Journée de rentrées des masters de la FMJH

IA et génomique : apprentissage régularisé pour la découverte de régions génomiques d'intérêt

Chloé-Agathe Azencott

Center for Computational Biology (CBIO)
Mines Paris – Institut Curie – INSERM U900
PSL Research University & PR[AI]RIE, Paris, France

31 août 2022

http://cazencott.info chloe-agathe.azencott@mines-paristech.fr @cazencott

Qu'est-ce que l'intelligence artificielle ?

- Dartmouth Conference, 1956

  The science and engineering of making computers behave in ways that, until recently, we thought required human intelligence.

- Reproduire **artificiellement** les comportements du vivant **perçus comme intelligents.**

Qu'est-ce que l'intelligence artificielle ?

- Dartmouth Conference, 1956

  The science and engineering of making computers behave in ways that, until recently, we thought required human intelligence.

- Reproduire **artificiellement** les comportements du vivant **perçus comme intelligents.**

- **Motivations :**
  - **Comprendre** l'intelligence naturelle
  - Créer des **outils** qui nous aident.

Quelques sous-domaines de l'IA

- **Robotique**
  - mécanique, électronique, contrôle
- **Systèmes experts**
  - utilisent des règles
  - logique des propositions, logique des prédicats
- **Traitement du langage naturel**
- **Vision par ordinateur**
- **Informatique affective**
  - reconnaître, modéliser, exprimer les émotions humaines
  - sciences cognitives, psychologie
- **Apprentissage automatique**

Quelques sous-domaines de l'IA

- **Robotique**
  - mécanique, électronique, contrôle
- **Systèmes experts**
  - utilisent des règles
  - logique des propositions, logique des prédicats
- **Traitement du langage naturel**
- **Vision par ordinateur**
- **Informatique affective**
  - reconnaître, modéliser, exprimer les émotions humaines
  - sciences cognitives, psychologie
- **Apprentissage automatique**

Apprentissage automatique

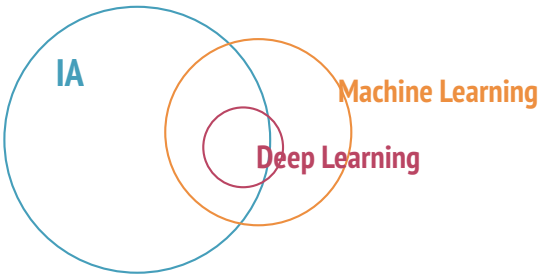
(apprentissage statistique, machine learning)

– **Apprendre :** acquérir une **compétence** par l'**expérience**, la pratique.

Apprentissage automatique

(apprentissage statistique, machine learning)

- **Apprendre :** acquérir une **compétence** par l'**expérience**, la pratique.
- Pour une machine :
  - **compétence** = algorithme
  - **expérience,** pratique = exemples, données



Apprentissage supervisé

– **But : apprendre** un modèle **prédictif**

Apprentissage supervisé

- **But : apprendre** un modèle **prédictif**
- **Formalisation :**
  - **Données :**
    n observations $\vec{x}_1, \ldots, \vec{x}_n$; $\vec{x}_i \in \mathcal{X}$ et leurs étiquettes $y_1, \ldots, y_n$; $y_i \in \mathcal{Y}$
    $\mathcal{Y} = \mathbb{R}$ (régression), ou $\{0, 1, \ldots, C-1\}$ (classification)
  - **Sortie :**
    modèle $f : \mathcal{X} \to \mathcal{Y}$ tel que $f(\vec{x}_i) \approx y_i$.

Exemples

- Exemples de **problèmes de régression :** $\mathcal{Y} = \mathbb{R}$
  - **Prédiction de clics :** combien de personnes vont cliquer sur ce lien ?
  - Prédiction de la **solubilité d'une molécule** dans l'éthanol en mg/mL

Exemples

- Exemples de **problèmes de régression :** $\mathcal{Y} = \mathbb{R}$
  - **Prédiction de clics :** combien de personnes vont cliquer sur ce lien ?
  - Prédiction de la **solubilité d'une molécule** dans l'éthanol en mg/mL
- Exemples de **problèmes de classification binaire :** $\mathcal{Y} = \{0, 1\}$
  - **Filtrage de spam :** cet email est-il un spam ?
  - **Reconnaissance faciale :** est-ce Y Le Cun sur cette photo ?

