Fondements du Machine Learning

Chapitre 1:
Mathématiques pour le Machine Learning

Mme BEDDAD Fatima
fatima.bedad@univ-temouchent.edu.dz

2025-2026

Les mathématiques, une compétence fondamentale pour travailler dans l'IA?

Derrière le Machine Learning…

Le terme machine learning, dont les traductions varient entre apprentissage machine, apprentissage automatique et apprentissage artificiel, fait partie d'un ensemble de mots-clés qui ont récemment gagné en popularité.

Parmi ceux-ci, on trouve également l'analyse de données **(data analysis),** la fouille de données **(data mining),** l'intelligence articielle **(artificial intelligence**, ou simplement **AI**), les masses de données **(Big Data),** etc.

Machine learning pour ce cours

Ensemble de **techniques** qui visent à extraire de l'information d'un jeu de **données.**

Principe 1: Grande quantité de données

- Pour avoir de l'information à extraire;
- Pour être représentatif.

Principe 2: Utilisation d'algorithmes

- Traitement systématique et efficace;
- Théorie mathématique+implémentation.

Pourquoi s'intéresser aux données ?

Essentiel pour les entreprises

- Modèle économique des GAFA(M);
- Service gratuit mais valeur dans l'exploitation des données.

Important pour la recherche

- Quantité massive de données générées en biologie, médecine,...
- Difficultés mathématiques et informatiques.

Approches guidées par les données

Ou data-driven, drivées par les données, etc.

- Pallie le manque de modèles formels;
- Pourrait remplacer la modélisation à terme.

Exemple 1.1 (Systèmes de recommandation)

Les plate-formes commerciales telles que Netflix ou Youtube suggèrent du contenu pertinent a leurs utilisateurs en fonction de leurs préférences.
Pour ce faire, elles disposent d'une matrice d'avis : il s'agit d'un tableau en deux dimensions, l'une représentant les clients et l'autre les produits. Les grandes questions qui se posent sont donc :

1) Quels sont les éléments principaux de nos préférences ?
2) Comment gérer un grand nombre d'avis ?
3) Les avis reflètent-ils vraiment la réalité ?

On considèrera 3 approches d'analyse de données :

Des modèles linéaires de l'information dans les données;
- Les données sont vues comme des réalisations de variables aléatoires (typiquement gaussiennes).
- Souvent efficace en pratique;
- Très souvent le premier cas considéré en recherche;
- Utilise des savoirs fondamentaux en algèbre linéaire, statistiques (et optimisation).

Fondamentaux de l'apprentissage (*machine learning)*

Modèle fonctionnel :

- Distribution (de probabilité) des données connues;
- Développement de modèles ad hoc;
- Apprentissage supervisé. (exemple : **Régression linéaire).**

Modèle prédictif:

- Pas de distribution connue;
- Extraction d'information des données;
- Fréquent en apprentissage non supervisé. (exemple : **Analyse en composante principales).**

En bref

Machine learning/Apprentissage

- Décrire le comportement de données;
- Prédire les propriétés de données futures.

Notre approche

- Le cas linéaire : souvent efficace et populaire en pratique.
- Le modèle linéaire : permet de présenter les enjeux et les outils fondamentaux.

Les techniques que nous emploierons reposent sur des algorithmes, c'est-à-dire des traitements systématiques à appliquer aux données. Comme on le verra, le développement d'un algorithme efficace repose à la fois sur des arguments mathématiques et sur une implémentation bien pensée.

Voici les domaines des mathématiques qui vont nous intéresser. Au menu :

Ces 6 domaines mathématiques constituent la base du Machine Learning. Chaque matière est entrelacée pour développer notre modèle de Machine Learning et atteindre le « meilleur » modèle afin de généraliser l'ensemble de données.

Algèbre Linéaire

Qu'est-ce que l'algèbre linéaire ? C'est une branche des mathématiques qui concerne l'étude des vecteurs et des règles permettant de manipuler les vecteurs. Lorsque nous formalisons des concepts intuitifs, l'approche commune consiste à construire un ensemble d'objets (symboles) et un ensemble de règles pour manipuler ces objets. C'est ce que nous connaissons sous le nom d'*algèbre*.

Si nous parlons d'algèbre linéaire dans le Machine Learning, elle est définie comme la partie des mathématiques qui utilise l'espace vectoriel et les matrices pour représenter les **équations linéaires**

Un exemple de l'utilisation de l'algèbre linéaire est l'équation linéaire. L'algèbre linéaire est un outil utilisé dans l'équation linéaire car de nombreux problèmes peuvent être présentés systématiquement de manière linéaire. L'équation linéaire typique est présentée sous la forme ci-dessous.

$$a_{m1}\,x_1 + \ldots + a_{mn}\,x_n = b_m$$

Pour résoudre le problème d'équation linéaire ci-dessus, nous utilisons l'algèbre linéaire pour présenter l'équation linéaire dans une représentation systématique. De cette façon, nous pouvons utiliser la caractérisation matricielle pour rechercher la solution la plus optimale.

