Université Abdelmalek Essaadi

Faculté des sciences de Tétouan

Licence d'Excellence

Machine Learning et Intelligence Artificielle

Intitulé du Module :
Unsupervised Learning Algorithms

Chapitre 2 :
Métriques de similarité

Pr. HARCHLI FIDAE

A.U : 2024 / 2025

Objectifs du chapitre

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Comprendre le concept de similarité et son importance en apprentissage non supervisé.
- Découvrir différentes métriques de similarité adaptées aux types de données.
- Analyser les avantages, inconvénients et applications des métriques de similarité.
- Apprendre à implémenter ces métriques dans des applications pratiques.

Plan

- Motivation
- Similarité et Dissimilarité
- Mesures pour les données numériques
- Mesures pour les données catégorielles
- Mesures pour les données mixtes
- Mesures de similarité et de dissimilarité entre les clusters

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Motivation (1)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Une plateforme de streaming (comme **NETFLIX**) utilise des métriques de similarité pour analyser les préférences des utilisateurs et recommander des films ou séries susceptibles de leur plaire.
- Cette plateforme veut recommander des films à ses utilisateurs en fonction de leurs préférences Les utilisateurs notent les films sur une échelle de 1 à 5. La tâche est de mesurer la similarité entre les utilisateurs ou les films pour générer des recommandations.

Données (simplifiées) :

Utilisateur\Film	Film A	Film B	Film C	Film D
Alice	5	4	0	0
Bob	5	0	4	0
Charlie	0	0	4	5

- `0` signifie que l'utilisateur n'a pas encore vu ce film.

Motivation (1)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Calculons la distance entre **Alice** et **Bob** et entre **Alice** et **Charlie** par deux mesures différentes :
  - **Distance Euclidienne : (Théorème de Pythagore)**

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

  - **Similarité Cosinus :**

$$\mathrm{Sim}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Motivation (1)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- $d(Alice, Bob) = 5.66$, $d(Alice, Charlie) = 9.06$
  - **Conclusion :**
    - Alice est plus proche de Bob que de Charlie (distance plus petite).
- $Sim(Alice, Bob) = 0.61$, $\mathrm{Sim}(\mathrm{Alice}, \mathrm{Charlie}) = 0$
  - **Conclusion :**
    - Selon la similarité cosinus, Alice et Charlie n'ont aucune similarité, tandis qu'Alice et Bob ont une similarité modérée (0.61).

Motivation (1)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Analyse du Résultat

- **Avec la distance Euclidienne :**
  - Alice semble plus proche de Bob que de Charlie, mais cette mesure est affectée par les zéros dans les données. Les films non notés augmentent artificiellement la distance.
- **Avec la similarité Cosinus :**
  - La similarité met en évidence que Bob et Alice ont des goûts similaires (ils ont donné une note élevée au **Film A**). Cependant, elle ne considère pas que Bob et Charlie ont également des goûts similaires sur des films différents (**Film C** et **Film D**).

