

# 4I-SI2 - Machine Learning

## Introduction

Romain Negrel  
[romain.negrel@esiee.fr](mailto:romain.negrel@esiee.fr)

ESIEE Paris

2018 - 2019

# Informations

## Note finale

- 1 ou 2 QCM
- 1 Projet noté

## Prérequis

- Algèbre linéaire
- Programmation Python
- Notebook jupyter

- 1 Introduction
- 2 Formulation & Optimisation
- 3 Mise en œuvre
- 4 Critères d'évaluations

# L'informatique

## Objectif majeur

Effectuer des **tâches** et résoudre des problèmes de manière **automatique**

## Comment ?

- ➊ Un ordinateur est capable d'exécuter des **milliards d'instructions par seconde**
- ➋ **Un humain écrit un algorithme** pour résoudre une problème

## Exemples

- Déterminer la plus petite valeur entre deux valeurs
- Calcul du PGCD de deux nombres (Euclide 300 avant J.-C.)
- Tri d'un tableau
- Chemin dans un graphe

# La prédition

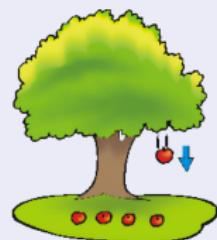
## Définition

Annonce d'événements futurs par la connaissance qu'on a de leurs causes

## Modèle mathématique

Un humain crée le modèle mathématique qui décrit le phénomène

Exemple :

