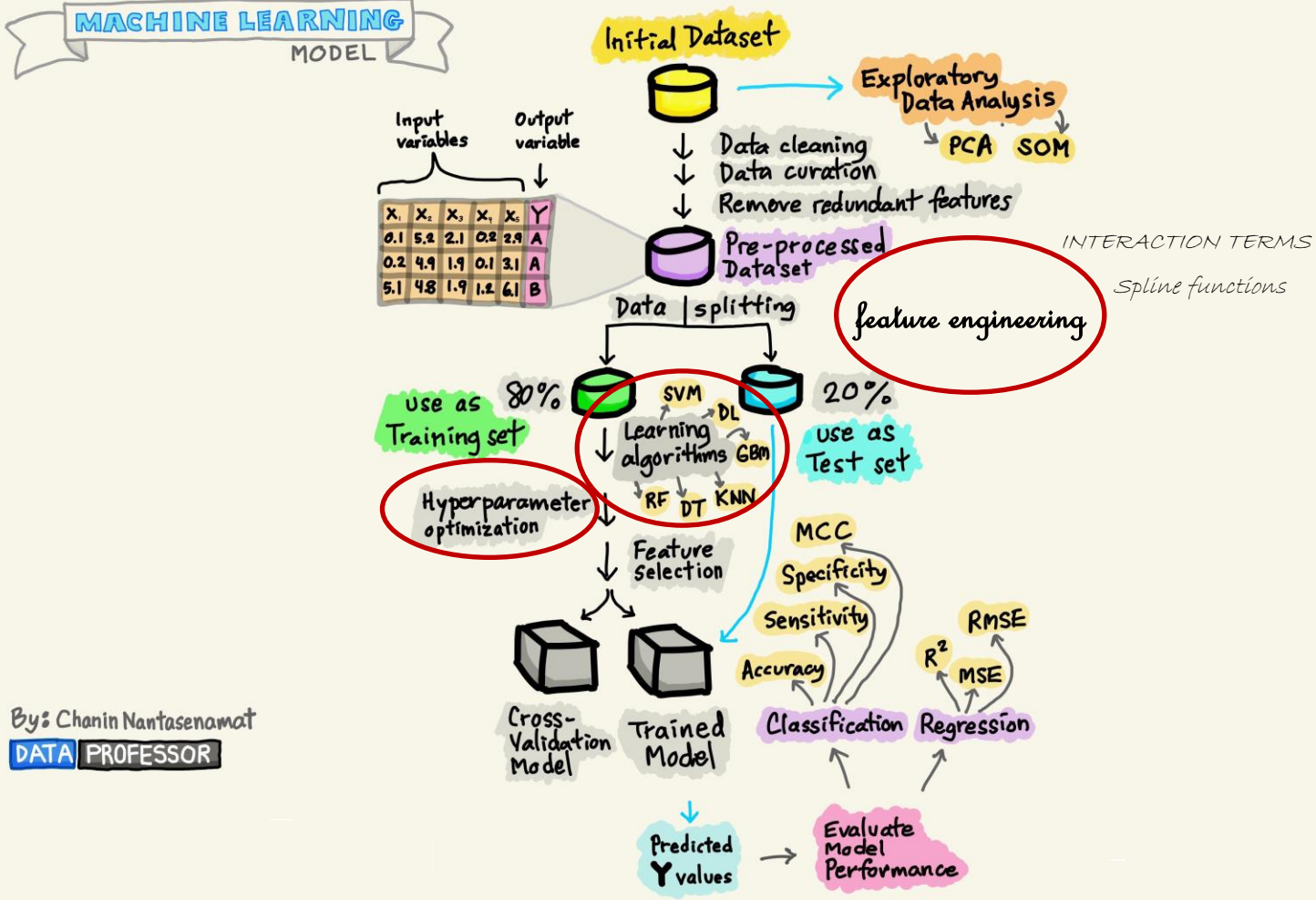
Model cross-validation

Différentes méthodes pour s'assurer que le modèle « apprend » bien, qu'il « généralise »