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Pour synthétiser le domaine de l'algèbre linéaire, il y a trois éléments que vous devriez connaître parfaitement comme point de départ :

- **vecteur**
- **matrice**
- **équation linéaire**

Rappels et compléments en algèbre linéaire matricielle

- Notations et résultats de base
- Valeurs propres et décomposition spectrale
- Décomposition en valeurs singulières

Des histoires de notation

Notations en algèbre linéaire

- Coefficients/Scalaires : α, β, γ;
- Vecteurs/Variables : x, y, z;
- Matrices : A, B, C.

Notations vectorielles

- $\mathbb{R}^n$: ensemble des vecteurs à $n \geq 1$ composantes réelles;
- Par convention, $\boldsymbol{x} \in \mathbb{R}^n$ est un vecteur colonne, et on note $\boldsymbol{x}^T$ le vecteur ligne correspondant.
- Pour tout $\boldsymbol{x} \in \mathbb{R}^n$ et tout $i \in \{1, \ldots, n\}$, on notera $x_i \in \mathbb{R}$ sa i-ème coordonnée (dans la base canonique de $\mathbb{R}^n$) $\Rightarrow \boldsymbol{x} = [x_i]_{1 \leq i \leq n}$.

Structure d'espace vectoriel normé

- *Addition dans* $\mathbb{R}^n$: $\boldsymbol{x} + \boldsymbol{y} := [x_i + y_i]_{1 \leq i \leq n}$;
- *Multiplication par un réel :*

$$\lambda \boldsymbol{x} \overset{n}{=} \lambda \cdot \boldsymbol{x} := [\lambda x_i]_{1 \leq i \leq n} ;$$

- *Norme euclidienne :*

$$\|\boldsymbol{x}\| := \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Notations vectorielles(suite)

Produit scalaire sur $\mathbb{R}^n$

Le produit scalaire induit par la norme euclidienne est défini pour tous vecteurs $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ par

$$\boldsymbol{x}^T \boldsymbol{y} := \sum_{i=1}^{n} x_i \, y_i.$$

C'est une forme bilinéaire symétrique définie positive (NB : $\boldsymbol{y}^T \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{y}$).

Sous-espace engendré

Soient $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ p vecteurs de $\mathbb{R}^n$. Le *sous-espace engendré par les vecteurs* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ est le sous-espace vectoriel

$$\text{vect}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) := \left\{ \boldsymbol{x} = \sum_{i=1}^{p} \alpha_i \boldsymbol{x}_i \;\middle|\; \alpha_i \in \mathbb{R}^n \; \forall i \right\}.$$

Ce sous-espace est de dimension au plus $\min\{n, p\}$.

Notations matricielles

- $\mathbb{R}^{m\times n}$: ensemble des matrices à m lignes, n colonnes à coefficients réels (on supposera toujours $m \geq 1$ et $n \geq 1$).
- NB : $\mathbb{R}^{m\times 1} \simeq \mathbb{R}^m$.

Coefficients, lignes et colonnes

Pour $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, on utilisera

- $[\boldsymbol{A}]_{ij}$ pour le coefficient (i,j) de $\boldsymbol{A}$;
- $\boldsymbol{a}_i^{\mathsf{T}}$ pour la i-ème ligne de $\boldsymbol{A}$;
- OU $\boldsymbol{a}_j$ pour la j-ème colonne de $\boldsymbol{A}$.

Les notations suivantes seront équivalentes à $\boldsymbol{A}$:

$$[\boldsymbol{A}_{ij}]_{\substack{1\leq i\leq m,\\ 1\leq j\leq n}} \quad \begin{bmatrix} \boldsymbol{a}_1^{\mathsf{T}} \\ \cdots \\ \boldsymbol{a}_m^{\mathsf{T}} \end{bmatrix} \quad \text{OU} \quad [\boldsymbol{a}_1 \cdots \boldsymbol{a}_n].$$

Transposée, symétrie

Définitions

Soit $\boldsymbol{A} = [\boldsymbol{A}_{ij}] \in \mathbb{R}^{m \times n}$ une matrice à m lignes et n colonnes.

- La *transposée de* $\boldsymbol{A}$, notée $\boldsymbol{A}^{\mathrm{T}}$, est la matrice à n lignes et m colonnes telle que

$$\forall i = 1, \ldots, m, \ \forall j = 1, \ldots, n, \qquad \left[\boldsymbol{A}^{\mathrm{T}}\right]_{ij} = \boldsymbol{A}_{ji}.$$

Cas des matrices carrées

- $\boldsymbol{A}^{\mathrm{T}} \in \mathbb{R}^{n \times n}$;
- $\boldsymbol{A}$ est dite *symétrique* si $\boldsymbol{A} = \boldsymbol{A}^{\mathrm{T}}$.

Ex) Matrices diagonales, de covariance, d'adjacence, etc.

Noyau, image, rang

Soit $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.