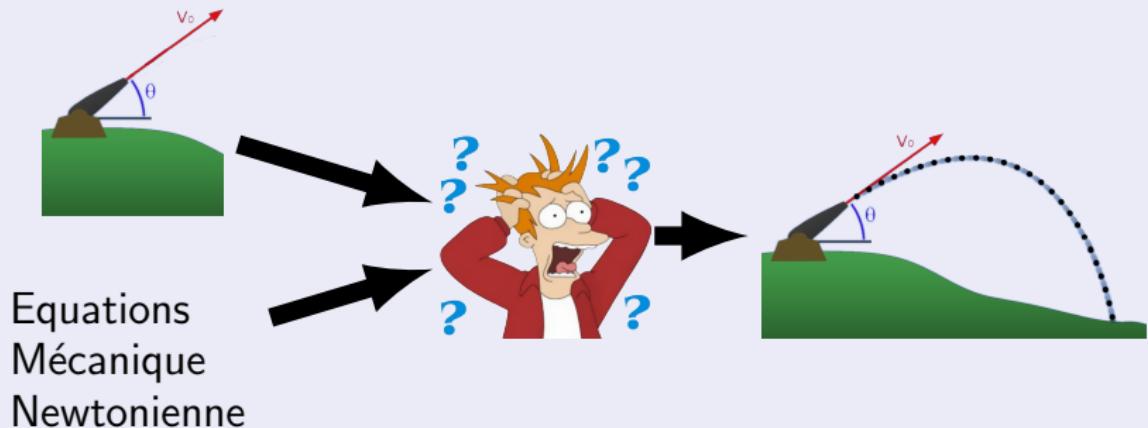


Équations  
Mécanique  
Newtonienne

# La prédition

## Exemple de prédition

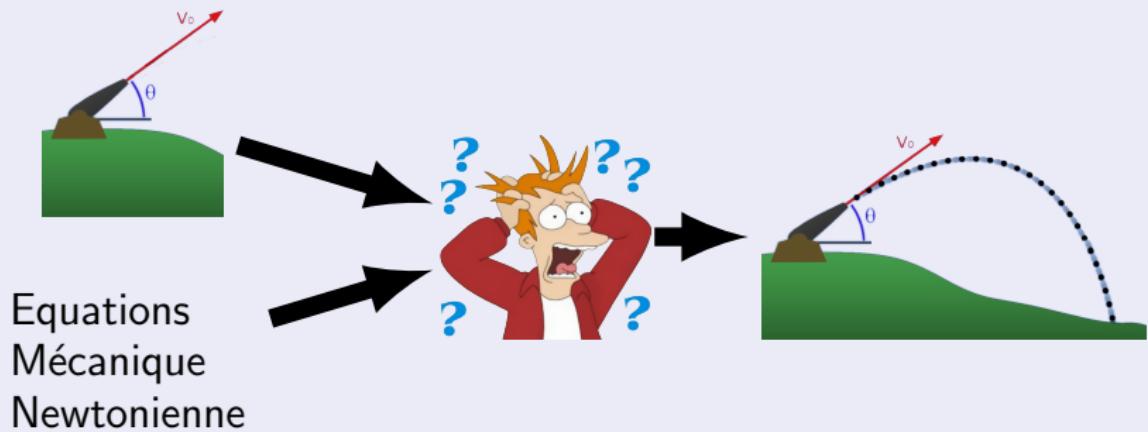
### La Balistique



# La prédition

## Exemple de prédition

### La Balistique



## Première problématique

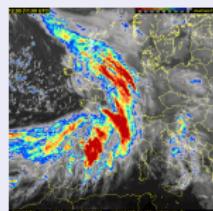
Comment faire quand **les calculs** sont **trop complexe** pour un Humain ?

## La prédition

## Solution

## Automatisé les calculs !

Exemple, la prédition météorologique



## Modèle météorologique

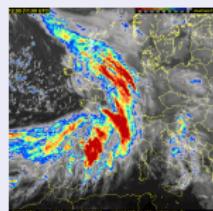


# La prédition

## Solution

Automatisé les calculs !

Exemple, la prédition météorologique



Modèle  
météorologique



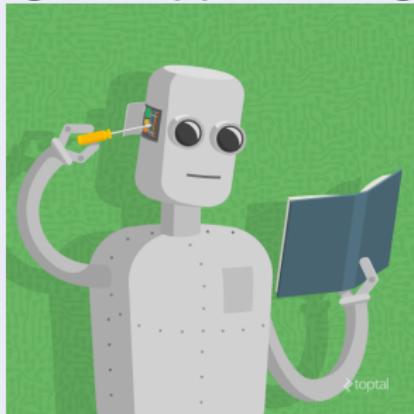
## Deuxième problématique

Comment faire quand le **modèle mathématique** sont **trop complexe** pour être trouvé par un Humain ?

# La prédition

## Solution

### Machin Learning ou Apprentissage Automatique



**Définition :** Ensemble des méthodes permettant à une machine à résoudre un **problème automatiquement** à partir d'un ensemble **d'exemples de données**

# Exemples d'application

liés à notre quotidien :

- trier des mails par thème, filtrer les spams;
- reconnaître le style musical d'un morceau;
- regrouper les acheteurs par types;
- savoir si un message est important, pertinent ou non;
- prédire le prix de l'immobilier, le nombre de ventes d'un nouveau produit;
- prédire la hausse ou la baisse d'un cours boursier.

plus techniques :

- prédire une caractéristique manquante d'un individu à partir d'autres caractéristiques connues;
- faire un diagnostic automatique à partir d'une analyse médicale;
- extraire et reconnaître du texte ou un visage dans une image.

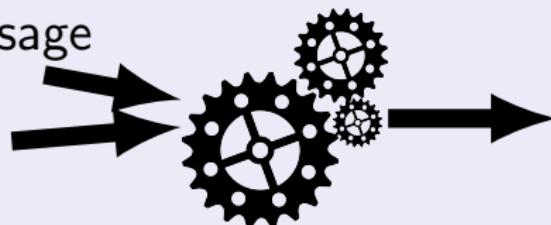
# Idée principale

Apprendre de la **connaissance** à partir de **données**

## Apprentissage

Données d'apprentisage

Modèle générique  
paramétrable



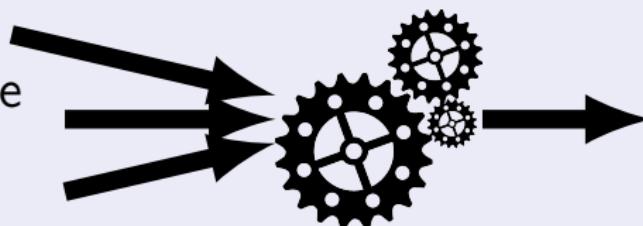
Paramètres  
du modèle

## Prédiction

Cas particulier

Modèle générique  
paramétrable

Paramètres du  
modèle



Résultat de  
prédiction

# Idée principale

## Données

- données quantitatives;
- données catégorielles;
- textes;
- images;
- sons;
- ...