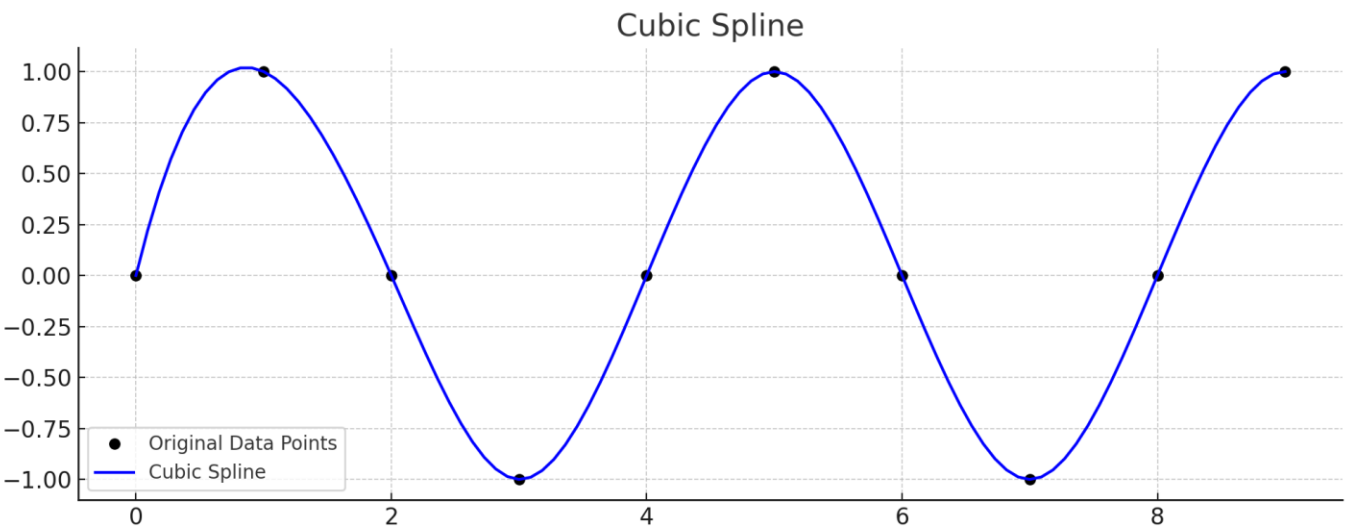
Idée de la validation croisée pour comparer la performance des modèles, indépendamment des « individus » utilisés pour apprendre (risque si l'on utilise 1 seul jeu d'apprentissage).
Habituellement, on utilise entre 5 ou 10 « *folds* ».

