Les mathématiques au coeur de l'Intelligence Artificielle pour la science des données massives

Jérémie Bigot

Institut de Mathématiques de Bordeaux
Université de Bordeaux

MidisMath de l'UFMI

Février 2020

Un petit sondage...

Qui a déjà entendu parler (dans les médias) de :

- Intelligence Artificielle (IA) ?
- Big Data (Données Massives) ?

Un petit sondage...

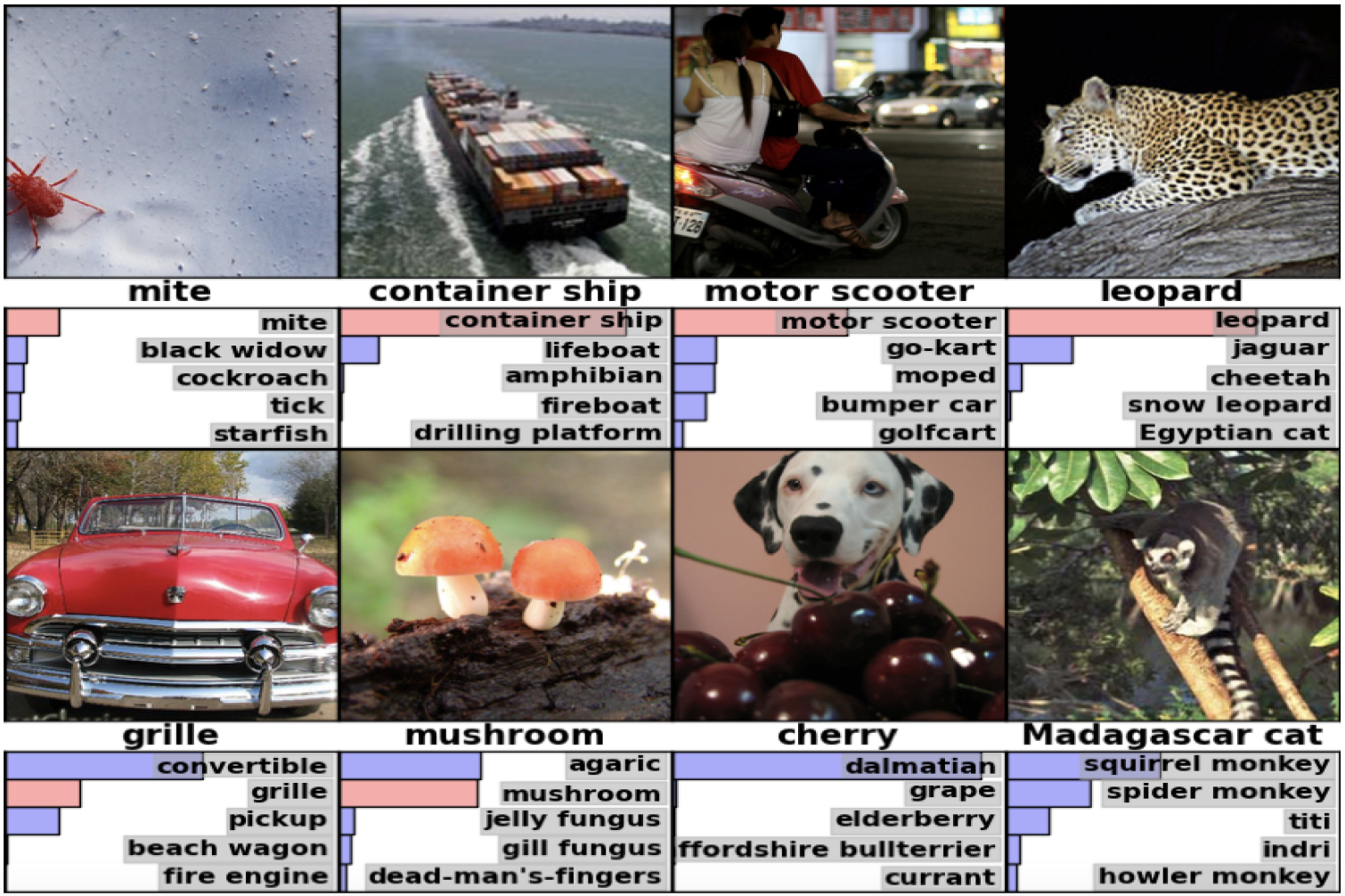
Qui a déjà entendu parler :

- des mathématiques au coeur des méthodes qui font le succès de l'IA (telle que médiatisée aujourd'hui) ?

Un peu de terminologie... pour se mettre d'accord ?

Mais c'est quoi l'IA ?

Reconnaissance automatique d'images [1]



1. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012)

Un peu de terminologie... pour se mettre d'accord ?

- Distinction entre IA forte et IA faible cf. Wikipedia [1]
  - "L'IA faible est une intelligence artificielle non-sensible qui se concentre sur une tâche précise"
  - "Tous les systèmes actuellement existants sont considérés comme des intelligences artificielles faibles"
- Cet exposé (et beaucoup des références actuelles dans les médias) = IA faible
- IA faible = apprentissage automatique à partir d'exemples en très grand nombre

1. fr.wikipedia.org/wiki/Intelligence_artificielle_faible

Classification d'images - ILSVRC Challenge (2010) [1]

- apprentissage : 1.2 million d'images labellisées (1000 classes)
- test : 150 000 images

1. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012)

Intelligence artificielle et génération d'images

Base de données d'images de célébrités : **CelebA Dataset** [1]

1. http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Intelligence artificielle et génération d'images

Question : peut-on apprendre à partir d'un ensemble de visages à en générer aléatoirement de nouveaux ?

Intelligence artificielle et génération d'images

Réponse : solution proposée par des chercheurs de la société Nvidia [1]

1. https://research.nvidia.com/publication/2017-10_Progressive-Growing-of

Intelligence artificielle et génération d'images

Question : quelles sont les vraies images et celles générées aléatoirement [1] ?

1. https://research.nvidia.com/publication/2017-10_Progressive-Growing-of

Intelligence artificielle et génération d'images

Réponse - 1ère ligne : génération aléatoire et lignes 2 à 5 : vraies images (les plus proches de l'image générée) [1] !

1. https://research.nvidia.com/publication/2017-10_Progressive-Growing-of

Succès récents et diffusion de l'IA

Raisons du succès de l'IA (faible)

- Raffinement des méthodes d'apprentissage
- Moyens de calculs
- Taille des bases d'apprentissage
- Popularisation par librairies de calcul facilement utilisables

Cet exposé ?

Les mathématiques en Licence à l'Université de Bordeaux à la base des méthodes d'apprentissage de l'IA ?

Concepts enseignés en Licence à l'UB (bases pour l'IA)

1. fonctions de plusieurs variables à valeurs réelles
2. géométrie euclidienne (Pythagore encore et toujours !)
3. calcul vectoriel et matriciel
4. continuité, dérivabilité (différentiabilité), composition de fonctions
5. convergence et limite des suites
6. variables aléatoires, probabilités et statistique

Débouchés à la portée des jeunes diplômés en mathématiques

L'offre croissante des métiers en science des données !

Eléments de modélisation mathématique

Cadre de base - Apprentissage supervisé à 2 classes en traitement d'images :

- soit $X_1, \ldots, X_n$ un ensemble d'images appartenant à 2 classes possibles notées 0 ou 1
- on connait les classes des images notées $Y_1, \ldots, Y_n$ avec $Y_i \in \{0, 1\}$ pour $1 \leq i \leq n$.

