MA4I23 - Machine Learning

Introduction

Romain Negrel
romain.negrel@esiee.fr

ESIEE Paris

Avril 2017

Définition

Machine Learning

Exemples d'application

liés à notre quotidien :

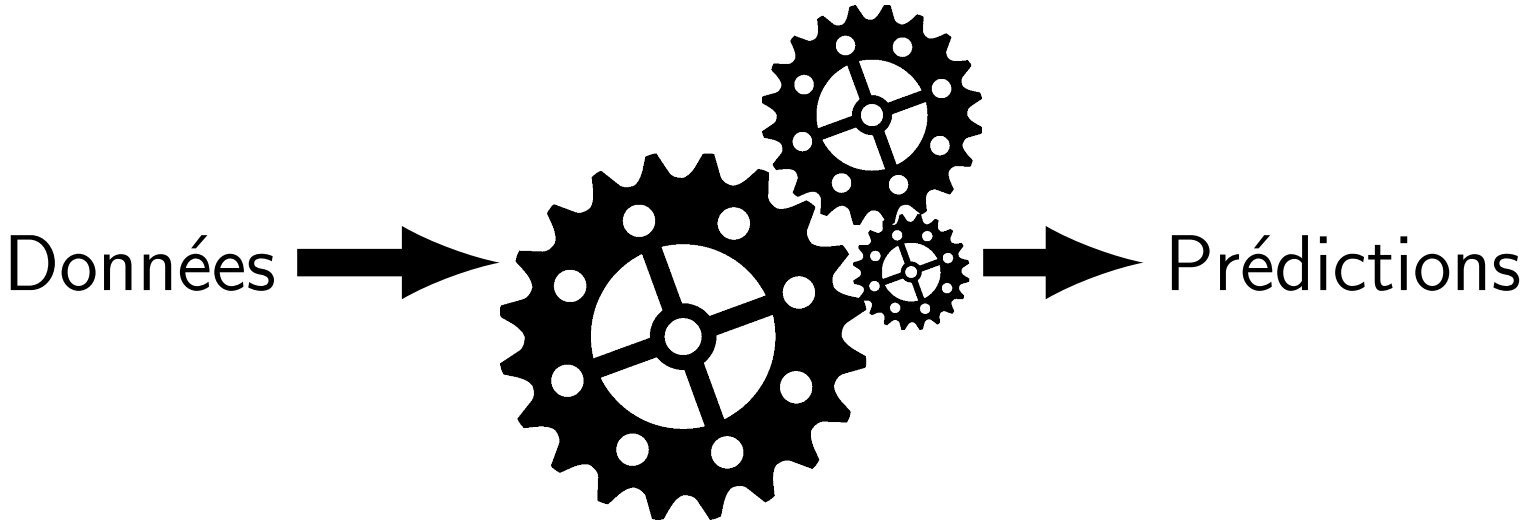
- trier des mails par thème, filtrer les spams;
- reconnaitre le style musical d'un morceau;
- regrouper les acheteurs par types;
- savoir si un message est important, pertinent ou non;
- prédire le prix de l'immobilier, le nombre de ventes d'un nouveau produit;
- prédire la hausse ou la baisse d'un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

Idée principale

Apprendre à prédire



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 prédiction attendu.

Apprentissage non supervisé (*unsupervised learning*) :

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

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'age** : son entrée brute est une photo d'identité et la sortie est l'age 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.

