UMAP : Uniform Manifold Approximation and Projection

Youness et Ivanhoé

28 octobre 2023

Introduction

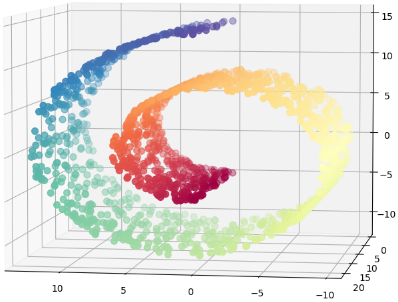

Figure – Représentation 3D du swiss roll dataset.

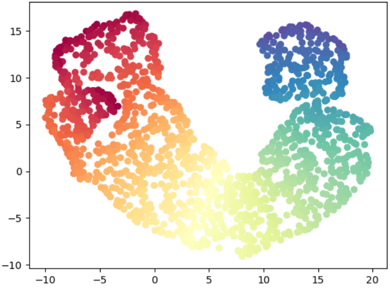

Figure – UMAP projection (n_neighbors=15, min_dist = 0.1).

Problématique : Comment fonctionne l'algorithme UMAP et pourquoi utiliser cette méthode sur un jeu de données ?

Les intérêts de la méthode UMAP sur un jeu de données

- Méthode de *réduction de dimension* utile pour la visualisation des données.
- Projection dans un espace de plus petite dimension pour des données non-linéaires donc non-séparables par un sous-espace vectoriel.
- Retransmettre dans un espace de plus petite dimension la structure topologique des données initiales.
- Utile pour faire du clustering et peut servir pour entraîner des modèles de machine learning.

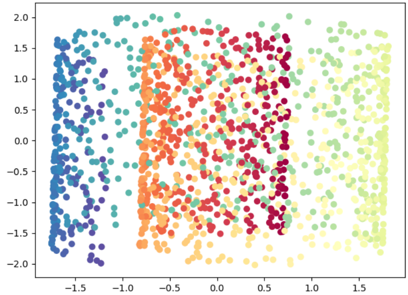

Figure – Scatter plot de l'ACP du dataset swiss roll avec les deux premières composantes principales

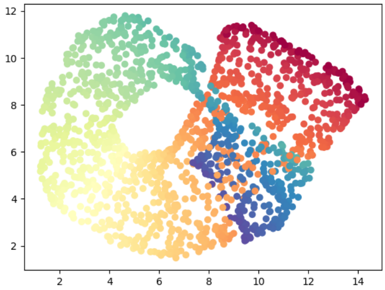

Figure – UMAP projection du nuage de points dans un plan. (n_neighbors=15, min_dist = 0.1)

Remarque

L'algorithme de UMAP ressemble à celui de T-SNE. Il est présenté comme plus rapide et pouvant s'adapter à une réduction de la dimension $d \geq 2$. UMAP préserve aussi davantage la structure globale du jeu de donnés par rapport à T-SNE.

Idée :

Pour un jeu de données $X \in \mathcal{M}_{n,p}(\mathbb{R})$ nous cherchons préserver les relations locales entre les individus dans un espace de dimension plus petit ($d \lll p$). Cela revient à trouver $Y \in \mathcal{M}_{n,d}(\mathbb{R})$ sous certaines contraintes.

Les étapes importantes :

1. Pour un nombre de voisins k fixé, donner pour tout couple d'individu (x_i, x_j) une probabilité $p_{i,j}$ d'appartenir à un même voisinage.
2. Initialiser la matrice Y et en déduire la similarité entre les couples d'individu (y_i, y_j) que l'on note $q_{i,j}$.
3. Minimiser sur l'espace d'arrivée une fonction objectif d'entropie croisée qui mesure la dissimilarité entre les deux distributions p et q. Ici q est une fonction de Y.

Matrice de probabilités jointes sur l'espace de départ

1. Fixons $k \geq 2$ le nombre de voisins à considérer. Soit x_i, x_j deux individus, la similarité $p_{i|j}$ entre deux observations est définie par :

$$p_{i|j} = \exp\left(-\frac{\|x_i - x_j\|_2 - \rho_i}{\sigma_i}\right)$$

   - On définit $\rho_i = \min_{i \neq j} \|x_i - x_j\|$.
   - On définit σ_i de telle sorte que $\sum_{j=1}^{k} p_{i|i_j} = \log_2(k)$

2. Symétrisation pour obtenir la probabilité d'appartenir à un même voisinage :

$$p_{i,j} = p_{i|j} + p_{j|i} - p_{i|j} \times p_{j|i}$$

Similarité entre les individus de l'espace d'arrivé

Fixons une distance minimale *"min_dist"*.

1. Nous aimerions définir la similarité entre deux individus y_i, y_j de la manière suivante :

$$q_{i,j} = \begin{cases} 1 & \text{si} \quad \|y_i - y_j\| \leq min\_dist \\ \exp(-\|y_i - y_j\|) & \text{si} \quad \|y_i - y_j\| > min\_dist \end{cases}$$

2. Pour cela on s'inspire d'une loi de Student

$$q_{i,j} = \left(1 + a\|y_i - y_j\|^{2b}\right)^{-1}$$

L'algorithme UMAP optimise les deux paramètres (a, b) pour coller au mieux à la fonctions précédente.

Fonction d'entropie croisée et minimisation

But : trouver la distribution des points y_i dans l'espace d'arrivée telles que les similarités $q_{i,j}$ s'approchent le plus possibles des similarités $p_{i,j}$.

- Introduction de la fonction d'entropie croisée :

$$\begin{aligned} CE(y_1, \cdots, y_n) = CE(p,q) &= \sum_{i=1}^{n} \sum_{j=1}^{n} p_{i,j} \ln\left(\frac{p_{i,j}}{q_{i,j}}\right) + (1 - p_{i,j}) \ln\left(\frac{(1-p_{i,j})}{(1-q_{i,j})}\right) \\ &= KL(p,q) + KL(1-p, 1-q) \end{aligned}$$

- Minimiser la fonction d'entropie croisée à l'aide d'une descente de gradient stochastique :

$$y_i := y_i - \alpha \nabla_i CE(y_1, \cdots, y_i, \cdot, y_n)$$

Pour la descente de Gradient Stochastique une seule observation est utilisée pour calculer le gradient à chaque itération et celle-ci est choisie de manière aléatoire.

Utiliser UMAP au quotidien et ses résultats

Conclusion et faiblesses de UMAP

- Pour des datasets qui présentent une structure topologique marquée dans chaque sous groupe, les rendus de l'algorithme UMAP sont visibles et permettent bien de visualiser les clusters.

- À l'inverse d'une méthode de réduction de dimension par projection linéaire type ACP, avec UMAP les coordonnées des observations dans l'espace d'arrivée sont moins interprétables physiquement.
- La recherche des bons paramètres **n_neighborgs** et **min_dist** n'est pas évidente et ne permets pas toujours de rendre une vision globales des données initiales.