Exemple - Détection de la présence d'une éolienne dans une image satellite [1]

$(X_1, Y_1 = 1), \quad (X_2, Y_2 = 0) \quad \cdots \quad (X_{n-1}, Y_{n-1} = 1), (X_n, Y_n = 0)$

1. https ://defi-ia.insa-toulouse.fr/

Eléments de modélisation mathématique

Représentation mathématique d'une image - chaque image est considéré comme un grand vecteur de dimension d dont les éléments sont les valeurs prises par les pixels

Données - ensemble de couples $(X_i, Y_i) \in \mathbb{R}^d \times \{0, 1\}$ pour $1 \leq i \leq n$, dit ensemble d'apprentissage

Problématique - determiner la classe d'une nouvelle image $X \in \mathbb{R}^d$?

Principe - trouver une fonction $f : \mathbb{R}^d \to [0, 1]$ tel que $f(X)$ représente la probabilité que X appartiennent à la classe 1.

Terminologie - la fonction f est appelée règle de classification

Principes de base de l'apprentissage automatique

Principe - trouver une fonction $f : \mathbb{R}^d \to [0,1]$ tel que $f(X)$ représente la probabilité que X appartiennent à la classe 1.

Choix d'une méthode d'apprentissage - recherche d'une fonction [1] dépendant d'un ensemble de paramètres $\theta \in \mathbb{R}^p$

$$\begin{cases} f : & \mathbb{R}^d \times \mathbb{R}^p & \to & [0,1] \\ & (x,\theta) & \mapsto & f(x,\theta) \end{cases}$$

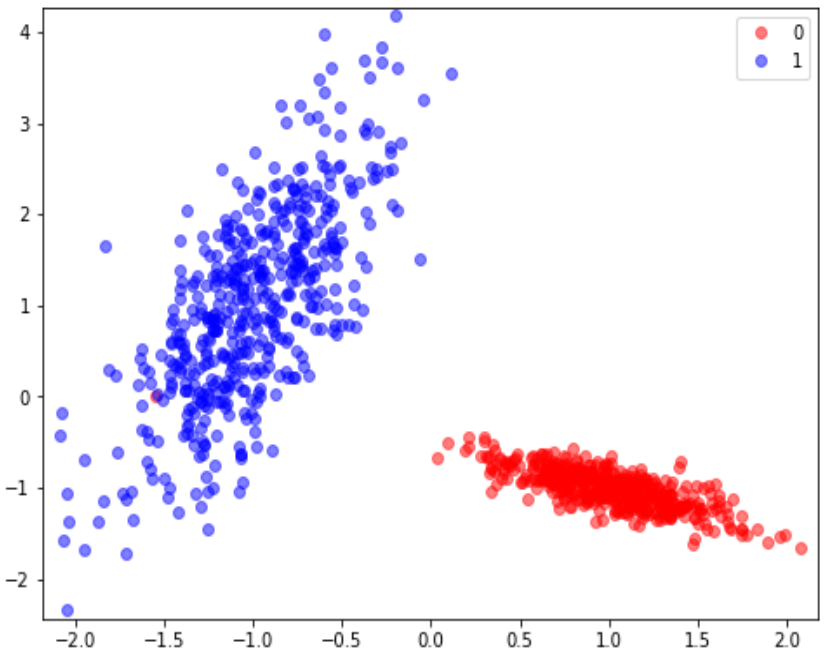
Recherche des meilleurs paramètres - minimisation de l'erreur d'apprentissage

$$\min_{\theta \in \mathbb{R}^p} F(\theta) \quad \text{avec} \quad F(\theta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i, \theta))^2$$

1. fonctions de plusieurs variables à valeurs réelles

Méthode d'apprentissage - brique de base

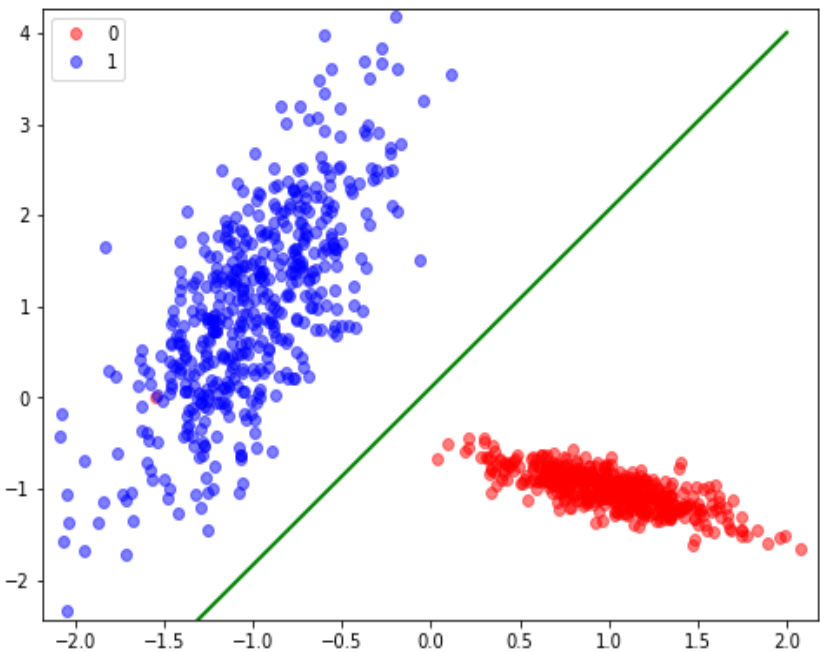
Choix d'une méthode d'apprentissage - séparation de classes par un hyperplan [2]



2. géométrie euclidienne

Méthode d'apprentissage - brique de base

Choix d'une méthode d'apprentissage - séparation de classes par un hyperplan [2]

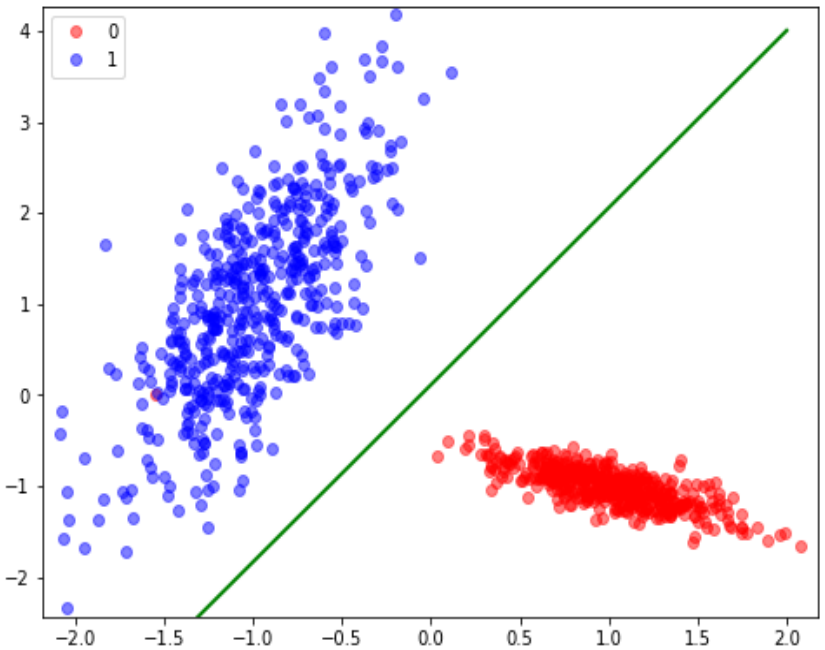


Equation d'un hyperplan $\{x \in \mathbb{R}^d \ : \ \langle x, w \rangle + b = 0\}$ où $\theta = (w, b) \in \mathbb{R}^d \times \mathbb{R}$ sont les paramètres de l'hyperplan (ici $d = 2$)

