Extended Abstract: Uniform Manifold Approximation and Projection (UMAP)

Victor Tuekam*

December 21, 2020

*Ludwig Maximilian University of Munich, t.victor@campus.lmu.de

Introduction

UMAP is a nonlinear dimensionality reduction technique with strong mathematical footing. Published in 2018 [McInnes et al., 2018], UMAP is a new graph layout algorithm (such as Local Linear Embedding [Roweis and Saul, 2000] and t-SNE [Maaten and Hinton, 2008]) built on the ideas of Riemannian geometry and Algebraic topology.

Theoretical foundation

UMAP is a two-fold algorithm. In a first step, a high-dimensional weighted neighborhood-graph representation of the data is built. The second step requires optimizing a low-dimensional (in euclidean space) representation of data to minimize the cross-entropy between the two topological representations.

We shall start by intuitively defining some of the theoretical building blocks and terms found in the UMAP literature [McInnes], [Andy and Adam]:

1. **Global and Local structure:** Global structure refers to the separation between clusters, while local structure would represent the separation within clusters.
2. **k-simplex:** A `k` dimensional simplex is built by connecting `k + 1` points, i.e., a data point is a 0-simplex, a 1-simplex is a line segment between 2 points, a 2-simplex is a triangle (with 1-simplices as faces).
3. **Open cover:** An open cover is a family of sets (set of simplices) whose union is the whole set (manifold).
4. **Simplicial complex:** A set of simplices glued together along faces. Mathematically the simplicial complex is equivalent to the union of the cover.
5. **Čech complex:** A way of building an abstract simplicial complex through the intersection of sets (cover).
6. **Vietoris-Rips Complex:** Just like the Čech complex but solely determined by 0- and 1-simplices. It is computationally easier to work with particularly for large datasets.

A way to approximate an open cover of a topological space is to create balls of fixed radius about each data point. The simplicial complex will then be formed by connecting intersecting balls with a straight line (e.g., two intersecting balls will form a 1-simplex, while 3 intersecting balls will form a 2-simplex, and so on). However, 0- and 1-simplices are the dominant constituents of the simplicial complex. Which motivates the use Vietoris-Rips complexes. See [Andy and Adam] for a visual representation of a simplicial complex.

One problem that arises is that of choosing a good radius for the cover. Another problem, is that in the real world, data is generally unevenly distributed across the manifold, with regions of the data space denser than others, which would again lead to unnecessary high-dimensional simplicies. To solve these problems, UMAP assumes that the data on the manifold are uniformly distributed [McInnes et al., 2018]. Under this assumption, a ball centered at a data point containing the k-nearest-neighbors of the data point should have the same fixed volume, independent of the point. This means that a unit ball about a point should contain exactly k-nearest-neighbors independent of the point. So when the data appears to be unevenly distributed, the assumption is not relaxed but a unique geodesic distance function is given to each point, so that the ball around it has the k-nearest-neighbors and a radius of one with respect to this distance metric. Thus each unit ball forms a separate metric space. Now, the problem of finding the radius is converted to that of finding a good value for `k`. The reason for this is that a good value for `k` is less dataset dependent than a good value for the radius, e.g., a value of `k = 10` might work well for most datasets. Because no point should be completely isolated, an additional assumption of UMAP is that the manifold is locally connected.

Because of the local distance metric defined by each unit ball, distances between points are incompatible (`d(a, b) != d(b, a)`). We can think of this a directed pair of edges between two points, weighted by a function of their distances, representing the probability that the edge exists. The two edges are then combined with a weight indicating the probability that at least one of the edges exists [McInnes]. The result is therefore a weighted neighborhood-graph representation of the data on the manifold.

Finally, in the second step of UMAP, spectral embedding techniques (e.g., Laplacian Eigenmaps) are usually used to get an initial low-dimenional representation of the data [McInnes]. This low-dimensional representation is then optimized to minimize the cross entropy loss with the high-dimensional neighborhood-graph representation. The use of cross entropy ensures that the two representations are topological close to each other (the neighborhoods in the higher-dimensional representation are preserved as much as possible in the embedding).

Hyperparameters

UMAP takes 4 hyperparameters [McInnes et al., 2018]:

1. `k`, the number of neighbors used to compute the neighborhood-graph. `k` represents the trade-off between local and global structure preservation, where small values of k preserve the local neighborhood structure, while large values preserve the global structure.
2. `d`, the target embedding dimension.
3. `min_dist`, the minimum distance allowed between points in the low-dimensional embedding space.
4. `n_epochs`, the number of training epochs to use when optimizing the low-dimensional representation.

Comparison with other techniques

[McInnes et al., 2018] provides comparisons of UMAP with other algorithms, notably t-SNE and LargeVis [Tang et al., 2016]. In terms of quality of the embedding, UMAP performs similarly to both t-SNE and LargeVis. UMAP however tends to preserve more of the global structure of the data. Computationally, the run time of UMAP is better than of t-SNE and LargeVis, and it scales better to embeddings in dimensions higher than two and larger datasets.

Weaknesses

The weaknesses of UMAP outlined in [McInnes et al., 2018] are summarized below:

1. The dimensions of the resulting embedding has no specific meaning and are thus not interpretable.
2. UMAP assumes that the data live on a manifold in an ambient space. If the data are noisy, UMAP would try to preserve the noisy structure of the data in its low-dimensional representation.
3. The assumption that the data are uniformly distributed on the manifold is less suitable, when one knows that the data are inherently unevenly distributed on the manifold.
4. Even though UMAP preserves the global topological structure of the data better than t-SNE, the algorithm is built on the premise that local distance is more important than global distances. Other algorithms such are Multi-dimensional scaling are better suited when the goal is to represent global distances with higher fidelity.

Applications

UMAP is widely used as a dimensionality reduction technique. Dimensionality reduction generally allows for visualization of high-dimensional data and subsequently in the identification of clusters in the data. [Diaz-Papkovich et al., 2021] reviews the application of UMAP in population genetics. A main characteristic of genomic data are its high-dimensionality. While techniques such as PCA allows for projections of data in low-dimensional spaces while preserving the variance in the dataset, these techniques often do not preserve the local pattern in the data very well. UMAP is particularly suitable for genomic data because the embeddings produced allows for closely related observations to be closer together in the embedded space while mildly preserving the global structure of the data. This makes it possible to uncover features of the data such as demographic history, covariations between genetics, etc. [Diaz-Papkovich et al., 2021]

References

Coenen Andy and Pearce Adam. A deeper dive into umap theory. URL https://pair-code.github.io/understanding-umap/supplement.html.

Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66, 1 2021. ISSN 1434-5161. doi: 10.1038/s10038-020-00851-4.

Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-sne, 2008. URL https://lvdmaaten.github.io/tsne/.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

Leland McInnes. How umap works — umap 0.5 documentation. URL https://umap-learn.readthedocs.io/en/latest/how_umap_works.html.

Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323. URL http://www.sciencemag.org/cgi/content/abstract/290/5500/2323.

Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualization large-scale and high-dimensional data. *CoRR*, abs/1602.00370, 2016. URL http://arxiv.org/abs/1602.00370.