A Brief Survey of UMAP[1]

Kokic Liu

In 2008, van der Maaten and Geoffrey Hinton introduced t-SNE [1][2] as a method for visualizing high-dimensional data. This visualization approach essentially implemented what we now call **dimensionality reduction**. While t-SNE's dimensionality reduction performance is now recognized as limited, it represented a significant advancement at the time, combining capabilities that were unmatched by existing alternatives[3]. The novel theoretical framework of t-SNE made profound contributions to data science, establishing it as a seminal work in the field. A decade later, McInnes, Healy and colleagues developed UMAP [2], which claimed to resolve t-SNE's two primary weaknesses - computational efficiency and memory requirements - through a fundamentally different mathematical approach. Within a year, UMAP gained widespread recognition due to its superior performance in biological applications[4]. This article systematically presents UMAP's core concepts and mathematical foundations to provide readers with a concise yet comprehensive understanding of this influential technique.

Both t-SNE and UMAP belong to manifold learning, a field that has achieved considerable theoretical advancements, though early works often exhibited limited practical effectiveness. The fundamental premise of manifold learning is that certain high-dimensional data are inherently embedded within a lower-dimensional manifold structure in the high-dimensional space. Consequently, dimensionality reduction can significantly mitigate the frequent "curse of dimensionality" problem encountered in high-dimensional data spaces. Thus, even if real-world data does not fully conform to the assumption of an underlying low-dimensional manifold structure, dimensionality reduction methods can still preserve as much original information as possible while reducing data dimensions.

---

[1]Posted on November 9, 2024.

[2]The name stands for t-Distributed Stochastic Neighbor Embedding.

[3]Popular linear methods included PCA and LDA, while nonlinear approaches consisted mainly of Isomap and LLE.

[4]UMAP has become the dominant method in single-cell genomics and sees extensive use in statistical genetics.

The proposed t-SNE algorithm achieved superior visualization quality compared to contemporary techniques including Sammon mapping, Isomap, and Locally Linear Embedding (LLE) across most benchmark datasets. However, its performance in general dimensionality reduction tasks proved inadequate, often requiring extensive data preprocessing. McInnes and Healy's critical analysis suggested t-SNE's limitations stemmed from problematic theoretical assumptions, particularly in its probability formulations that failed to capture true data relationships. This is exemplified in the key probability definition p_{ij}, where the Gaussian-based pairwise distance modeling exhibits fundamental constraints.

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}, \qquad p_{j|i} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-\|x_i - x_k\|^2}{2\sigma_i^2}\right)}$$

Consequently, McInnes et al. emphasized that their employment of mathematical tools like fuzzy logic could better address the complexities arising from assumptions that more accurately reflect real-world data characteristics. They further maintained that, in the long term, decisions grounded in rigorous theoretical foundations would enable the development of more scalable and generalizable algorithms. The UMAP algorithm operates under three fundamental assumptions about the data:

1. There exists a manifold on which the data is uniformly distributed.
2. The underlying manifold is locally connected.
3. The primary objective is to preserve the topological structure of this manifold.

In subsequent sections, we will examine why these three points must serve as axiomatic foundations for the entire algorithm. First, we introduce several definitions that will be utilized in later discussions.

Definition 1.1. A **simplex** is the n-dimensional generalization of geometric concepts such as triangles and tetrahedrons, where $n \in \mathbb{N}$. The n-dimensional simplex is defined as the topological space $\mathbb{\Delta}^n$, which forms a specific subspace of $\mathbb{R}^{n+1}$:

$$\mathbb{\Delta}^n = \left\{(x_0, \cdots, x_n) \in \mathbb{R}^{n+1} \,\middle|\, x_0, \cdots, x_n \geq 0, \textstyle\sum_{0 \leq i \leq n} x_i = 1\right\}$$

equipped with the subspace topology from $\mathbb{R}^{n+1}$. This particular set is referred to as the standard n-simplex. Simplices serve as fundamental building blocks for constructing more complex geometric structures. In topology, simplices can be assembled into simplicial complexes, while in combinatorics they form simplicial sets.

Definition 1.2. The **simplex category** $\mathbb{\Delta}$ is defined as follows:

- Its objects are all sets of the form $[n] = \{0, \cdots, n\}$ where $n \geq 0$.
- The morphism set $\Delta([m], [n])$ consists of all order-preserving maps from $[m]$ to $[n]$ (i.e., maps that preserve the $\leq$ relation).

The category $\mathbb{\Delta}$ can also be viewed as the category of standard simplices with ordered vertices, where morphisms are linear maps between simplices that map vertices to vertices while preserving their ordering.

