Analyse de Données

Antoine Lavault[1,2]

[1]Apeira Technologies

[2]UMR CNRS 9912 STMS, IRCAM, Sorbonne Université

19 janvier 2024

Contenu

1. Regroupement(Clustering)
   - K-Means
   - Clustering Hiérarchique
   - DBSCAN

2. Réduction de dimension
   - Généralités
   - Analyse en composantes principales
   - IsoMap

Contenu

1 Regroupement(Clustering)
- K-Means
- Clustering Hiérarchique
- DBSCAN

2 Réduction de dimension

Story time !

Résumé : différents fabriquants, regrouper les produits similaires.

Définition

Clustering
Le clustering est une technique d'apprentissage non supervisé qui consiste à diviser un ensemble de données en groupes distincts (ou clusters) de sorte que les données au sein d'un groupe soient similaires les unes aux autres et différentes des données dans les autres groupes.
Le clustering peut être utilisé dans de nombreux domaines, tels que la segmentation de clients, la détection d'anomalies, l'analyse des réseaux sociaux, etc.

Remarque

Point de vocabulaire : la principale différence entre les méthodes supervisées et non-supervisées dans le contexte de l'analyse de données réside dans l'existence d'un ensemble de données étiqueté pour les méthodes supervisées, tandis que les méthodes non-supervisées cherchent à découvrir des modèles et des structures inhérents dans les données non étiquetées.

Contenu

1 Regroupement(Clustering)
- K-Means
- Clustering Hiérarchique
- DBSCAN

K-Means

Une des méthodes les plus simples et les plus utilisées est le K-Means (littéralement K-Moyennes).
L'algorithme K-Means est un algorithme de clustering non-supervisé qui vise à partitionner un ensemble de données en un certain nombre de clusters en fonction de leur similarité.

Algorithme

Voici les étapes de l'algorithme K-Means, pour un jeu de données $(x_n)_{n\in\{0,1,\cdots,N\}}$

1. Initialisation : choisir aléatoirement K centres de cluster parmi les données.
2. Attribution : attribuer chaque point de données au centre de cluster le plus proche.
3. Réaffectation : recalculer les centres de cluster en utilisant la moyenne des points qui lui sont attribués.
4. Répéter les étapes 2 et 3 jusqu'à convergence (lorsque les centres de cluster ne changent plus ou que le nombre maximal d'itérations est atteint).

Et maintenant avec 18% de maths en plus !

1. Initialisation : Sélectionnez le nombre k de clusters à trouver, et autant de points initiaux comme centres de chaque cluster (généralement au hasard)
2. Attribution des points aux clusters : Pour chaque point dans les données, calculez la distance à chaque centre de cluster et assignez le point au cluster le plus proche.
   La distance euclidienne entre le point x_i et le centre du cluster c_j est définie comme suit :

$$dist(x_i, c_j) = \sum_{p=1}^{d} (x_{i,p} - c_{j,p})^2 \tag{1}$$

ce qui est équivalent à :

$$||x_i - c_j||^2 = (x_i - c_j)^T (x_i - c_j) \tag{2}$$

Où x_i représente un point dans l'espace de données, c_j représente le centre du cluster j, d est le nombre de dimensions, et $x_{i,p}$ et $c_{j,p}$ sont les valeurs de la dimension p pour le point x_i et le centre c_j respectivement.

Cont.

3. Mise à jour des centres de cluster : Pour chaque cluster, calculez le nouveau centre en prenant la moyenne des points dans le cluster. Le nouveau centre c_j pour le cluster j est défini comme suit :

$$c_j = \frac{1}{n_j} \sum_{i=1}^{n} x_i$$

Où n_j est le nombre de points assignés au cluster j, x_i est le i-ème point dans le cluster j, et n est le nombre total de points dans les données.

4. Répétez les étapes 2 et 3 jusqu'à ce que les centres de cluster convergent (c'est-à-dire qu'ils ne changent plus) ou que le nombre d'itérations limite soit dépassé.

Demo Time !

