Comparaison de méthodes de Machine Learning pour l'analyse de données spectroscopiques

Jordane Poulain-Lallemand, Sylvie Roussel

Ondalys - 4 rue Georges Besse, 34830 Clapiers, France

Tel : +33 (0)4 67 67 97 87

jlallemand@ondalys.fr - sroussel@ondalys.fr

Comparaison de méthodes de Machine Learning pour l'analyse de données spectroscopiques

Le Machine Learning : Qu'es aquò ??

Jordane Poulain-Lallemand, Sylvie Roussel

Ondalys - 4 rue Georges Besse, 34830 Clapiers, France

Tel : +33 (0)4 67 67 97 87

jlallemand@ondalys.fr - sroussel@ondalys.fr

Hélio SPIR

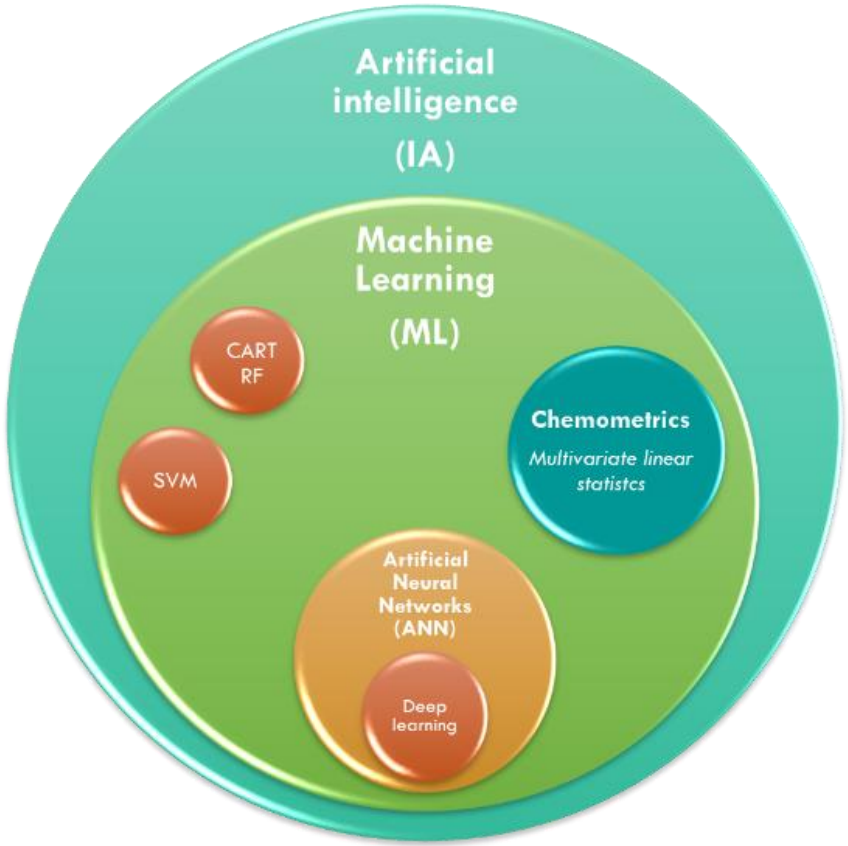
ondalys

Introduction – Le « Machine Learning » : Qu'es aquò ?

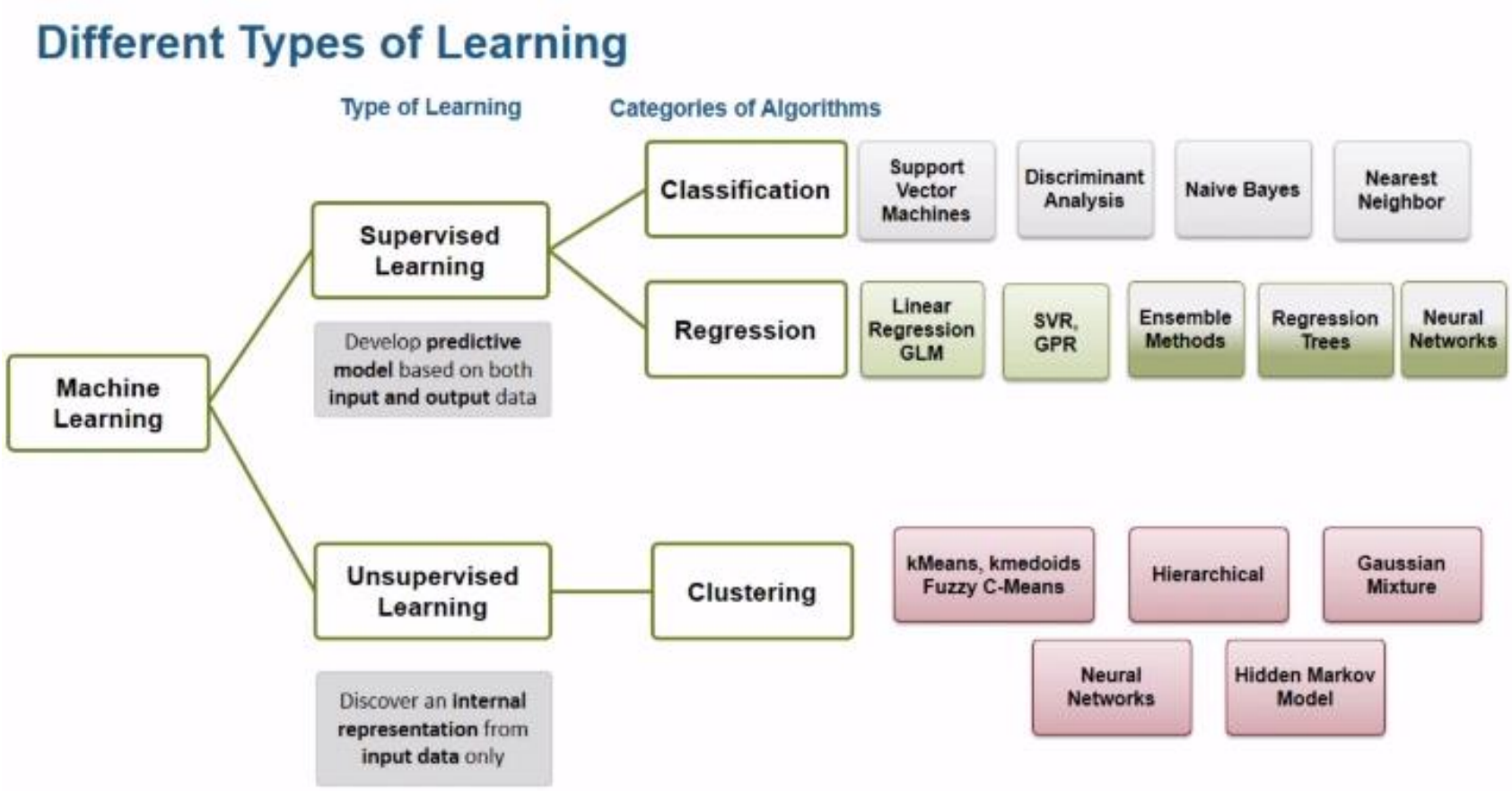
« Machine learning » : apprentissage automatique par ordinateur, grâce à des algorithmes basés sur l'expérience / les données