Definition 1.3. A **simplicial set** is a functor from $\mathbb{\Delta}^{\mathrm{op}}$ to Set, that is, a contravariant functor from the simplex category $\mathbb{\Delta}$ to Set, or equivalently, a presheaf on $\mathbb{\Delta}$. The category of simplicial sets is denoted $\mathsf{sSet} = \mathsf{Fun}\big(\mathbb{\Delta}^{\mathrm{op}}, \mathsf{Set}\big)$.

Definition 1.4. For the simplex category $\mathbb{\Delta}$, let $\mathbb{\Delta}_{\leq n}$ denote its full subcategory on objects $[0], [1], \cdots, [n]$. The inclusion map $\mathbb{\Delta}|_{\leq n} \hookrightarrow \mathbb{\Delta}$ induces a truncation functor $\mathsf{tr}_n : \mathsf{sSet} = \left[\mathbb{\Delta}^{\mathrm{op}}, \mathsf{Set}\right] \to \left[\mathbb{\Delta}^{\mathrm{op}}_{\leq n}, \mathsf{Set}\right] = \mathsf{sSet}_{\leq n}$ which restricts a simplicial set to degrees $\leq$ n. This functor has a fully faithful left adjoint, which can be given by right Kan extension: $\mathsf{sk}_n : \mathsf{sSet}_{\leq n} \to \mathsf{sSet}$ This is called the **n-skeleton**.

Ideally, the authors of UMAP aim for the low-dimensional representation to possess a fuzzy topological structure as similar as possible to the original. This objective raises two key issues: first, how to determine the fuzzy topological structure of the low-dimensional representation; and second, how to identify an optimal fuzzy topological structure.

Definition 1.5. Let $X = \{X_1, \cdots, X_n\}$ be a dataset in $\mathbb{R}^n$, and $\{(X, d_i)\}_{1 \leq i \leq n}$ be a family of extended pseudometric spaces with common carrier set X, where

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

Here, ρ represents the distance to the nearest neighbor of X_i, and $d_{\mathcal{M}}$ denotes the geodesic distance on the manifold $\mathcal{M}$. Both distances can be either

precomputed or approximated. The **fuzzy topological representation** of X is defined as

$$\bigcup_{i=1}^{n} \mathsf{FinSing}((X, d_i))$$

This construction yields a fuzzy simplicial set that serves as a global representation of the manifold, formed by piecing together numerous local representations. Given that such topological structures can be constructed either from known manifolds or by learning the metric structure of the manifold, dimensionality reduction can be achieved by finding a low-dimensional representation whose topological structure closely matches that of the source data. The remaining challenge is to determine how to find such an optimal low-dimensional representation.

UMAP follows a workflow similar to t-SNE as its successor, both employing graph layout algorithms to arrange data in low-dimensional space. The process involves first constructing a high-dimensional graph representation of the data, then optimizing a low-dimensional graph to maximize structural similarity. Although the original UMAP paper presents the high-dimensional graph construction and proves its properties using sophisticated mathematical tools with specialized prerequisites, the underlying intuition is quite straightforward. For instance, the **fuzzy simplicial complex** defined above can be intuitively understood[5] as a **representation of a weighted graph**.

Simplicial complexes provide a reasonable approach to initially capturing the fundamental topology of datasets. Importantly, most of the work is actually accomplished by 0-simplices and 1-simplices, which are computationally more tractable - in the sense of nodes and edges, they simply form a graph. This observation leads to the Vietoris-Rips complex, which is similar to the Čech complex but completely determined by 0-simplices and 1-simplices. The Vietoris-Rips complex is computationally more feasible to use, particularly for large datasets, and serves as one of the primary tools in topological data analysis.

The method described above explains why neighborhood graph-based approaches should capture manifold structures during dimensionality reduction. However, when attempting to implement this theory in practice, the

[5]Naturally, this intuition primarily holds for 0-simplices and 1-simplices.

first obvious difficulty encountered is selecting an appropriate radius for the balls that form the open cover. If the chosen radius is too small, the resulting simplicial complex will fragment into many disconnected components. If the radius is too large, the simplicial complex will collapse into a few very high-dimensional simplices and their faces, thereby failing to capture the manifold structure.

This dilemma arises partly because the Nerve Theorem provides theoretical justification for this topology-capturing process. Specifically, the theorem shows that the simplicial complex will be homotopy equivalent to the union of the cover. When working with finite data, for certain radii, the cover may not encompass the entire underlying manifold we assume for our data - this is precisely what happens when radius r is too small, leading to insufficient coverage and disconnected components. Similarly, when points are too densely packed, the resulting cover will include excessive data, producing higher dimensionality than desired. However, if the data is uniformly distributed on the manifold, selecting an appropriate radius becomes straightforward, as the average distance between points will be properly characterized. Moreover, uniform distribution ensures that the cover will span the entire manifold without gaps or unnecessary disconnected components, while also avoiding those problematic clustering effects that lead to undesirably high-dimensional simplifications.