Figure – Enseignant ou étudiant moyen quand le cours commence après 8h

Contenu

1 Regroupement(Clustering)
- K-Means
- Clustering Hiérarchique
- DBSCAN

Définition

Le clustering hiérarchique est une méthode de clustering qui permet de regrouper les données en fonction de leur similarité.
Contrairement au clustering K-Means, le clustering hiérarchique ne nécessite pas la spécification préalable du nombre de clusters. Il crée une structure de clustering en forme d'arbre, également appelée dendrogramme, qui peut être visualisée pour aider à déterminer le nombre de clusters approprié.

Remarque

Le clustering hiérarchique peut être divisé en deux approches différentes : agglomérative et divisive. L'approche agglomérative commence par considérer chaque point de données comme un cluster et fusionne ensuite les clusters similaires en un seul cluster à chaque étape. L'approche divisive commence par considérer tous les points comme un seul cluster et divise ensuite le cluster en sous-clusters à chaque étape.

Algorithme

L'algorithme de clustering hiérarchique agglomératif peut être décrit par les étapes suivantes :

1. Initialisation : Considérez chaque point de données comme un cluster séparé.
2. Calcul de la similarité : Calculez la matrice de similarité entre tous les paires de clusters. La mesure de similarité dépend de la méthode de liaison utilisée. Les méthodes les plus courantes sont la liaison simple, la liaison complète et la liaison moyenne.
3. Fusion de clusters : Fusionnez les deux clusters les plus similaires en un seul cluster, en utilisant la mesure de similarité calculée à l'étape 2. Cette étape est répétée jusqu'à ce qu'un seul cluster reste.
4. Construction du dendrogramme : Enregistrez la séquence des fusions dans un dendrogramme. Les fusions sont représentées par des branches qui se rejoignent, et les points de données sont représentés par les feuilles.

Mesurer la similarité

La mesure de similarité la plus couramment utilisée est la distance euclidienne, déjà vue précédemment. Mais il est possible d'utiliser d'autres critères de similarité. On citera la méthode de Ward à titre d'exemple, qui cherche à minimiser la variance au sein d'un cluster (dans ce cas, c'est la similarité statistique qui est considérée).
Le clustering hiérarchique peut être utilisé pour trouver des groupes de points de données similaires dans des ensembles de données multidimensionnels. La structure hiérarchique des dendrogrammes peut également fournir des informations supplémentaires sur la relation entre les clusters et aider à identifier les sous-groupes à différents niveaux de granularité.

Demo Time 2

Figure – Résumé du cours d'analyse de données

Contenu

1 Regroupement(Clustering)
- K-Means
- Clustering Hiérarchique
- DBSCAN

DBSCAN

Définition

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering qui utilise la densité pour identifier les clusters dans un ensemble de données. Il est particulièrement utile pour identifier des clusters de formes arbitraires dans des données de haute dimension.

Remarque

L'algorithme DBSCAN fonctionne en examinant la densité de points dans l'espace de données. Il identifie les points centraux et les points frontières qui sont proches d'un point central, ainsi que les points bruités. Les points centraux sont ceux qui ont un nombre minimum de points à une distance spécifiée, appelé le rayon de voisinage. Les points frontières sont ceux qui ne satisfont pas les critères pour être des points centraux, mais qui sont proches d'un point central. Les points bruités sont les points qui ne sont pas des points centraux et ne sont pas proches d'un point central.

Algorithme

L'algorithme DBSCAN utilise deux paramètres principaux : le rayon de voisinage ϵ et le nombre minimum de points *MinPts*.
Voici les étapes principales de l'algorithme DBSCAN :

1. Sélectionnez un point p au hasard qui n'a pas encore été visité.
2. Trouvez tous les points dans un rayon de voisinage ϵ autour de p.
3. Si le nombre de points dans le rayon de voisinage est inférieur à *MinPts*, marquez p comme un point bruité et passez à l'étape suivante. Sinon, marquez p comme un point central et créez un nouveau cluster contenant p.
4. Ajoutez tous les points dans le rayon de voisinage ϵ de p au cluster. Si un point est marqué comme un point central, ajoutez également tous les points dans son rayon de voisinage ϵ au cluster.
5. Répétez les étapes 1 à 4 pour tous les points du cluster jusqu'à ce qu'il n'y ait plus de points à visiter.
6. Répétez les étapes 1 à 5 jusqu'à ce que tous les points soient marqués comme appartenant à un cluster ou comme points bruités.