Motivation (1)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Conclusion

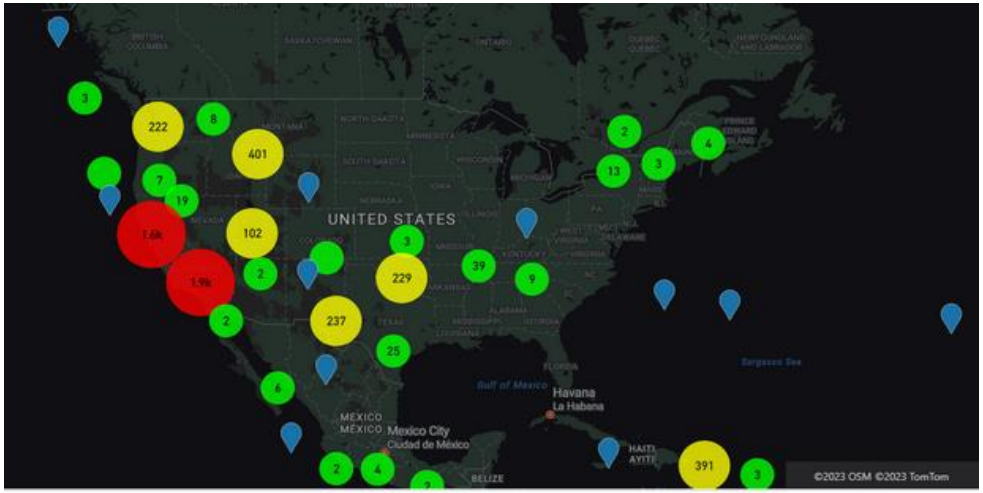
- Le choix de la mesure dépend du **contexte** :
  - Si l'objectif est de comparer les notes globales (peu importe quels films ont été vus), la **distance Euclidienne** peut suffire.
  - Si l'objectif est de comparer des **schémas de notation indépendamment de l'intensité des notes,** la **similarité Cosinus** est plus appropriée.

Dans ce cas, la similarité Cosinus est meilleure car elle ignore les zéros et se concentre sur les préférences communes.

Motivation (2)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Une entreprise de livraison veut trouver le **dépôt le plus proche** pour livrer une commande. Les rues de la ville sont disposées en **grille orthogonale** (comme beaucoup de grandes villes). Les dépôts et les points de livraison sont repérés par leurs coordonnées (x,y).
- **Quelle mesure peut-on utiliser pour identifier les points de livraison les plus proches ?**
- Comparant deux distance sur un petit exemple : la distance euclidienne et la distance de Manhattan.



Motivation (2)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- **Distance de Manhattan :**

$$d_{\text{Manhattan}}(A, B) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Points donnés :**
  - Position des **dépôts** :
    - Dépôt 1 : (2,3)
    - Dépôt 2 : (5,1)
  - Position du **client** :
    - Client : (4,4)

- **Calcul des distances :**

Distance	Client ↔ Dépôt 1	Client ↔ Dépôt 2
Euclidienne	2.24	3.16
Manhattan	3	4

Motivation (2)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Analyse du Résultat

- **Manhattan est souvent plus réaliste** dans un environnement urbain où :
  - Les déplacements sont contraints par des rues en grille.
  - Les véhicules doivent suivre les routes (pas de ligne droite possible comme en Euclidienne).
- Les rues en grille obligent les véhicules à se déplacer uniquement horizontalement et verticalement.
- La distance Euclidienne suppose un trajet direct, ce qui n'est pas possible dans ce contexte.
- Manhattan tient compte des obstacles et des limitations réelles (bâtiments, intersections).

Plan

- Motivation
- Similarité et Dissimilarité
- Mesures pour les données numériques
- Mesures pour les données catégorielles
- Mesures pour les données mixtes
- Mesures de similarité et de dissimilarité entre les clusters

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Similarité et dissimilarité

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- ***Rôle des métriques de similarité en apprentissage non supervisé :***
  - Identifier des groupes homogènes dans les données (clustering).
  - Comparer des éléments (documents, images, séquences ADN).
  - Réaliser des tâches de recommandation.

Dissimilarité

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- ***Définition*** *:*
  - *La fonction de* ***dissimilarité****, également appelée mesure de dissimilarité, est une fonction qui calcule la différence ou la* ***distance*** *entre deux objets, éléments ou points dans un espace. Elle est utilisée pour quantifier à quel point deux objets sont différents les uns des autres.*
  - Dissimilarité est un terme utilisé pour décrire l'absence de similitude ou de ressemblance entre deux objets.
  - Une valeur élevée indique une grande dissimilarité.
- Il existe plusieurs types de coefficients de dissimilarité :
  - Distance euclidienne, distance de Manhattan, distance de Minkowski, etc.
- Le choix du coefficient de dissimilarité dépend de la nature des données et des exigences spécifiques de similitude de l'application.