Exemples

- Exemples de **problèmes de régression :** $\mathcal{Y} = \mathbb{R}$
  - **Prédiction de clics :** combien de personnes vont cliquer sur ce lien ?
  - Prédiction de la **solubilité d'une molécule** dans l'éthanol en mg/mL
- Exemples de **problèmes de classification binaire :** $\mathcal{Y} = \{0, 1\}$
  - **Filtrage de spam :** cet email est-il un spam ?
  - **Reconnaissance faciale :** est-ce Y Le Cun sur cette photo ?
- Exemples de **problèmes de classification multiclasse :**
  - **OCR**, reconnaissance de chiffres manuscrits : $\mathcal{Y} = \{0, 1, 2, \ldots, 9\}$
  - Reconnaissance de **plantes** ou de **chants d'oiseaux**

Minimisation du risque empirique

Données : n observations $\vec{x}_1, \ldots, \vec{x}_n$ et leurs étiquettes $y_1, \ldots, y_n$

– **Apprendre** = **trouver** un modèle **dans la famille considérée** qui **minimise** l'**erreur** en moyenne sur les données

$$\hat{f} = \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \text{Erreur}(y_i, f(\vec{x}_i))$$

$\Rightarrow$ problème **d'optimisation**

Minimisation du risque empirique

Données : n observations $\vec{x}_1, \ldots, \vec{x}_n$ et leurs étiquettes $y_1, \ldots, y_n$

– **Apprendre** = **trouver** un modèle **dans la famille considérée** qui **minimise** l'**erreur** en moyenne sur les données

$$\hat{f} = \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \text{Erreur}(y_i, f(\vec{x}_i))$$

$\Rightarrow$ problème **d'optimisation**…plus ou moins facile à résoudre

Régression linéaire

Données : n observations $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^p$ et leurs étiquettes $y_1, \ldots, y_n \in \mathbb{R}$

– **Famille de modèles :**

$$\mathcal{F} = \left\{ \vec{x} \mapsto \langle \vec{\beta}, \vec{x} \rangle = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \, ; \vec{\beta} \in \mathbb{R}^{p+1} \right\}$$

Régression linéaire

Données : n observations $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^p$ et leurs étiquettes $y_1, \ldots, y_n \in \mathbb{R}$

- **Famille de modèles :**

$$\mathcal{F} = \left\{ \vec{x} \mapsto \langle \vec{\beta}, \vec{x} \rangle = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \, ; \vec{\beta} \in \mathbb{R}^{p+1} \right\}$$

- **Erreur :** perte quadratique Erreur$(y, f(\vec{x})) = (y - f(\vec{x}))^2$

Régression linéaire

Données : n observations $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^p$ et leurs étiquettes $y_1, \ldots, y_n \in \mathbb{R}$

- **Famille de modèles :**

$$\mathcal{F} = \left\{ \vec{x} \mapsto \langle \vec{\beta}, \vec{x} \rangle = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \, ; \vec{\beta} \in \mathbb{R}^{p+1} \right\}$$

- **Erreur :** perte quadratique Erreur$(y, f(\vec{x})) = (y - f(\vec{x}))^2$

- **Minimisation du risque empirique**

$$\hat{\vec{\beta}} = \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j \right) \right)^2$$

Régression linéaire

Données : n observations $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^p$ et leurs étiquettes $y_1, \ldots, y_n \in \mathbb{R}$

- **Famille de modèles :**

$$\mathcal{F} = \left\{ \vec{x} \mapsto \langle \vec{\beta}, \vec{x} \rangle = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \, ; \vec{\beta} \in \mathbb{R}^{p+1} \right\}$$