- Basées sur des données sans nécessiter une équation théorique prédéterminée
- Pas d'hypothèse forte sur la distribution des variables.

➔ Les méthodes classiques utilisées en chimiométrie font partie du « machine learning »



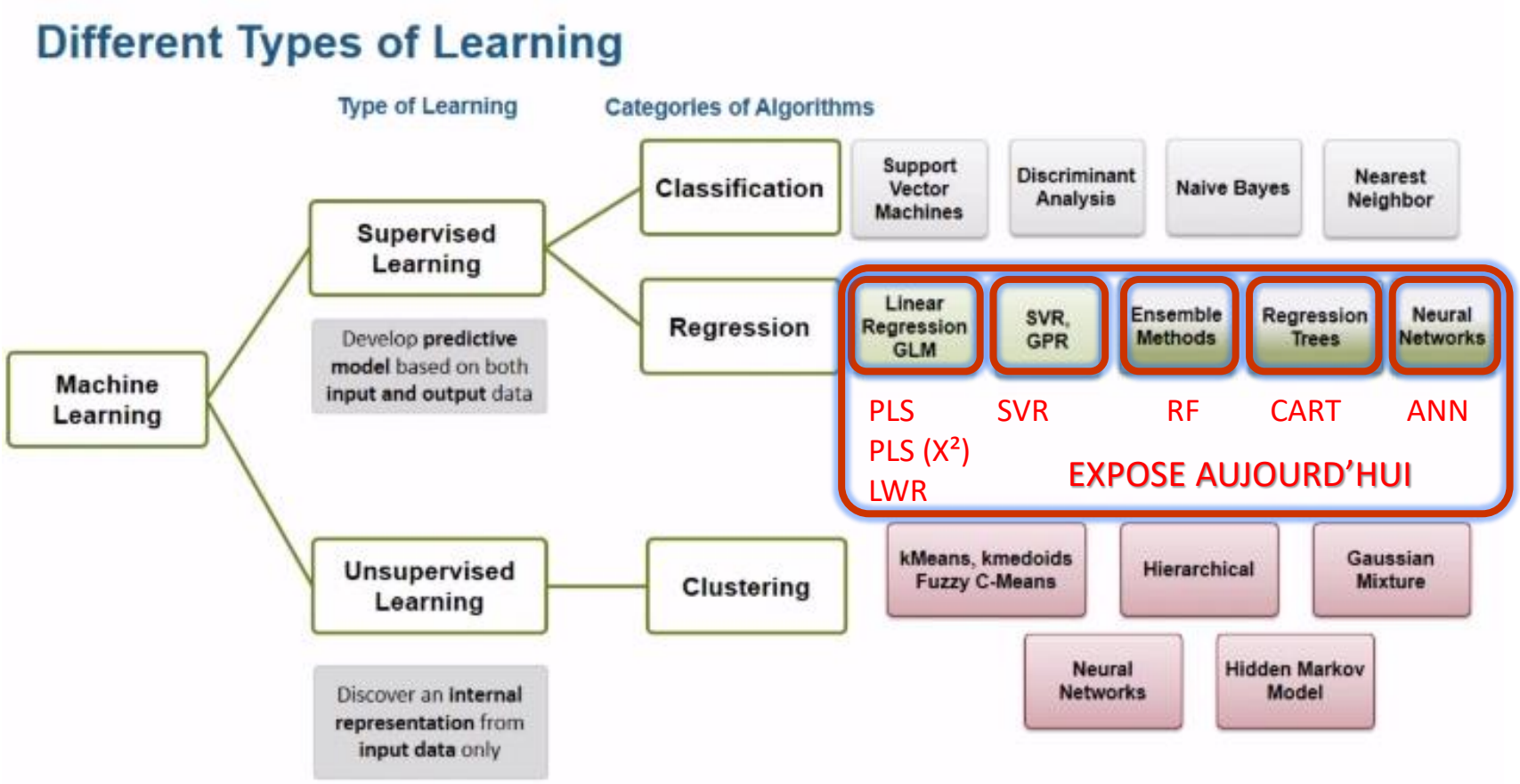
Introduction – Le « Machine Learning » : Qu'es aquò ?



Source : MathWorks

ondalys

Introduction – Le « Machine Learning » : Qu'es aquò ?