## Prédictions

- valeurs quantitatives;
- catégories;
- textes;
- images;
- sons;
- ...

# Deux grandes catégories d'apprentissage

## Apprentissage supervisé (*supervised learning*) :

pour lequel, les **prédictions attendu sont prédéterminées**. Nous avons une base d'apprentissage avec des données complètes (ou couples d'information)

- Chaque donnée d'apprentissage est associé à la prédition attendu  
On parle alors de base d'apprentissage avec une **vérité terrain** (*ground truth* en anglais) ou que les exemples ont des **labels**

## Apprentissage non supervisé (*unsupervised learning*) :

pour lequel, les **prédictions attendu sont inconnues**. Nous avons une base d'apprentissage avec uniquement les données

# Apprentissage supervisé (*supervised learning*)

## Applications :

- **La régression** : le paramètre de sortie est de type quantitatif;
- **La discrimination** (*Classification*) : le paramètre de sortie est de type catégoriel.
- ...

## Exemples :

- **Prédiction de l'âge** : son entrée brute est une photo d'identité et la sortie est l'âge en année (c'est de la régression);
- **Filtre de SPAMS** : son entrée brute est un email et la sortie désirée est la décision SPAM ou NON-SPAM (c'est de la discrimination).

# Apprentissage non supervisé (*unsupervised learning*)

## Applications :

- **La catégorisation** (*Clustering*) : les données sont regroupées en groupe homogènes;
- **La détection d'anomalie** : détection des données qui sont anormales par rapport à l'ensemble d'apprentissage;
- ...

## Exemples :

- **Gestion de la relation client** : regrouper les clients en groupe homogènes pour offrir un service client personnalisé en fonction du groupe;
- **Alarme de vidéo surveillance** : détection d'événement rare (accident, mouvement de foule, agressions, etc) à partir du flux vidéo d'une caméra de surveillance.

# Notation mathématique

Formulation de prédicteur :

Fonction de prédiction :

$$\tilde{\mathbf{y}} = f_{\mu, \lambda}(\mathbf{x})$$

avec

- $\mathbf{x} \in \mathbb{R}^N$  : le vecteur de données;
- $\tilde{y} \in \mathbb{R}$  ou  $\in \{-1, 1\}$  : la valeur prédite;
- $\mu$  : les paramétrés d'apprentissage;
- $\lambda$  : les hyper-paramétrés.

Exemple : régression polynomial

$$\tilde{y} = f_{\mu, \lambda}(x) = \sum_{k=0}^D a_k x^k$$

avec  $\mathbf{x} \in \mathbb{R}$ ,  $\tilde{\mathbf{y}} \in \mathbb{R}$ ,  $\mu = \{a_1, \dots, a_D\} \in \mathbb{R}^D$  et  $\lambda = \{D \in \mathbb{N}\}$ .

# Notation mathématique

Formulation de l'apprentissage :

Fonction d'apprentissage :

$$g(f_{\mu,\lambda}, \{\mathcal{X}, \mathcal{Y}\})$$

avec

- $f_{\mu,\lambda}$  : la fonction de prédiction;
- $\mathcal{X}$  : ensemble des données d'apprentissages;
- $\mathcal{Y}$  : ensemble des prédictions attendues.

Exemple : régression polynomial (Erreur quadratique moyenne)

$$g(f_{\mu,\lambda}, \{\mathcal{X}, \mathcal{Y}\}) = \sum_{i=1}^P (f_{\mu,\lambda}(x_i) - y_i)^2$$

avec  $\mathcal{X} = \{x_1, \dots, x_P\}$ ,  $\mathcal{Y} = \{y_1, \dots, y_P\}$  et P le nombre d'exemple de la base d'apprentissage.