- **Erreur :** perte quadratique Erreur$(y, f(\vec{x})) = (y - f(\vec{x}))^2$

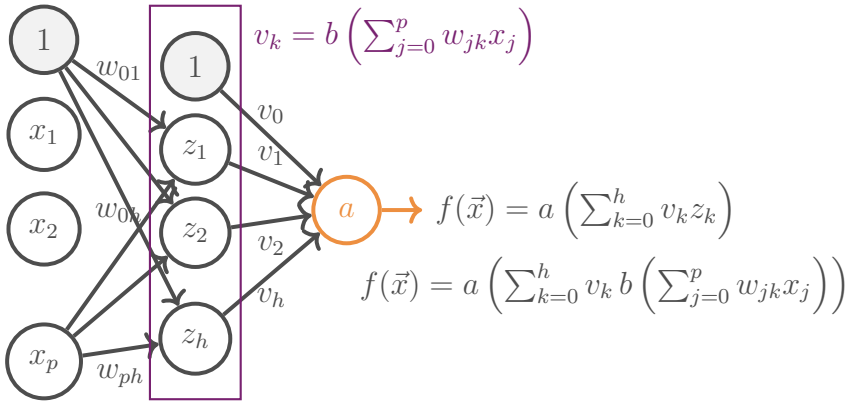
- **Minimisation du risque empirique**

$$\hat{\vec{\beta}} = \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j \right) \right)^2$$

= moindres carrés (Gauss / Legendre) ; système de n équations à $p+1$ inconnues

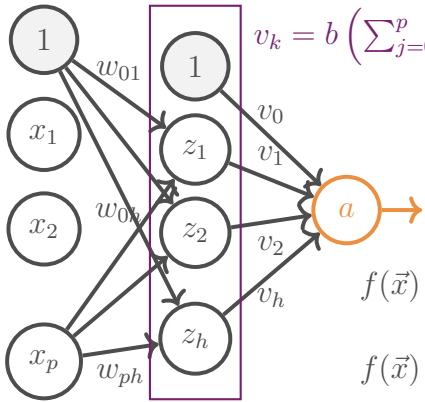
Perceptron multicouches

– Une **famille de modèles** plus complexe…et un problème d'optimisation plus ardu !



Perceptron multicouches

– Une **famille de modèles** plus complexe…et un problème d'optimisation plus ardu !



$$v_k = b\left(\sum_{j=0}^{p} w_{jk} x_j\right)$$

$$f(\vec{x}) = a\left(\sum_{k=0}^{h} v_k z_k\right)$$

$$f(\vec{x}) = a\left(\sum_{k=0}^{h} v_k \, b\left(\sum_{j=0}^{p} w_{jk} x_j\right)\right)$$

$$f(\vec{x}) = \frac{1}{1+e^{-\left(\sum_{k=0}^{h} v_k \frac{1}{1+e^{-\left(\sum_{j=0}^{p} w_{jk} x_j\right)}}\right)}}$$

Quelques applications en génomique

- Aide au **diagnostic** / **pronostic** / **traitement**

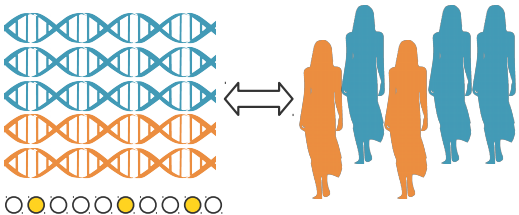
  R. Dias & A. Torkamani (2019). **Artificial intelligence in clinical and genomic diagnostics,** Genome Medicine.