Dissimilarité

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Une fonction de distance (dissimilarité) f doit respecter quatres conditions pour être une métrique :

1. nonnegativity: $f(\mathbf{x}, \mathbf{y}) \geq 0$;
2. reflexivity: $f(\mathbf{x}, \mathbf{y}) = 0 \Longleftrightarrow \mathbf{x} = \mathbf{y}$;
3. commutativity: $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$;
4. triangle inequality: $f(\mathbf{x}, \mathbf{y}) \leq f(\mathbf{x}, \mathbf{z}) + f(\mathbf{y}, \mathbf{z})$

Similarité

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- ***Définition*** *:*
  - *Le coefficient de similarité est une mesure statistique qui quantifie le degré de similitude entre deux ensembles de données ou entre deux individus ou objets dans un espace donné.*
  - Plus deux points de données se ressemblent, plus le coefficient de similarité est grand.
- Il existe plusieurs types de coefficients de similarité :
  - Similarité de Jaccard, la similarité de Cosinus, le coefficient de corrélation de Pearson, etc.
- Le choix du coefficient de similarité dépend de la nature des données et des exigences spécifiques de similitude de l'application.

Similarité

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Une fonction de similarité doit respecter généralement les propriétés suivantes :

- Symétrie (souvent mais pas toujours) : $S(x, y) = S(y, x)$

La similarité entre deux objets est souvent indépendante de leur ordre. Cependant, certaines mesures, comme l'indice d'inclusion, n'est pas symétrique.

- Valeurs normalisées : $S(x, y) \in [0,1]$ ou parfois $S(x, y) \in [-1,1]$

Une valeur proche de 1 indique une forte similarité positive, une valeur proche de 0 indique une très faible similarité, et une valeur proche de -1 indique une forte similarité opposée

- Identité des indiscernables : $S(x, x)\ est\ maximale$

Deux objets identiques ont la similarité maximale, qui vaut souvent $S(x, x) = 1$

- Non-négativité (pas obligatoire) :

Certaines mesures de similarité, comme le coefficient de corrélation de Pearson, peuvent prendre des valeurs négatives (ce qui indique une dissimilarité ou une relation inverse).

Différences

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- ***Différences entre similarité et distance :***
  - Contrairement aux distances, les similarités ne respectent pas toujours l'inégalité triangulaire.
  - Une valeur de similarité élevée indique une proximité dans le sens conceptuel, alors qu'une distance faible indique une proximité géométrique.
    - La similarité peut être basée sur des concepts ou des relations abstraites, pas nécessairement sur une position physique ou spatiale. Par exemple, la similarité cosinus mesure l'angle entre deux vecteurs, pas leur distance géométrique. Deux vecteurs de longueurs très différentes peuvent être très similaires (angle proche de 0°).
    - La distance est souvent basée sur une norme ou une métrique géométrique stricte, comme la distance Euclidienne, qui est influencée par la magnitude et la position dans l'espace.

Matrice de proximité

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- ***Définition*** *:*
  - *Une matrice de proximité est une matrice qui est Directement liée aux métriques de similarité ou dissimilarité contenant les indices de proximité par paires d'un ensemble de données.*
  - *Un indice de proximité fait référence soit à un indice de similarité, soit à un indice de dissimilarité.*
- La matrice de proximité est une représentation explicite des relations entre les paires d'objets.
- Dans le clustering hiérarchique ou les graphes (partitionnement spectral), la matrice de proximité joue un rôle clé.
- Étant donné un ensemble de données $D = (X_1, X_2, \ldots, X_n)$ dont chaque objet est décrit par un vecteur de caractéristiques de dimension d, les matrices de proximité sont données par :

Matrice de proximité

- la matrice de distance pour D est définie par :

$$M_{dist}(D) = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

$d_{ij} = d(Xi, X_j)$ par rapport à une fonction de distance d(·, ·)

- La matrice de similarité pour D est définie comme suit :

$$M_{sim}(D) = \begin{pmatrix} 1 & s_{12} & \cdots & s_{1n} \\ s_{21} & 1 & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & 1 \end{pmatrix}$$