- Le **noyau** (*kernel/null space*) de $\boldsymbol{A}$ est le sous-espace vectoriel

$$\ker(\boldsymbol{A}) := \{\boldsymbol{x} \in \mathbb{R}^n \mid \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}_{\mathbb{R}^m}\}.$$

- L'**image** (*range space*) de $\boldsymbol{A}$ est le sous-espace vectoriel

$$\mathrm{Im}(\boldsymbol{A}) := \{\boldsymbol{y} \in \mathbb{R}^m \mid \exists \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}\}.$$

- Le **rang** (*rank*) de $\boldsymbol{A}$, noté $\mathrm{rang}(\boldsymbol{A})$, est la dimension du sous-espace vectoriel $\mathrm{Im}(\boldsymbol{A})$. On a $\mathrm{rang}(\boldsymbol{A}) \leq \min\{m, n\}$.

Théorème du rang

Pour toute matrice $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, on a

$$\dim(\ker(\boldsymbol{A})) + \mathrm{rang}(\boldsymbol{A}) = n.$$

Matrices inversibles, définies, positives

Une matrice $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ est dite *inversible* s'il existe $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ telle que $\boldsymbol{BA} = \boldsymbol{AB} = \mathbf{I}_n$, où $\mathbf{I}_n$ est la matrice identité de $\mathbb{R}^{n \times n}$.
La matrice $\boldsymbol{B}$ est alors unique : elle est appelée l'*inverse de* $\boldsymbol{A}$ et se note $\boldsymbol{A}^{-1}$.

Une matrice $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ est dite *semi-définie positive* si

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0.$$

Elle est dite *définie positive* lorsque $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$ pour tout vecteur $\boldsymbol{x}$ non nul.

Valeurs propres et vecteurs propres

Definition

Soit $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. On dit que $\lambda \in \mathbb{R}$ est une *valeur propre de* $\boldsymbol{A}$ si

$$\exists \boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{v} \neq \vec{0}, \qquad \boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}.$$

Le vecteur $\boldsymbol{v}$ est alors un *vecteur propre associé à la valeur propre* λ.
L'ensemble des valeurs propres de $\boldsymbol{A}$ s'appelle le *spectre*.

- Toute matrice de $\mathbb{R}^{n \times n}$ possède n valeurs propres complexes, mais pas nécessairement n valeurs propres réelles.
- Les valeurs propres réelles d'une matrice semi-définie positive (resp. définie positive) sont positives (resp. strictement positives).
- Le noyau de $\boldsymbol{A}$ est engendré par les vecteurs propres associés à la valeur propre 0.

Application linéaire

- Généralités

DÉFINITION (1) *Soit $(E,+,\cdot)$ et $(F,+,\cdot)$ deux $\mathbb{K}-$ espaces vectoriels et soit f une application de E dans F, on dit que f est une application linéaire si et seulement si :*

$$\forall x,y\in E,\forall\lambda\in\mathbb{K},f(x+y)=f(x)+f(y)etf(\lambda\cdot x)=\lambda\cdot f(x),$$

où d'une manière équivalente :

$$\forall x,y\in E,\forall\lambda,\mu\in\mathbb{K},f(\lambda x+\mu y)=\lambda f(x)+\mu f(y).$$

(2) *Si de plus f est bijective, on dit alors que f est un isomorphisme de E dans F.*

(3) *Une application linéaire de $(E,+,\cdot)$ dans $(E,+,\cdot)$ est dite un endomorphisme.*

(4) *Un isomorphisme de $(E,+,\cdot)$ dans $(E,+,\cdot)$ est aussi appelé un automorphisme de E dans E.*

EXEMPLE (1) *L'application*

$$f_1 : \mathbb{R}^2 \longmapsto \mathbb{R}$$
$$(x, y) \longmapsto x - y$$

est une application linéaire, car : $\forall (x, y), (x', y') \in \mathbb{R}^2, \forall \lambda, \mu \in \mathbb{R},$

$$f_1(\lambda(x, y) + \mu(x', y')) = f_1(\lambda x + \mu x', \lambda y + \mu y') = \lambda x + \mu x' - (\lambda y + \mu y')$$
$$\Rightarrow f_1(\lambda(x, y) + \mu(x', y')) = \lambda(x - y) + \mu(x' - y') = \lambda f_1(x, y) + \mu f_1(x', y').$$

(2) *L'application*

$$f_2 : \mathbb{R}^3 \longmapsto \mathbb{R}^3$$
$$(x, y, z) \mapsto (x \mid y, x \quad 5z, y)$$

est une application linéaire, car : $\forall (x, y, z), (x', y', z') \in \mathbb{R}^3, \forall \lambda, \mu \in \mathbb{R},$

$$f_2(\lambda(x, y, z) + \mu(x', y', z')) = f_2(\lambda x + \mu x', \lambda y + \mu y', \lambda z + \mu z')$$
$$\Leftrightarrow f_2(\lambda(x, y, z) + \mu(x', y', z')) = (-\lambda x - \mu x' + \lambda y + \mu y', \mu x' + \mu x' - 5\lambda y - 5\mu y', \lambda y + \mu y')$$
$$\Leftrightarrow f_2(\lambda(x, y, z) + \mu(x', y', z')) = (-\lambda x + \lambda y, \lambda x - 5\lambda z, \lambda y) + (-\mu x' + \mu y', \mu x' - 5\mu z', \lambda y').$$
$$\Leftrightarrow f_2(\lambda(x, y, z) + \mu(x', y', z')) - \lambda(-x + y, x - 5z, y) + \mu(-x' + y', x' - 5z', y') - \lambda f_2(x, y, z) + \mu f_2(x', y', z').$$