Algorithme Cont.

Voici l'équation pour calculer la densité locale d'un point p :

$$\rho_p = \sum_{q \in D} I(||x_p - x_q|| < \epsilon) \tag{3}$$

Où D est l'ensemble des points de données, x_p et x_q sont les vecteurs de caractéristiques des points p et q respectivement, et I est une fonction indicatrice qui vaut 1 si la distance entre x_p et x_q est inférieure à ϵ et 0 sinon. Un point p est considéré comme un point central s'il a au moins *MinPts* voisins dans un rayon de voisinage ϵ autour de lui, i.e :

$$\rho_p \geq MinPts \tag{4}$$

Un point frontière est un point qui est dans un rayon de voisinage ϵ autour d'un point central, mais qui n'a pas suffisamment de voisins pour être considéré comme un point central.
On considère qu'un point est une donnée aberrante quand son epsilon voisinage ne contient pas de points dont l'epsilon voisinage contient MinPts points ou plus.

Illustration

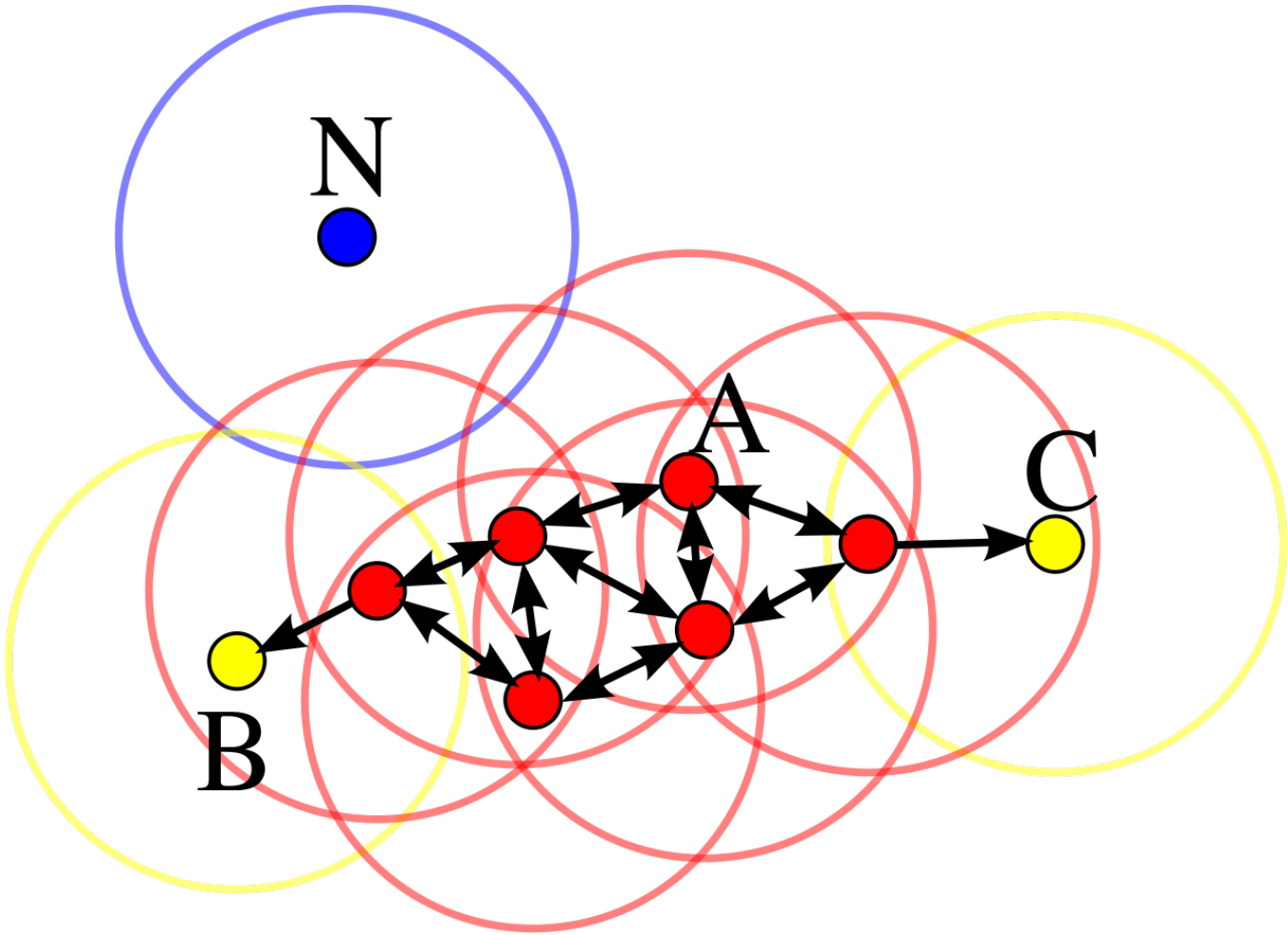

Figure – Illustration du principe de DBSCAN

Avec A est un point central, B et C sont des points frontières, N est considéré comme bruité.

D.E.M.O. Time.

Figure – En fait c'est bien les démos, ça prend du temps.

Contenu

1 Regroupement(Clustering)

2 Réduction de dimension
- Généralités
- Analyse en composantes principales
- IsoMap

Contenu

2 Réduction de dimension
- Généralités
- Analyse en composantes principales
- IsoMap

Malédiction de la dimension

Aussi connue sous le nom de *Curse of Dimensionality*. Peut prendre plusieurs formes :

- Explosion combinatoire : prenons comme exemple le problème du voyageur de commerce.
  Pour n villes, on aura $(n-1)!$ chemins possibles pour un point de départ fixé (ce qui ne change pas le résultat). De la même manière, les chemins peuvent être parcourus dans les deux sens sans modifier la longueur, ce qui réduit le nombre de cas à $\frac{(n-1)!}{2}$
  Par exemple, pour 20 villes, on obtient un nombre de possibilités égal à 6.082×10^{16} ce qui est du même ordre de grandeur que le nombre de mètres dans un Parsec.

Malédiction de la dimension

