Introduction

Sabeur Elkosantini
Email: Sabeur.Elkosantini@fsegn.u-carthage.tn
Website: http://sabeur.elkosantini.me

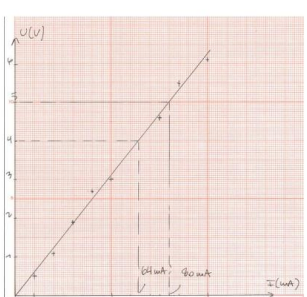
1

Machine Learning

- Machine Learning : Une discipline de l'informatique (intégrée dans l'intelligence artificielle) destinée à modéliser les relations entre les données.
- Dans un autre domaine, on parlerait de modélisation statistique, ou de méthodes de data mining, ou encore d'analyse de données.
- On retrouve bien – quelle que soit l'appellation utilisée – les grands thèmes du traitement statistique des données

2

Machine Learning



On sait que la relation entre intensité et tension est linéaire :

$$U = RI$$

On a collecté des données $(I_j, U_j)_{j \in \{1,\dots,n\}}$.

On cherche à estimer la résistance R inconnue.

3

La statistique et ML

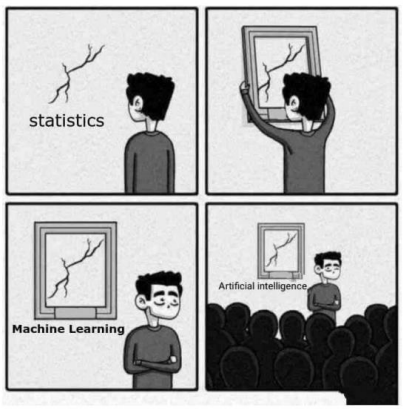
- On suppose la relation entre intensité et tension linéaire : $U = RI$.
  $\rightarrow$ C'est **un modèle statistique**.
- On peut calculer explicitement le paramètre $\hat{R}$ estimé à partir des données. On peut même parfois obtenir une mesure d'incertitude.
  $\rightarrow$ C'est **l'apprentissage statistique**.
- Une fois $\hat{R}$ calculé, on peut l'utiliser pour prédire la tension à une nouvelle intensité I_{new} :

$$U_{new} = \hat{R} I_{new}$$

$\rightarrow$ C'est **l'inférence statistique**.

4

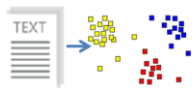
La statistique et ML



Pourquoi ML ?

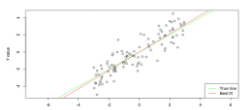
Classification

Anomaly Detection
Sequence labeling
…



Clustering

http://mbjoseph.github.io/2013/11/27/measure.html



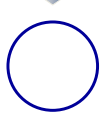
Regression

Applications de ML

- Champs d'application
  - Reconnaissance de :
    - ✓ Formes (Identification des signatures, partie d'images)
    - ✓ Caractères (Identification des signatures)
    - ✓ Signaux acoustiques (bruits sous-marins, …)
  - Classification
  - ✓ Mémoire Associative :
    - ✓ Restituer une donnée à partir d'informations incomplètes et/ou bruitées.
  - Aide à la décision (domaine médical, bancaire, management, …)

Applications de ML

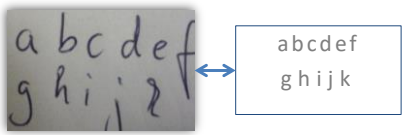
- Champs d'application



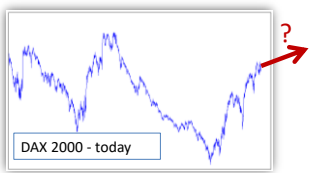
Applications de ML

☞ Champs d'application

- Pattern recognition:
  - Face recognition
  - Character recognition
  - Sound recognition



- Predicting the future
  - Stock market
  - Weather



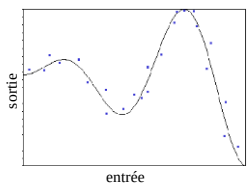
- Signal processing

Applications de ML

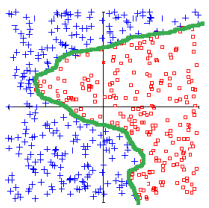
☞ Champs d'application

- Pattern recognition:

Régression
(approximation)



Classification
(y_i = « étiquettes »)



entrée = position point

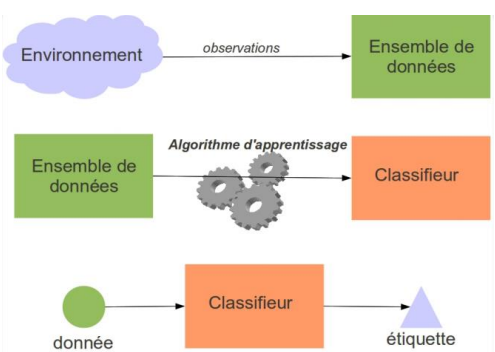
sortie désirée = classe (□ =-1,+=+1)

↓

Fonction étiquette=f(x) (et frontière de séparation)

points = exemples ➔ courbe = régression

Schéma global



Comment créer ces modèles ML ?