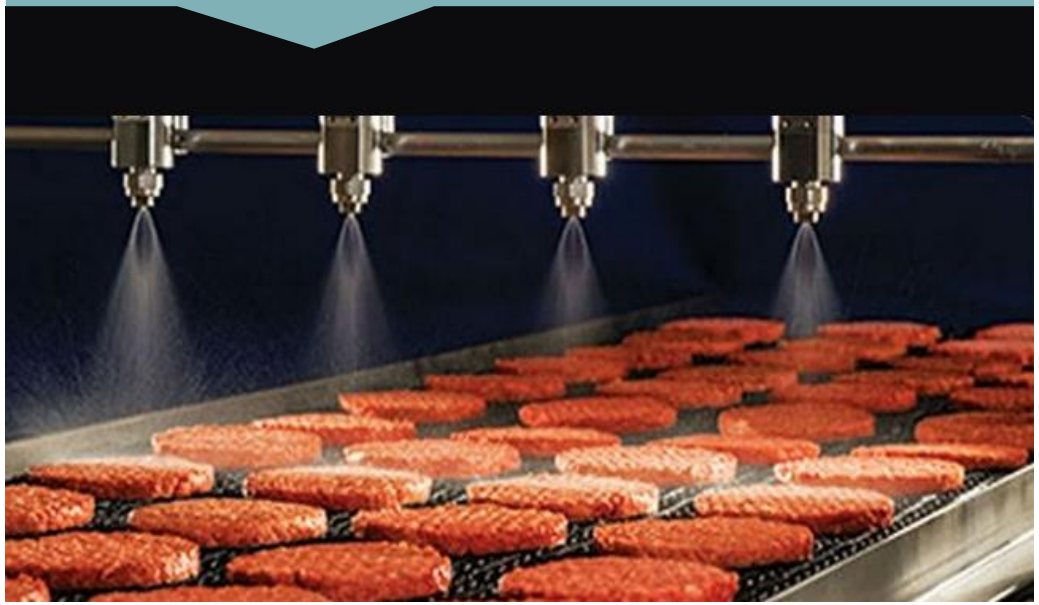
Source : MathWorks

ondalys

Description du cas concret - Proof of Concept (PoC)

Composition chimique de la viande

- Prédictions les plus précises
- Prédictions les plus robustes



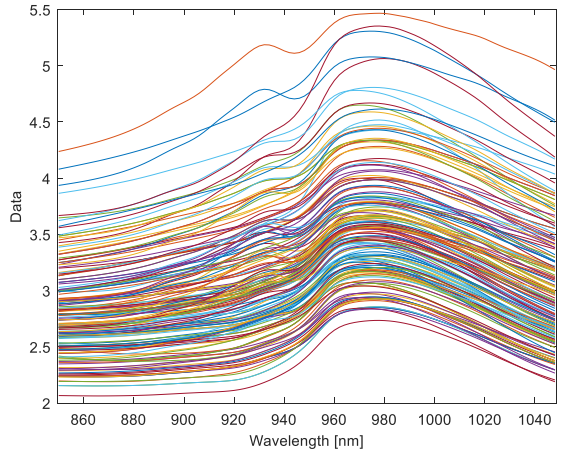
Source : http://lib.stat.cmu.edu/datasets/tecator

Echantillons de viande hachée

- 172 échantillons d'étalonnage
- 43 échantillons de validation
- Composition : humidité, protéines et matières grasses

Spectroscopie proche infrarouge (NIR)

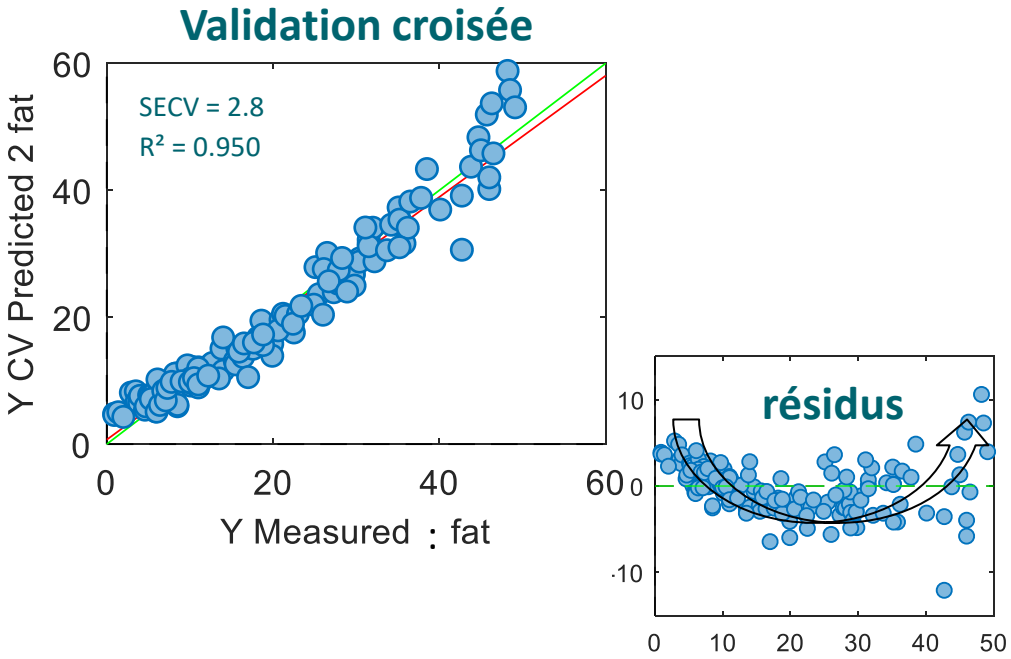
- FOSS Tecator Infratec Food and Feed Analyzer
- Mesures en Transmittance
- Longueurs d'onde : [850-1050nm]