2. géométrie euclidienne

Méthode d'apprentissage - brique de base

Choix d'une méthode d'apprentissage - séparation de classes par un hyperplan [2]

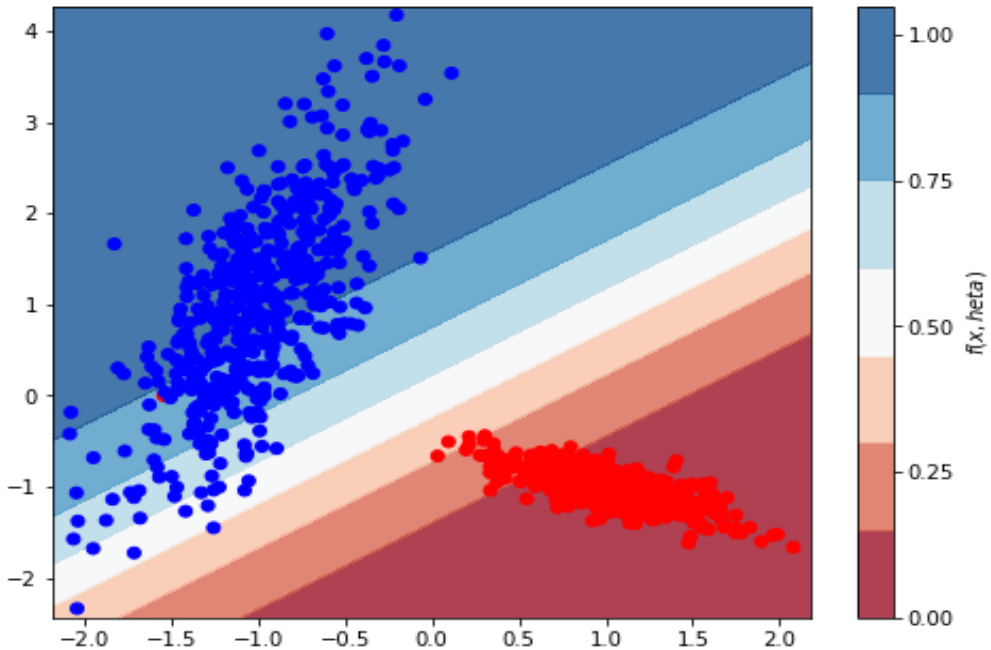


- Points au-dessus de l'hyperplan $\{x \in \mathbb{R}^d \ : \ \langle x, w \rangle + b > 0\}$
- Points au-dessous de l'hyperplan $\{x \in \mathbb{R}^d \ : \ \langle x, w \rangle + b < 0\}$

2. géométrie euclidienne

Méthode d'apprentissage - brique de base

Choix d'une méthode d'apprentissage - séparation de classes par un hyperplan [2]

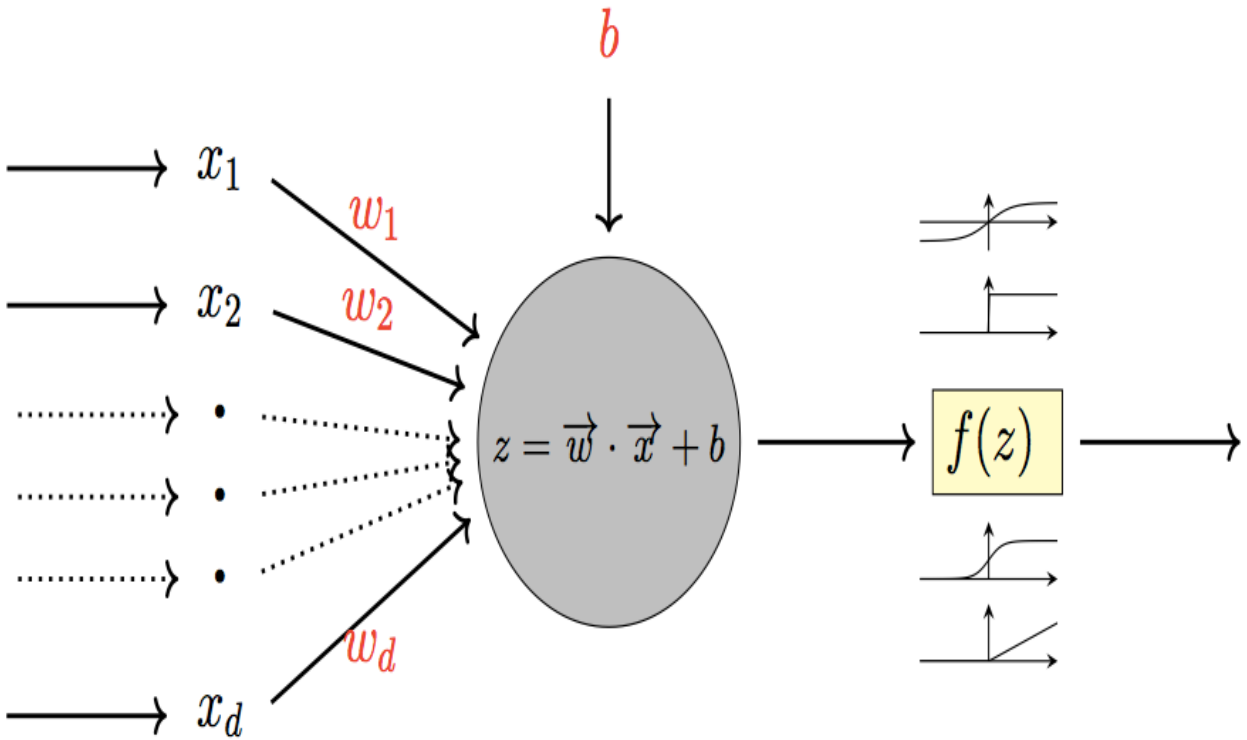


Règle de classification $f(x, \theta) = \sigma\left(\langle x, w\rangle + b\right)$ avec $\sigma(z) = \mathbb{1}_{\mathbb{R}^+}(z)$ ou $\sigma(z) = \frac{1}{1+\exp(-z)}$ avec $\theta = (w, b) \in \mathbb{R}^d \times \mathbb{R}$

2. géométrie euclidienne

Construction d'un réseau de neurones

Neurone de base : modèle du **Perceptron** (Rosenblatt, 1957)



