

Manifold Learning and UMAP

Paul Snopov

MIPT & IITP

November, 29, 2021

Overview

- 1 What is Manifold Learning
 - Different Approaches
 - What is UMAP
- 2 UMAP: theoretical side
 - Local distance
 - Review for simplicial sets
 - Fuzzy simplicial sets
 - Extended-pseudo-metric spaces
 - Returning to the problem
- 3 Optimizing a low dimensional representation
 - Layout
- 4 UMAP: computational side
 - Basic Structure
 - Graph Construction
 - Graph Layout
- 5 Examples
 - MNIST

What is Manifold Learning

- 1 An approach to non-linear dimensionality reduction
- 2 The input data is assumed to be sampled from a low dimensional manifold embedded in some \mathbb{R}^n for sufficiently large n
- 3 Manifold Learning methods look for that low dimensional representation of original data

What is Manifold Learning

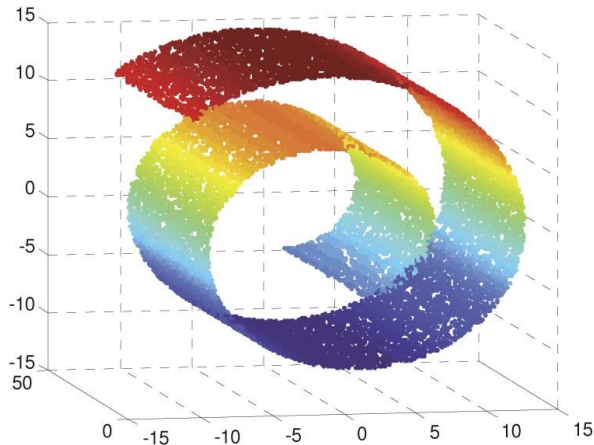


Figure The image is taken from https://www.researchgate.net/figure/Swiss-roll-data-set-Fig-11-Three-dimensional-clusters-data-set_fig7_224189094

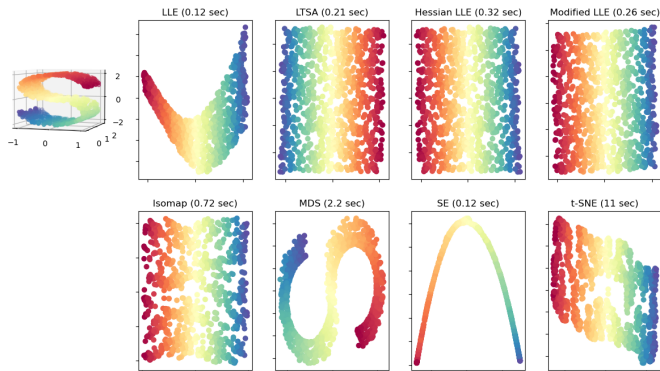


Figure The image is taken from [3]

The idea behind UMAP

Let's suppose our data $X = \{x_1, \dots, x_N\}$ lies on a manifold in some ambient metric space.

The idea of UMAP is sort of borrowed from TDA: having finite sets of points, we can consider balls about each point and construct Čech complex via these balls. By Nerve theorem, this complex will give us a representation of the manifold the data is sampled from.

The idea behind UMAP

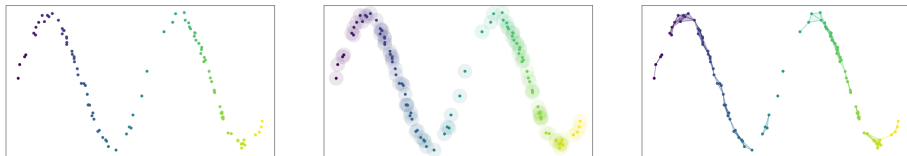


Figure Images are taken from [3]

The idea behind UMAP

If the data were uniformly distributed, the appropriate radius would have been chosen easily ($\frac{d(x_i, x_j)}{2} + \varepsilon$).