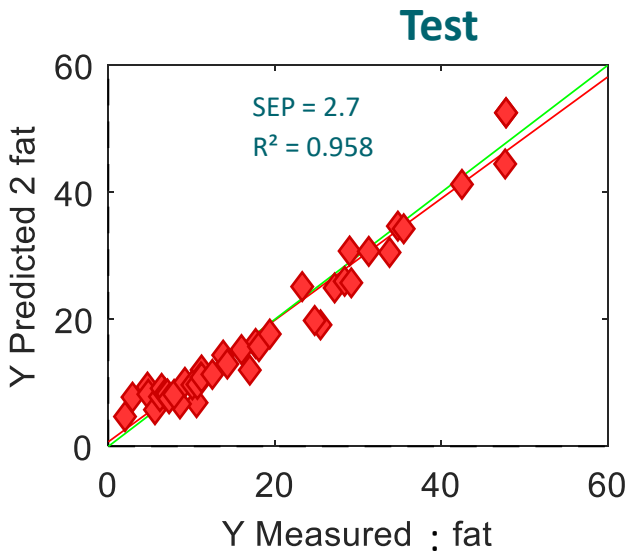


ondalys

1. Résultats PLS

- Non-linéarité résiduelle bien visible
- Performances insuffisantes





Méthode	SEC	SECV	SEP
1. PLS	2.6	2.8	2.7

Software : PLS_Toolbox

Problématique des relations non-linéaires

Parfois, la relation entre les spectres et la propriété à prédire n'est pas linéaire

- La méthode PLS gère bien de petites non-linéarités,
  mais elle est insuffisante dans les cas plus complexes

- 2 solutions :

  (i) Linéariser la relation entre les spectres et le paramètre à prédire

  Ou

  (ii) Utiliser des méthodes de Machine Learning capables de modéliser la non-linéarité

Transformation des variables

Une solution simple consiste à transformer les variables

- Transformer y : log(y), sqrt(y)
  - ☹ Souvent la transformation inverse fait ressortir les erreurs

ou

- Transformer X : ajout de carrés, de termes croisés, …
  - ☹ augmente fortement le nombre de variables
  - ☹ certaines combinaisons ne sont pas prédictives et ajoutent du bruit

- Pour des données spectroscopiques (X) :
  - Faire une ACP et sélectionner k scores
  - Appliquer la transformation sur les scores
  - Sélection de variables les plus prédictives

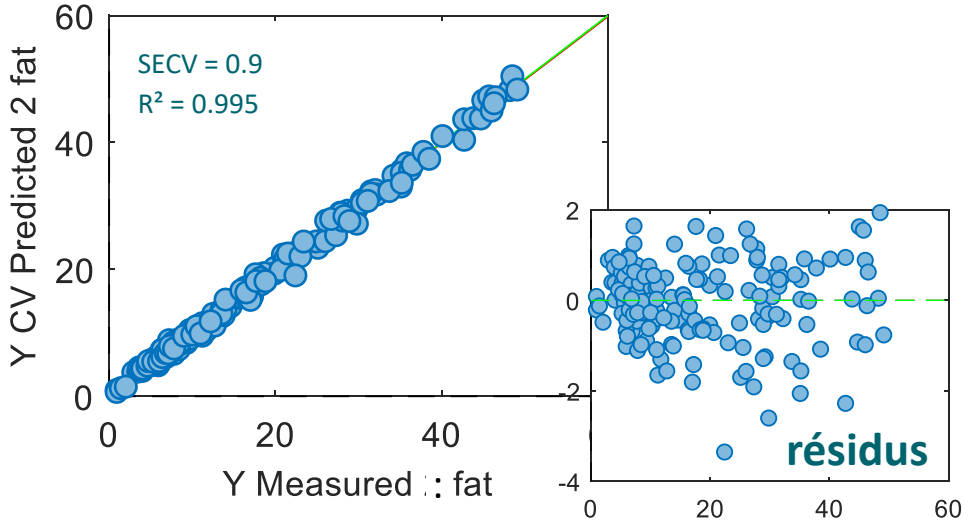
Il est nécessaire de centrer et réduire les scores, surtout après transformation

2. Résultats PLS avec transformation des variables X

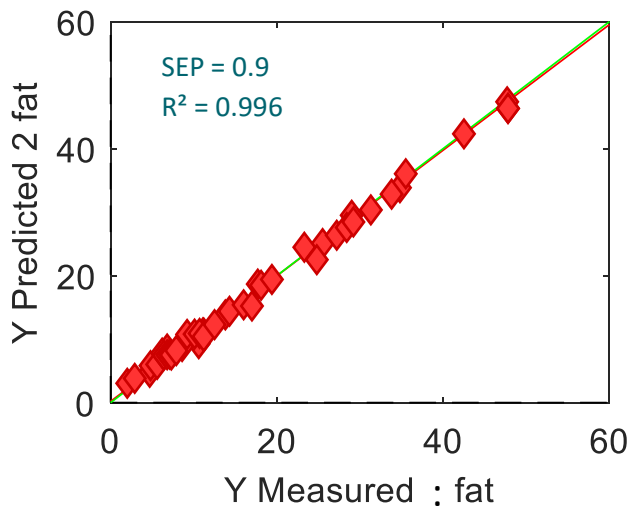
Sur [10 scores ACP + termes au carré + termes croisés] et sélection de 14 variables

- ➔ La non-linéarité a été prise en compte
- ➔ Les performances sont très nettement améliorées

Validation croisée



Test



Méthode	SEC	SECV	SEP
1. PLS	2.6	2.8	2.7
2. PLS sur X²	0.7	0.9	0.9

Software : PLS Toolbox

EIGENVECTOR RESEARCH INCORPORATED

ondalys

Sommaire

- Introduction – Le « Machine Learning » : Qu’es aquò ?
- Description du cas concret
- Comparaison des résultats

3. Modèles locaux

- Les modèles locaux ont plusieurs utilités
  - Regrouper dans une même base des spectres d'échantillons très divers
  - Réaliser des modèles « linéaires par morceaux » en fonction de la gamme du paramètre à prédire