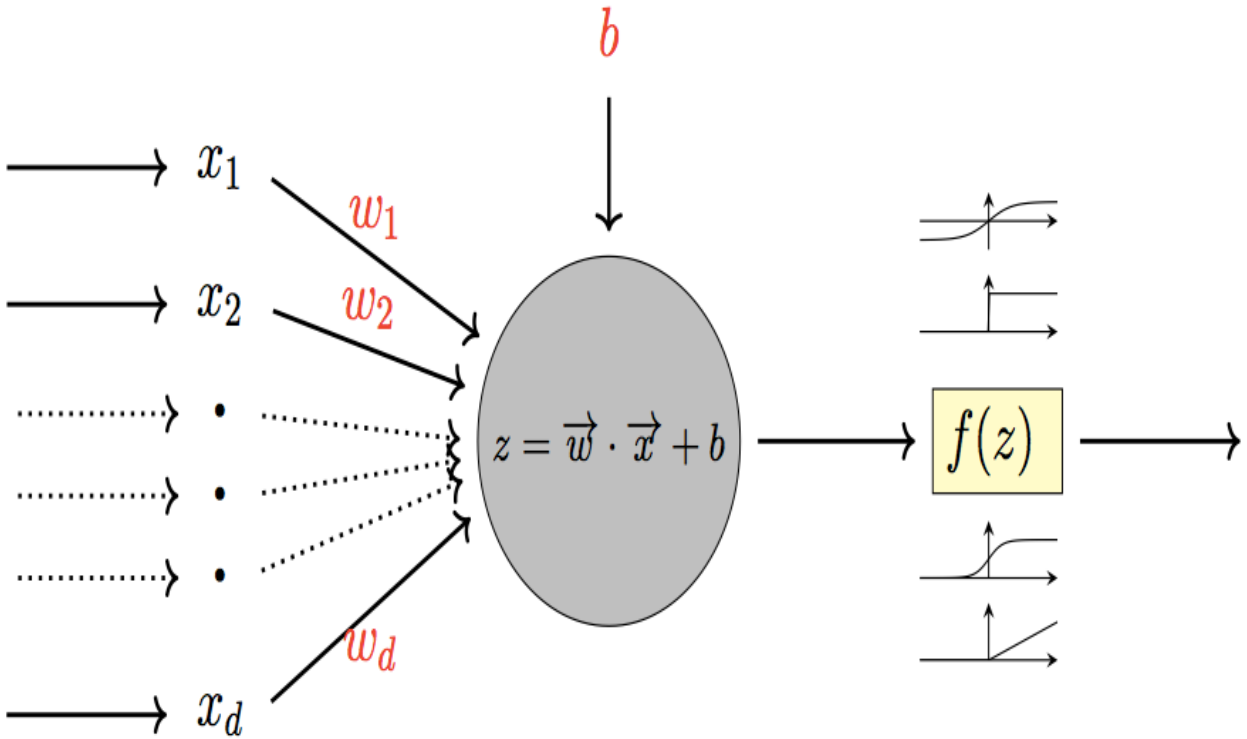
Source : `https://stats385.github.io/`

Combinaison linéaire de $x \in \mathbb{R}^d$ avec $\omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$ et $b \in \mathbb{R}$

Fonction non-linéaire d'activation $f(z) = \sigma(z) = \mathbb{1}_{\mathbb{R}^+}(z)$

Construction d'un réseau de neurones

Neurone de base : modèle du **Perceptron** (Rosenblatt, 1957)



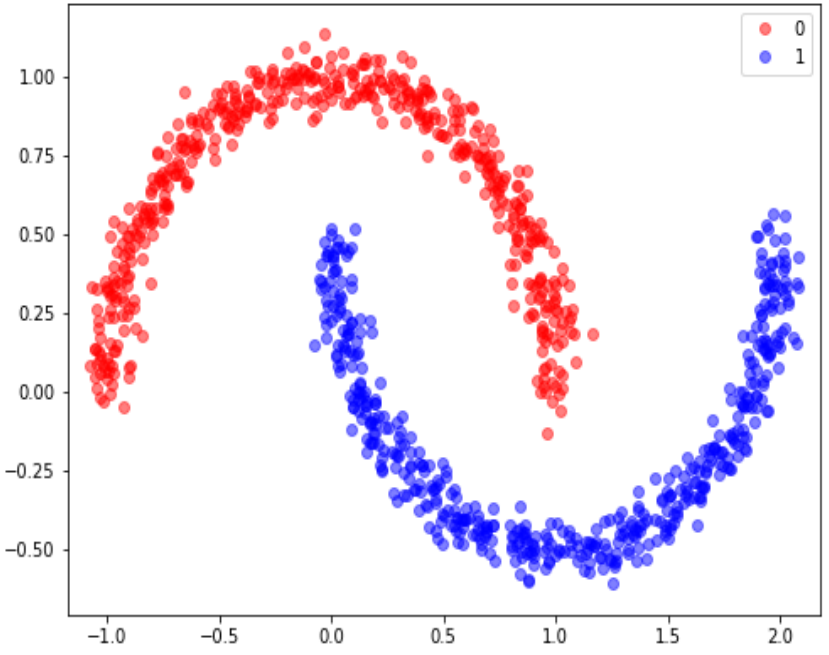
Source : `https://stats385.github.io/`

Combinaison linéaire de $x \in \mathbb{R}^d$ avec $\omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$ et $b \in \mathbb{R}$

Autre choix $\sigma(z) = \frac{1}{1+\exp(-z)}$ (sigmoïde) ou $\sigma(z) = \max(0, z)$ (ReLU)

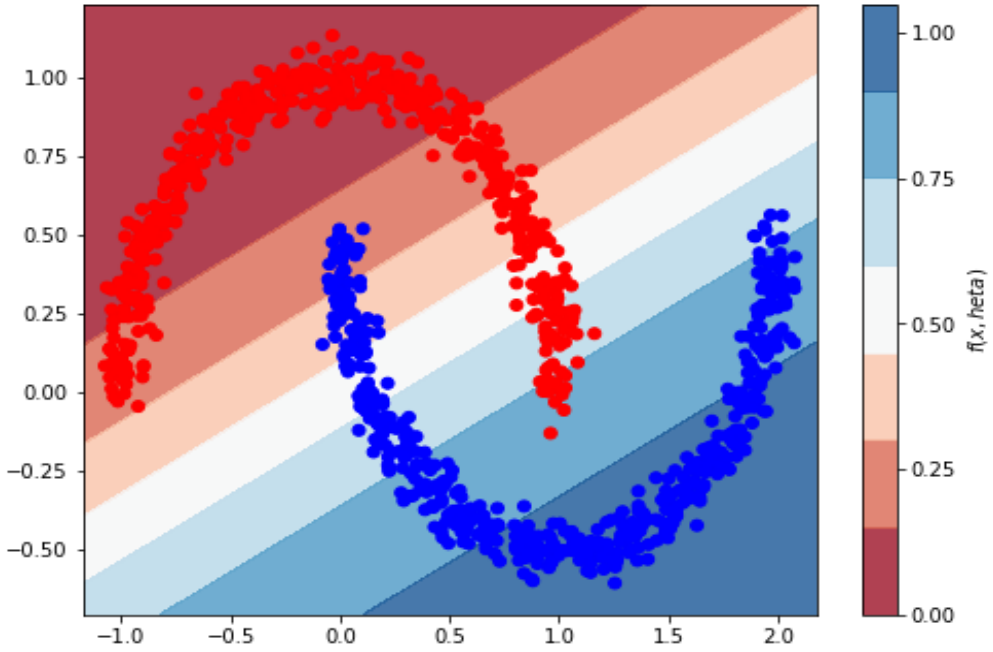
Apprentissage par combinaison de briques de base

Cas de données non-linéairement séparables ?