- Explosion des distances : les espaces de grandes dimensions sont "vides". Une façon d'illustrer l'"immensité" d'un espace euclidien à haute dimension est de comparer la proportion d'une hypersphère inscrite de rayon r et de dimension d, à celle d'un hypercube dont les arêtes sont de longueur $2r$

  Le volume d'une telle sphère est $\frac{2r^d \pi^{d/2}}{d\,\Gamma(d/2)}$, où Γ est la fonction gamma, tandis que le volume du cube est $(2r)^d$. Lorsque la dimension augmente, de l'espace, l'hypersphère devient un volume insignifiant par rapport à celui de l'hypercube. On le voit bien en comparant les proportions lorsque la dimension d passe à l'infini :

$$\frac{V_{\text{hypersphere}}}{V_{\text{hypercube}}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0, \text{ quand } d \to \infty. \tag{5}$$

Changement de point de vue

Une autre manière de voir le problème est que chaque nouvelle dimension ajoute un terme non négatif à la somme, de sorte que la distance augmente avec le nombre de dimensions pour des vecteurs distincts.
En d'autres termes, lorsque le nombre de caractéristiques augmente pour un nombre donné d'observations, l'espace des caractéristiques devient de plus en plus clairsemé, c'est-à-dire moins dense, ou plus vide.
En revanche, la faible densité des données nécessite davantage d'observations pour que la distance moyenne entre les points de données reste la même.

Contenu

2 Réduction de dimension
- Généralités
- Analyse en composantes principales
- IsoMap

Définition

L'Analyse en Composantes Principales (ACP) est une méthode de réduction de dimension qui permet de représenter un ensemble de données multivariées dans un espace de dimension inférieure tout en conservant la variance maximale des données. Elle permet également d'identifier les variables qui contribuent le plus à la variation des données.

Remarque

L'ACP fonctionne en transformant les données initiales en un nouveau système de coordonnées dans lequel les axes sont ordonnés en fonction de la quantité de variance expliquée par chaque axe.
Les premiers axes représentent donc les directions de plus grande variance dans les données.

SVD

L'ACP utilise la décomposition en valeurs singulières (Singular Value Decomposition, SVD) pour calculer les composantes principales des données.

