

CNRS - IRSN

MÉTROLOGIE ET CAPTEURS

Data Science : Apprentissage automatique

PSN-RES/SEMIA/LSMA

Eric Chojnacki

Apprentissage automatique

Données : (facteurs , réponse) = ($X = x_i$, $Y = y_i$)

Objectif : prédire la réponse y'_i pour de nouveaux facteurs $X = x'_i$



Condition nécessaire : l'existence d'une généricité :

1. dans l'observation des facteurs,
2. dans la dépendance de la réponse aux facteurs

Condition implicite :

le modèle prédictif est difficile à écrire précisément

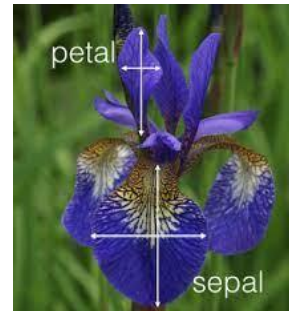
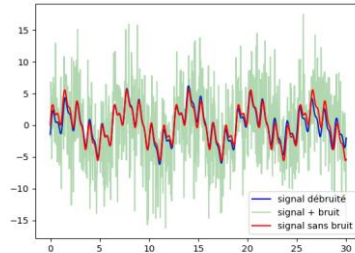
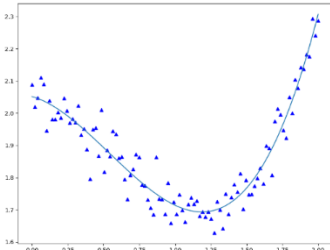
Apprentissage automatique : pour quelles applications ?

L'intérêt de l'apprentissage automatique dépend de la difficulté à identifier le modèle de dépendance existant entre les facteurs et la réponse

Intérêt faible

Intérêt fort

Difficulté à expliciter le modèle prédictif



Apprentissage automatique : généricité & cadre théorique

Génération des Y sachant X
 $Y = f(X) + \sigma$

Génération des X
 $P(X=x)$

Observation des données:
 x_i, y_i

Fonction d'erreur :
 $\text{Erreur} = E(h(x_i), y_i)$

Espace d'hypothèses H
dans lequel on recherche f
 $h=H(\text{paramètres})$

Algorithme d'apprentissage



Apprentissage / mémorisation

Jeu pile ou face : estimer la probabilité μ d'avoir pile à la précision ε en fonction du nombre de lancers N ayant observé v = nombre de piles/ N

Inégalité de Hoeffding : $P(|\mu - v| \leq \varepsilon) \leq 2 * \exp(-2 \frac{N}{\varepsilon^2})$

Si M joueurs $P(|\mu - v| \leq \varepsilon) \leq 2 * M * \exp(-2 \frac{N}{\varepsilon^2})$



La complexité du domaine d'hypothèse est comparable au nombre de joueurs.

Un modèle qui a un bon accord avec l'observation (= nombre de piles / nombre de lancers) ne signifie pas que le modèle peut se généraliser (si le nombre de joueurs est important, un joueur peut être arbitrairement chanceux)

Bonne généralisation : Apprendre le signal et limiter la mémorisation du bruit

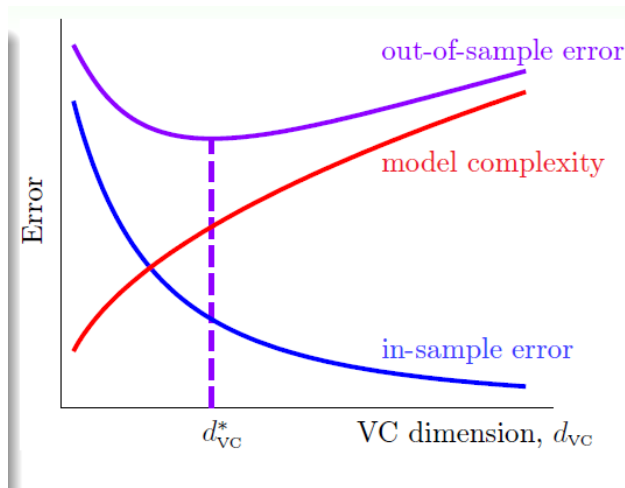


Théorie de la généralisation
Vapnik - Chervonenkis



Espace d'hypothèses H
dans lequel on recherche f
 $h(\text{paramètres})$

$$P(|\mu - \nu| \leq \varepsilon) \leq 2 * \mathbf{M} * \exp\left(-2 \frac{N}{\varepsilon^2}\right)$$