  ➔ Permet de gérer les non-linéarités

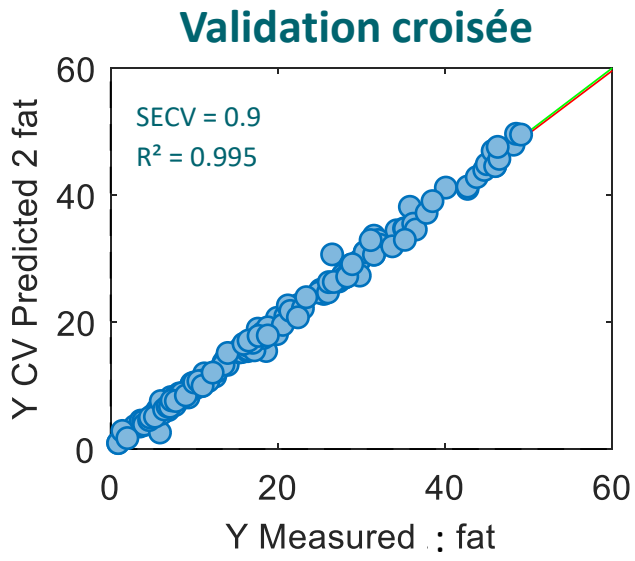
- Il existe divers algorithmes de régression locale (LOCAL, LWR, etc.)
- Principe de la Locally Weigthed Regression (LWR) :
  - 1 modèle différent est établi pour chaque échantillon à prédire
  - Sélection de k échantillons les plus proches spectralement (X) de cet échantillon
  - Possibilité de prendre en compte la distance à y en utilisant itérativement la prédiction
  - Réalisation d'un modèle linéaire (PLS) sur ces k échantillons

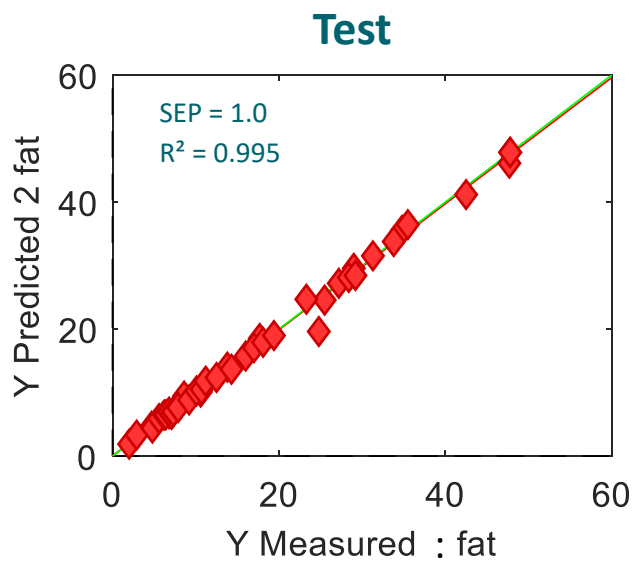
Software : PLS Toolbox EIGENVECTOR RESEARCH INCORPORATED

ondalys

3. Modèles locaux : LWR

- La non-linéarité a été prise en compte
- Les performances sont proches du modèle PLS sur variables transformées





Méthode	SEC	SECV	SEP
1. PLS	2.6	2.8	2.7
2. PLS sur X^2	0.7	0.9	0.9
3. LWR	0.5	0.9	1.0

Software : PLS Toolbox

Sommaire

- Introduction - Le « Machine Learning » : Qu’es aquò ?
- Description du cas concret
- Comparaison des résultats

4. Support Vector Machines Regression (SVR)

Principe des Support Vector Machines (SVM)

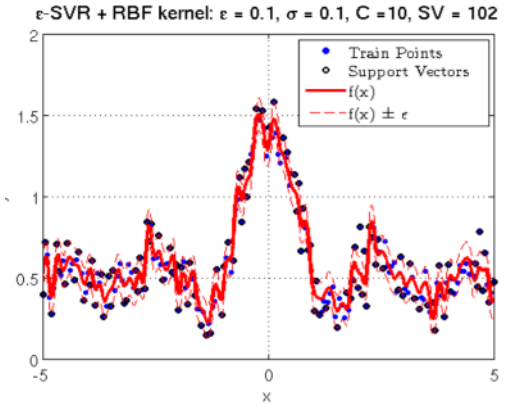
- Méthodes basées sur des frontières (SVM : Séparateurs à Vaste Marge)
- Seuls quelques échantillons participent à la construction du modèle final
  = échantillons définissant les frontières = vecteurs supports

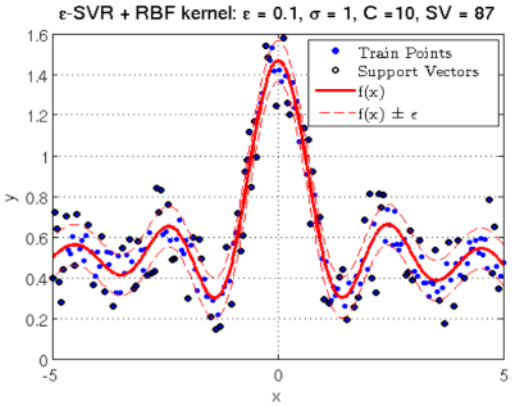
Gestion des non-linéarités

- Les non-linéarités peuvent être modélisées grâce une transformation des données via un noyau (kernel) exprimant la similarité entre échantillons
  - Kernel Gaussien : $K(x_i, x_j) = e^{-\gamma \|x_i x_j\|^2}$

Risque de sur-apprentissage : comment l'éviter ?