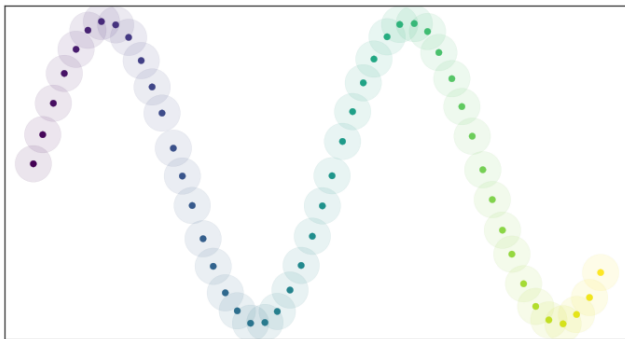


Figure The image is taken from [3]

The idea behind UMAP

Assuming the data is uniformly distributed on \mathcal{M} , any ball of fixed volume should contain approx. the same number of points of X regardless of where on the manifold it's centered. And conversely, a ball centered at X_i that contains exactly k neighbors should have approx. fixed volume regardless of the choice of $X_i \in X$.

The idea behind UMAP

But real world data simply isn't that nicely behaved. How to fix this?

A: let's suppose that the data *is* uniformly distributed on the manifold, but if the data looks like it isn't uniformly distributed that must simply be because the notion of distance is varying across the manifold – space itself is warping: stretching or shrinking according to where the data appear sparser or denser.

By assuming that the data is uniformly distributed we can actually compute an approx. of a local notion of distance for each point.

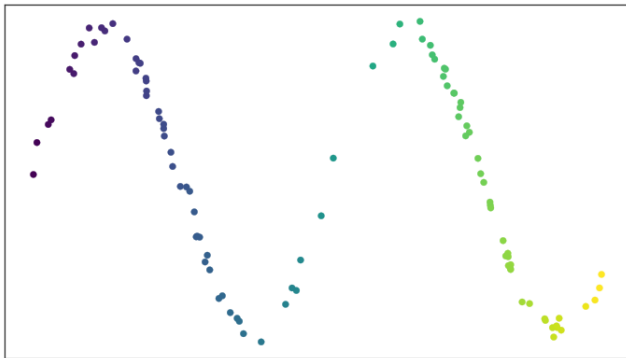


Figure The image is taken from [3]

Local distance

Let \mathcal{M} be the manifold on which the data lie on and g be the Riemannian metric on it.

Lemma

Let (\mathcal{M}, g) be in ambient \mathbb{R}^n and $p \in \mathcal{M}$. If g is locally constant about p in an open nbhd U s.t. g is constant diagonal matrix in ambient coordinates, then in a unit ball $B(p) \subseteq U$ (i.e. with volume $\frac{\pi^{n/2}}{\Gamma(n/2+1)}$ w.r.t. g), the geodesic distance from p to any point $q \in B(p)$ is $\frac{1}{r} d_{\mathbb{R}^n}(p, q)$, where r – the radius of $B(p)$ in the ambient space.

Local distance

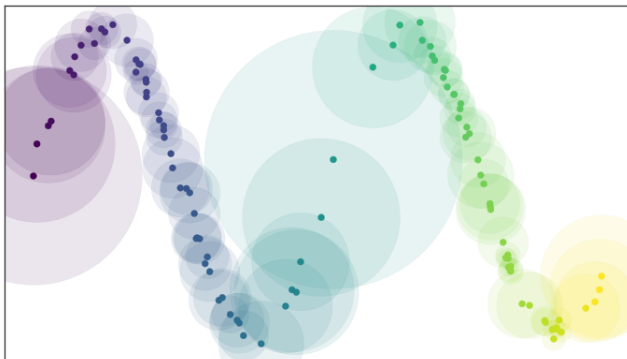


Figure The image is taken from [3]

Local distance

By previous lemma, we can approx. geodesic distance from X_i to its neighbors by normalising distances w.r.t. the distance of the farthest neighbor, i.e. creating custom distances for each X_i .

Thus now we have a family of discrete metric spaces for each X_i . And we need to merge it into something global. This can be done by converting the metric spaces into fuzzy simplicial sets.