Apprentissage automatique

- Supervisé: Apprentissage avec des données labellisés

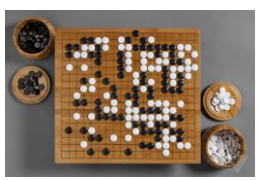
Exemple: classification des e-mails déjà étiquetés (spam ou non)

- Non supervisé: découvrez des modèles (des patterns) dans des données non étiquetées

Exemple: regrouper des documents similaires en fonction du texte

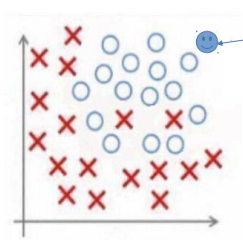
- Apprentissage par renforcement: apprenez à agir en fonction des retours / récompenses

Exemple: apprendre à jouer à Go



Apprentissage automatique

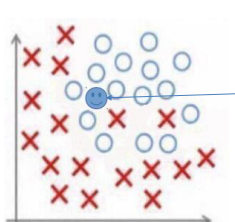
- Apprentissage Supervisé



Quelle classe?
Cercle ou croix ?

Apprentissage automatique

- Apprentissage Supervisé



Et pour cette donnée ?
Quelle classe?
Cercle ou croix ?

Apprentissage automatique

- Apprentissage Supervisé

Comment créer un modèle en utilisant l'apprentissage supervisé

Apprentissage automatique

- Apprentissage Supervisé: exemple de régression linéaire

Régression linéaire en dimension d :

- Modèle : $y = \langle w, x \rangle + \varepsilon$, où $x \in \mathbb{R}^d, y \in \mathbb{R}$ et $w \in \mathbb{R}^d$.
- Données d'apprentissage : $(x_i, y_i)_{i=1,\dots,n} \in (\mathbb{R}^d \times \mathbb{R})^n$.
- Minimisation de l'erreur d'apprentissage :

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2$$

- Solution : $\hat{w} = (X^\top X)^{-1} X^\top y$

où $X = \begin{pmatrix} x_1^1 & \dots & x_1^d \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^d \end{pmatrix}$ et $y = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix}$

$\hookrightarrow$ **Algèbre linéaire**

Apprentissage automatique

- Apprentissage Supervisé: exemple de régression linéaire

Données d'apprentissage : $(X_i, y_i)_{i=1,\dots,n} \in (\mathcal{X} \times \mathbb{R})^n$.
Modèle : $y = f(x)$, pour un certain $f \in \mathcal{F}$.
Problème à résoudre pour l'apprentissage :

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega(f)$$

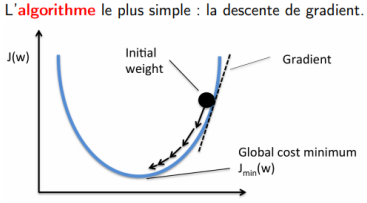
Optimisation Erreur Régularisation

Trouver un modèle qui fait peu d'erreurs, et le plus simple possible.

Apprentissage automatique

- Apprentissage Supervisé: exemple de régression linéaire

L'**algorithme** le plus simple : la descente de gradient.



Répéter jusqu'à convergence : $w \leftarrow w - \eta J'(w)$

Apprentissage automatique

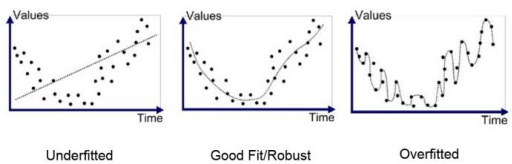
- Apprentissage Supervisé: exemple de régression linéaire

Autres algorithmes d'optimisations:

- Gradient descent
- Stochastic gradient descent
- Coordinate gradient descent
- Accelerated gradient descent
- Averaged gradient descent
- Subgradient descent
- Proximal gradient descent
- Conjugate gradient descent
- Conditional gradient descent
- Newton method
- Quasi Newton methods
- Alternative direction method of multipliers
- Douglas-Rachford
- ...