(3) *L'application*

$$f_3 : \mathbb{R} \longmapsto \mathbb{R}$$
$$x \longmapsto -3x$$

est isomorphisme, en effet, f_3 *est linèaire car :*

$$\forall x, y \in \mathbb{R}, \forall \lambda, \mu \in \mathbb{R}, f_3(\lambda x + \mu y) = -3\lambda x - 3\mu y = \lambda f_3(x) + \mu f_3(y),$$

Application linéaire

- Généralités

Définition *Soit f une application linéaire de E dans F.*

(1) *On appelle image de f et on note Imf l'ensemble défini comme suit*

$$Imf = \{y \in F / \exists x \in E : f(x) = y\} = \{f(x) / x \in E\}.$$

(2) *On appelle noyau de f et on note $\ker f$ l'ensemble défini comme suit :*

$$\ker f = \{x \in E / f(x) = O_F\},$$

On note parfois $\ker f$, par $f^{-1}(\{0\})$.

Proposition 5.6. *Si f est une application linéaire de E dans F, alors si $dimImf = n < +\infty$, alors n est appelé rang de f et on note $rg(f)$.*
Imf et $\ker f$ sont des sous espaces vectoriels de E.

EXEMPLE (1) *Déterminons le noyau de l'application* f_1,

$$\ker f = \{(x,y) \in \mathbb{R}^2 / f(x,y) = 0\} = \{(x,y) \in \mathbb{R}^2 / x+2y = 0\} = \{(x,y) \in \mathbb{R}^2 / x = -2y\}$$

ainsi

$$\ker f = \{(-2y, y)/y \in \mathbb{R}\} = \{y(-2,1)/y \in \mathbb{R}\}$$

donc le $\ker f$ *est un sous espace vectoriel engendré par* $u = (-2,1)$ *donc il est de dimension* 1, *et sa base est* $\{u\}$.

(2) *Cherchons l'image de*

$$f_2 : \mathbb{R}^3 \longmapsto \mathbb{R}^3$$

$$(x,y,z) \longmapsto (-x+y, x-z, y)$$

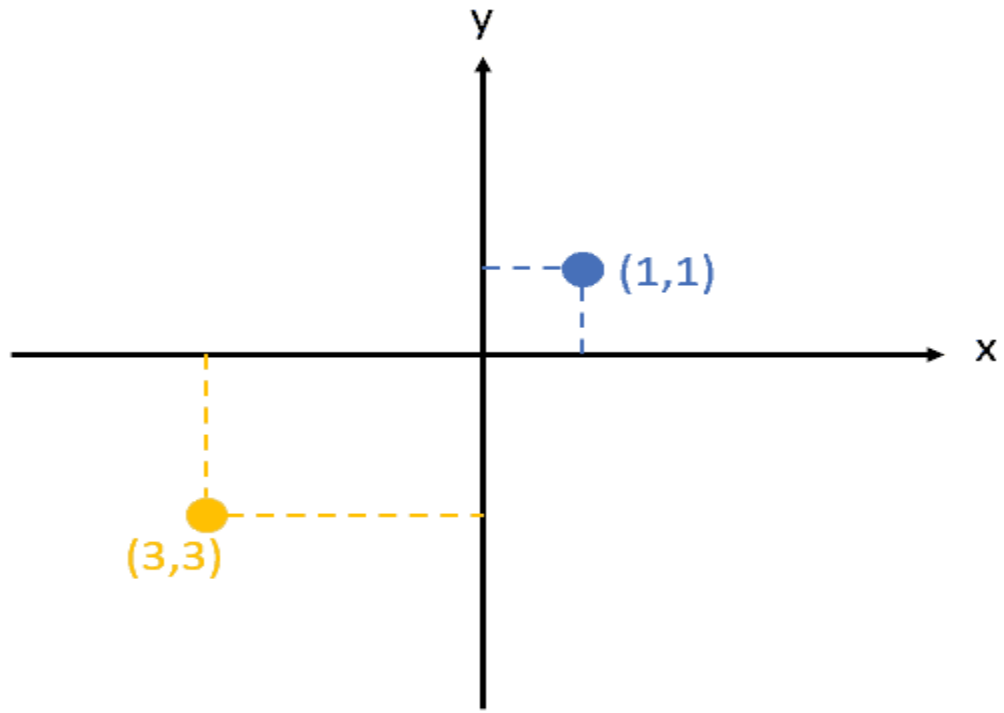
$$Imf_2 = \{f(x,y,z)/(x,y,z) \in \mathbb{R}^3\} = \{(-x+y, x-z, y)/(x,y,z) \in \mathbb{R}^3\}$$

$$Imf_2 = \{x(-1,1,0) + y(1,0,1) + z(0,-1,0)/(x,y,z) \in \mathbb{R}^3\}$$

donc Imf_2 *est un s.e.v de* $\mathbb{R}^3$ *engendré par* $\{(-1,1,0),(1,0,1),(0,-1,0)\}$ *il est facile de montrer que cette famille est libre et donc il forment une base de* $\mathbb{R}^3$ *donc* $dimImf_2 = 3, rg(f_2) = 3, Imf = \mathbb{R}^3$.