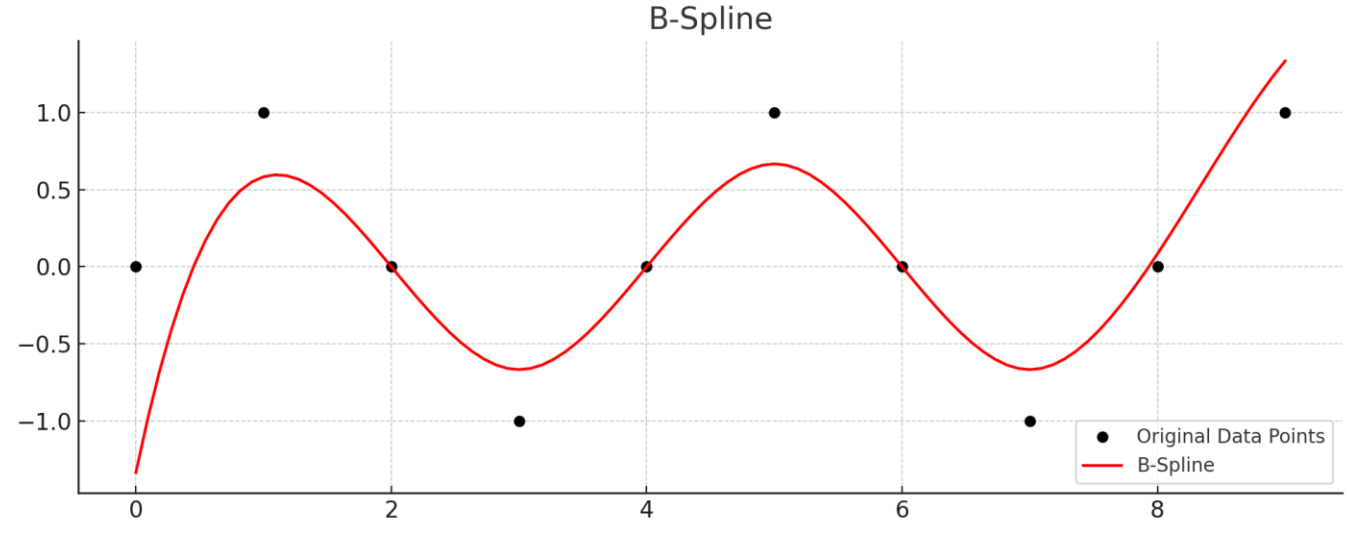
BUILDING THE MACHINE LEARNING MODEL



X_1	X_2	X_3	X_4	X_5	Y
0.1	5.2	2.1	0.2	2.9	A
0.2	4.9	1.9	0.1	3.1	A
5.1	4.8	1.9	1.2	6.1	B

By: Chanin Nantasenamat
DATA PROFESSOR


Cubic Spline
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
Original Data Points
Cubic Spline


B-Spline
1.0
0.5
0.0
−0.5
−1.0
0
2
4
6
8
Original Data Points
B-Spline

B-spline de degré 3 (cubic B-spline)

Si vous avez des nœuds $t = [1, 2, 3, 4, 5]$, une cubic B-spline est définie par la formule récursive :

$$B_{i,3}(x) = \frac{x - t_i}{t_{i+3} - t_i} B_{i,2}(x) + \frac{t_{i+4} - x}{t_{i+4} - t_{i+1}} B_{i+1,2}(x),$$

où $B_{i,2}(x)$ est une B-spline de degré 2.

Cubic spline (non-B-spline)

Pour les mêmes nœuds, une cubic spline s'écrit explicitement comme :

$$f(x) = a_i(x - t_i)^3 + b_i(x - t_i)^2 + c_i(x - t_i) + d_i, \quad \text{si } t_i \leq x \leq t_{i+1}.$$

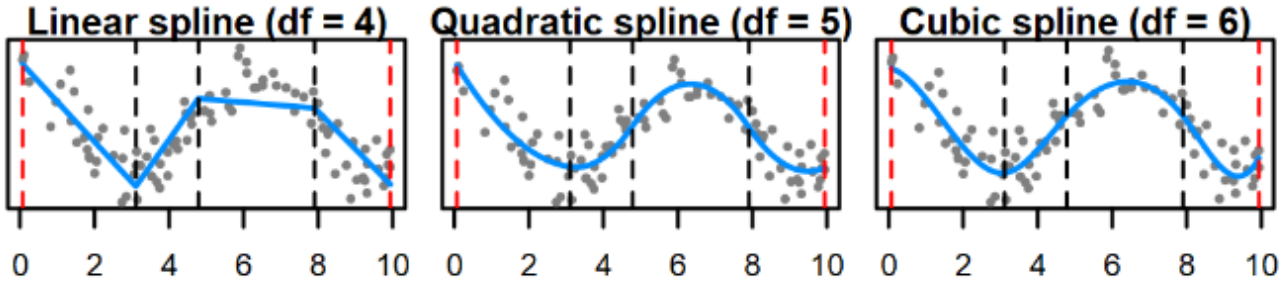
Les coefficients a_i, b_i, c_i, d_i sont calculés en résolvant un système linéaire basé sur les conditions de continuité.