Formulation

Fonction de prédiction :

$$\tilde{\mathbf{y}} = \mathrm{f}_{\boldsymbol{\mu},\boldsymbol{\lambda}}(\mathbf{x})$$

avec

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

Exemple : régression polynomial

$$\tilde{y} = \mathrm{f}_{\boldsymbol{\mu},\boldsymbol{\lambda}}(x) = \sum_{k=0}^{D} a_k x^k$$

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

Formulation

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

Régression polynomial (D=4)

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

$$f_{\boldsymbol{\mu},\boldsymbol{\lambda}}(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$$

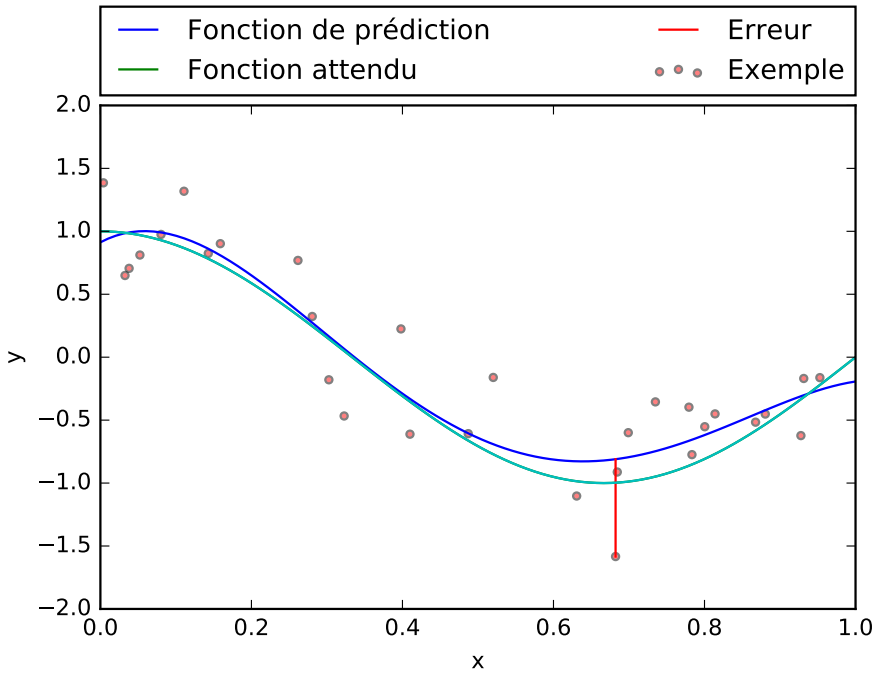


Figure: Erreur quadratique moyenne en régression polynomial

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:

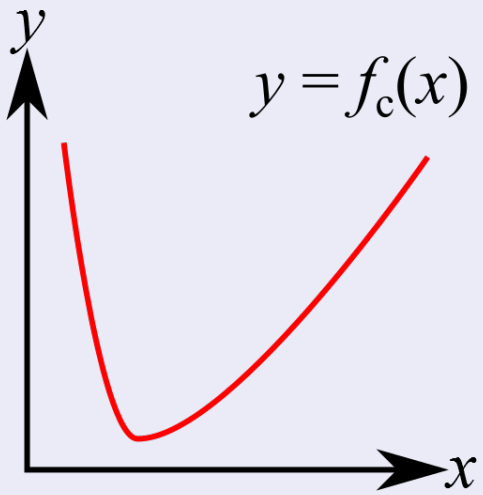
$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{P} (\mathrm{f}_{\boldsymbol{\mu},\boldsymbol{\lambda}}(x_i) - y_i)^2$$

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

Deux grandes catégories de problème d'optimisation

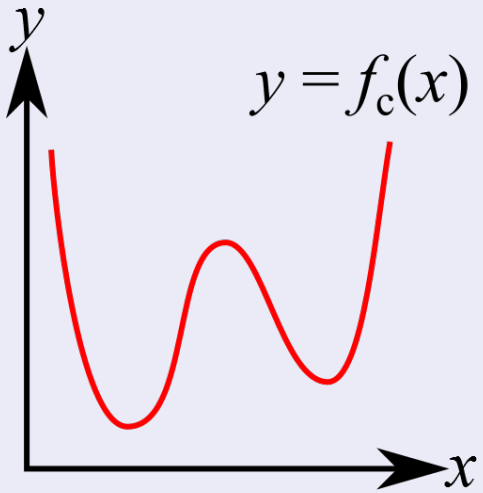
L'optimisation convexe

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



L'optimisation non convexe

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



Composantes d'un système d'apprentissage automatique

L'extraction de paramètres

Extraire des données brutes des valeurs numériques appelées paramètres ou attributs ou variables 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
  Exemple :

