1 Organisation

Moodle

Toutes les informations sont sur le Moodle. À consulter régulièrement.

Ouvrage de référence

- **Foundations of Data Science** par Blum, Hopcroft, Kannan, Cambridge University Press, 2020. Disponible à la librairie.
- **Data Mining and Machine Learning: Fundamental Concepts and Algorithms** par Mohammed J. Zaki, Wagner Meira, Gr. Cambridge University Press, 2nd Ed., 2020. Disponible à la librairie.

2 Algèbre linéaire

Concepts généraux

- Qu'est-ce qu'un **vecteur / espace vectoriel** ?
- Qu'est ce que le **produit scalaire** représente ?
- Pourquoi la **projection** est un concept important ?
- Qu'est ce qui est critique par rapport au **age things** ?

2.1 Vecteurs et espace vectoriel

Définition d'un espace vectoriel

Soit un champ $K(\mathbb{Q}$ ou $\mathbb{C})$ et un ensemble V. Un espace vectoriel sur K, est un ensemble E, dont les éléments sont appelés vecteurs muni de deux lois :

- Une loi de composition interne «+»: $E^2 \to E$, appelée addition ou somme vectorielle
- Une loi de composition externe à gauche «·»: $K \times E \to E$, appelée multiplication par un scalaire

tel que les propriétés suivantes soient vérifiées :

1. $(E, +)$ est un groupe abélien, autrement dit :
   - la loi «+»est commutative,
   - elle est associative,
   - elle admet un élément neutre $\vec{0_E}$ appelé vecteur nul et,
   - tout vecteur v a un opposé, noté $-v$.

   C'est-à-dire que pour tous vecteurs $\vec{u}$, $\vec{v}$ et $\vec{w}$ de E:

$$\begin{aligned} \vec{u} + \vec{v} &= \vec{v} + \vec{u} \\ \vec{u} + (\vec{v} + \vec{w}) &= (\vec{u} + \vec{v}) + \vec{w} \\ \vec{0_E} + \vec{v} &= \vec{v} \\ \vec{u} + (-\vec{u}) &= \vec{0_E} \end{aligned} \tag{1}$$

2. La loi «·»vérifie les propriétés suivantes :
   - elle est distributive à gauche par rapport à la loi «+de E et à droite par rapport à l'addition du corps K,
   - elle vérifie une associativité mixte (par rapport à la multiplication dans K),
   - l'élément neutre multiplicatif du corps K, noté $\vec{1_K}$, est neutre à gauche pour $\cdot$.

   C'est-à-dire que pour tous vecteur $\vec{u}$, $\vec{v}$ et $\vec{w}$ de E, et tous scalaire λ, μ :

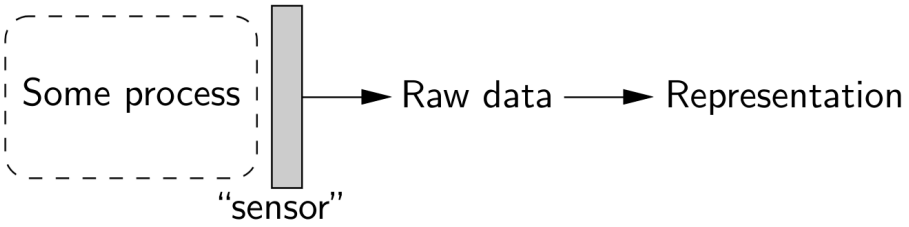
$$\begin{aligned} \lambda \cdot \vec{u} + \vec{v} &= (\lambda \cdot \vec{u}) + (\lambda \cdot \vec{v}) \\ (\lambda + \mu) \cdot \vec{u} &= (\lambda \cdot \vec{u}) + (\mu \cdot \vec{u}) \\ (\lambda\mu) \cdot \vec{u} &= \lambda \cdot (\mu \cdot \vec{u}) \\ 1_K \cdot \vec{u} &= \vec{u} \end{aligned} \tag{2}$$

Ces axiomes impliquent que E est non vide et pour tout vecteur $\vec{u}$ de E et tout scalaire λ :

$$\begin{aligned} \lambda \cdot \vec{u} = \vec{0_E} \Leftrightarrow (\lambda = \vec{0_K} \text{ ou } \vec{u} = \vec{0_E}) \\ (-\lambda) \cdot \vec{u} = -(\lambda \cdot \vec{u}) = \lambda \cdot (-\vec{u}) \end{aligned} \tag{3}$$

3 Représentation des espace à haute dimensionnalité

Flot de données et représentation Le flot typique des données vient d'un capteur, qui produit des données brutes qui doivent ensuite être représentée d'une façon ou d'une autre.



Capteur	Données brutes	Représentation
Caméra	Pixels de l'image	Matrice
Population	Résultats de vote	Matrice
Document texte	Fréquence de mots	Matrice
Réseaux sociaux	Relations	Matrice
Environnement	Mesures	Matrice

On obtient donc des données $\mathcal{X} = \{x_1, \cdots, x_n\}$ où $x_i \in \Omega \subseteq \mathbb{R}^D \quad \forall i \in [\mathbb{N}]$.

Vue statistique $\mathcal{X}$ est un échantillon de l'ensemble de taille N à D-dimensions où x_i correspond à une mesure de la variable aléatoire associée (p.e. Température, humidité). Chaque x_i peut être associé à une PDF, comme par exemple un lancé d'un dé-6 étant limité aux valeurs $[1, 6]$ où chaque face à une probabilité de $\frac{1}{6}$.
Il est donc possible d'associé une probabilité d'apparition pour chaque valeur de l'ensemble de donnée.

Vue algébrique Soit $\mathcal{X}$, on forme une matrice $X = \begin{pmatrix} | & & | \\ x_1 & \cdots & x_N \\ | & & | \end{pmatrix} \in \mathbb{R}^{D \times N}$ dont les colonnes $X_{\cdot j} = x_j$ sont les vecteurs de données.

Un espace de représentation a des propriété bien définies qui sont les suivantes :

- L'espace de représentation Ω est généralement un sous-ensemble de $\mathbb{R}^D$
- Ω est un **espace vectoriel** et peut être créer à partir d'un **espace de produit scalaire**. Dans ce cas, le produit scalaire $\langle \cdot, \cdot \rangle$ induit une norme $\| \cdot \|$, qui est elle-même induit une function de distance $d(\cdot, \cdot)$
- C'est donc le cas favorable où (Ω, d) est un **espace métrique** sur lequel un **apprentissage** peut être fait.

Cependant, certaines questions surviennent :

- Que ce passe-t-il lorsque D et/ou N augmente ?
- Comment cela impacte la modélisation de données et comment y faire face ?
- Comment cela impacte l'apprentissage ? La fameuse **Curse of Dimensionality**.

Le cours va porter sur les différentes vues possible (statistique et algébrique), et aussi comprendre ce qu'il se passe quand le nombre de dimension augmente (D et N dans $\mathbb{R}^{D\times N}$)

quand on a un nombre N fix (samples) et que l'on veut augmenter le nombre de bins, la stabilité de la représentation se dégradera de façon exponentielle (slide 9)

D (nombre de dimension ou features) doit être très contrôlé, car si ce nombre augmenter, les données deviennent exponentiellement instable

KNN c'est le fait de choisir les K voisins les plus proches, εNN c'est le fait de choisir tout les voisin dans un périmètre ε.