Review for simplicial sets

Definition

The category Δ is called *simplicial category*. It's the category with:

- 1 Objects are finite order sets $[0], [1], \dots, [n] = \{0, 1, \dots, n\}$
- 2 Morphism $[n] \rightarrow [m]$ is order-preserving map

Definition

A *simplicial set* is a presheaf on Δ

Review for simplicial sets

Standard n -simplex Δ^n is defined as the functor $\text{hom}(-, [n])$. By the Yoneda lemma, for simplicial set X , $X([n]) \simeq \text{Nat}(\Delta^n, X)$, moreover, the next theorem holds

Theorem (Density theorem for simplicial set)

Let X be a presheaf, then $X \simeq \lim_{\rightarrow} \Delta^n$

Review for simplicial sets

There is a standard covariant functor geometric realization $|\cdot| : \Delta \rightarrow \mathbf{Top}$ that sends $[n]$ to the standard n -simplex $|\Delta^n| \subseteq \mathbb{R}^{n+1}$

$$|\Delta^n| = \{(t_0, \dots, t_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^n t_i = 1, t_i \geq 0\}$$

For simplicial set, one can construct $|X|$ as the colimit

$$|X| = \lim_{\rightarrow} |\Delta^n|$$

Review for simplicial sets

Conversely, for topological space Y one can construct associated simplicial set $\text{Sing}(Y)$ – singular set of Y , by defining:

$$\text{Sing}(Y) : [n] \mapsto \text{hom}_{\mathbf{Top}}(|\Delta^n|, Y)$$

It's known from homotopy theory that $|\cdot| \dashv \text{Sing}$

Fuzzy sets

Classical theory of singular sets and topological realization can be extended to fuzzy singular sets and metric realization(due to Spivak [2]).

Classically, a *fuzzy set* is defined in terms of a carrier set A and a map $\nu : A \rightarrow [0, 1]$ – *membership function*. The value $\nu(x)$ is sort of the membership strength of x to the set A .

Fuzzy sets

Let $I = (0, 1]$ with topology given by intervals of the form $(0, a)$ for $a \in (0, 1]$. Let's consider the category of open sets of **$Op(I)$** .

Definition

A fuzzy set is a presheaf \mathcal{P} on **$Op(I)$** such that all maps $\mathcal{P}(a \leq b)$ are injections.

To link to the classical approach to fuzzy sets one can think of a section $\mathcal{P}((0, a))$ as the set of all elements with membership strength at least a .

Category of fuzzy sets

Under the Grothendieck topology on $Op(I)$, presheaves are trivially sheaves, and then

Definition

The category **Fuzz** of fuzzy sets is the full subcategory of sheaves on I spanned by fuzzy sets

Category of fuzzy simplicial sets

Definition

The category of fuzzy simplicial sets ***sFuzz*** is the category of presheaves on Δ with value in ***Fuzz***

We will use $\Delta_{<a}^n$ to denote the sheaf given by representable functor of the object $([n], (0, a))$. The importance of this fuzzy version of simplicial sets is their relation to metric spaces.

Category of extended-pseudo-metric spaces

Definition

An extended-pseudo-metric space (e.p. metric space) (X, d) is a metric space in which the distance between two distinct points can be either 0, or ∞ .

The category of e.p. metric spaces ***EPMet*** has non-expansive maps as morphisms. The subcategory of finite e.p. metric spaces is ***FinEPMet***.

Extensions of classical constructions

In [2] Spivak constructs a pair of adjoint functors $Real \dashv Sing$ between categories **sFuzz** and **EPMet**, which are just the natural extension of the classical functors from algebraic topology.