Types de données "problématique" :

- Données catégorielles
- Textes
- Images

Conversion de type

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	$x = (1, 0, 0, 0, 0)$
X	$x = (0, 1, 0, 0, 0)$
M	$x = (0, 0, 1, 0, 0)$
L	$x = (0, 0, 0, 1, 0)$
XL	$x = (0, 0, 0, 0, 1)$

Évaluation des performances

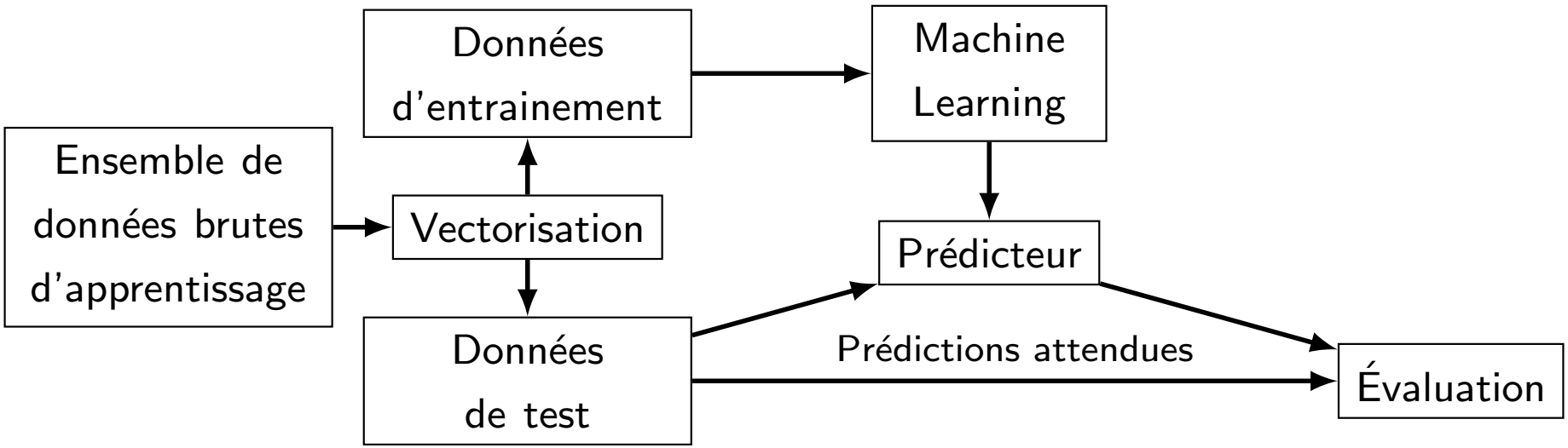
Apprentissage supervisé

- 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

Pour évalué les performances, nous divisons l'ensemble des données disponible en deux sous-ensembles disjoints : un ensemble de données d'entrainement et un ensemble de données de test

Apprentissage supervisé



Note

Les sous-ensembles d'entrainement 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.