$$B_i(x) = a_i(x - k_i)^3 + b_i(x - k_i)^2 + c_i(x - k_i) + d_i, \quad \text{si } k_i \leq x \leq k_{i+1}.$$

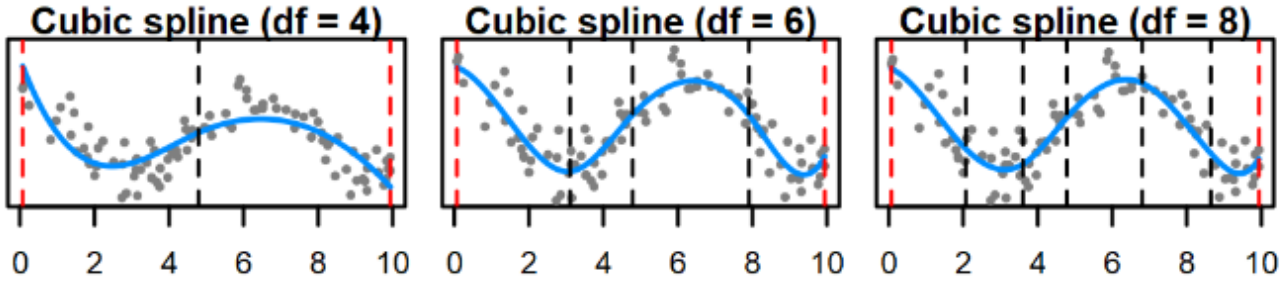
For B-splines: $\mathrm{df} = k + \mathrm{degree}$ if you specify the knots or $k = \mathrm{df} - \mathrm{degree}$ if you specify the degrees of freedom and the degree. For natural (restricted) cubic splines: $\mathrm{df} = k - 1$ if you specify the knots or $k = \mathrm{df} + 1$ if you specify the degrees of freedom.

As an example: A cubic spline ($\mathrm{degree} = 3$) with 4 (internal) knots will have $\mathrm{df} = 4 + 3 = 7$ degrees of freedom. Or: A cubic spline ($\mathrm{degree} = 3$) with 5 degrees of freedom will have $k = 5 - 3 = 2$ knots.

Let's see some illustrations. In the scatterplots below you see some artifical data together with the spline fits of different degrees but the same amount of knots ($k = 3$). The knots are indicated by dashed vertical lines (Boundary knots by red dashed lines) and are placed at the 25th, 50th and 75th percentile of x. The first plot shows a linear spline ($\text{degree} = 1$), the second one a quadratic spline ($\text{degree} = 2$) and the third is a cubic spline with $\text{degree} = 3$.



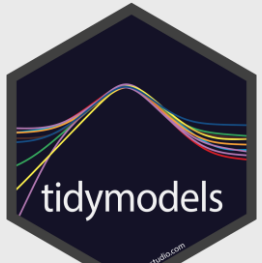
In the next plot, you see three cubic splines with different degrees of freedom. As before, the knots are shown as dashed vertical lines. With increasing degrees of freedom, the number of knots gets larger (from 1 to 3 to 5). The spline gets wigglier although the difference is only really noticeable between the first and second plot.



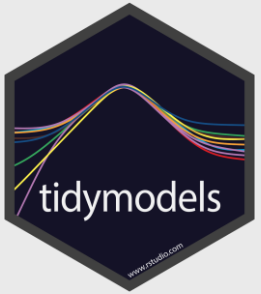
spline function procedures

A cubic spline function, with three knots (τ_1, τ_2, τ_3) will have 7 degrees of freedom. Using representation given in Eq. 2, the function can be written as:

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X - \tau_1)^3 + \beta_5 (X - \tau_2)^3 + \beta_6 (X - \tau_3)^3$$