- Différents paramètres à ajuster : ε, σ, γ
- En spectroscopie :
  - Prétraitements ++ 
  - Compression des données

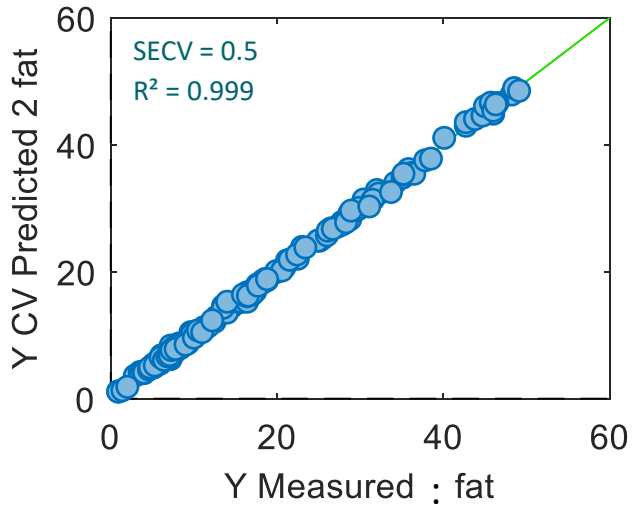




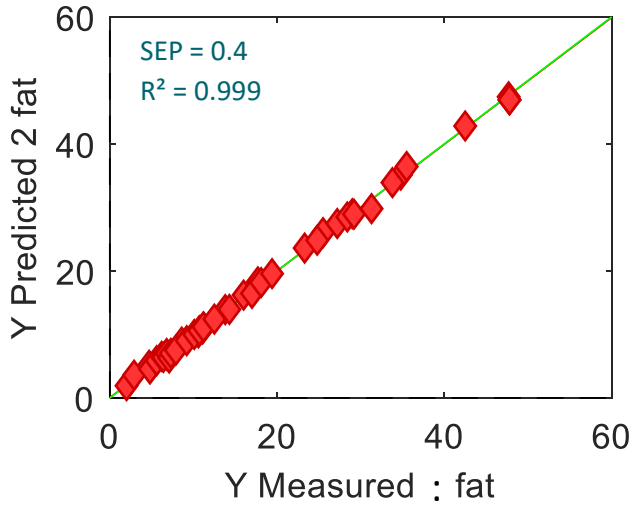
4. Résultats des SVR

- La non-linéarité a été modélisée
- Les performances sont améliorées par rapport aux modèles précédents

Validation croisée



Test



Méthode	SEC	SECV	SEP
1. PLS	2.6	2.8	2.7
2. PLS sur X^2	0.7	0.9	0.9
3. LWR	0.5	0.9	1.0
4. SVR	0.4	0.5	0.5

Software : PLS Toolbox

EIGENVECTOR RESEARCH INCORPORATED

ondalys

Sommaire

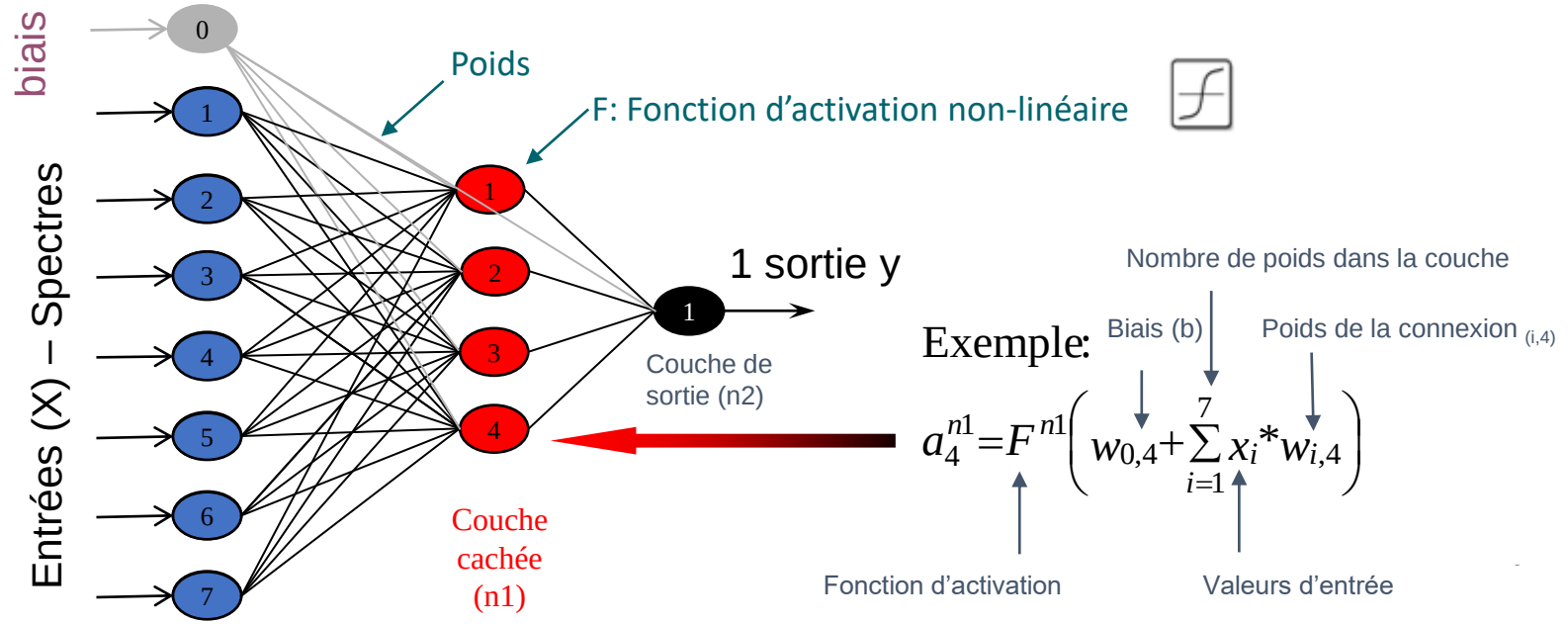
- Introduction - Le « Machine Learning » : Qu’es aquò ?
- Description du cas concret
- Comparaison des résultats

1. PLS
2. PLS avec transformation des données
3. Modèle local (LWR)
4. Support Vector Machine Regression (SVR)
5. **Artificial Neural Networks (ANN)**
6. Classification and Regression Tree (CART) / Random Forests

5. Réseaux de Neurones Artificiels (*Artificial Neural Networks* : ANN)

➔ « *Shallow networks* » vs « *Deep learning* »

- Principe du « perceptron multi-couches » (*Multilayer Perpectron* MLP) :
  - Basé sur la biologie (système nerveux) : réseau de neurones connectés en couches
  - Apprentissage des poids (*weights*) entre les neurones connectés
  - Fonctions d'activation non linéaires pour les neurones de la couche cachée



Risque important de surapprentissage. Comment l'éviter ?

