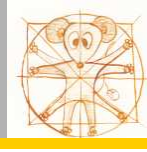
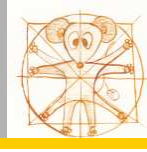


Apprendre en grande dimension: quel est le problème?

et débuts de solutions...

Christelle REYNES

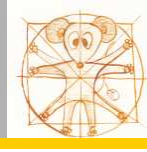
Université de Montpellier - Institut de Génomique Fonctionnelle
UMR5203 CNRS - U1191 INSERM



Qu'est-ce que l'intelligence artificielle ?

C'est l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence
(Encyclopédie Larousse)

Ici, nous parlerons d'apprentissage automatique (*machine learning*):
apprendre aux ordinateurs à apprendre



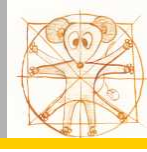
Qu'est-ce que l'intelligence artificielle ?

C'est l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence
(Encyclopédie Larousse)

Ici, nous parlerons d'apprentissage automatique (*machine learning*):
apprendre aux ordinateurs à apprendre

Examples:

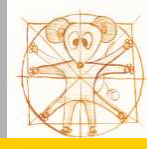
- quel est le trajet optimal pour relier un certain nombre de villes ?
- quel le meilleur traitement pour ce groupe de patients ?
- peut-on prédire la structure 3D de cette protéine ?
- cette molécule est-elle une bonne candidate pour cibler cette protéine ?
- ...



Qu'est-ce que le Big Data ?

désigne

- des **données** avec beaucoup de lignes et/ou de colonnes
- les **méthodes** qui permettent de les analyser.

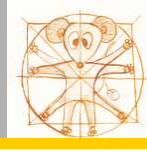


Qu'est-ce que le Big Data ?

désigne

- des **données** avec beaucoup de lignes et/ou de colonnes
- les **méthodes** qui permettent de les analyser.

Mais à partir de quand les données deviennent-elles BIG ?



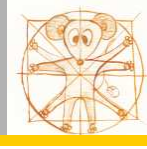
Qu'est-ce que le Big Data ?

désigne

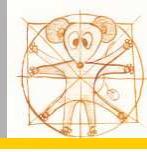
- des **données** avec beaucoup de lignes et/ou de colonnes
- les **méthodes** qui permettent de les analyser.

Mais à partir de quand les données deviennent-elles BIG ?

Quand l'un ou plusieurs des problèmes suivants se posent...



QUELS SONT LES PROBLEMES POUR APPRENDRE EN GRANDE DIMENSION ?

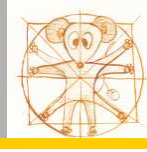


Acquisition et manipulation des données

Ex.: signature définie à partir de 1000 gènes

$=>$ nécessité de mesurer leur expression lors de chaque diagnostic

$=>$ potentiellement coûteux



Acquisition et manipulation des données

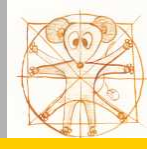
Ex.: signature définie à partir de 1000 gènes

$=>$ nécessité de mesurer leur expression lors de chaque diagnostic

$=>$ potentiellement coûteux

Acquérir de gros volumes de données

$=>$ problèmes de stockage



Acquisition et manipulation des données

Ex.: signature définie à partir de 1000 gènes

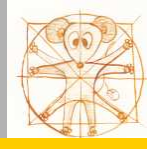
$=>$ nécessité de mesurer leur expression lors de chaque diagnostic

$=>$ potentiellement coûteux

Acquérir de gros volumes de données

$=>$ problèmes de stockage

Impossibilité d'utiliser certaines méthodes statistiques classiques



Acquisition et manipulation des données

Ex.: signature définie à partir de 1000 gènes

$=>$ nécessité de mesurer leur expression lors de chaque diagnostic

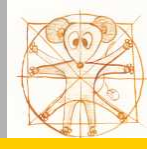
$=>$ potentiellement coûteux

Acquérir de gros volumes de données

$=>$ problèmes de stockage

Impossibilité d'utiliser certaines méthodes statistiques classiques

Temps de calculs potentiellement longs (y compris pour opérations simples)