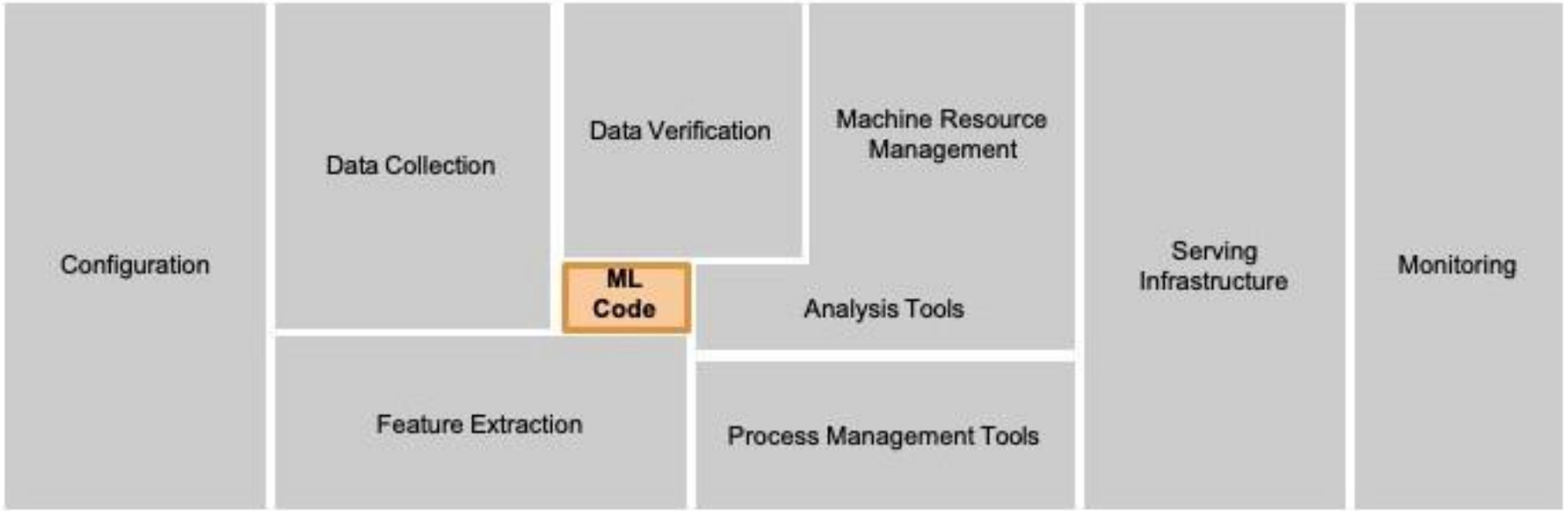
Detecting Interaction Effects



REF :
https://bookdown.org/max/FES/detecting-interaction-effects.html

Excercice

The Requirements Surrounding ML Infrastructure



Expressions régulières (1/1)

- ion : recherche les mots qui contiennent la chaine “ion”, dans n’importe quelle position
- ion$: mots se terminant par “ion” ($=fin de mot)
- ^anti : recherche tous les mots commençant par “anti” (^=début de mot)
- ^maison$: recherche exactement le mot “maison”
- p.r : recherche les mots qui contiennent un “p”, suivi d’une lettre quelconque, puis d’un “r” (le point correspond à n’importe quel caractère)
- ^p...r$: mots commençant par “p”, suivi de trois lettres quelconques, et finissant par “r” (le symbole . dans une regex correspond à n’importe quel caractère)
- ^p.*r$: mots commençant par “p” et finissant par “r” (*= répétitions – 0 ou plusieurs fois – du caractère précédent, ici ‘.’, donc n’importe quel caractère)
- oid|ion|ein : recherche les mots qui contiennent (au moins) une des trois chaînes “iod”, “ion” ou “ein” (| = ou).