- Déterminer la **fonction d'un gène**

  A. G. Fraser & E. M. Marcotte (2004). **A probabilistic view of gene function,** Nature Genetics.

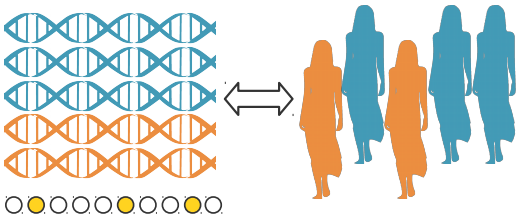
- Déterminer les **régions du génome** impliquées dans une maladie

Du génotype au phénotype



Quelles variables génomiques expliquent le phénotype ?

Du génotype au phénotype

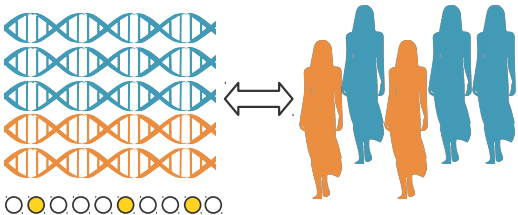


Quelles variables génomiques expliquent le phénotype ?

$p = 10^5 - 10^7$ **variables**

$n = 10^3 - 10^5$ **échantillons**

Du génotype au phénotype



Quelles variables génomiques expliquent le phénotype ?

$p = 10^5 - 10^7$ **variables**

$n = 10^3 - 10^5$ **échantillons**

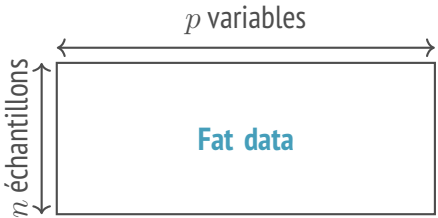
Peu d'échantillons (n faible) **grande dimension** (p grand).

Ceci n'est pas **du Big Data**

Magritte

Peu d'échantillons en grande dimension

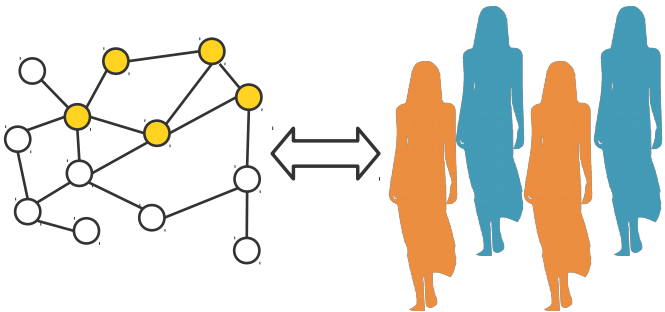
Peu d'échantillons en grande dimension



Idée 1 : Utiliser des connaissances a priori structurées

Hypothèse : Les variables pertinentes **respectent vraisemblablement une certaine structure.**

Exemple : Les variables sont **connectées sur un graphe donné**
Les **réseaux biologiques**, dont les arêtes représentent des relations biologiques entre les sommets/gènes, permettent de mieux comprendre les maladies.



C.-A. Azencott (2016). **Network-guided biomarker discovery,** Lecture Notes in Computer Science.

Minimisation du risque empirique régularisée

- **Régularisation :** imposer une contrainte sur le modèle
- Pour un modèle linéaire :

$$\vec{\beta}^* = \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \text{Erreur}(y_i, \langle \vec{\beta}, \vec{x}_i \rangle) + \lambda \, \Omega(\vec{\beta})$$

Minimisation du risque empirique régularisée

- **Régularisation :** imposer une contrainte sur le modèle
- Pour un modèle linéaire :

$$\vec{\beta}^* = \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \text{Erreur}(y_i, \langle \vec{\beta}, \vec{x}_i \rangle) + \lambda \, \Omega(\vec{\beta})$$

- **Régularisation** par graphe : $\Omega(\vec{\beta})$ est tel que
  - $\beta_j = 0$ pour de nombreuses valeurs de j
  - si les sommets j et k sont connectés, alors β_j et β_k ont des valeurs proches

Minimisation du risque empirique régularisée

- **Régularisation :** imposer une contrainte sur le modèle
- Pour un modèle linéaire :

$$\vec{\beta}^* = \underset{\vec{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \text{Erreur}(y_i, \langle \vec{\beta}, \vec{x}_i \rangle) + \lambda \, \Omega(\vec{\beta})$$