The functor $Real$ is defined for standard fuzzy simplices $\Delta_{<a}^n$ as:

$$Real(\Delta_{<a}^n) = \{(t_0, \dots, t_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^n t_i = -\log(a)\},$$

similarly to the classical realization functor $|\cdot|$. The metric on $Real(\Delta_{<a}^n)$ is inherited from ambient space. A morphism $\Delta_{<a}^n \rightarrow \Delta_{<b}^m$ exists iff $a \leq b$ and is determined by a Δ morphism $\sigma : [n] \rightarrow [m]$. The action of $Real$ on such morphism is a map

$$(x_0, \dots, x_n) \mapsto \frac{\log(b)}{\log(a)} \left(\sum_{i_0 \in \sigma^{-1}(0)} x_{i_0}, \dots, \sum_{i_0 \in \sigma^{-1}(m)} x_{i_0} \right)$$

Extensions of classical constructions

For general simplicial set X , we can define

$$Real(X) = \lim_{\rightarrow} Real(\Delta_{<a}^n)$$

Analogously to the classical case, we can define functor $Sing$:

$$Sing(Y) : ([n], (0, a)) \mapsto \text{hom}_{\mathbf{EPMet}}(Real(\Delta_{<a}^n), Y)$$

But in our case, we are interested in finite metric spaces. We can consider subcategory of bounded fuzzy simplicial sets **Fin** – **sFuzz** and therefore use the analogous adjoint pair $FinReal$ and $FinSing$.

Definition

Define $FinReal : \mathbf{Fin} - \mathbf{sFuzz} \rightarrow \mathbf{FinEPMet}$ by setting

$FinReal(\Delta_{<a}^n) = (x_1, \dots, x_n, d_a)$, where

$$d_a(x_i, x_j) = \begin{cases} -\log(a) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

Extensions of classical constructions

Similarly, for bounded simplicial set X $FinReal(X) = \lim_{\rightarrow} FinReal(\Delta_{<a}^n)$.
 Similar to Spivak's construction, for map $\Delta_{<a}^n \rightarrow \Delta_{<b}^m$, the action of $FinReal$ is given by

$$(x_1, \dots, x_n, d_a) \mapsto (x_{\sigma(1)}, \dots, x_{\sigma(n)}, d_b)$$

Definition

Define $FinSing : FinEPMet \rightarrow Fin - sFuzz$ by

$$FinSing(Y) : ([n], (0, a)) \mapsto \text{hom}_{FinEPMet}(FinReal(\Delta_{<a}^n), Y)$$

Extensions of classical constructions

The main theorem:

Theorem

The functors $FinReal$ and $FinSing$ form adjunction $FinReal \dashv FinSing$

Returning to our problem

Now we can handle the family of incompatible metric spaces described above. Each metric space can be translated into a fuzzy simplicial set via the fuzzy singular set functor. We get the family of fuzzy simplicial sets for which we can just take a fuzzy union across the entire family. The result is a single fuzzy simplicial set which captures the relevant topological and metric structure of the manifold \mathcal{M} .

Returning to our problem

But fuzzy singular functor applies to e.p. metric spaces. The result of Lemma 1 only provide accurate approx. of geodesic distance local to X_i for distances measured from X_i – the geodesic distances between other pairs of points within the neighborhood of X_i are not well defined. So due to the lack of information, let $d_i(X_j, X_k) = \infty$ for $i \neq j$ and $i \neq k$.

Returning to our problem

Definition

Let $X = \{X_1, \dots, X_N\}$ be a dataset in \mathbb{R}^n and let $\{(X, d_i)\}$ be a family of e.p. metric spaces such that

$$d_i(X_j, X_k) = \begin{cases} d_{\mathcal{M}}(X_j, X_k) - \rho & \text{if } i = j \text{ or } i = k \\ \infty & \text{otherwise} \end{cases}$$

where ρ is the distance to the nearest neighbor of X and $d_{\mathcal{M}}$ is geodesic distance on the manifold.

The fuzzy topological representation of X is

$$\bigcup_{i=0}^n \text{FinSing}((X, d_i))$$

Local connectivity

Why we subtract ρ there? It's because we assume that our manifold is locally connected. This assumption is natural in topological sense and actually helps us to prevent the curse of dimensionality [4].