Source : http://www.lexique.org/?page_id=101
https://buzut.net/la-puissance-des-regex/

Expressions régulières (2/2)

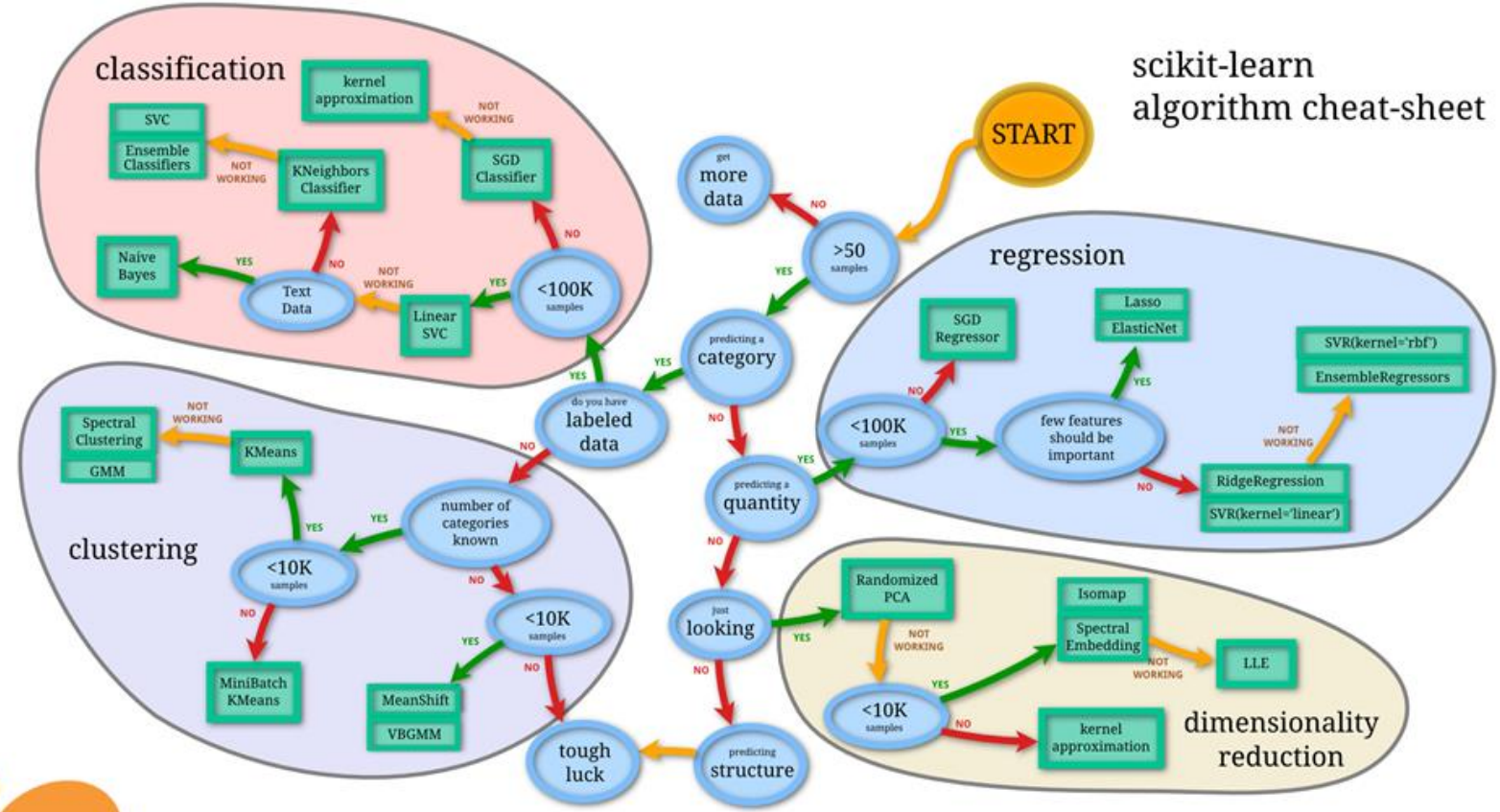
- [A-Za-z] : n'importe quoi comme caractère, majuscule ou minuscule
- + : 1 ou plus
- () : contenu à extraire
- \w : un caractère alphanumérique ou un tiret de soulignement (tiret de 8).