Acquisition et manipulation des données

Ex.: signature définie à partir de 1000 gènes

$=>$ nécessité de mesurer leur expression lors de chaque diagnostic

$=>$ potentiellement coûteux

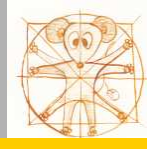
Acquérir de gros volumes de données

$=>$ problèmes de stockage

Impossibilité d'utiliser certaines méthodes statistiques classiques

Temps de calculs potentiellement longs (y compris pour opérations simples)

Problèmes éthiques

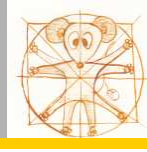


Problèmes d'interprétation

Construction de modèles incluant de nombreuses variables

$=>$ modèles difficiles voire impossibles à interpréter

$=>$ boîtes noires et rupture du sens



Problèmes d'interprétation

Construction de modèles incluant de nombreuses variables

$=>$ modèles difficiles voire impossibles à interpréter

$=>$ boîtes noires et rupture du sens

Example:

On a collecté l'expression de 1000 gènes lors d'une étude de pharco-génomique

Question :

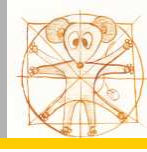
Quel est le meilleur traitement pour chaque patient ?

Solution possible :

utiliser le jeu de données complet pour construire un modèle

Problème :

SI il fonctionne, difficile de revenir à l'importance de chaque gène...

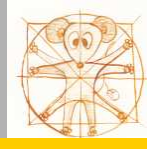


Difficultés pour l'apprentissage statistique

Le risque de sur-ajustement

En général, deux étapes en apprentissage statistique :

- construction d'un modèle (jeu d'apprentissage)
- validation du modèle sur d'autres données (jeu test)



Difficultés pour l'apprentissage statistique

Le risque de sur-ajustement

En général, deux étapes en apprentissage statistique :

- construction d'un modèle (jeu d'apprentissage)
- validation du modèle sur d'autres données (jeu test)

bon modèle = modèle qui apprend les généralités dans le jeu d'apprentissage et pas les détails spécifiques.