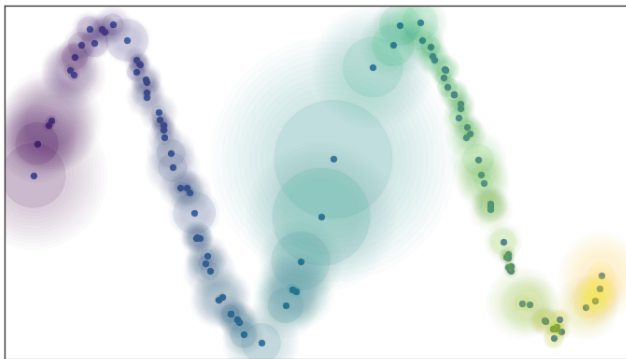


Figure The image is taken from [3]

Layout

Let $Y = \{Y_1, \dots, Y_N\} \subseteq \mathbb{R}^d$ be a low dim. representation of X . Since we know apriori a manifold on which Y lies on (usually just \mathbb{R}^d) we can compute fuzzy topological representation directly.

We will use fuzzy set cross entropy to compare 2 fuzzy sets. For this, we need to revert to classical fuzzy set notation, i.e. to a reference set A and function $\nu : A \rightarrow [0, 1]$. Having fuzzy simplicial set \mathcal{P} , we can let

$$A = \bigcup_{a \in (0,1]} \mathcal{P}([0, a)) \text{ and } \nu(x) = \sup\{a \in (0, 1] | x \in \mathcal{P}([0, a))\}$$

Layout

Definition

The cross entropy C of two fuzzy sets (A, ν) and (A, μ) is

$$C((A, \nu), (A, \mu)) = \sum_{a \in A} \left(\nu(a) \log \frac{\nu(a)}{\mu(a)} + (1 - \nu(a)) \log \frac{1 - \nu(a)}{1 - \mu(a)} \right)$$

We can optimize embedding Y w.r.t. fuzzy set cross entropy using gradient descent. Fuzzy singular set functor is then approximated via differentiable functions.

Basic structure for approximating manifold

- Graph Construction

- ① Construct a weighted k -neighbour graph.
- ② Apply some transform on the edges to ambient local distance.
- ③ Deal with the inherent asymmetry of the k -neighbour graph.

- Graph Layout

- ① Define an objective function that preserves desired characteristics of this k -neighbour graph.
- ② Find a low dimensional representation which optimizes this objective function.

Graph Construction

From now, let $X = \{x_1, \dots, x_N\}$ be the input dataset with a metric (or dissimilarity measure) $d : X \times X \rightarrow \mathbb{R}_{\geq 0}$.

Given the hyperparameter k , we compute the set $\{x_{i1}, \dots, x_{ik}\}$ of the k nearest neighbors for each $x_i \in X$ under the metric d .

Graph Construction

For each x_i we define 2 parameters: ρ_i and σ_i s.t.

$$\rho_i = \min\{d(x_i, x_j) | 1 \leq j \leq k, d(x_i, x_j) > 0\}$$
$$\sum_{j=1}^k \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = \log_2 k$$

- ρ_i – local connectivity
- σ_i – smoothed normalisation factor, defining the Riemannian metric local to x_i

Graph Construction

Define a weighted directed graph $\bar{G} = (V, E, w)$, where:

- Vertices V are just points X
- $(x_i, x_j) \in E$ if x_j is a k -neighbor of x_i
- $w((x_i, x_j)) = \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right)$

This graph is 1-skeleton of the fuzzy simplicial set associated to the metric space local to x_i where the local metric is defined in terms of ρ_i and σ_i . The weight – membership strength of the corresponding 1-simplex within the fuzzy simplicial set (or one can think of the weight as the probability that the edge exists).

Unified topological representation

Let A be the weighted adjacency matrix of \bar{G} , and consider

$$B = A + A^T - A \circ A^T,$$

where \circ is the Hadamard product. If A_{ij} – probability that the directed edge (x_i, x_j) exists, then B_{ij} is the probability that either (x_i, x_j) or (x_j, x_i) exists.

Then UMAP graph G is an undirected weighted graph whose adjacency matrix is B .