# Régression polynomial (D=4)

$$g(f_{\mu,\lambda}, \{\mathcal{X}, \mathcal{Y}\}) = \sum_{i=1}^P (f_{\mu,\lambda}(x_i) - y_i)^2$$

$$f_{\mu,\lambda}(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$$

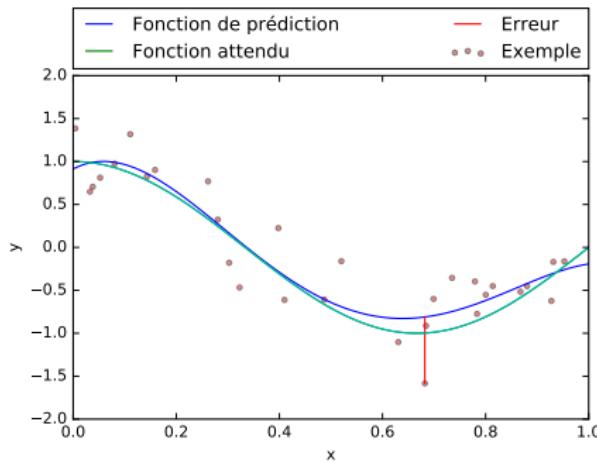


Figure: Erreur quadratique moyenne en régression polynomial

# Optimisation

## Définition

L'optimisation est une branche des mathématiques cherchant à trouver le minimum ou maximum d'une fonction

## Application en apprentissage :

Apprendre consiste alors à recherché le minimum (ou maximum) de la fonction d'apprentissage en fonction de ces paramètres d'apprentissage.

Exemple pour la régression polynomial:

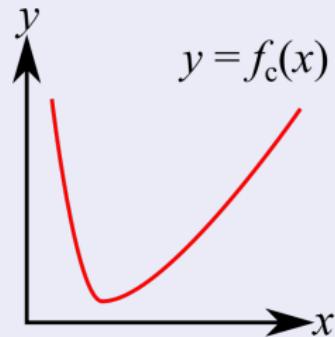
$$\boldsymbol{\mu}^* = \arg \min_{\boldsymbol{\mu}} \sum_{i=1}^P (f_{\boldsymbol{\mu}, \lambda}(x_i) - y_i)^2$$

$\boldsymbol{\mu}^*$  sont les paramètres d'apprentissage optimum

# Deux grandes catégories de problème d'optimisation

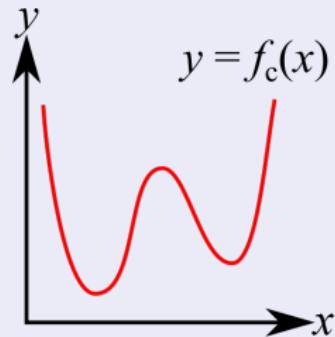
## L'optimisation convexe

- **Simples** à analyser et à résoudre
- Algorithme **générique efficace** pour trouver la solution
- unique minimum global  
⇒ **unique** solution



## L'optimisation non convexe

- **Difficile** à analyser et à résoudre
- **Pas** d'algorithme générique efficace
- Multiple minimum globaux et locaux  
⇒ **Plusieurs** solutions possibles !



# Composantes d'un système d'apprentissage automatique

## L'extraction de paramètres

Extraire des données brutes des **paramètres** et regroupées le plus souvent en **vecteurs**

En général les **données brutes sont inexploitables.**

## Le choix de la technique utilisée

Pour une même tache, il existe plusieurs techniques pour la résoudre.

## L'évaluation de l'apprentissage.

Dans le cas d'apprentissage supervisé, l'évaluation permet de mesurer des **performances** et la **capacité de généralisation** de l'apprentissage.

Dans le cas d'apprentissage non-supervisé, il est plus difficile de mesurer des performances.