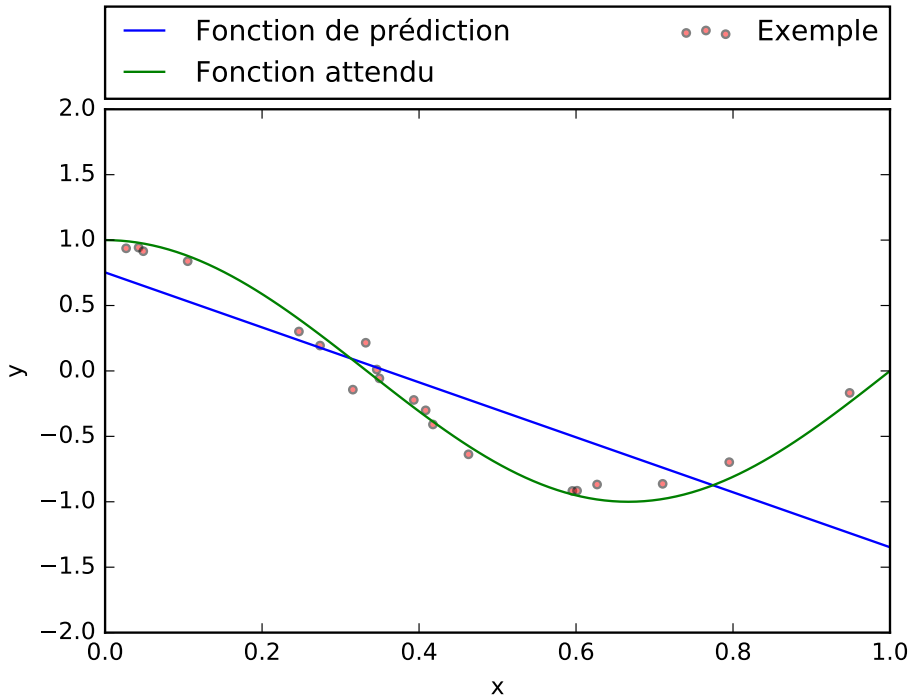


Figure: Exemple : régression linéaire de donnée non-linéaire

Sous-apprentissage et Sur-apprentissage

Il y a Sur-apprentissage quand

la méthodes d'apprentissage apprend par cœurs les données d'entrainement et ce trompe sur les donnée de teste

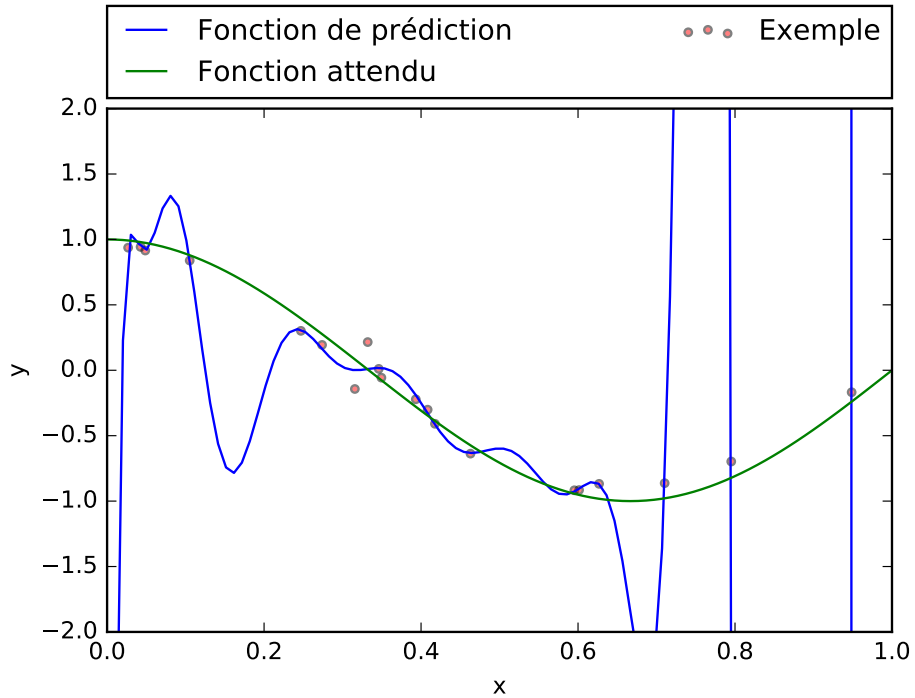


Figure: Exemple : régression polynomiale (degré 15) avec un petit nombre d'exemple d'apprentissage

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 !

