Algèbre linéaire et applications aux sciences des données

28 novembre 2024

Cette semaine

CM : Applications

TD : (demain) : Correction CC + TD Espaces euclidiens

Vecteurs de $\mathbb{R}^n$, normes et produits scalaires

<u>Cadre de travail</u> : $\mathbb{R}^n$ en tant qu'espace euclidien

Produit scalaire : $x^T y = \sum_{i=1}^{n} x_i y_i$ $\quad \forall x \in \mathbb{R}^n$, $\forall y \in \mathbb{R}^n$

Forme quadratique associée $x^T x = \sum_{i=1}^{n} x_i^2$

Norme associée : $\|x\| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{n} x_i^2}$

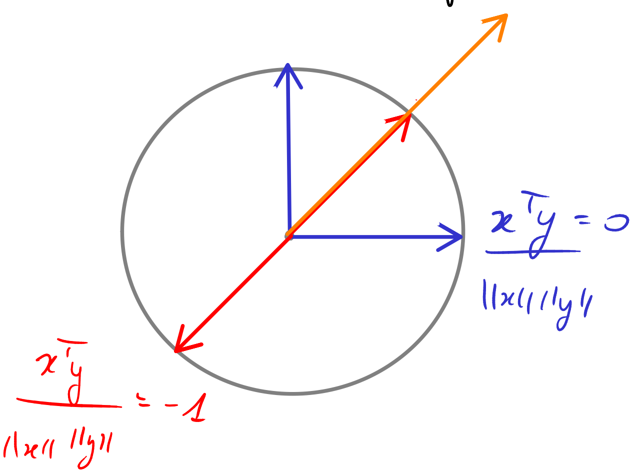
Etant donnés deux vecteurs x et y de $\mathbb{R}^n$, on peut utiliser pour les comparer

a) $\|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \geq 0$ avec égalité ssi $x = y$

b) $\dfrac{x^T y}{\|x\| \, \|y\|} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{j=1}^{n} y_j^2}} \in [-1, 1]$ $\quad \left(\begin{array}{l} = 0 \text{ ssi } x \text{ et } y \text{ sont orthogonaux} \\ = \pm 1 \text{ ssi } x \text{ et } y \text{ sont colinéaires} \end{array} \right.$

$\left(\dfrac{x}{\|x\|}\right)^T \left(\dfrac{y}{\|y\|}\right)$

vecteur y "normalisé"



<u>Def</u> : La similarité cosinus ou la distance angulaire entre deux vecteurs non nuls x et y de $\mathbb{R}^n$ est définie par

$$\frac{x^T y}{\|x\| \|y\|} \in [-1, 1]$$

On dit que la valeur $\underbrace{\arccos\left(\frac{x^T y}{\|x\|\,\|y\|}\right)}_{\in [0, \pi] \text{ par convention}}$ est l'angle entre les vecteurs x et y.

- Comparer des vecteurs en utilisant la distance euclidienne $\|x-y\|$

↳ Approche dite des plus proches voisins

$(x_1, \dots, x_N)$ N points de $\mathbb{R}^n$

__But__ : Regrouper en k groupes (clustering)

__Processus__ : Partant de $(z_1, \dots, z_k)$ k points de $\mathbb{R}^n$

- Pour chaque x_i, définir le plus proche voisin comme $z_{c_i} \in \{z_1, \dots, z_k\}$ tel que

$$\|x_i - z_{c_i}\| \leq \|x_i - z_j\| \quad \forall j = 1..k$$

- Mettre à jour les z_j en fonction des x_i dont ils sont les plus proches voisins

$$\forall j = 1..k, \quad z_j \leftarrow \frac{1}{m_j} \sum_{i \in I_j} x_i$$

avec $I_j = \{ i \in \{1, \dots, N\} \mid z_{c_i} = z_j \}$ $\quad |I_j| = m_j$

- Comparer des vecteurs en utilisant la distance angulaire

→ Technique de base en analyse de texte

<u>Cadre général</u> : Corpus de p documents (textes
- article Wikipédia
- syllabus de cours)

Dictionnaire de n mots

Pour chaque document, on définit le vecteur $x_i \in \mathbb{R}^n$ qui contient les occurrences des mots du dictionnaire dans le document i $(x_i \in \mathbb{N}^n)$

- Pour évaluer la similarité entre des documents, on regarde $\dfrac{x_i^T x_j}{\|x_i\| \|x_j\|} \in [0,1]$ car les x_i sont à coefficients positifs

$\dfrac{x_i^T x_j}{\|x_i\| \|x_j\|} \simeq 1 \Rightarrow$ Documents similaires

$\dfrac{x_i^T x_j}{\|x_i\| \|x_j\|} \simeq 0 \Rightarrow$ Documents sont "orthogonaux" (traitent de sujets différents au sens du dictionnaire)

- On peut aussi effectuer des requêtes
Étant donné $q \in \mathbb{N}^n$ et $\tau \in (0,1]$, les documents du corpus les plus similaires à celui représenté par q sont les vecteurs x_i tels que $\dfrac{x_i^T q}{\|x_i\| \|q\|} \geqslant \tau$.