Définition

La SVD est une factorisation matricielle qui décompose une matrice X en trois matrices : une matrice de gauche U dite de sortie, une matrice diagonale Σ et une matrice de droite V^T dite d'entrée. Mathématiquement, la SVD s'écrit comme suit :

$$X = U \cdot \Sigma \cdot V^T$$

Où U est une matrice orthogonale de taille $n \times n$, Σ est une matrice diagonale de taille $n \times d$ contenant les valeurs singulières de X, et V^T est une matrice orthogonale de taille $d \times d$.
Les valeurs singulières dans Σ sont généralement ordonnées par ordre décroissant, de sorte que les premières valeurs représentent les directions de plus grande variance dans les données.

Reformulation

On va dans un premier temps reformuler le problème à résoudre :

Définition (ACP réformulée)

Soit X une matrice $n \times d$ avec μ le vecteur des moyennes empiriques.
Alors il existe une famille $\{v_h\}, 1 \leq h \leq d$ telle que les vecteurs sont unitaires et mutuellement orthogonaux, et telle qu'avec les points centrés associés

$$Y_h = (X - \mu)v_h \tag{6}$$

on ait :

- $Var[Y_h]$ est maximisée,
- $Cov(Y_h, Y_l) = 0$ quand $h \neq l$ i.e pas de corrélation entre les composantes principales différentes.

Remarque

$X - \mu$, avec $\mu = (\mu_1 \cdots \mu_j)$, $\frac{1}{n} 1_n^T X$

Algorithme

1. Calculer la matrice centrée $X - \mu$ (réduction possible)
2. Calculer sa SVD (supposée décroissante) : $(X - \mu) = U\Sigma V^T$
3. Obtenir les points centrés : $E = (X - \mu)V = U\Sigma$
4. Réduction de dimension : $R = E \cdot I_{d\times k} = U\Sigma \cdot I_{d\times k}$

Remarque

Pour une réduction de dimension d à k, on a $(X - \mu)V \cdot I_{d\times k}$, multiplication de dimensions $(n \times d)(d \times d)(d \times k)$.