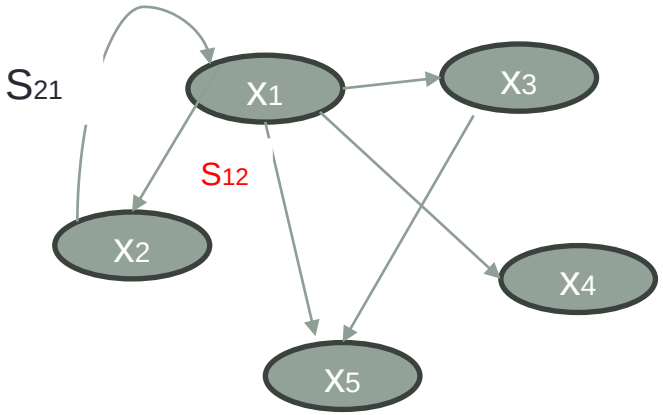
$s_{ij} = s(x_i, x_j)$ par rapport à une mesure de similarité s(·, ·)

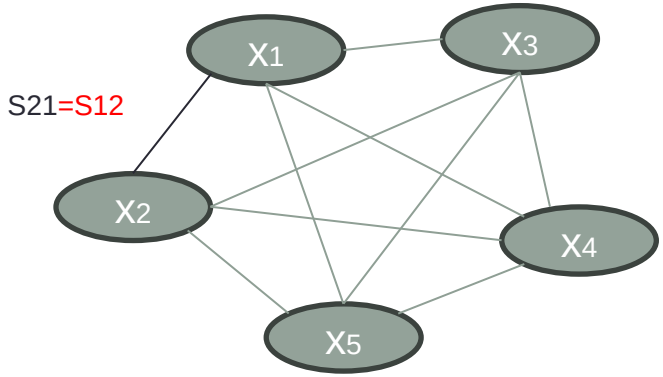
- *Si la fonction de distance et la fonction de similarité sont symétriques, alors les deux matrices de proximité sont symétriques.*

Graphe de proximité

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- ***Définition*** *:*
  - *Un graphe de proximité $G = (S, A, V)$ est un graphe pondéré, tel que :*
    - *S est l'ensemble des nœuds représentés par les points de données*
    - *A est l'ensemble des connexions entre les nœuds (arêtes ou arcs)*
    - *V est l'ensemble des valuations sur les connexions représentant les indices de proximité entre les points,*
  - *Un graphe orienté correspond à une matrice de proximité asymétrique,*
  - *Un graphe non orienté correspond à une matrice de proximité symétrique.*





Matrice de dispersion

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- ***Définition*** *:*
  - *Une matrice de dispersion est une matrice qui mesure la répartition des points de données dans l'espace, souvent en termes de distances par rapport à un barycentre ou entre les points eux-mêmes.*
  - *C'est une mesure globale ou locale qui peut influencer l'interprétation des similarités.*
- Une dispersion élevée indique une faible similarité globale entre les points (ou entre clusters). Une dispersion faible correspond souvent à des groupes compacts, indiquant une plus grande similarité intra-cluster.
- Les matrices de dispersion ne sont pas directement des matrices de similarité, mais elles influencent l'interprétation des distances dans l'espace des données.
- Dans des méthodes comme k-means, la dispersion intra- et inter-cluster est souvent mesurée pour évaluer la qualité des regroupements.