- **Régularisation** par graphe : $\Omega(\vec{\beta})$ est tel que
  - $\beta_j = 0$ pour de nombreuses valeurs de j
  - si les sommets j et k sont connectés, alors β_j et β_k ont des valeurs proches

  Exemple :

$$\Omega(\vec{\beta}) = \sum_{j=0}^{p} |\beta_j| + \eta \sum_{j \sim k} (\beta_j - \beta_k)^2$$

C. Li and H. Li (2008). **Network-constrained regularization and variable selection for analysis of genomic data,** Bioinformatics, 24, 1175–1182.

Pertinence régularisée

Remplacer la minimisation du risque empirique par la **maximisation d'une mesure de pertinence** (corrélation, test statistique, etc.)

$$\mathcal{S}^* = \underset{\mathcal{S}\subset\{1,2,\ldots,p\}}{\arg\max} \ \text{Pertinence}(\mathcal{S}) - \lambda\, \Omega(\mathcal{S})$$

Exemple :

$$\Omega(\mathcal{S}) = |\mathcal{S}| + \eta \sum_{j\sim k} 1_{j\in\mathcal{S} \text{ et } k\notin\mathcal{S}}$$

☺ Peut être résolu très efficacement comme un problème de **coupe minimale / flot maximal**.

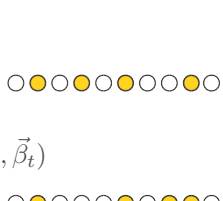
C.-A. Azencott et al. (2013) **Efficient network-guided multi-locus association mapping with graph cuts,** Bioinformatics.

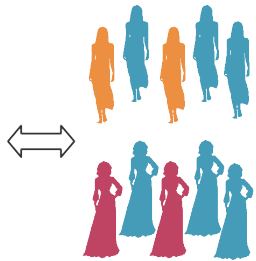
H. Climente-González et al. (2021). **Boosting GWAS using biological networks: A study on susceptibility to familial breast cancer,** PLoS Comp Bio.

Idée 2 : Résoudre ensemble plusieurs problèmes liés

Utiliser la **régularisation** pour encourager des tâches **similaires** à **partager les mêmes variables**

$$\underset{\vec{\beta}_1,\ldots,\vec{\beta}_T \in \mathbb{R}^{p+1}}{\arg\min} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n} \text{Erreur}(y_{ti}, \langle \vec{\beta}_t, \vec{x}_{ti} \rangle) + \lambda\, \Omega(\vec{\beta}_1, \ldots, \vec{\beta}_t)$$

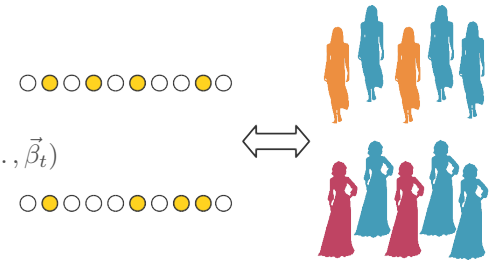




Idée 2 : Résoudre ensemble plusieurs problèmes liés

Utiliser la **régularisation** pour encourager des tâches **similaires** à **partager les mêmes variables**

$$\underset{\vec{\beta}_1,\ldots,\vec{\beta}_T \in \mathbb{R}^{p+1}}{\arg\min} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n} \text{Erreur}(y_{ti}, \langle \vec{\beta}_t, \vec{x}_{ti} \rangle) + \lambda\, \Omega(\vec{\beta}_1, \ldots, \vec{\beta}_t)$$



– Exemple 1 : utiliser des **descripteurs** de tâches $z_{t1}, \ldots, z_{td}$ pour que les tâches **les plus similaires** partagent le plus de variables