Source : http://www.lexique.org/?page_id=101
https://buzut.net/la-puissance-des-regex/

Python vs. R: What's the Difference?

What is the difference between Anaconda Prompt and Anaconda Powershell Prompt?

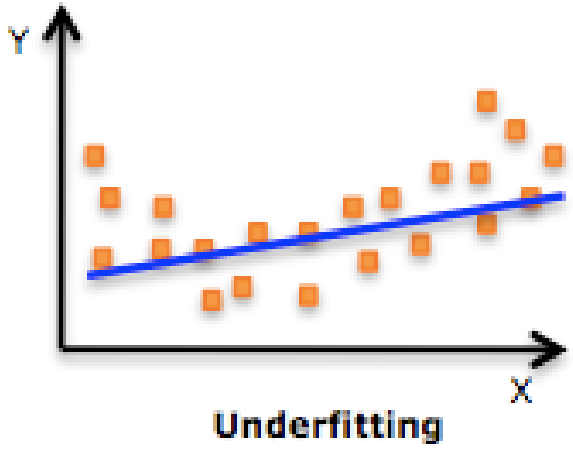
Choisir son algorithme



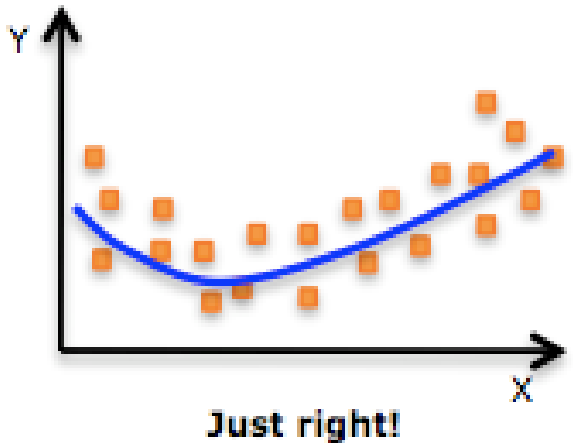
THANKS

Does anyone have any questions?

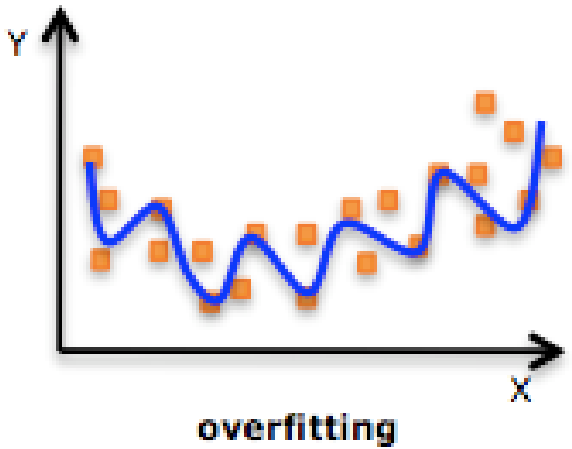
CREDITS: Fei GAO (EHESP)