# Les données d'entrée

La majorité des algorithmes de Machine Learning utilisent une **représentation vectorielle** des données ( $x \in \mathbb{R}^N$ )

## Problème de représentation

Types de données "faciles" :

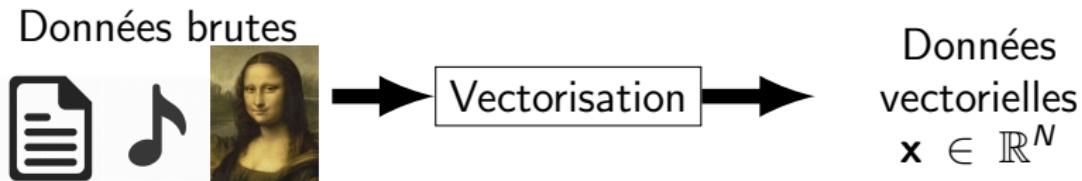
- Données quantitatives

Types de données "problématique" :

- Données catégorielles
- Textes
- Images

# Extraction de caractéristiques

Le but est alors de convertir des données brutes en données vectorielles.  
On parle alors de vectorisation (ou *embedding* en anglais)



## Exemple avec des données catégorielles

Les tailles de T-shirts:

Catégorie	Représentation vectorielle
XS	$x = (1, 0, 0, 0, 0)$
S	$x = (0, 1, 0, 0, 0)$
M	$x = (0, 0, 1, 0, 0)$
L	$x = (0, 0, 0, 1, 0)$
XL	$x = (0, 0, 0, 0, 1)$

# Extraction de caractéristiques

## Exemple pour le text

- Nombre de mots
- Fréquence des mots
- Nombre moyen de lettres par mot
- *etc*

## Exemple pour l'image

- Histogramme des couleurs
- Histogramme du gradient
- *etc*

## Exemple pour le son

- Spectre des fréquences
- *etc*

# Évaluation des performances

## Apprentissage supervisé

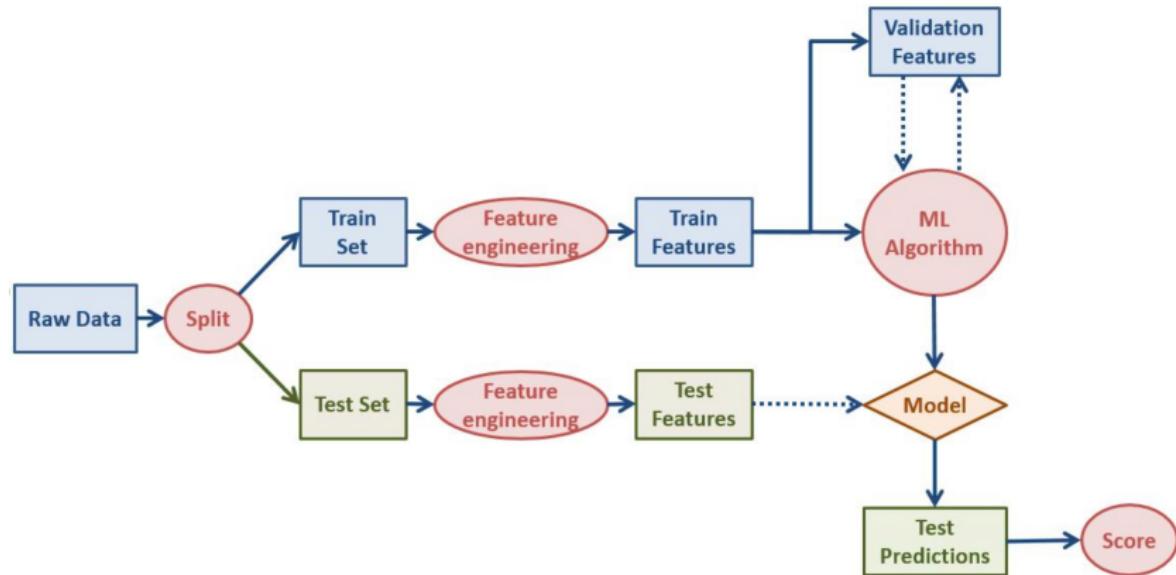
- Déterminer la **capacité de prédiction** obtenu par l'apprentissage
  - ▶ taux d'erreur
  - ▶ erreur quadratique moyenne
  - ▶ probabilité d'erreur
  - ▶ etc
- Déterminer la **capacité de généralisation** obtenu par l'apprentissage
  - ▶ Évaluation de performance sur des exemples qui n'ont pas servi à l'apprentissage.