Algèbre Linéaire
Géométrie Analytique
Décomposition Matricielle
Calcul Vectoriel
Probabilité et Distributions
Optimisation

La **géométrie analytique** (ou géométrie des coordonnées) est un domaine dans lequel nous utilisons la position des (points de) données à l'aide d'une paire ordonnée de coordonnées. Ce domaine porte sur la définition et la représentation numérique des formes géométriques et sur l'extraction d'informations quantifiables à partir de ces formes. De façon plus simple, nous projetons des données dans un plan, et puis nous recevons des informations numériques à partir de cela.



L’exemple ci-dessus montre comment nous avons obtenu des informations à partir du point de données en projetant l’ensemble des données dans le plan. La façon dont nous acquérons les informations à partir de cette représentation est le cœur de la géométrie analytique.

Pour vous aider à commencer à apprendre ce sujet, voici quelques termes importants dont vous pourriez avoir besoin.

Fonction distance

Une **fonction distance** est une fonction qui fournit des informations numériques sur la distance entre les éléments d'un ensemble. Si la distance est nulle, alors les éléments sont équivalents. Dans le cas contraire, ils sont différents les uns des autres.

Un exemple de fonction distance est la distance euclidienne qui calcule la distance linéaire entre deux points de données.

$$\sqrt{(q_2 - q_1)^2 + (p_2 - p_1)^2}$$

Produit scalaire

Le produit scalaire est un concept qui introduit des notions géométriques intuitives, telles que **la longueur d'un vecteur** et **l'angle** ou **la distance** entre deux vecteurs. Il est souvent noté ⟨x,y⟩ (ou parfois (x,y) ou ⟨x|y⟩).


Algèbre Linéaire
Géométrie Analytique
Décomposition Matricielle
Calcul Vectoriel
Probabilité et Distributions
Optimisation

La **décomposition matricielle** est le domaine qui concerne la manière de réduire une matrice. La décomposition matricielle vise à simplifier les opérations matricielles plus complexes sur la matrice décomposée plutôt que sur sa matrice d'origine.

Une analogie courante pour la **décomposition matricielle** est la factorisation de nombres, comme la factorisation de 8 en 2 x 4. C'est pourquoi la décomposition matricielle est synonyme de **factorisation de matrice**. Il existe de nombreuses façons de décomposer une matrice, et donc un ensemble de techniques de décomposition matricielle varié.

Un exemple est la décomposition LU ci-dessous. Il s'agit d'une méthode de décomposition d'une matrice comme produit d'une matrice triangulaire inférieure *L* (comme *lower*, inférieure en anglais) par une matrice triangulaire supérieure *U* (comme *upper*, supérieure).

$$\begin{bmatrix} 5 & 7 \\ 8 & 2 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}$$

Décomposition en valeurs propres

Théorème

Toute matrice $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ symétrique admet une décomposition dite **spectrale** de la forme :

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1},$$

avec

- $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ matrice orthogonale ($\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{P}^{-1}$), dont les colonnes $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ forment une *base orthonormée* de vecteurs propres.
- $\boldsymbol{\Lambda}$ matrice diagonale, qui contient les n valeurs propres de $\boldsymbol{A}$ $\lambda_1, \ldots, \lambda_n$ sur la diagonale.

- Il n'y a pas unicité de la décomposition spectrale;
- Aux permutations près, l'ensemble des valeurs propres est unique.

Sur la décomposition spectrale

Importance de la décomposition

Chaque valeur propre λ_i caractérise l'effet de $\boldsymbol{A}$ sur le vecteur $\boldsymbol{p}_i$:

- $|\lambda_i| >> 1 \Rightarrow \|\boldsymbol{A}\boldsymbol{p}_i\| >> \|\boldsymbol{p}_i\|$;
- $|\lambda_i| << 1 \Rightarrow \|\boldsymbol{A}\boldsymbol{p}_i\| << \|\boldsymbol{p}_i\|$.

Interprétation géométrique

L'action de la matrice $\boldsymbol{A}$ sur un vecteur $\boldsymbol{x}$

- Allonge les composantes de $\boldsymbol{x}$ dans la base des vecteurs propres associées aux plus grandes valeurs propres en valeur absolue;
- Réduit les composantes de $\boldsymbol{x}$ associées aux valeurs propres les plus faibles en valeur absolue;
- Cas extrême : si $\ker(\boldsymbol{A}) \neq \{\boldsymbol{0}\}$, toute composante de $\boldsymbol{x}$ selon un vecteur du noyau est réduite à zéro par $\boldsymbol{A}$.