Difficultés pour l'apprentissage statistique

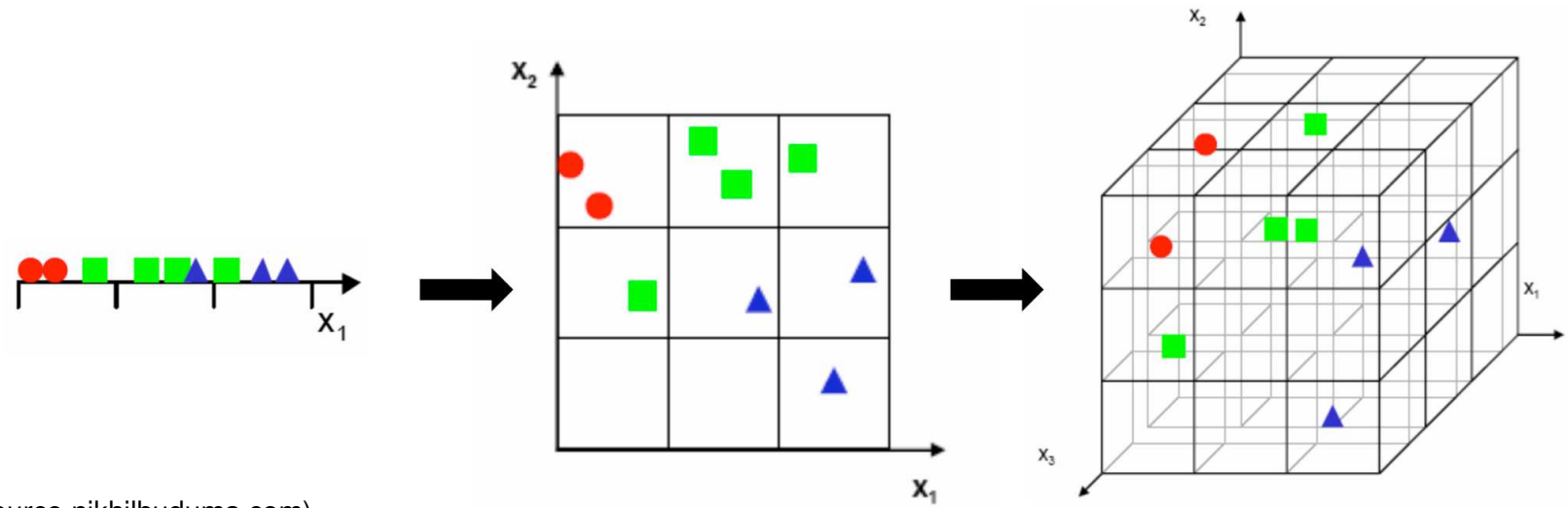
Le risque de sur-ajustement

En général, deux étapes en apprentissage statistique :

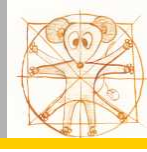
- construction d'un modèle (jeu d'apprentissage)
- validation du modèle sur d'autres données (jeu test)

bon modèle = modèle qui apprend les généralités dans le jeu d'apprentissage et pas les détails spécifiques.

Mais en très grande dimension : risque de sur-ajustement



(source nikhilbuduma.com)

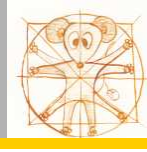


Difficultés pour l'apprentissage statistique

Le rapport signal/bruit

Peu de variables pertinentes noyées dans des milliers de variables sans intérêt

$=>$ très difficile de trouver un bon modèle



Difficultés pour l'apprentissage statistique

Le rapport signal/bruit

Peu de variables pertinentes noyées dans des milliers de variables sans intérêt

$\Rightarrow$ très difficile de trouver un bon modèle

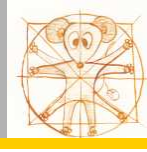
Illustration:

Objectif : prédire le volume à l'équilibre (importante propriété thérapeutique) pour $n = 128$ molécules décrites par $p = 1532$ descripteurs physico-chimiques (1D, 2D and 3D). Un problème difficile.

$p > n$

$\Rightarrow$ impossible d'utiliser la régression linéaire classique

$\Rightarrow$ utilisation de la régression PLS (Partial Least Squares)



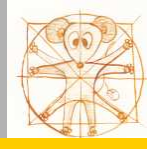
Difficultés pour l'apprentissage statistique

Le rapport signal/bruit

Illustration:

Application de PLS à différents (sous-)ensembles de variables

- jeu complet ($p = 1532$ descripteurs)
- élimination des fortes redondances ($=>$ $p = 524$)
- utilisation d'une méthode appropriée de sélection de variables ($=>$ $p = 25$)



Difficultés pour l'apprentissage statistique

Le rapport signal/bruit

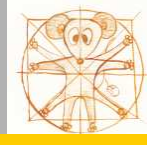
Illustration:

Application de PLS à différents (sous-)ensembles de variables

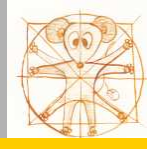
- jeu complet ($p = 1532$ descripteurs)
- élimination des fortes redondances ($=>$ $p = 524$)
- utilisation d'une méthode appropriée de sélection de variables ($=>$ $p = 25$)

Résultats

Data	R^2	R^2_{CV}	R^2_{test}	A	Phénomène
$p = 1532$	0.536	0.282	0.175	12	Bruit
$p = 524$	0.834	0.392	0.375	18	Sur-ajustement
$p = 25$	0.624	0.550	0.493	7	Bon modèle



UNE SOLUTION :
LA REDUCTION DE DIMENSION

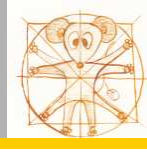


Extraction de caractéristiques

Définition : consiste à construire de nouvelles variables à partir des variables initiales.