## Division de l'ensemble de données disponible

Divisons l'ensemble des données en deux sous-ensembles disjoints :

- ensemble **d'entraînement**
- ensemble **de test**

# Apprentissage supervisé



## Note

Les sous-ensembles d'entraînement et de test sont **tirés aléatoirement**, il est préférable de construire **plusieurs sous-ensembles** pour évaluer des performances

# Sous-apprentissage et Sur-apprentissage

Il y a sous-apprentissage quand

la méthodes d'apprentissage n'a pas la capacité d'apprendre correctement par rapport à la complexité des données.

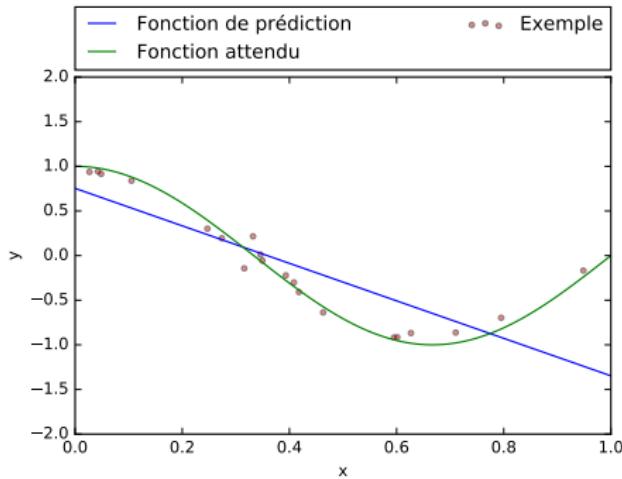


Figure: Exemple : régression linéaire de donnée non-linéaire

# Sous-apprentissage et Sur-apprentissage

Il y a Sur-apprentissage quand

la méthodes d'apprentissage apprend par cœur les données d'entraînement et ce trompe sur les données de test

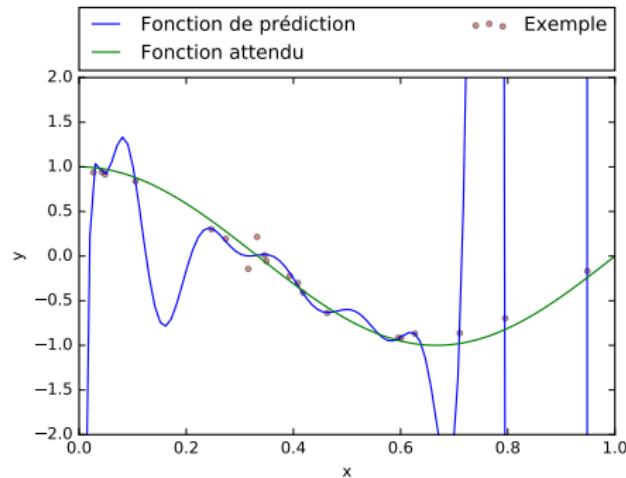


Figure: Exemple : régression polynomiale (degré 15) avec un petit nombre d'exemple d'apprentissage

# Sous-apprentissage et Sur-apprentissage

## Comment choisir ?

Pour choisir la bonne méthode d'apprentissage, il faut étudié ces performances sur des exemples qui n'ont pas servi a l'apprentissage !