Exemples : Soient 4 résumés de cours disponibles sur le site de Dauphine

2. Fondements du $\overset{\text{Machine Learning}}{\text{ML}}$: (L3)
1. Algèbre linéaire et applications aux Sciences des données (DL2)
3. Mathématiques pour les sciences des données (M1)
4. Programmation stochastique (M2 MODO)

$p = 4$

Dictionnaire : { "données", "algèbre linéaire", "statistiques", "optimisation" }

$m = 4$

→ Pour construire les vecteurs x_i, on compte le nombre d'occurrences de chacun des mots du dictionnaire dans le résumé du cours

	1 Algèbre linéaire et applications aux SD	2 Fondements du ML	3 Maths pour les SD	4 Programmation stochastique
"données"	2	3	3	0
"algèbre linéaire"	2	2	1	0
"statistiques"	0	2	3	1
"optimisation"	0	1	4	3

$$x_1 = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$

$\Rightarrow$ on obtient $x_1 = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix}$ $x_2 = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 1 \end{bmatrix}$, $x_3 = \begin{bmatrix} 3 \\ 1 \\ 3 \\ 4 \end{bmatrix}$, $x_4 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 3 \end{bmatrix}$

On calcule l'ensemble des $\dfrac{x_i^T x_j}{\|x_i\| \|x_j\|}$ pour $\{i,j\} \in \{1,2,3,4\}^2$

$\Rightarrow$ On présente en général cela sous la forme d'une matrice

$\dfrac{x_1^T x_1}{\|x_1\| \|x_1\|}$ $\qquad$ $\dfrac{x_1^T x_4}{\|x_1\| \|x_4\|}$

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ \end{array}$$

$$\begin{array}{c|cccc} & x_1 & x_2 & x_3 & x_4 \\ \hline x_1 & 1 & 0.83 & 0.48 & 0 \\ x_2 & 0.83 & 1 & 0.84 & 0.37 \\ x_3 & 0.48 & 0.84 & 1 & 0.80 \\ x_4 & 0 & 0.37 & 0.80 & 1 \end{array}$$

NB : Par symétrie du produit scalaire, $\dfrac{x_i^T x_j}{\|x_i\| \|x_j\|} = \dfrac{x_j^T x_i}{\|x_j\| \|x_i\|}$

Interprétation de la 1re ligne

$\rightarrow$ x_2 est le vecteur le plus "semblable" (au sens de la similarité cosinus) à x_1. On interprète cela comme "Le (descriptif du) cours Fondements du ML est le plus semblable à celui du cours Algèbre linéaire et applications aux sciences des données" (c'est vrai !)

$\rightarrow$ x_4 est orthogonal à x_1. On en déduit que "Le (descriptif du) cours Programmation stochastique ne couvre par les mêmes sujets que celui du cours d'Algèbre linéaire " (c'est vrai !)

NB :

$$\|x_1 - x_2\| = 2.45$$
$$\|x_1 - x_3\| = 5.2$$
$$\|x_1 - x_4\| = 4.24$$

⟹ Au sens de la norme, x_2 est le vecteur le plus proche de x_1 mais x_4 est plus proche que x_3

⟹ Clairement moins pertinent

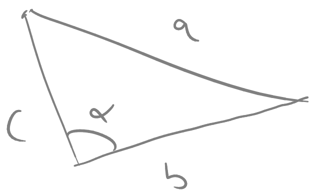
⟹ Perte de l'information d'orthogonalité

A retenir

La distance euclidienne et la similarité cosinus ne permettent pas toujours les mêmes interprétations et donc il peut arriver que l'une de ces mesures soit plus utile que l'autre.

⟹ Il y a quand même entre les deux mesures

$$\forall (x,y) \in \mathbb{R}^n, \quad \|x-y\|^2 = \|x\|^2 - 2\|x\|\|y\| \frac{x^T y}{\|x\|\|y\|} + \|y\|^2$$



$$a^2 = b^2 + c^2 - 2bc(\cos\alpha)$$

(Al-Kashi)

Remarques

- La similarité cosinus est invariante par rapport à la norme

$\forall (x,y) \in \mathbb{R}^n, \quad \forall \lambda > 0,$

$$\frac{(\lambda x)^T(\lambda y)}{\|\lambda x\| \, \|\lambda y\|} = \frac{x^T y}{\|x\| \, \|y\|}$$

(alors que $\|\lambda x - \lambda y\| = \lambda \|x-y\|$)

A venir : Pour $x \in \mathbb{R}^n$ fixé, comment projeter x sur un sous-espace de $\mathbb{R}^n$?

Réponse : Trouver le y du sous-espace qui minimise $\|x-y\|$