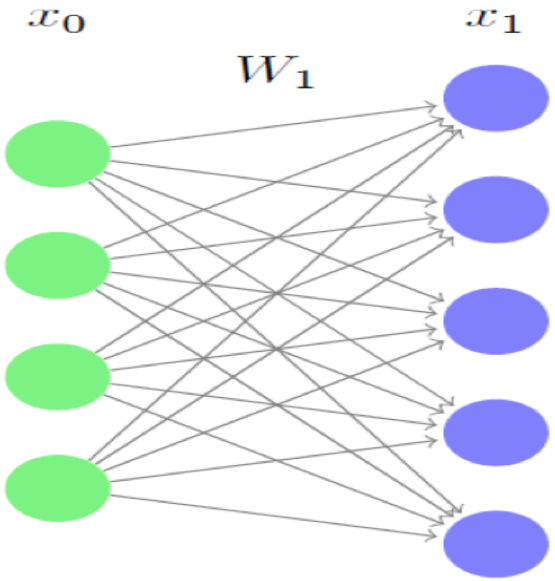
Apprentissage par combinaison de briques de base

Limitations d'une règle de classification basée sur **un seul hyperplan** !



Construction d'un réseau de neurones

Utilisation de plusieurs hyperplans - Neurones cachés !



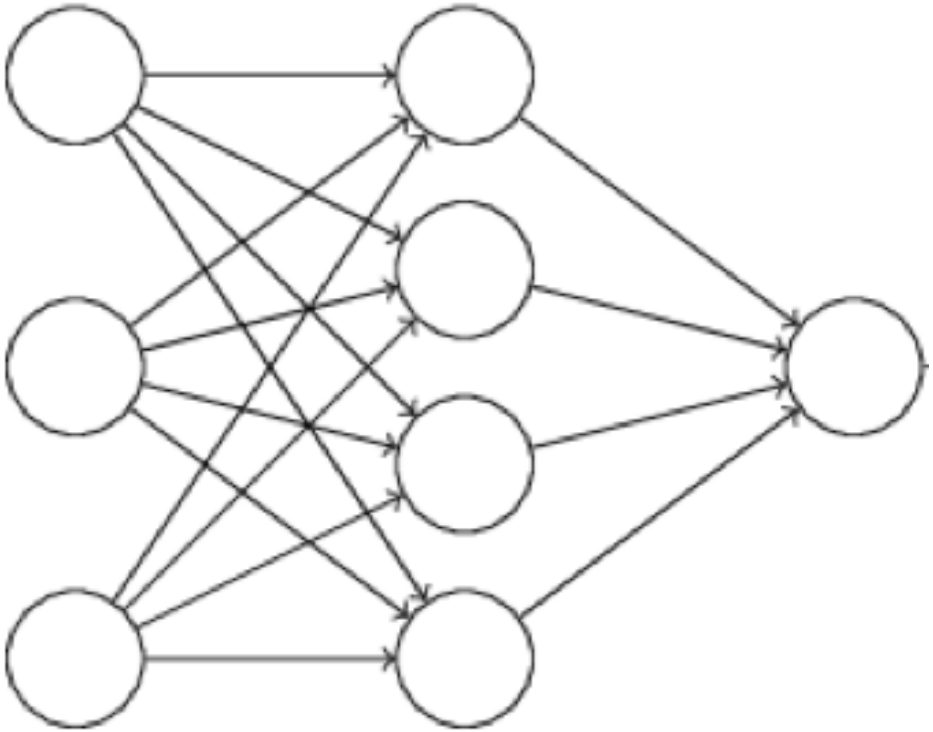
Source : `https://stats385.github.io/`

Ecriture condensée : $x_1 = \sigma_1 (W_1 x_0 + b_1)$, avec $x_0 = x$, où

- $\sigma_1 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1}$ est une **fonction non-linéaire entrée par entrée**
- $W_1 \in \mathbb{R}^{d \times d_1}$ (poids) $b_1 \in \mathbb{R}^{d_1}$ (biais)
- $\theta = (W_1, b_1)$: paramètres du réseau

Construction d'un réseau de neurones

Combinaisons linéaires des "**décisions de chaque hyperplan**" !



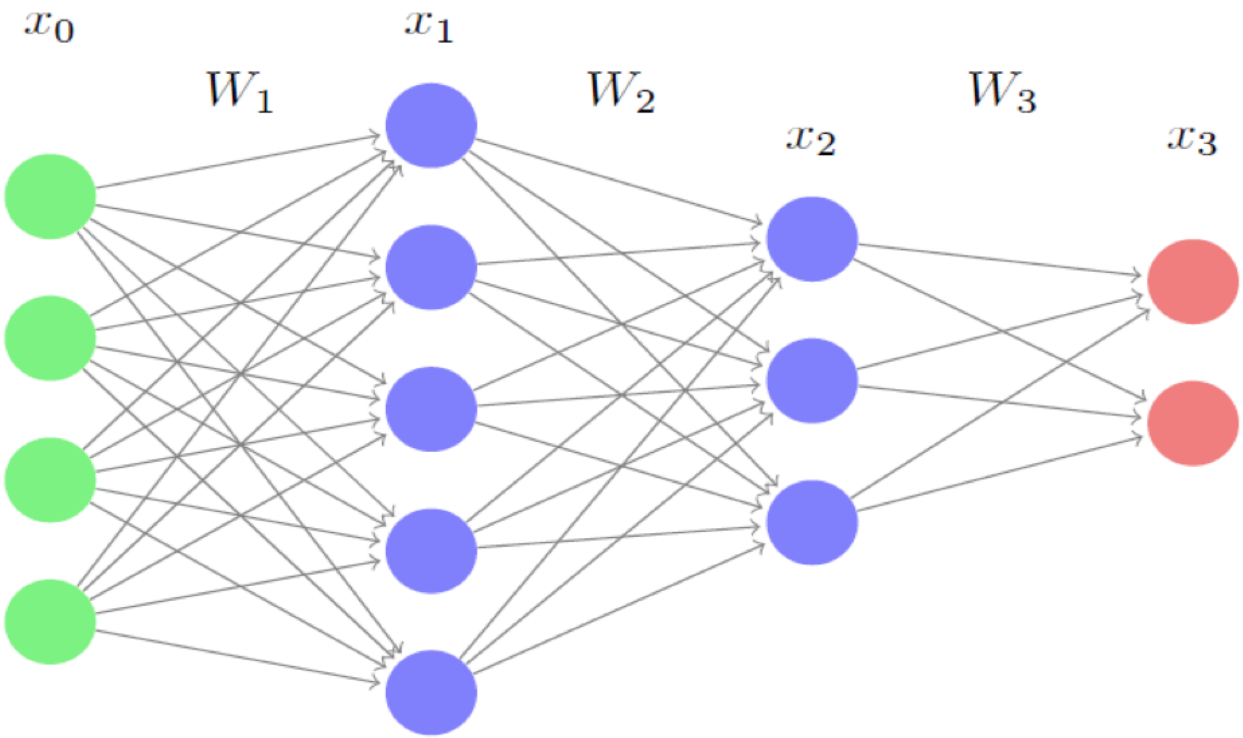
Source : `http://neuralnetworksanddeeplearning.com/index.html`

Ecriture condensée : $f(x,\theta) = \sigma(W_2\sigma_1(W_1x + b_1) + b_2)$,

avec $W_1 \in \mathbb{R}^{d\times d_1}$, $b_1 \in \mathbb{R}^{d_1}$, $W_2 \in \mathbb{R}^{d_1\times 1}$, $b_2 \in \mathbb{R}$ et $\theta = (W_1, b_1, W_2, b_2)$