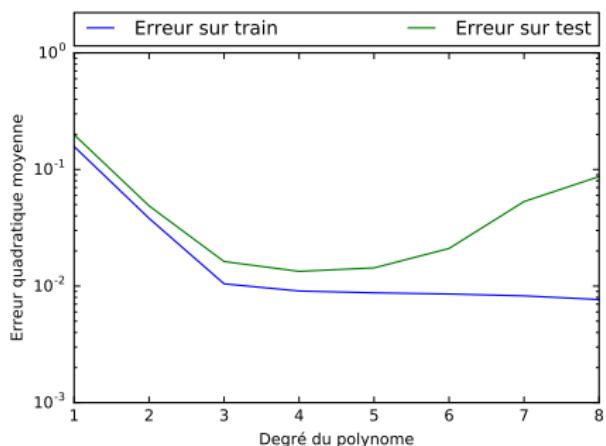


Figure: Évolution de l'erreur en fonction du degré du polynôme

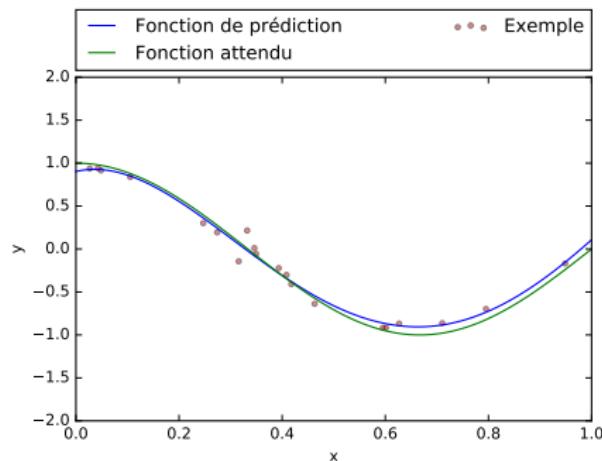


Figure: Régression polynomiale optimal (degré 4)

## Erreur quadratique moyenne

L'erreur quadratique moyenne est un critère de performance très souvent utilisé quand l'on cherche à estimé une **quantitative** :

$$\text{EQM}(\hat{y}) = \mathbb{E} [(\hat{y} - y)^2]$$

Elle se décompose en deux termes :

$$\mathbb{E} [(\hat{y} - y)^2] = \underbrace{\mathbb{E} [(\mathbb{E}(\hat{y}) - y)^2]}_{\text{Biais}(\hat{y})^2} + \underbrace{\mathbb{E} [(\mathbb{E}(\hat{y}) - \hat{y})^2]}_{\text{Var}(\hat{y})}$$

Ces deux critères sont également intéressante à étudier :

- Une **biais important** signifie que le modèle sous-jacent est trop simple (**Sous-apprentissage**)
- Une variance importante signifie que le modèle sous-jacent est trop complexes (**Sur-apprentissage**)

# Erreur quadratique moyenne

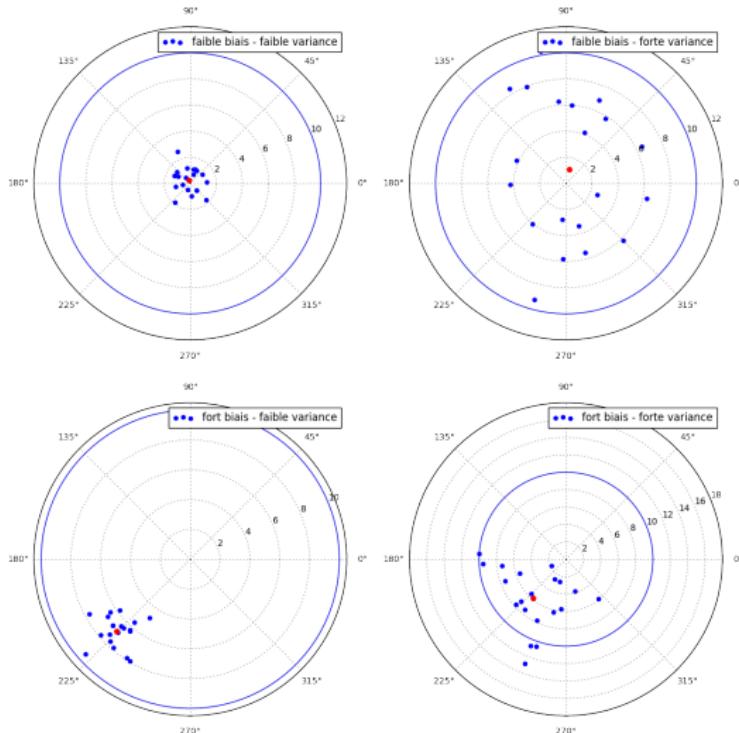


Figure: Illustration la signification des termes de biais et variance

# Taux d'erreur - Matrice de Confusion

## Critère de performance pour le catégorielles

Le critère utilisé le plus courant est le **taux d'erreur** :