Matrice de dispersion

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Étant donné un ensemble de données $D = (X_1, X_2, \ldots, X_n)$ dont chaque objet est décrit par un vecteur de caractéristiques de dimension d, la matrice de dispersion est donnée par :

$$\mathbf{M_D} = \boldsymbol{X_c^t}.\boldsymbol{X_c} = \left(\boldsymbol{X} - \mathbf{1}\overline{\boldsymbol{X}}\right)^{\boldsymbol{t}}\left(\boldsymbol{X} - \mathbf{1}\overline{\boldsymbol{X}}\right) \textbf{ Avec: } \quad \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

dispersion intra et inter cluster

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Pour un cluster *C* de *D* données, $M_t(C)$ est aussi appelée matrice intra-dispersion de C.
- La matrice de dispersion intra-cluster est définie comme suite

$$M_w(\mathcal{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{z}_i)^T (\mathbf{x} - \mathbf{z}_i)$$

où z_i est la moyenne du cluster C_i $\quad \mathbf{z}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$

- La matrice de dispersion inter-clusters est définie comme

$$M_b(\mathcal{C}) = M_t(D) - M_w(\mathcal{C})$$

Matrice de covariance

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- ***Définition*** *:*
  - *La matrice de covariance décrit les relations linéaires entre les dimensions (ou variables) des données. Chaque entrée quantifie dans quelle mesure deux variables varient ensemble.*
  - *Une covariance élevée entre deux dimensions indique une relation linéaire forte, souvent associée à une similarité structurelle.*
  - *Mesure des relations entre dimensions, utilisée pour ajuster ou calculer des similarités ou dissimilarités.*
- En pratique, la matrice de covariance sert à ajuster ou transformer les données pour une meilleure évaluation des relations entre points, notamment dans des espaces multi-dimensionnels.

Matrice de covariance

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Matrice de covariance

- Soit D un ensemble de données avec n objets, dont chacun est décrit par *d* attributs $v_1, v_2, \ldots, v_d$.
- Les attributs $v_1, v_2, \ldots, v_d$ sont également appelés variables.
- La covariance entre deux variables v_r et v_s est définie comme étant le rapport de la somme des produits de leur écart à la moyenne sur le nombre d'objets.

$$c_{rs} = \frac{1}{n}\sum_{i=1}^{n}(x_{ir} - \bar{x}_r)(x_{is} - \bar{x}_s)$$

- où x_{ij} est la $j^{ème}$ composante du point de données x_i et $\bar{x}_j$ est la moyenne de tous les points de données dans la $j^{ème}$ variable.

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j = 1, 2, \ldots, d$$

Matrice de covariance

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- La matrice de covariance est une matrice d×d dans laquelle l'entrée (r, s) contient la covariance entre la variable v_r et v_s .

$$\Sigma = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1d} \\ c_{21} & c_{22} & \cdots & c_{2d} \\ \vdots & \vdots & & \vdots \\ c_{d1} & c_{d2} & \cdots & c_{dd} \end{pmatrix}$$

- Cette matrice est symétrique

Exemple d'application

- Supposons que nous avons un jeu de données avec 3 observations (points) et 2 variables :

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \\ 4 & 8 & 12 \end{bmatrix}$$

- Calculons la matrice de dissimilarité, de similarité, de dispersion et de covariance :

$$S = \begin{bmatrix} 5.0 & 10.0 & 15.0 \\ 10.0 & 20.0 & 30.0 \\ 15.0 & 30.0 & 45.0 \end{bmatrix}$$

$$\Sigma = \frac{1}{3} \begin{bmatrix} 5.0 & 10.0 & 15.0 \\ 10.0 & 20.0 & 30.0 \\ 15.0 & 30.0 & 45.0 \end{bmatrix} = \begin{bmatrix} 1.6667 & 3.3333 & 5.0 \\ 3.3333 & 6.6667 & 10.0 \\ 5.0 & 10.0 & 15.0 \end{bmatrix}$$

Comparaison

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- La **matrice de proximité** est la plus directement associée aux métriques de similarité.
- Les matrices de dispersion et de covariance jouent des rôles complémentaires, influençant la manière dont ces similarités sont interprétées, ajustées ou représentées dans un espace multidimensionnel.
- La matrice de dispersion est une version non normalisée de la matrice de covariance.
- Les matrices de proximité (similarité ou dissimilarité) sont utilisées pour des comparaisons entre points.
- Les matrices de covariance sont centrées sur les relations entre les variables.