Le cas des matrices rectangulaires

Notion de valeur propre

Soit $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. On peut parler :

- des valeurs propres de $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} \in \mathbb{R}^{n \times n}$;
- des valeurs propres de $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} \in \mathbb{R}^{m \times m}$.

Peut-on s'en servir pour obtenir une décomposition de $\boldsymbol{A}$?

Observations concernant $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$

- $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ est semi-définie positive;
- $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ est symétrique.
- $\ker(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \ker(\boldsymbol{A})$;
- $\mathrm{Im}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \mathrm{Im}(\boldsymbol{A}^{\mathrm{T}})$;
- $\mathrm{rang}(\boldsymbol{A}^{T}\boldsymbol{A}) = \mathrm{rang}(\boldsymbol{A})$.

(On a des résultats similaires pour $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$.)

Décomposition en valeurs singulières

Théorème

Toute matrice $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ admet une **décomposition en valeurs singulières** (*SVD*) de la forme

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}.$$

- $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ est orthogonale;
- $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ est orthogonale;
- $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ est diagonale par blocs, avec des coefficients nuls sauf ceux de la diagonale $\{[\boldsymbol{\Sigma}]_{ii}\}_i$ qui sont positifs (ou nuls). Ces éléments, notés $\{\sigma_i\}$, s'appellent les **valeurs singulières de** $\boldsymbol{A}$.

- $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \Leftrightarrow \boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{\Sigma}$;
- Une *SVD* n'est pas définie de façon unique mais ses valeurs singulières le sont.

Le principal intérêt de la décomposition en valeurs singulières est de permettre de compresser la représentation de données matricielles.

la décomposition en valeurs singulières, une technique visant à extraire de l'information d'un jeu de données (exemple matrice) : ce paradigme est celui de l'apprentissage non supervisé.

Ce chapitre aborde un autre paradigme, celui de l'apprentissage supervisé, dans lequel il s'agira de déterminer un modèle (une fonction linéaire pour les besoins de ce cours) décrivant une relation entre différents éléments d'un jeu de données, pour potentiellement prédire (on parle également d'inférer) le comportement de données futures.


Algèbre Linéaire
Géométrie Analytique
Décomposition Matricielle
Calcul Vectoriel
Probabilité et Distributions
Optimisation

Calcul vectoriel

Le **calcul** est un domaine mathématique qui concerne la variation continue, qui consiste principalement en des fonctions et des limites. Le **calcul vectoriel** lui-même s'intéresse à la différenciation et à l'intégration des **champs vectoriels**. Le calcul vectoriel est souvent appelé **calcul multi varié,** bien que son cas d'étude soit légèrement différent. Le calcul multi varié traite des fonctions d'application du calcul des multiples variables indépendantes.

Il y a quelques termes importants que vous devez connaître lorsque vous commencez à apprendre le calcul vectoriel :

Dérivée et différentiation

La **dérivée** est une fonction de nombres réels qui mesure la variation de la valeur de la fonction (valeur de sortie) concernant une variation de son argument (valeur d'entrée).
La **différentiation** est l'action de calculer une dérivée.

$$m = \frac{\text{variation de } y}{\text{variation de } x} = \frac{\Delta y}{\Delta x}$$

Dérivée partielle

La **dérivée partielle** est une fonction dérivée où plusieurs variables sont calculées à l'intérieur de la fonction dérivée par rapport à l'une de ces variables qui peut varier, sachant que les autres variables sont maintenues constantes (par opposition à la dérivée totale, dans laquelle toutes les variables peuvent varier).

Gradient

Le **gradient** est un mot lié à la dérivée ou au taux de variation d'une fonction (vous pourriez considérer le gradient comme est un mot fantaisiste pour dérivée). Le terme gradient est généralement utilisé pour les fonctions ayant plusieurs entrées et une seule sortie (scalaire). Le **gradient a une direction pour se déplacer** à partir de son emplacement actuel, par exemple, vers le haut, le bas, la droite, la gauche…


Algèbre Linéaire
Géométrie Analytique
Décomposition Matricielle
Calcul Vectoriel
Probabilité et Distributions
Optimisation

Probabilité et Distributions

La **probabilité** est l'étude de l'incertitude (au sens large). La probabilité peut être considérée ici comme le moment où l'événement se produit ou le degré de croyance de l'occurrence d'un événement. La distribution de probabilité est une fonction qui mesure la probabilité d'un résultat particulier (ou d'un ensemble de résultats) qui se produirait, associé à la variable aléatoire. La **fonction de distribution de probabilité** courante est illustrée dans l'image ci-dessous.

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi * \sigma^2}} * e^{-\frac{1}{2} *(\frac{x-\mu}{\sigma})^2}$$

La **théorie des probabilités** et les statistiques sont souvent associées à une chose similaire, mais elles concernent des aspects différents de l'incertitude :

- En **mathématiques**, nous définissons la probabilité comme un modèle où les variables aléatoires capturent l'incertitude sous-jacente, et nous utilisons les règles de probabilité pour résumer ce qui se passe.

- En **statistiques**, nous essayons de comprendre le processus sous-jacent qui observe quelque chose qui s'est produit et tente d'expliquer les observations.