~~Demo~~ Nightmare Time

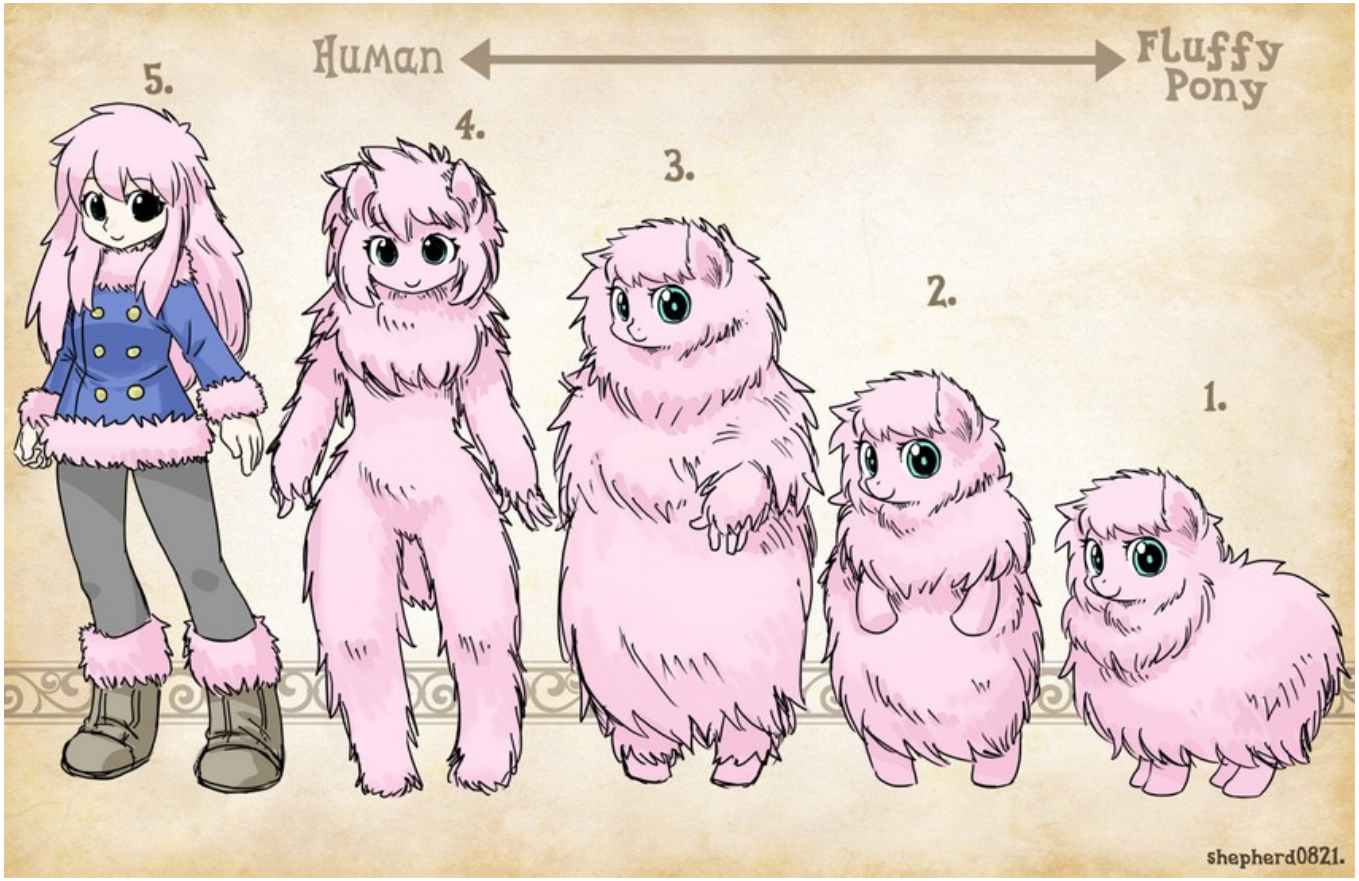

Figure – Can't get any worse than this. Right ?

Contenu

2 Réduction de dimension
- Généralités
- Analyse en composantes principales
- IsoMap

Limites de l'ACP

L'ACP a une limite claire : elle n'est pas adaptée à des jeux de données non-linéaires et corrélées.
Il serait bien sûr possible d'utiliser une méthode à noyau avec l'ACP pour la rendre compatible avec les datasets non-linéaires, mais on présente ici une autre méthode, à titre d'illustration.

Devinez c'est l'heure de quoi ? De la démo !

Figure – Chat

IsoMap

La méthode Isomap (Isometric Mapping) est une technique de réduction de dimensionnalité non-linéaire utilisée pour représenter des données multidimensionnelles de haute dimension en des dimensions inférieures, tout en préservant la structure géométrique des données.
La méthode Isomap est basée sur la notion de géodésiques, qui sont des chemins les plus courts dans l'espace des données, mesurés à partir de la distance euclidienne entre les points voisins.
Pour cela, la méthode utilise un algorithme appelé algorithme de Floyd-Warshall (un algorithme de calcul de plus court chemin dans un graphe) , qui permet de trouver les chemins les plus courts entre toutes les paires de points du graphe.

Algorithme

Plus précisément, la méthode Isomap procède en plusieurs étapes :

1. Construction du graphe de voisinage : on construit un graphe reliant les points les plus proches les uns des autres, en utilisant la distance euclidienne. Et un seuil de proximité, soit en nombre de plus proches voisins, soit avec un rayon fixé.
2. Calcul des distances géodésiques : on calcule les distances géodésiques (les plus courts chemins dans le graphe) entre chaque paire de points à l'aide de l'algorithme de Floyd-Warshall (non décrit).
3. Réduction de dimension : on applique ensuite une technique de réduction de dimensionnalité linéaire (comme l'ACP) sur la matrice des distances géodésiques pour obtenir des coordonnées en dimensions inférieures.
4. Interpolation des coordonnées manquantes : enfin, on peut utiliser les coordonnées obtenues pour interpoler les coordonnées manquantes des données originales et obtenir une représentation en dimensions inférieures de l'ensemble des données.

L'avantage de la méthode Isomap est qu'elle est capable de capturer les relations non-linéaires entre les données, tout en conservant leur structure géométrique. Cela permet de mieux comprendre et visualiser les données complexes en dimensions élevées, et peut également faciliter leur analyse et leur traitement par des algorithmes de machine learning.

Demo Time

Figure – Faces in the Wild