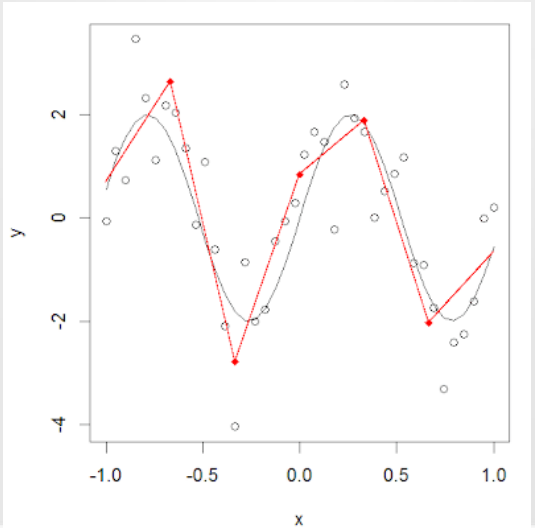
Underfitting

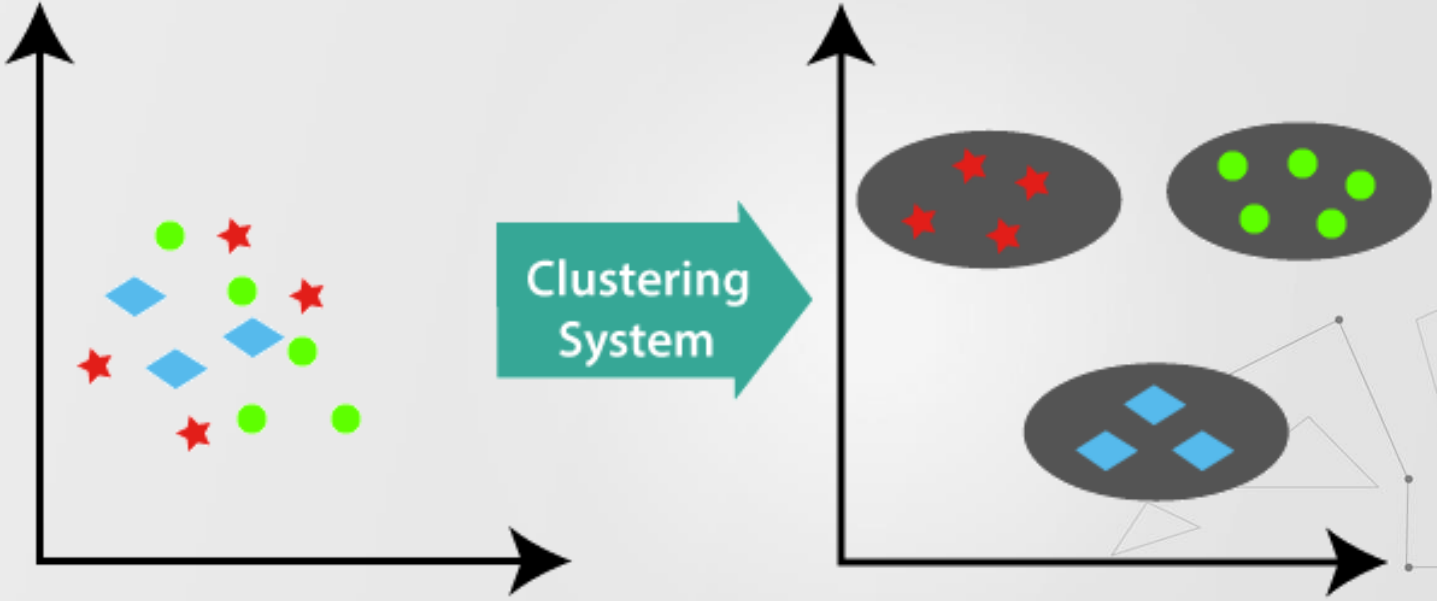


Just right!



overfitting




Clustering
System

Créez votre compte Databricks

1/2

Prénom

Fei

Nom

GAO

E-mail professionnel

constance.fei.gao.pro@gmail.com

Entreprise

hdh

Intitulé de poste

DS

Numéro de téléphone (facultatif)

0661213981

Pays

France

Oui, je souhaite recevoir les communications marketing concernant les services, les événements et les produits open source de Databricks. Je suis conscient(e) que je peux modifier mes préférences à tout moment.

Continuer

Sélectionnez un fournisseur de cloud

2/2

Amazon Web Services

Microsoft Azure

Google Cloud Platform

Continuer

En cliquant sur « Essayer », vous acceptez la Politique de confidentialité et les Conditions de service .

Vous n'avez pas de compte cloud ?

Community Edition est un environnement Databricks limité destiné à un usage personnel et à la formation.

Essayer Community Edition →

En cliquant sur « Essayer Community Edition », vous acceptez la Politique de confidentialité et les Conditions de service undefined.