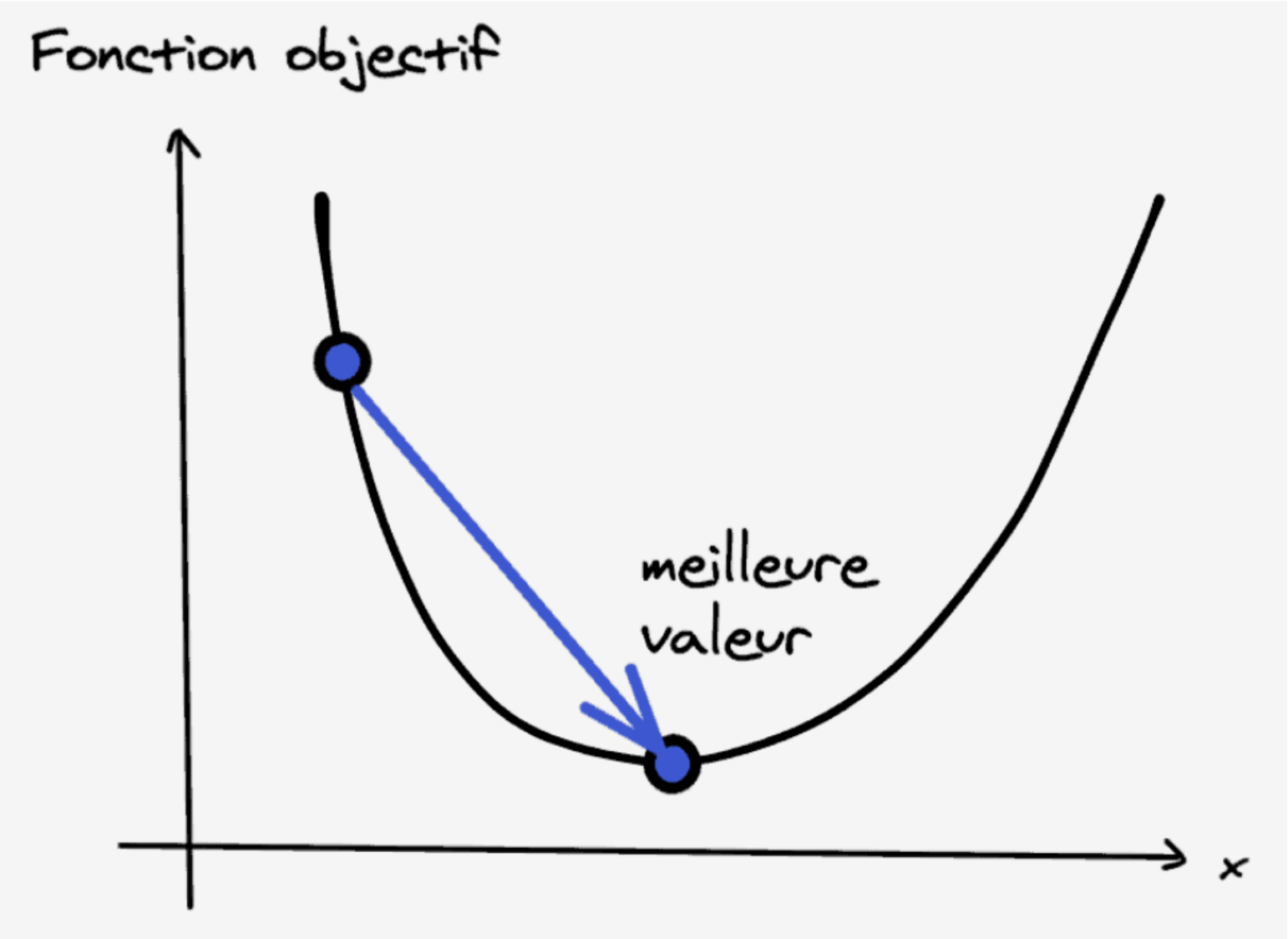
Lorsque l'on parle de **Machine Learning**, il est proche des statistiques car son objectif est de construire un modèle qui représente de manière adéquate le processus qui a généré les données.


Algèbre Linéaire
Géométrie Analytique
Décomposition Matricielle
Calcul Vectoriel
Probabilité et Distributions
Optimisation

Optimisation

Dans l'objectif d'apprentissage, l'entraînement d'un modèle de **Machine Learning** consiste à trouver un bon ensemble de paramètres. Ce que nous considérons comme **« bon »** est déterminé par la **« fonction objectif »** ou les modèles probabilistes. C'est à cela que servent les algorithmes d'optimisation (étant donné **une fonction objectif**, nous essayons de trouver la meilleure valeur).