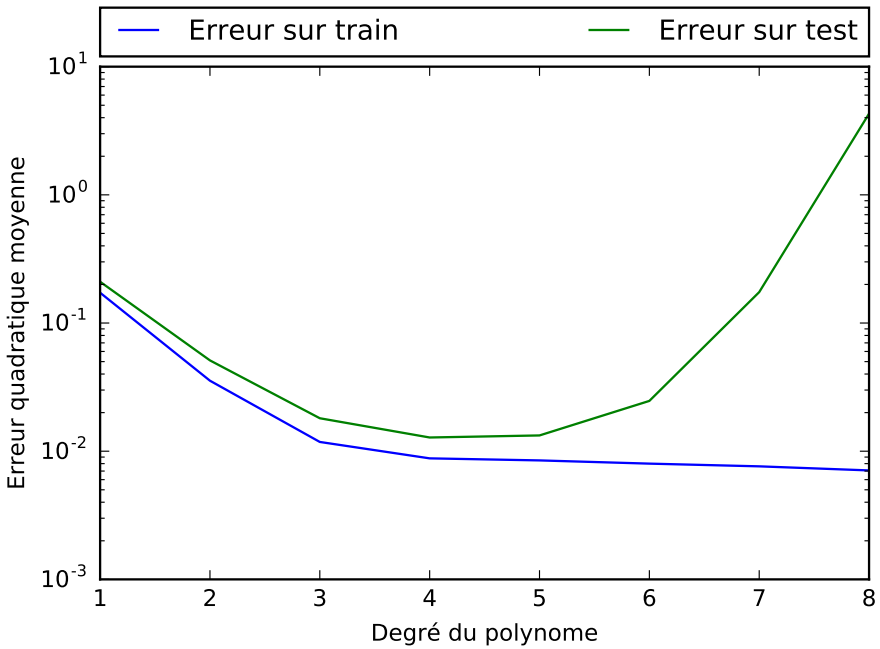


Figure: Évolution de l'erreur en fonction du degré du polynôme

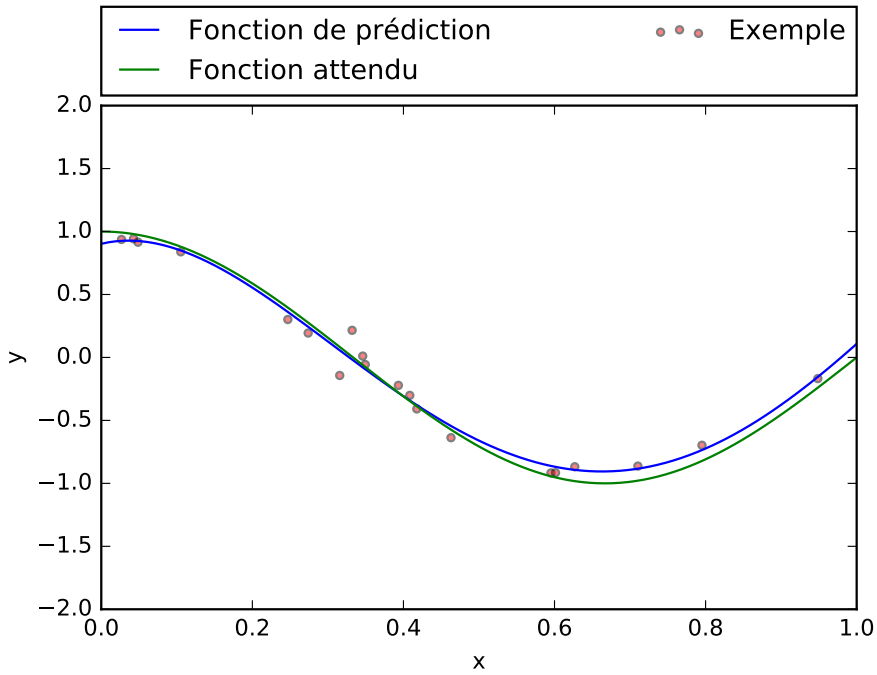


Figure: Régression polynomiale optimal (degré 4)

Erreur quadratique moyenne

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

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

Elle se décompose en deux termes :

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

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)