université PARIS-SACLAY

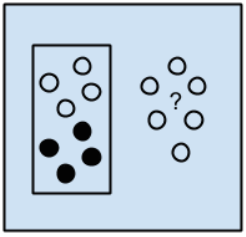
LE MACHINE-LEARNING EN PRATIQUE

Vincent Guigue
vincent.guigue@agroparistech.fr

Introduction

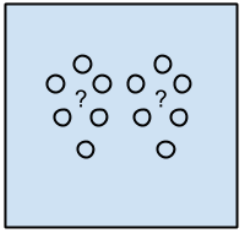
Différents cadres de machine learning

Supervisé



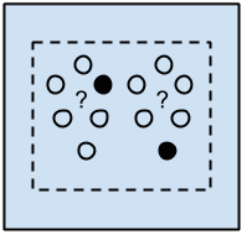
Supervised Learning Algorithms

Non-supervisé



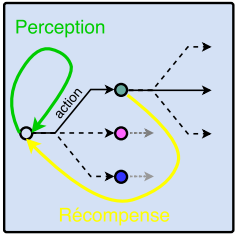
Unsupervised Learning Algorithms

Semi-supervisé



Semi-supervised Learning Algorithms

Renforcement

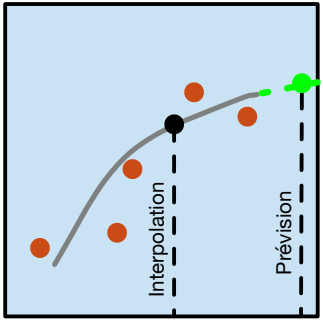


- Différents algorithmes... ... et différentes évaluations
- Différentes **données**, différents **coûts**...

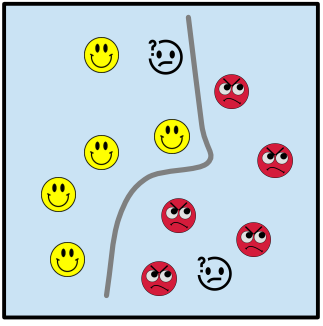
Et une nouvelle donne avec *Amazon Mechanical Turk*