Construction d'un réseau de neurones

Perceptron multi-couches : calcul matriciel + composition de fonctions



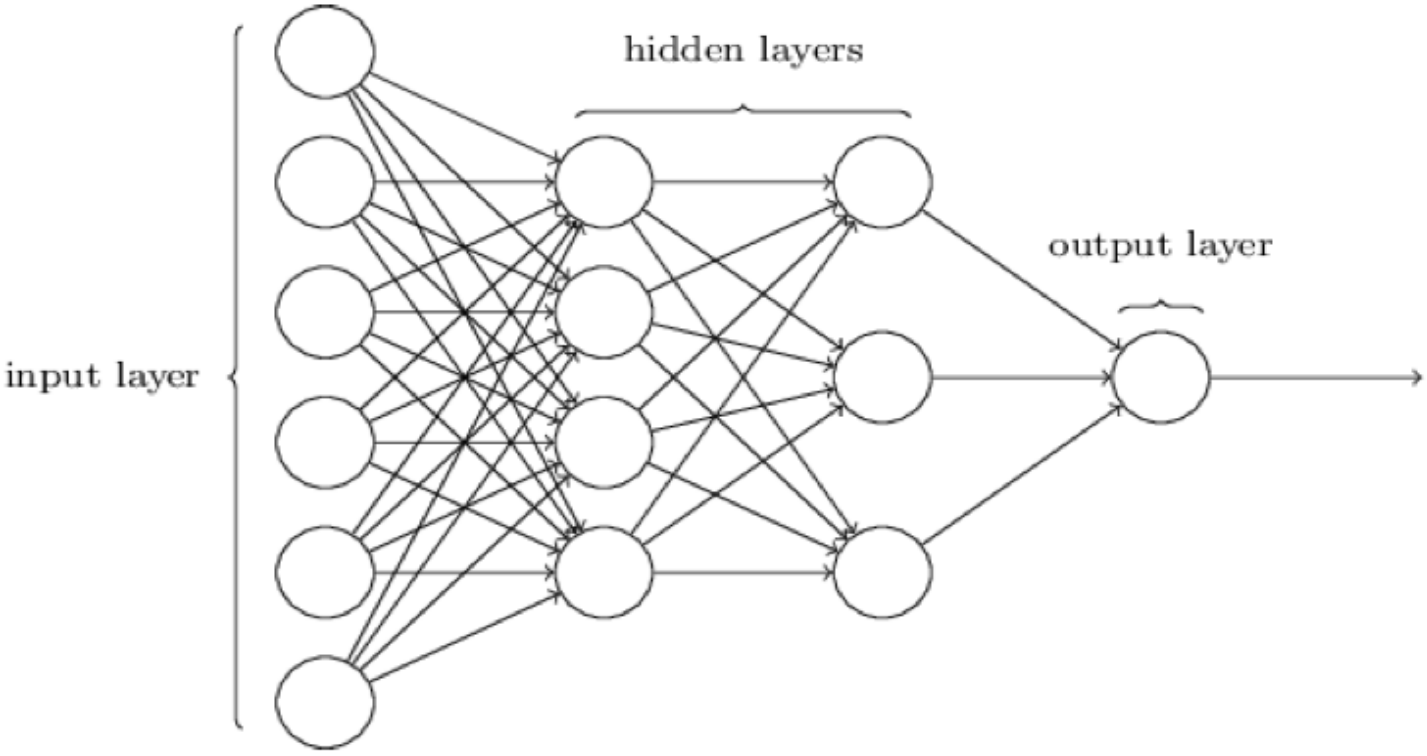
Source : `https://stats385.github.io/`

Ecriture condensée - entrée $x_0 = x \in \mathbb{R}^d$, et pour $\ell = 1, \ldots, L$,

faire $x_\ell = \sigma_\ell \left(W_\ell x_{\ell-1} + b_\ell \right)$ avec σ_L fonction sigmoïde.

Construction d'un réseau de neurones

Perceptron multi-couches : calcul matriciel + composition de fonctions



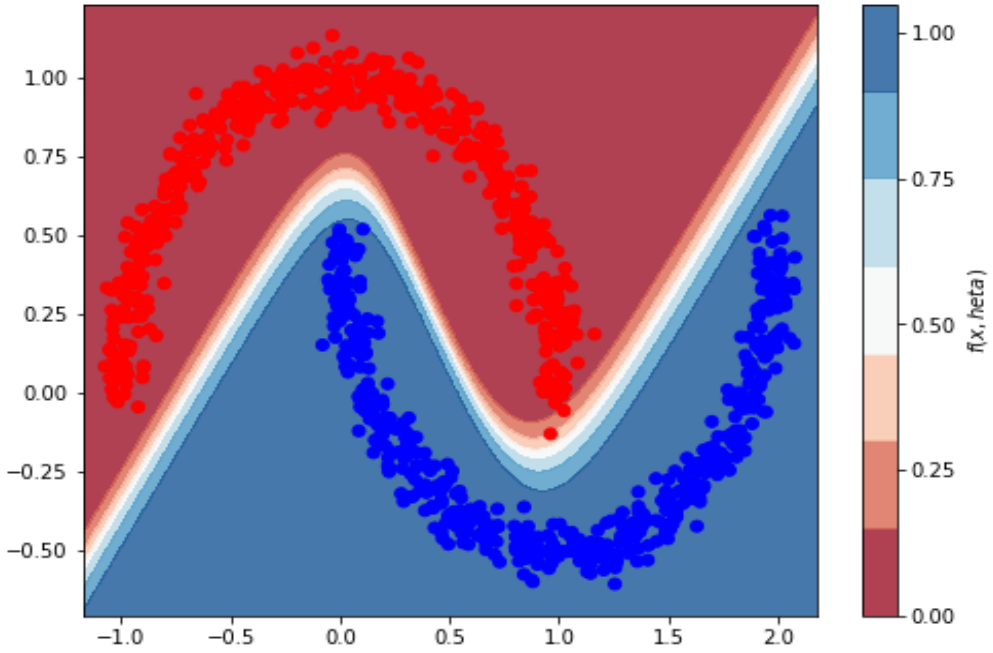
Source : `http://neuralnetworksanddeeplearning.com/index.html`

Réseau de neurones profonds : "nombreuses" couches cachées

Apprentissage par combinaison de briques de base

Réseau de neurones à deux couches cachées (sortie dans $[0, 1]$) :

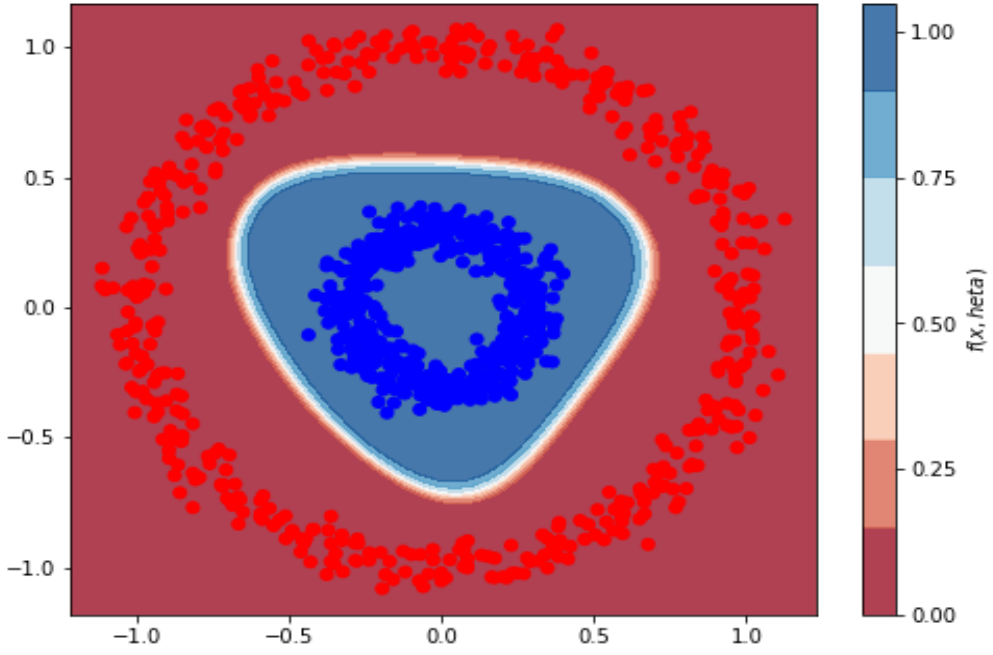
- entrées $d = 2$
- 2 couches cachées $d_1 = 4$ et $d_2 = 2$ (activation par tangente hyperbolique $\sigma_1(z) = \sigma_2(z) = \frac{1-\exp(2z)}{1+\exp(2z)}$)