Unified topological representation

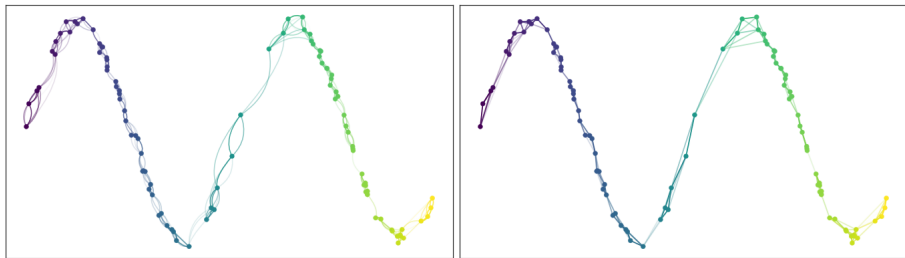


Figure Images are taken from [3]

Force directed graph layout algorithm

UMAP uses a force directed graph layout algorithm in low dimensional space. Force directed layout algorithm requires a description of both the attractive and repulsive forces.

In UMAP the attractive force between i and j is

$$\frac{-2ab \|y_i - y_j\|_2^{2(b-1)}}{1 + \|y_i - y_j\|_2^2} w((x_i, x_j))(y_i - y_j)$$

The repulsive force is given by:

$$\frac{2b}{(\varepsilon + \|y_i - y_j\|_2^2)(1 + a \|y_i - y_j\|_2^{2b})} (1 - w((x_i, x_j)))(y_i - y_j)$$

Force directed graph layout algorithm

The algorithm applies iteratively the forces at each edge. Slowly decreasing the forces, it converges to a local minima.

The forces are derived from gradients optimising the cross-entropy between UMAP graph G and an equivalent weighted graph H with points y_i .

We are looking for the positions of y_i such that H closely approximates G . Since G captures the topology of the source data, H matches the topology as closely as the optimization allows.

MNIST

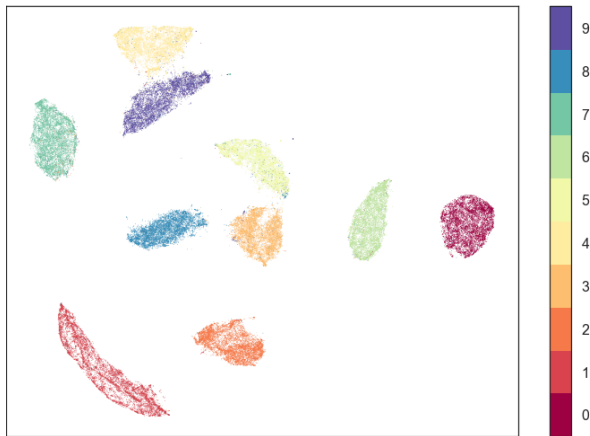


Figure The image is taken from [1]

FashionMNIST

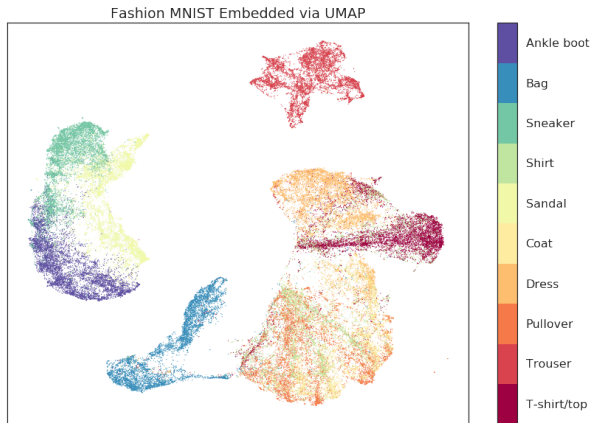






Figure The image is taken from [3]

References

-  L. McInnes et. al. – UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020
-  David I Spivak. Metric realization of fuzzy simplicial sets. Self published notes, 2012.
-  "How UMAP works", UMAP documentation. https://umap-learn.readthedocs.io/en/latest/how_umap_works.html
-  UMAP Uniform Manifold Approximation and Projection for Dimension Reduction — SciPy 2018 — <https://youtu.be/nq6iPZVUxZU>