- Réduire la taille du réseau / le nombre de poids à entrainer
  - Couche d'entrée : compresser les spectres (scores d'ACP ou PLS par ex)
  - 1 seule couche cachée généralement suffisante

➔ Principe de parcimonie : plus le réseau est simple, plus il est robuste

- *Stopped-learning* : limiter le nombre d'itérations d'apprentissage grâce à un jeu de validation externe

- Autres techniques : apprentissage par régularisation (*weight decay – Bayesian Regularization training* : minimisation de l'erreur et de l'amplitude des poids), élagage de neurones (*pruning*), introduction de bruit, etc

5. Résultats des ANN

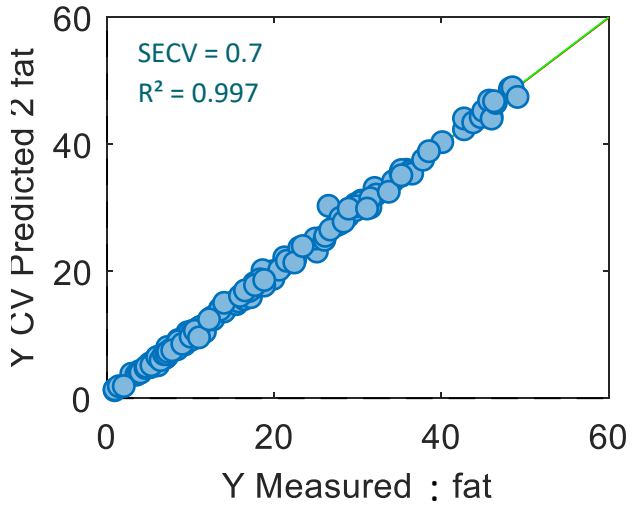
Structure du réseau de neurones :

- Inputs : 10 scores d'ACP et 3 neurones dans la couche cachée
- Taille : (10+1 biais)*3 + (3+1 biais) * 1 = 37 poids à modéliser avec 172 échantillons d'apprentissage

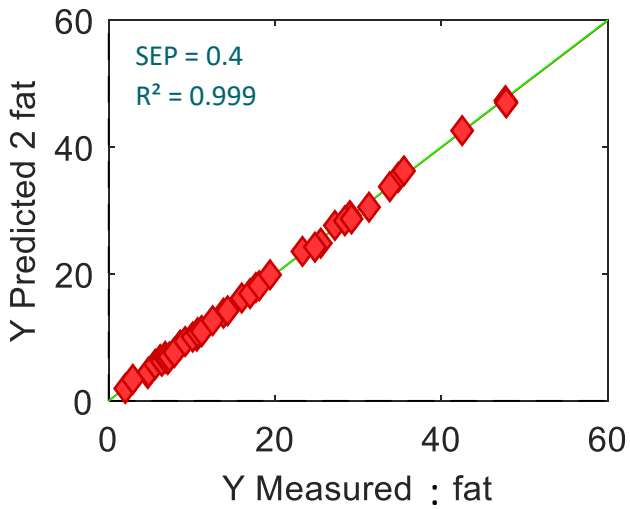
➔ 5 échantillons / poids à entrainer. **Attention au surapprentissage !**

- La non-linéarité est modélisée
- Les performances sont similaires aux SVM

Validation croisée



Test



Méthode	SEC	SECV	SEP
1. PLS	2.6	2.8	2.7
2. PLS sur X^2	0.7	0.9	0.9
3. LWR	0.5	0.9	1.0
4. SVR	0.4	0.5	0.5
5. ANN	0.5	0.6	0.4

Software : PLS_Toolbox EIGENVECTOR RESEARCH INCORPORATED ou Deep Learning Toolbox MathWorks

Sommaire

- Introduction - Le « Machine Learning » : Qu’es aquò ?
- Description du cas concret
- Comparaison des résultats

1. PLS
2. PLS avec transformation des données
3. Modèle local (LWR)
4. Support Vector Machine Regression (SVR)
5. Artificial Neural Networks (ANN)
6. **Classification and Regression Tree (CART) / Random Forests**

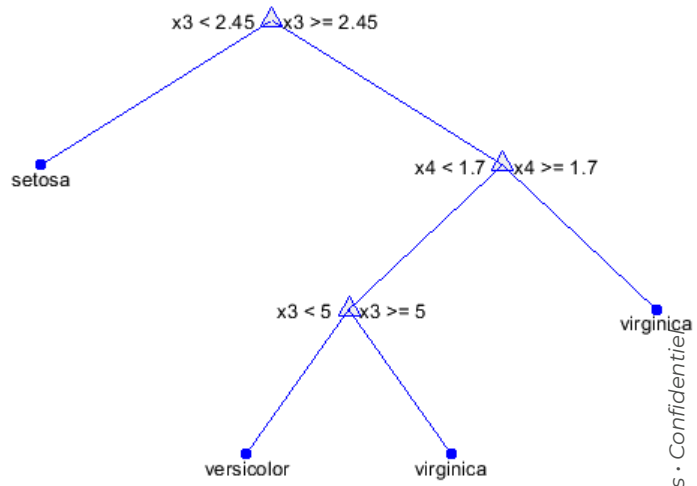
6. Classification and Regression Tree (CART) / Random Forests

- ## Principe de Classification and Regression Tree (CART) :
  - Arbre fait de séparations dichotomiques successives des échantillons
    - A chaque nœud, 1 variable (1 longueur d'onde) est choisie pour séparer en 2 les données en fonction d'un seuil

    ➔ ☺ Méthode simple et modèle interprétable
  - Un arbre complet est réalisé = 1 échantillon par feuille
  - Elagage des feuilles pour éviter le **fort risque de sur-apprentissage**
  - Application de CART : un nouvel échantillon est soumis à l'arbre
    - Valeur prédite = moyenne des échantillons de la feuille terminale