Apprentissage par combinaison de briques de base

Réseau de neurones à deux couches cachées (sortie dans $[0, 1]$) :

- entrées $d = 2$
- 2 couches cachées $d_1 = 4$ et $d_2 = 2$ (activation par tangente hyperbolique $\sigma_1(z) = \sigma_2(z) = \frac{1-\exp(2z)}{1+\exp(2z)}$)



Principes de base de l'apprentissage automatique

Réseau de neurones comme méthode d'apprentissage

- choix d'une fonction $f(x,\theta)$ dépendant d'un ensemble de paramètres $\theta \in \mathbb{R}^p$
- f est définie (récursivement) par composition d'applications affines et d'applications non-linéaires

$$f(x,\theta) = \sigma_L\left(W_L x_{L-1} + b_L\right) \in [0,1],$$

avec $x_\ell = \sigma_\ell\left(W_\ell x_{\ell-1} + b_\ell\right)$ pour $\ell = 1,\dots,L-1$ et $x_0 = x \in \mathbb{R}^d$

Recherche des meilleurs paramètres - minimisation de l'erreur d'apprentissage

$$\min_{\theta\in\mathbb{R}^p} F(\theta) \quad \text{avec} \quad F(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f(X_i,\theta)\right)^2$$

Question : comment trouver un minimum de F ?

Vers les mathématiques computationnelles

Réponse : un minimum θ^* de F vérifie

$$\nabla(F(\theta^*)) = 0 \tag{1}$$

où

$$\nabla(F(\theta)) = \left(\frac{\partial}{\partial \theta_1} F(\theta), \cdots, \frac{\partial}{\partial \theta_p} F(\theta) \right)$$

est le gradient de F i.e. le vecteur dont les coordonnées sont les dérivées partielles de F [4]

Question : comment résoudre l'équation (2) ?

Solution : méthode numérique basée sur l'algorithmique !

4. continuité, dérivabilité (différentiabilité), composition de fonctions

Vers les mathématiques computationnelles

Réponse : un minimum θ^* de F vérifie

$$\nabla(F(\theta^*)) = 0 \tag{1}$$

où

$$\nabla(F(\theta)) = \left(\frac{\partial}{\partial \theta_1} F(\theta), \cdots, \frac{\partial}{\partial \theta_p} F(\theta) \right)$$

Solution : ré-écriture de l'équation (2) sous la forme

$$\theta^* = G(\theta^*) \quad \text{avec} \quad G(\theta) = \theta - \gamma \nabla(F(\theta)) \text{ et } \gamma > 0,$$

et donc θ^* est un point fixe de G !

Algorithme itératif [5] **:** $\theta^{(k+1)} = G(\theta^{(k)}) = \theta^{(k)} - \gamma \nabla(F(\theta^{(k)}))$ pour $k = 0, 1, 2, \ldots$, et on considère que $\theta^{(K)} \approx \theta^*$ pour K assez grand !

5. convergence et limite des suites - Algorithme de descente du gradient

Vers les mathématiques computationnelles

Réponse : un minimum θ^* de F vérifie

$$\nabla(F(\theta^*)) = 0 \tag{2}$$

où

$$\nabla(F(\theta)) = \left(\frac{\partial}{\partial \theta_1} F(\theta), \cdots, \frac{\partial}{\partial \theta_p} F(\theta) \right)$$

et $F(\theta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i, \theta))^2$.

Algorithme itératif [5] **:** $\theta^{(k+1)} = G(\theta^{(k)}) = \theta^{(k)} - \gamma \nabla(F(\theta^{(k)}))$ pour $k = 0, 1, 2, \ldots$, et on considère que $\theta^{(K)} \approx \theta^*$ pour K assez grand !

Etape fondamentale pour les réseaux de neurones : possibilité d'un calcul rapide du gradient de $\theta \mapsto f(X_i, \theta)$ (pour déterminer $\nabla F(\theta)$) à l'aide de la formule de dérivation des fonctions composées !

5. convergence et limite des suites - Algorithme de descente du gradient

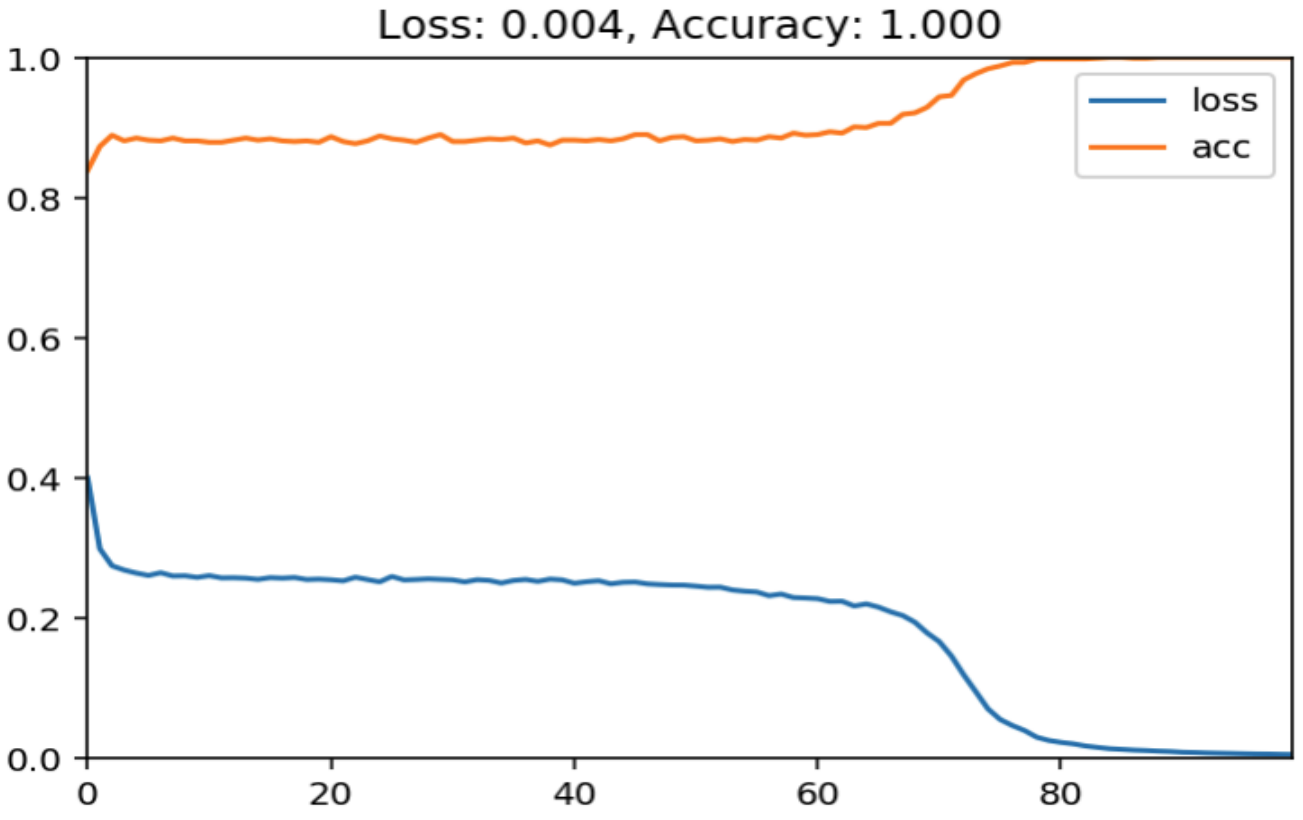
Vers les mathématiques computationnelles