Sous-, Sur-, Correcte Généralisation

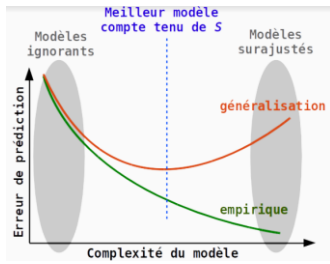
Choisir le bon niveau de complexité :



Sous-, Sur-, Correcte Généralisation

Erreur empirique : erreur sur l'échantillon d'apprentissage S

Erreur en généralisation : Erreur sur toutes les données (inconnue)

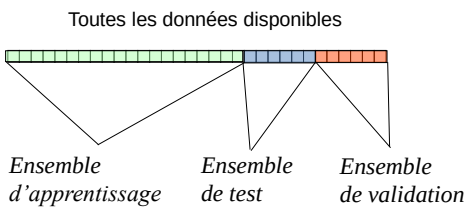


Plateformes pour le développement de ML

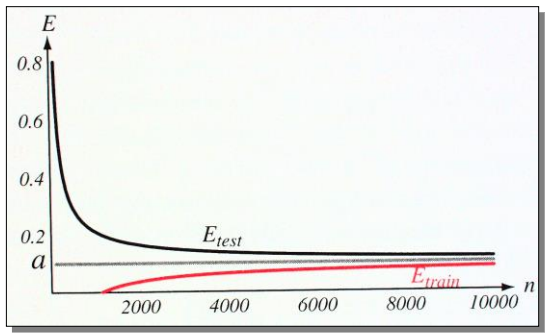
L'implémentation se fait majoritairement en **Python**, où la plupart des outils sont en *open source* et faciles d'utilisation.

Comment valider les modèles ?

Validation des modèles

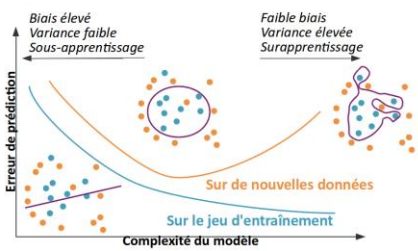


Validation des modèles



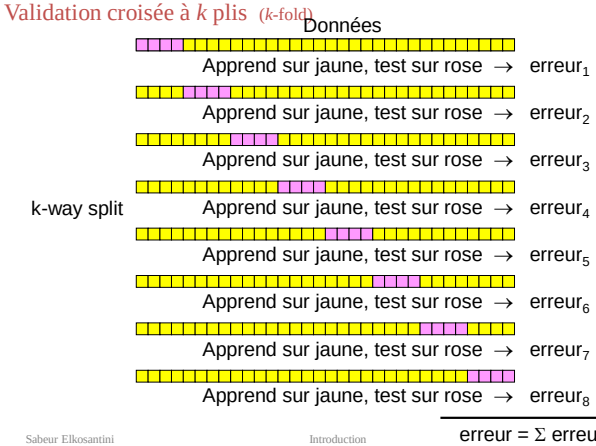
Validation des modèles

Le **sur-apprentissage** (*overfitting*)



Validation des modèles

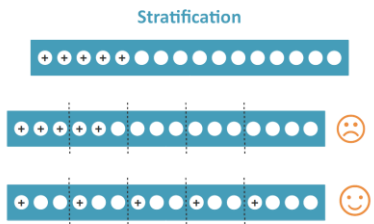
Validation croisée à *k* plis (*k*-fold)



Validation des modèles

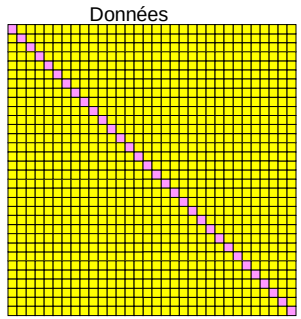
Validation croisée à *k* plis (*k*-fold)

- Dans le cas d'un problème de classification, il faut créer les k folds de sorte à ce qu'elles contiennent à peu près les mêmes proportions d'exemples de chaque classe que le jeu de données complet.
- On cherche à éviter qu'un jeu d'entraînement ne contienne que des exemples positifs et que le jeu de test correspondant ne contienne que des exemples négatifs, ce qui va affecter négativement la performance du modèle !



Validation des modèles

Procédure "leave-one-out"



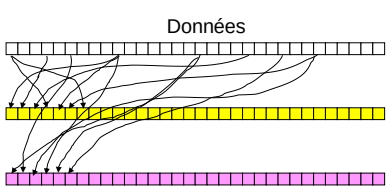
- Faible biais
- Haute variance
- Tend à sous-estimer l'erreur

Pourquoi ?

Nettement plus coûteux en calcul que la K validation croisée sans être meilleur

Validation des modèles

Le Bootstrap



- Apprend sur jaune, test sur rose → erreur
- Répéter et faire la moyenne

Conclusion

Quels outils pour le machine learning ?

- des statistiques ;
- de l'algèbre linéaire ;
- de l'optimisation ;
- de l'algorithmique ;
- du Python ;

Références

Eloise Berthier, « Une introduction à l'apprentissage automatique », support de cours, 2019.