Lemma 1.6. Let $(\mathcal{M}, g)$ be a Riemannian manifold in $\mathbb{R}^n$, and $p \in M$ a point. If g remains locally constant in an open neighborhood U containing p, making g a constant diagonal matrix in the ambient coordinates, then within a ball $B \subseteq U$ centered at p with volume $\pi^{\frac{n}{2}}\Gamma\left(\frac{n}{2}+1\right)^{-1}$, for any $q \in B$, the geodesic distance from p to q is $\frac{1}{r}d_{\mathbb{R}^n}(p,q)$, where r is the radius of the ball in ambient space, and $d^n_{\mathbb{R}}$ is the existing metric in ambient space.

Assuming the set of all possible 1-simplices is E[6], where $w_h(e)$ is the weight of 1-simplex e in high dimensions and $w_l(e)$ is its weight in low dimensions, the cross-entropy is:

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$

[6]Do not take this literally - the definition of simplicial sets was provided earlier.

This weighted sum actually has clear physical significance. Returning to the graph representation of data, minimizing cross-entropy becomes a force-directed graph layout algorithm. For each term in the sum: $w_h(e)\log\left(\frac{w_h(e)}{w_l(e)}\right)$ provides attraction between the endpoints of edge e when there's large weight in high dimensions, because this term is minimized when $w_l(e)$ is maximized (occurring when points are as close as possible). Conversely, $(1-w_h(e))\log\left(\frac{1-w_h(e)}{1-w_l(e)}\right)$ provides repulsion between endpoints when $w_h(e)$ is small, as minimizing this term requires minimizing $w_l(e)$.

This constitutes a non-convex optimization problem, where convergence to local minima is ensured by slowly reducing attraction and repulsion forces, similar to the approach used in simulated annealing. In UMAP, the attractive force between vertices i and j with coordinates y_i and y_j is determined by:

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

where a and b are hyperparameters. Due to computational constraints, repulsion is implemented through sampling. Thus, whenever attraction is applied to an edge, one of its vertices is repelled by sampled other vertices. The repulsive force is given by:

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

Here ε is chosen as a small number to prevent division by zero when $y_i = y_j$.

Generally, the cross entropy C between two fuzzy sets $(A,\mu),(A,\nu)$ can be defined analogously as:

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

This serves as UMAP's cost function. The key advantage is that it enables direct optimization of embeddings by minimizing fuzzy set cross entropy. Specifically, note that:

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

Let's denote each term of the first sum as $C_\mu(a)$, which depends only on the fixed μ values during optimization. Thus minimizing cross entropy depends solely on the second sum. We only need to minimize:

$$
-\sum_{a\in A} (\mu(a)\log(\nu(a)) + (1-\mu(a))\log(1-\nu(a)))
$$

Denoted as S, basic estimates show: $S \geq \sum_{a\in A} \mu(a) + \nu(a) - 2\mu(a)\nu(a)$

Similar to t-SNE[7], we can optimize the embedding Y by minimizing C using stochastic gradient descent. This requires introducing a differentiable fuzzy singular set functor. UMAP constructs topological representations by **approximating the manifold and piecing together local fuzzy simplicial set representations**, then optimizes the low-dimensional layout to minimize error between topological representations.

Focusing on 1-skeletons of fuzzy simplicial sets, let X_i be the fuzzy set of i-simplices of X, with λ_i as weights. The cost function C_ℓ can be defined as:

$$
C_\ell(X,Y) \quad \stackrel{\text{def}}{=} \quad \sum_{1\leq i\leq \ell} \lambda_i C_\ell(X_i, Y_i)
$$

For implementation, we sample 1-simplices with probability $\mu(a)$ and update according to $\nu(a)$. Negative sampling handles the $(1-\mu(a))\log(1-\nu(a))$ terms by randomly sampling potential 1-simplices as negative examples. This yields:

$$
P(x_i) \quad = \quad \frac{\sum_{\{a\in A\,:\,d_0(a)=x_i\}} 1-\mu(a)}{\sum_{\{b\in A\,:\,d_0(b)\neq x_j\}} 1-\mu(b)}
$$

For large datasets, uniform distribution provides reasonable approximation for negative sampling. Thus we can apply gradient descent optimization once we find a differentiable approximation $\nu(a)$ for a given 1-simplex a.

[7]The difference being t-SNE uses KL divergence as cost function.

References

1. Maaten, L. Van der, Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research. 9, (2008)
2. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. (2018)