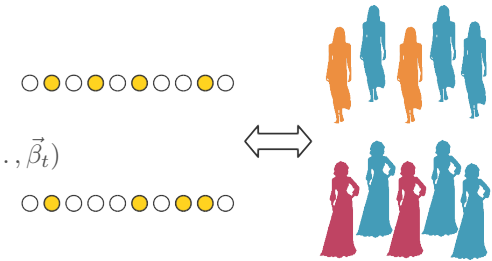
V. Bellón et al. (2016). **Multitask feature selection with task descriptors**, Pacific Symposium on Biocomputing.

$$\beta_{tj} = \theta_j \sum_{m=1}^{d} \alpha_{jd} z_{tm} \qquad \Omega(\vec{\beta}_1, \ldots, \vec{\beta}_t) = \sum_{j=1}^{p} |\theta_j| + \eta \sum_{m=1}^{d} \sum_{j=1}^{p} |\alpha_{jd}|$$

Idée 2 : Résoudre ensemble plusieurs problèmes liés

Utiliser la **régularisation** pour encourager des tâches **similaires** à **partager les mêmes variables**

$$\underset{\vec{\beta}_1,\ldots,\vec{\beta}_T \in \mathbb{R}^{p+1}}{\arg\min} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n} \text{Erreur}(y_{ti}, \langle \vec{\beta}_t, \vec{x}_{ti} \rangle) + \lambda\, \Omega(\vec{\beta}_1, \ldots, \vec{\beta}_t)$$



– Exemple 2 : utiliser **une tâche par population** + G **groupes** $\mathcal{G}_1, \ldots \mathcal{G}_G$ de variables corrélées

A. Nouira & C.-A. Azencott (2022). **Multitask group lasso for genome-wide association studies in diverse populations**, Pacific Symposium on Biocomputing.

$$\Omega(\vec{\beta}_1, \ldots, \vec{\beta}_t) = \sum_{g=1}^{G} \sqrt{p_g} \sum_{j \in \mathcal{G}_g} \sum_{t=1}^{T} \beta_{tj}^2$$

Difficultés

- **Volume des données**

  Peu de données, en (très) grande dimension

- Manque d'**intuitions**

  Des problèmes que les humains ne savent pas résoudre non plus !

- **Interprétabilité** et **explicabilité**

  Comprendre comment/pourquoi on prédit

CBIO: Víctor Bellón (now eDreams Odigeo), **Héctor Climente-González** (now RIKEN-AIP), Vivien Goepp, **Asma Nouira**, Lotfi Slim (now NVIDIA), Daniel Rovera, Véronique Stoven, Athénaïs Vaginay (now Lorraine University), Jean-Philippe Vert (now Owkin).

MLCB Tübingen: **Karsten Borgwardt** (now ETH Zürich), **Dominik Grimm** (now Weihenstephan), Barbara Rakitsch (now Bosch AI), Damian Roqueiro (now Roche), Nino Shervashidze (now Sancare), **Mahito Sugiyama** (now NII, Japan).

Liège University: Diane Duroux, Kristel van Steen

U900 Epidemio: Nadine Andrieu, Fabienne Lesueur, Christine Lonjou.

Sanofi-Adventis: Clément Chatelain, Hélène de Foucauld

MPI for Psychiatry: André Altmann (now UCL), Tony Kam-Thong (now Roche), Bertram Müller-Myhsok, Benno Pütz.

Osaka University: **Yoshinobu Kawahara.**

source: http://www.flickr.com/photos/wwworks/

Financement

- **Agence Nationale pour la Recherche**

  SCAPHE (JCJC ANR-18-CE45-0021-01) 2019–2021.
  PR[AI]RIE Springboard Chair 2020–2023.

- **Alexander von Humboldt Stiftung**

  Postdoctoral fellowship 2011–2013.

- **Deutsche Forschungsgemeinschaft**

  LaLa (DFG 164930347) 2013.

- **European Research Council**

  IC-3i-PhD (H2020-MSCA-COFUND-2014 666003) 2016–2019.

- **Sanofi-Adventis**