Exemples de méthodes :

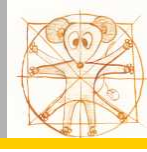
Objectif	Méthodes
Description/résumé	Analyse en Composantes Principales (ACP) Positionnement Multidimensionnel (MDS),...
Classification supervisée	Analyse Linéaire Discriminante (LDA), Forêts Aléatoires (RF), Séparateurs à Vastes Marges (SVM)
Régression	PLS, réseaux de neurones,...



Extraction de caractéristiques

Avantages :

- assez simples à mettre en œuvre
- possibilité d'interpréter les nouvelles caractéristiques en fonction des variables initiales



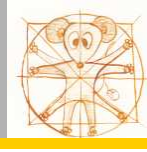
Extraction de caractéristiques

Avantages :

- assez simples à mettre en œuvre
- possibilité d'interpréter les nouvelles caractéristiques en fonction des variables initiales

Inconvénients :

- souvent *boîtes noires*
- choix de la dimension important et délicat
- risque important de sur-ajustement
- risque de ne pouvoir apprendre à cause d'un ratio signal/bruit défavorable

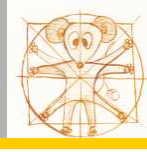


Sélection de caractéristiques (multivariée)

Définition : consiste à sélectionner un sous-ensemble des variables initiales pour construire un modèle.

Exemples de méthodes *embarquées* :

- arbres de classification et de régression (CART,...)
- classifieur naïfs bayésiens
- méthodes sparses



Sélection de caractéristiques (multivariée)

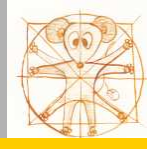
Définition : consiste à sélectionner un sous-ensemble des variables initiales pour construire un modèle.

Exemples de méthodes *embarquées* :

- arbres de classification et de régression (CART,...)
- classifieur naïfs bayésiens
- méthodes sparses

Exemples de méthodes *enveloppantes* :

- exploration exhaustive des sous-ensembles
- sélection forward/backward
- méta-heuristiques (algorithmes génétiques,...)



Sélection de caractéristiques (multivariée)

Définition : consiste à sélectionner un sous-ensemble des variables initiales pour construire un modèle.

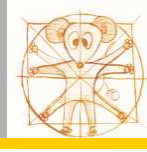
Exemples de méthodes *embarquées* :

- arbres de classification et de régression (CART,...)
- classifieur naïfs bayésiens
- méthodes sparses

Exemples de méthodes *enveloppantes* :

- exploration exhaustive des sous-ensembles
- sélection forward/backward
- méta-heuristiques (algorithmes génétiques,...)

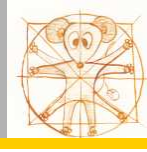
$=>$ sélection et extraction peuvent être combinées



Sélection de caractéristiques

Avantages :

- limitent le sur-ajustement (modèle plus parcimonieux)
- bonne interprétabilité du modèle
- choix de la dimension souvent automatisé à l'intérieur de la méthode
- permet d'aller chercher "l'aiguille dans la meule de foin"



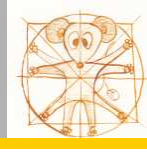
Sélection de caractéristiques

Avantages :

- limitent le sur-ajustement (modèle plus parcimonieux)
- bonne interprétabilité du modèle
- choix de la dimension souvent automatisé à l'intérieur de la méthode
- permet d'aller chercher "l'aiguille dans la meule de foin"

Inconvénients :

- si les méthodes simples ne fonctionnent pas, temps de calculs souvent longs et méthodes pouvant être plus difficiles à mettre en œuvre



CONCLUSION

Intelligence Artificelle et Big Data sont faits pour s'entendre !

MAIS nécessité d'une bonne connaissance des problèmes inhérents sinon, modèles sans intérêt

Nécessité de collaborations **étroites** entre biologistes/cliniciens et biostatisticients.