$$\text{taux d'erreur}(\hat{y}) = \frac{\text{nombre de fausse prediction}}{\text{nombre de prediction}}$$

Attention critère trop générique et trompeur !

## Problèmes du taux d'erreur

- Base d'exemple déséquilibrer (99% d'exemple positif)
  - ▶ Que signifie un taux d'erreur de 1% ?
- Il y a 1000 catégories
  - ▶ Est-ce bien d'avoir un taux d'erreur de 50% ?
- Application un diagnostic maligne/bénigne pour une tumeur
  - ▶ Est-ce suffisant de savoir que l'on a un taux d'erreur de 5% ?

# Taux d'erreur - Matrice de Confusion

## Matrice de Confusion

La matrice des erreurs de prédiction de chaque catégorie de chaque catégorie

Pour deux catégories (Vrai/Faux)

		Predit	
		Faux	Vrai
Vérité	Faux	TN	FP
	Vrai	FN	TP

- TP : Vrai Positif (*True Positive*)
- TN : Vrai Négatif (*True Negative*)
- FP : Faux Positif (*False Positive*)
- FN : Faux Négatif (*False Negative*)

Exemple : Deux diagnostics de tumeur

	maligne	bénigne
maligne	98	2
bénigne	50	850

	maligne	bénigne
maligne	55	45
bénigne	5	895

Quel est le meilleur test ?

# Taux d'erreur - Matrice de Confusion

MNIST - [yann.lecun.com/exdb/mnist/](http://yann.lecun.com/exdb/mnist/)

Base de données de chiffres écrits à la main, elle regroupe 60000 images d'apprentissage et 10000 images de test



Figure: Exemple de la base MNIST

0	1661	16	10	6	4	3	15	3	7	17
1	23	4950	23	38	35	10	12	65	9	6
2	9	21	3994	32	12	17	4	28	5	34
3	4	34	44	2648	14	59	18	17	9	10
4	2	29	11	5	2442	5	7	6	6	6
5	7	5	28	27	8	2253	27	3	4	9
6	25	7	8	14	3	27	1874	2	38	11
7	2	51	21	8	6	5	2	1887	2	
8	5	7	5	22	2	5	20	3	1570	10
9	11	1	16	88	5	8	7	6	15	1497

Figure: Matrice de confusion

# Taux d'erreur - Matrice de Confusion

Pour deux catégories (Vrai/Faux)

		Predit	
		Faux	Vrai
Vérité	Faux	$TN$	$FP$
	Vrai	$FN$	$TP$

- TP : Vrai Positif (*True Positive*)
- TN : Vrai Négatif (*True Negative*)
- FP : Faux Positif (*False Positive*)
- FN : Faux Négatif (*False Negative*)

## Critères dérivés

- L'exactitude (accuracy) :  $\frac{TP+TN}{TP+TN+FP+FN}$
- La précision (precision) :  $\frac{TP}{TP+FP}$
- La sensibilité (ou taux de vrais positifs) :  $\frac{TP}{TP+FN}$
- La spécificité (ou taux de vrais négatifs) :  $\frac{TN}{TN+FP}$

## Limitation

Ne convient pas pour la classification multi-label (une donnée peut appartenir à plusieurs catégories)

# Plus de critères

## Classification binaire :

- Coefficient de corrélation de Matthews
- Courbe rappel-précision
- Courbe ROC

## Classification multi-labels :

- Le score  $F_1$
- Le score  $F_\beta$
- Précision par label
- Rappel par label

## Classification binaire et multi-labels mais pas multi-classes:

- Précision moyenne