Utilisation de `Python` et des libraires `TensorFlow` et `Keras`

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_15 (Dense)             (None, 4)                 12
_________________________________________________________________
dense_16 (Dense)             (None, 2)                 10
_________________________________________________________________
dense_17 (Dense)             (None, 1)                 3
=================================================================
Total params: 25
Trainable params: 25
Non-trainable params: 0
_________________________________________________________________
```

Vers les mathématiques computationnelles

Utilisation de `Python` et des libraires `TensorFlow` et `Keras`



Hasard et modèles stochastiques au coeur de l'IA [6]

Deux sources de hasard (processus aléatoires) se glissent dans l'apprentissage d'une règle de classification :

- A chaque itération, la mise à jour de $\theta^{(k)}$ vers $\theta^{(k+1)}$ se base sur une sous partie I_k des données, choisie de façon aléatoire,

$$\theta^{(k+1)} = \theta^{(k)} - \gamma \sum_{i \in I_k} \nabla F_i(\theta), \quad \text{avec} \quad F_i(\theta) = (Y_i - f(X_i, \theta))^2$$

  afin de limiter le coût calculatoire sur des données massives

- Les données de l'ensemble d'apprentissage $(X_1, Y_1), \ldots, (X_n, Y_n)$ sont des réalisations d'un couple de variables aléatoires et donc la règle de classification par minimisation de l'erreur d'apprentissage est également un objet de nature aléatoire !

6. variables aléatoires, probabilités et statistique

Complexité des réseaux de neurones profonds

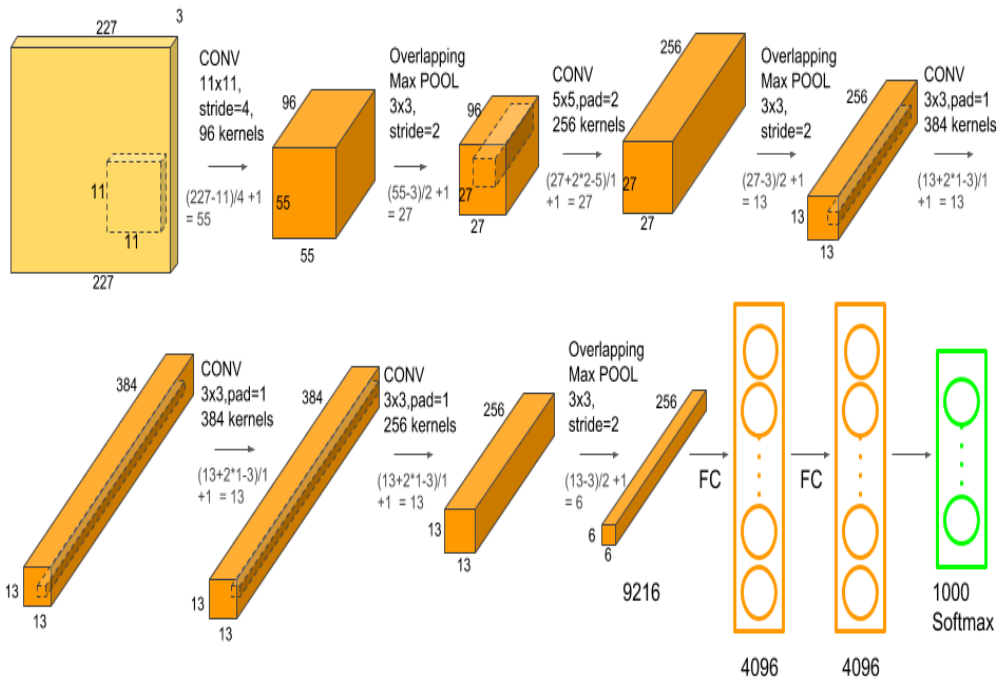
Classification d'images - ILSVRC Challenge (2010) [1]

- apprentissage : 1.2 million d'images labellisées (1000 classes)
- test : 150 000 images

1. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012)

Complexité des réseaux de neurones profonds

Deep Neural Network `AlexNet` [1]



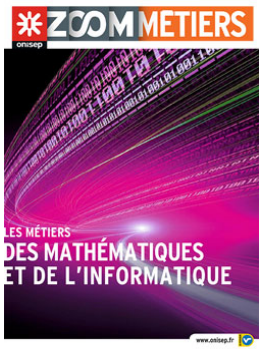
Quelle compréhension des décisions prises par l'IA ?

1. `https://www.learnopencv.com/understanding-alexnet/`

Ressources documentaires

https://www.sfds.asso.fr/fr/group/formations_et_metiers/470-zoom_sur_les_metiers/

- Brochure sur les métiers des mathématiques et de l'informatique (ONISEP / SFDS)



- Brochure sur les métiers de la statistique (ONISEP / SFDS)

Ressources documentaires

Le métier de Data Scientist !

LE *BIG DATA* : UN SECTEUR D'AVENIR ?

Le *big data* (grande quantité de données) constitue un immense réservoir d'emplois pour les années à venir. Quelque 10 000 emplois pourraient être créés dans ce domaine d'ici 2018, selon la commission « Innovation 2030 »*. Les métiers de *data scientist* (expert en données) ou *data miner* (fouilleur de données) nécessitent une grande expertise en informatique (pour récolter, stocker, indexer et sécuriser les données), mais aussi en mathématiques et statistique pour les analyser. « *Comme l'orpailleur, qui cherche une pépite dans le sable d'une rivière, le data miner traque, parmi des milliers d'informations, celle qui sera décisive pour l'avenir d'une entreprise* », explique Philippe Chabault, enseignant en IUT STID.

* Mise en place en 2013, cette commission devait déterminer les secteurs et technologies où la France serait susceptible d'occuper une position de leader à l'horizon 2030.

Tristan Launay, 31 ans, statisticien chez Google
➜ p. 12

Fabien Poulard, 30 ans, créateur et dirigeant de Dictanova
➜ p. 24

Ressources documentaires

Le métier de Data Scientist !

TRISTAN LAUNAY,
31 ANS

STATISTICIEN CHEZ GOOGLE

C'est grâce à mon master en ingénierie mathématique option statistique et probabilités que j'ai pris goût à la statistique, alors qu'au départ je voulais être prof ! J'ai poursuivi par une thèse en entreprise (3 ans chez EDF, puis dans une *start-up*). Chez Google, je réalise des études statistiques

Votre avis nous intéresse !

Les mathématiques en Licence à l'Université de Bordeaux à la base des méthodes d'apprentissage de l'IA !

Etes-vous convaincu ?

Merci de votre attention !