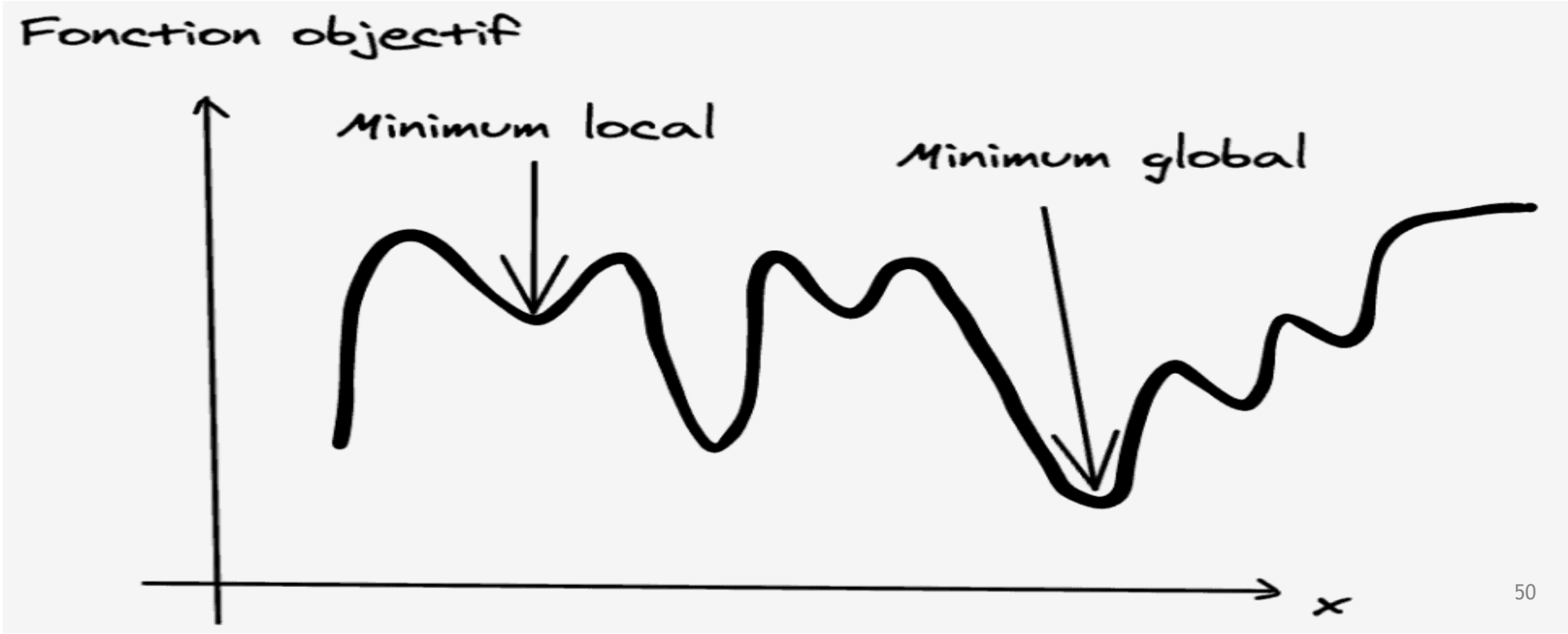
En général, **les fonctions objectifs** dans le Machine Learning essaient de minimiser la fonction. Cela signifie que la meilleure valeur est la valeur minimale. Intuitivement, si nous essayons de trouver la meilleure valeur, cela reviendrait à trouver les vallées de la fonction objectif là où les gradients nous font remonter. C'est pourquoi nous voulons nous déplacer en descendant (à l'opposé du gradient) et espérer trouver le point le plus bas (le plus profond). C'est le concept de **descente de gradient**.



Il existe quelques termes à connaître pour apprendre l'optimisation :

Minima locaux et minima globaux

Le point où une fonction prend la valeur minimale est appelé le **minimum global**. Cependant, lorsque l'objectif est de minimiser la fonction et de la résoudre à l'aide d'algorithmes d'optimisation tels que la **descente de gradient**, la fonction pourrait avoir une valeur minimale en différents points. Ces différents points qui semblent être des minima mais qui ne sont pas le point où la fonction prend réellement la valeur minimale sont appelés des **minima locaux**.



Optimisation sans contrainte et optimisation avec contrainte

L'**optimisation sans contrainte** est une fonction d'optimisation dans laquelle nous trouvons le minimum d'une fonction en supposant que les paramètres peuvent prendre n'importe quelle valeur possible (aucune limitation des paramètres). L'optimisation avec contraintes limite simplement la valeur possible en introduisant un ensemble de contraintes.

La **descente de gradient** est une optimisation sans contrainte s'il n'y a pas de limitation des paramètres. Si nous fixons une certaine limite, par exemple, $x > 1$, il s'agit d'une optimisation sans contrainte.

Conclusion

Le **Machine Learning** est un outil quotidien que les spécialistes des données ou data scientists utilisent pour obtenir de précieux modèles dont nous avons besoin. Apprendre les mathématiques du Machine Learning peut vous donner un réel avantage dans votre travail. Il existe de nombreux domaines mathématiques, mais il y a **6 domaines** qui comptent le plus lorsque nous commençons à apprendre les mathématiques du Machine Learning, qui sont :

- Algèbre Linéaire
- Géométrie Analytique
- Décomposition Matricielle
- Calcul Vectoriel
- Probabilité et Distributions
- Optimisation