Bonne généralisation : « réduction » de la dimension de l'espace d'hypothèse

40 observations et 200 facteurs possibles

Sélection de 3 facteurs parmi 200

Régression **LASSO** vs **Ridge**

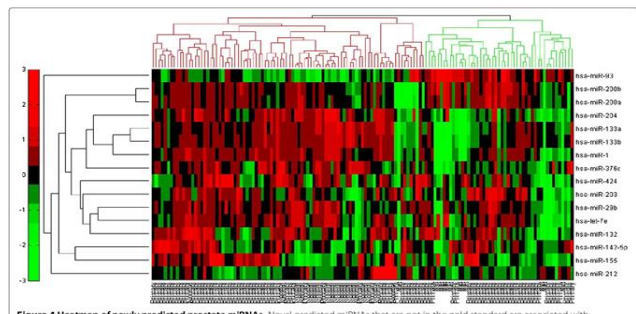
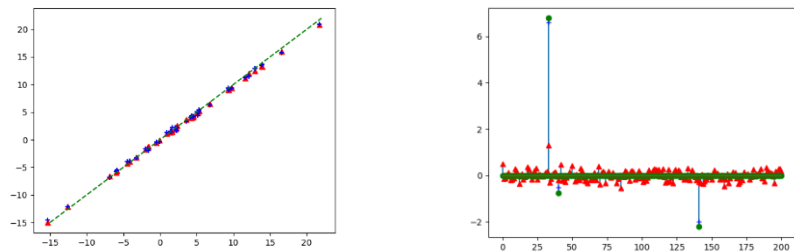
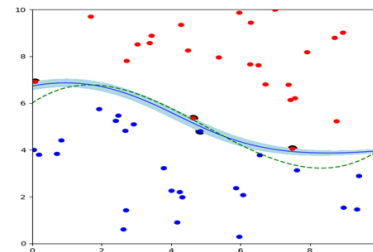


Figure 4 Heatmap of newly credited nonstate miRNAs. Novel predicted miRNAs that are not in the miRBase database are associated with

50 observations et 5 points de séparation

(recherchés dans un espace
éventuellement de dimension infini)



Etapes de l'apprentissage automatique

Données

Apprentissage/mémorisation

test

Cross-folding

1

2

3

4

5

Apprentissage

Validation

1

2

3

4

5

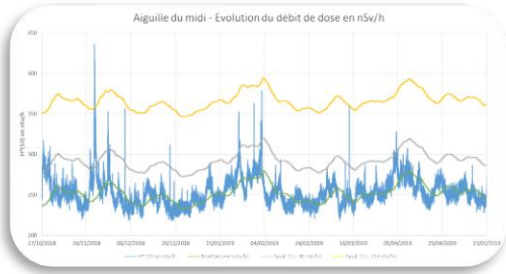
Évaluation finale

Sélection du meilleur modèle (E-out min)

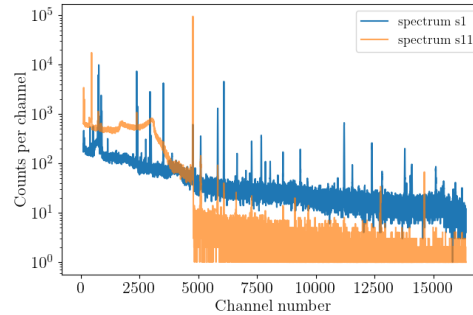
E-out

IRSN & Applications

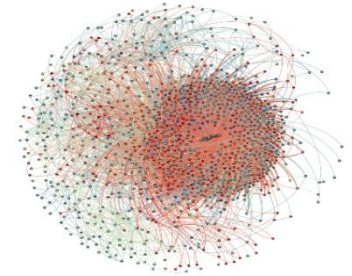
TELERAY & Détection d'anomalies



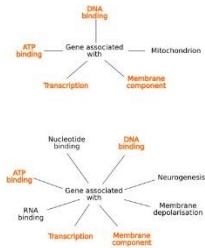
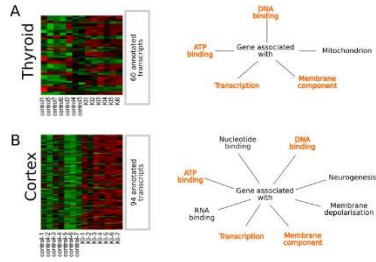
Spectrométrie & démixage spectral