Grande familles de problématiques supervisées

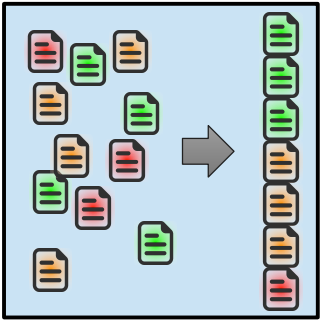
Régression



Classification

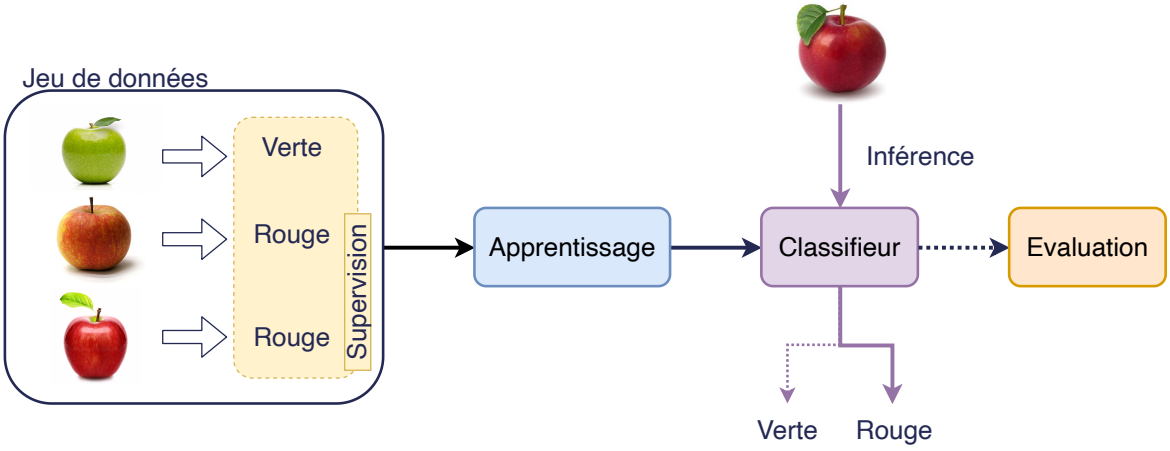


Ordonnancement



Chaine de traitement

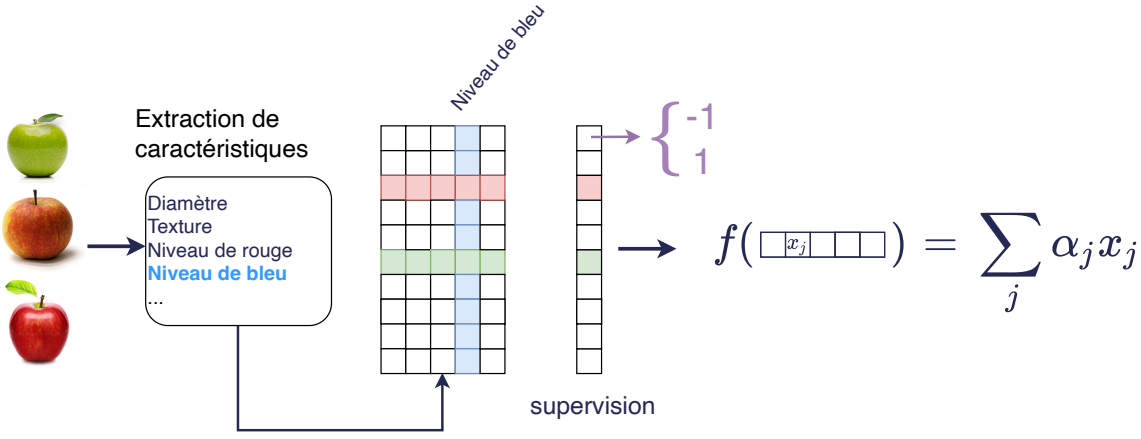
Identifier les entrées / sorties + évaluation



... En version abstraite

Chaine de traitement

En plus concret :



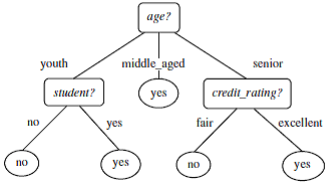
- Sélection des bonnes colonnes
- Ajout de colonnes intéressantes (calculs, sources de données externes, ...)

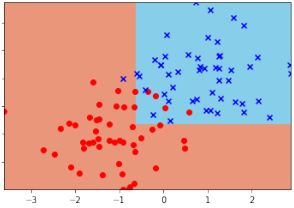
Classes de modèles

Modèles de ML : références historiques

Arbre de décision : entre IA symbolique & apprentissage statistique

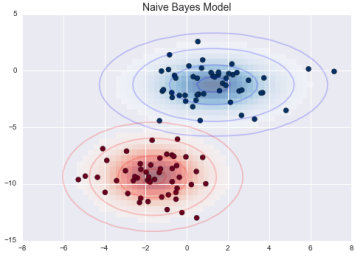
- Ensemble de règles
- Interprétable
  (selon la profondeur)
- Apprenable
  (sur critère entropique)

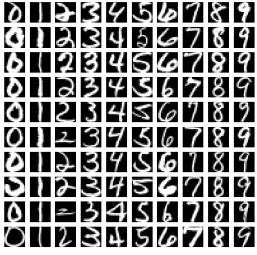




Modélisation bayesienne

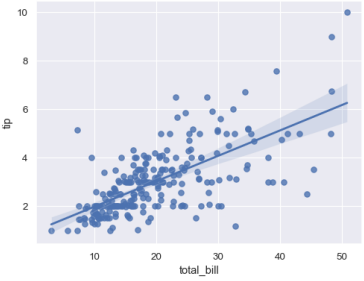
- Lois de probabilité
- Max. de vraisemblance
- Naive Bayes
- A priori des experts

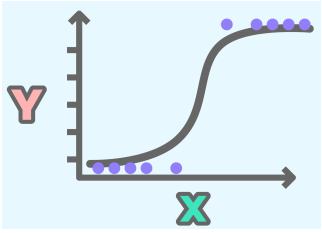




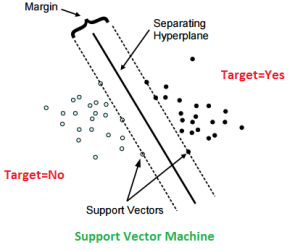
Modèles de ML : les bonnes affaires

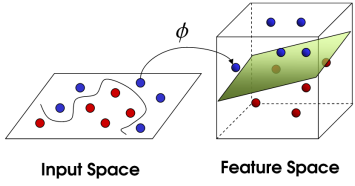
- **Modèles linéaires : Moindre carrés (MSE), régression logistique, ...**
  - Formulation simple & efficace
  - Classif, régression
  - Références très solides / modèle discriminant
  - Descente de gradient





- **SVM, noyaux et méthodes discriminantes**
  - Perceptron
  - Régularisation
  - SVM
  - Projection non linéaire