    ➔ ☹ prédictions en « escalier » = moyenne d'une feuille

    ➔ ☹ Minoration des prédictions dans les valeurs hautes

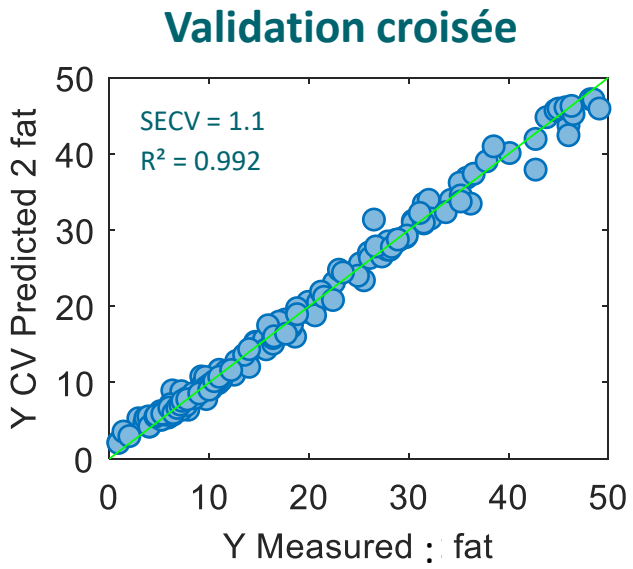


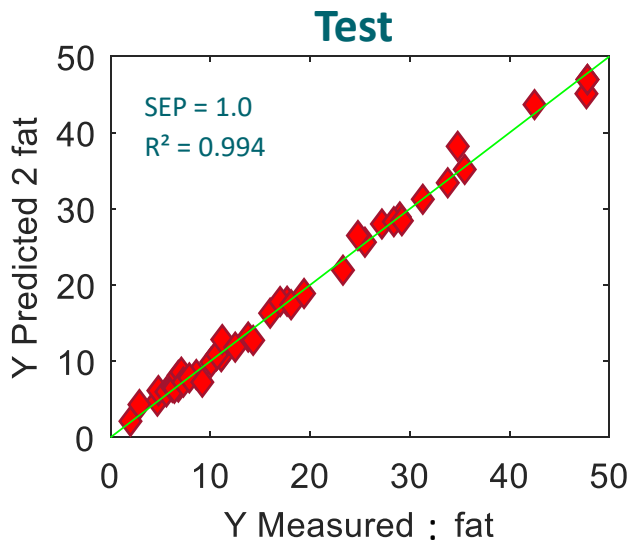
- ## « Ensemble methods » : Forêts aléatoires / *Random Forests*
  - Créer plusieurs prédicteurs CART « faibles »
  - Les combiner pour augmenter la performance et la robustesse des modèles (moins surappris)

6. Résultat des Random Forests

Sur une sélection de longueurs d'onde

- La non-linéarité a pu être partiellement modélisée
- Les performances sont moins élevées que pour les SVM et ANN
- Elles sont similaires à la transformation des variables et à la méthode locale





Méthode	SEC	SECV	SEP
1. PLS	2.6	2.8	2.7
2. PLS sur X^2	0.7	0.9	0.9
3. LWR	0.5	0.9	1.0
4. SVR	0.4	0.5	0.5
5. ANN	0.5	0.6	0.4
6. RF	0.6	1.1	1.0

Software : Statistics and Machine Learning Toolbox MathWorks

ondalys

Take-home message !
Si vous voulez appliquer du ML à la spectroscopie …

		Etalonnage				Validation croisée				Test			
Méthode	Nb LVs	SEC	SEC (%)	R^2	RPD	SECV	SECV (%)	R^2	RPD	SEP	SEP (%)	R^2	RPD
1. PLS	5	2.6	18.5%	0.958	5	2.8	20.2%	0.950	4	2.7	19.0%	0.958	5
2. PLS sur X transformé	5	0.7	5.3%	0.997	17	0.9	6.2%	0.995	15	0.9	6.6%	0.996	14
3. LWR	3	0.5	3.6%	0.998	25	0.9	6.5%	0.995	14	1.0	6.9%	0.995	13
4. SVR	-	0.4	2.6%	0.999	35	0.5	3.6%	0.998	25	0.5	3.4%	0.999	27
5. ANN	-	0.5	3.9%	0.998	23	0.6	4.5%	0.998	20	0.4	2.7%	0.999	34
6. RF	-	0.6	4.3%	0.998	21	1.1	8.2%	0.992	11	1.0	7.5%	0.994	12

➔ En présence de non-linéarités

Méthode	Gestion non-linéarité	Performance	Complexité de mise en œuvre	Risques de sur-apprentissage
1. PLS	-	-	-	-
2. PLS sur X transformé	+	+	-	+
3. LWR	+	+	+	+
4. SVR	++	++	++	++
5. ANN	++	++	+++	+++
6. RF	+	+	+	++

ondalys

Merci pour votre attention...

WITH ondalys, MAKE SENSE OF YOUR DATA !

18 Years

FOUNDED IN 2003

- **French Leaders** in chemometrics
- Located in Montpellier (34), France

Core business

- Consulting / R&D
  > 200 projects
- Training
  > 1000 trainees
- Software retail
  > 4 main chemometrics software

Multidisciplinary Team

- Diverse application Backgrounds
- Instrumentation / Data
- Machine Learning

Customers

- 50% Pharma-biotech
  SANOFI, gsk, SERVIER
- 25% Chemistry
  TOTAL, SOLVAY, ARKEMA
- 15% Food/Agri
  VALRHONA, MOËT & CHANDON, Bonduelle
- 10% Optical Equipment
  SARTORIUS, bertin TECHNOLOGIES, envea

Chemometrics - Machine learning

- Data mining
- Spectroscopic calibrations
- Model robustness
- Sensor fusion
- Process monitoring (MSPC, BSPC)
- Design of Experiments (DoE / QbD)

Software partners

EIGENVECTOR RESEARCH INCORPORATED
MathWorks
StatEase
aspentech
SARTORIUS