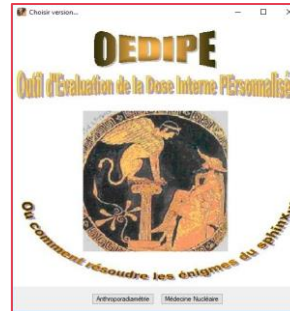
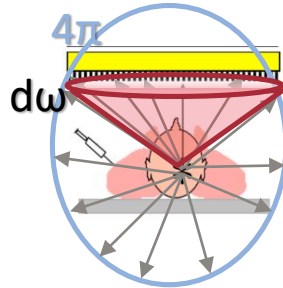
Composition isotopique



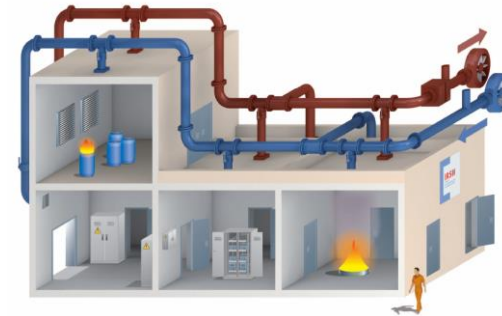
Biologie des systèmes



Dosimétrie interne



Incendie



Apprentissage automatique : synthèse

Apprendre est-ce possible ? Existence d'un modèle d'apprentissage

A-t-on les bons « features » ? La qualité et **la quantité des données** sont-elles suffisantes ?

Comment apprendre ? Détermination du 'meilleur' modèle.

Choix de l'espace H , de la fonction d'erreur et de l'algorithme d'apprentissage .

Quelle qualité de prédiction ? Performance attendue.

Erreur de généralisation (complexité & régularisation) , entraînement, validation et test

MERCI DE VOTRE ATTENTION

Références

E. Davesne, N. Blanchin, E. Chojnacki, L. Toure

Collective dosimetry to distinguish occupational exposure to natural uranium from alimentary uranium background in bioassay measurements
International Journal of Radiation Biology 90(11 2014

David P A Cohen , Mohamed Amine Benadjaoud, Phillipe Lestaevel , Dalila Lebsir, Marc Benderitter, Maâmar Soudi

Effects of repetitive Iodine thyroid blocking on the foetal brain and thyroid in rats: a systems biology approach
Sci Rep 10, 10839 (2020). <https://doi.org/10.1038/s41598-020-67564-8>

Jiaxin Xu, Jerome Bobin, Anne de de Vismes Ott, Christophe Bobin

.Spectral unmixing for activity estimation in Gamma-Ray Spectrometry 2019
<https://hal.archives-ouvertes.fr/hal-02060476>

Mirotta, P. Querre , J. Baccou, A. Gerbaud , T. Gerbaud,

Use of Artificial Intelligence algorithms for Hodoscope measurement interpretations, Nuclear Inst. and Methods in Physics Research, A, à paraître, 2021.

C Damon, A Liutkus, A Gramfort, S Essid.

Non-negative matrix factorization for single-channel EEG artifact rejection. *ICASSP*, 2013, Vancouver, Canada.

E Chojnacki, W Plumecocq, Laurent Audouin

An expert system based on a Bayesian network for safety analysis in nuclear area.
Fire Safety Journal, (105):28–40, feb 2019.

Benjamin Dechenaux

On the Use of Graph Theory to Interpret the Output Results from a Monte-Carlo Depletion Code
Nuclear Science and Engineering volume 195, 2021 - Issue 5