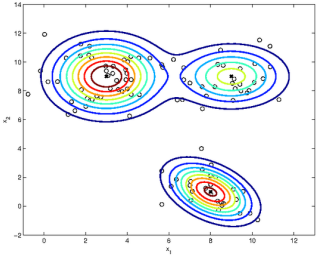


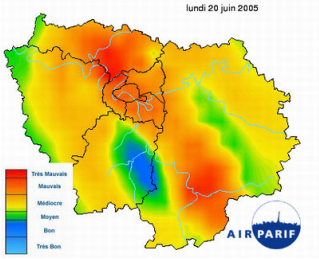


Modèles de ML : approches non-supervisées

Estimation de densité

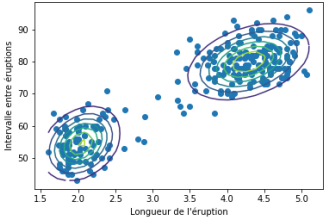
- Parzen
- Nadaraya-Watson
- Détour par les Knn
- EM

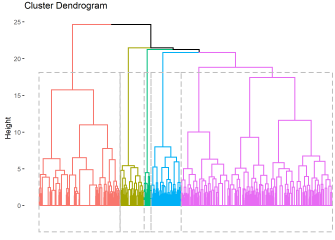




Clustering

- clustering hierarchique
- k-means / C-EM
- Clustering spectral
- A Priori

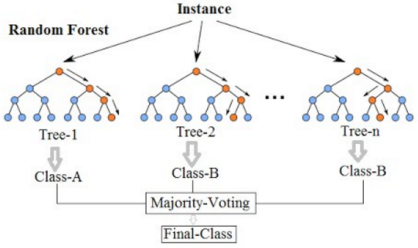




Modèles de ML : l'état de l'art

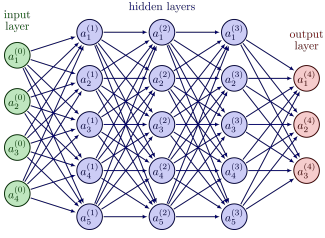
- **Approches ensemblistes**

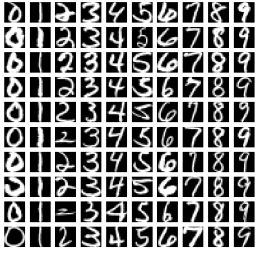
- Bagging
- Boosting
- Forêt, forêt aléatoire
- XGBoost



- **Réseaux de neurones (⇒ `pytorch`)**

- Perceptron
- Réseaux de neurones
- Rétropropagation du gradient
- Différentes architecture



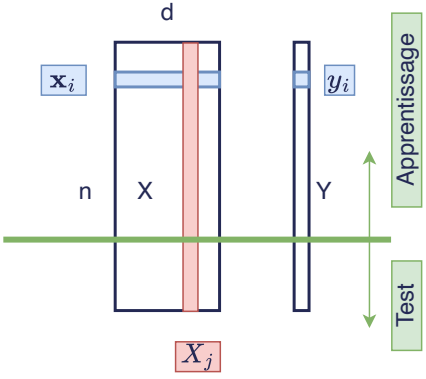


Evaluation

Evaluation du modèle / Sélection de modèle

!! L'évaluation est aussi importante que l'apprentissage !!

- Evaluer sur les données d'apprentissage (=qui ont servi à régler les paramètres)
  $\Rightarrow$ **Tricherie, surestimation des performances**
- Evaluer sur des données vierges = OK



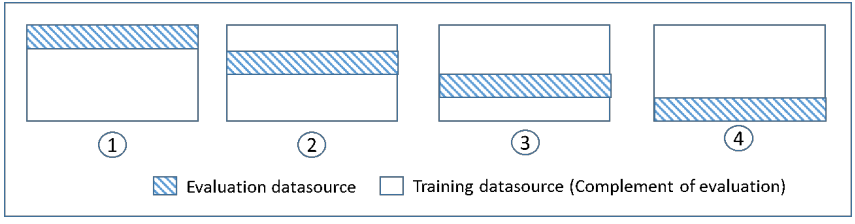
Problème de la répartition entre apprentissage et test

- La validation croisée

Evaluation du modèle / Sélection de modèle

!! L'évaluation est aussi importante que l'apprentissage !!

- Evaluer sur les données d'apprentissage (=qui ont servi à régler les paramètres)
  $\Rightarrow$ **Tricherie, surestimation des performances**
- Evaluer sur des données vierges = OK
- La validation croisée



Le cas déséquilibré

- Anomalies,
- Fraudes,
- Entités dans les textes
- ...