SEPTEMBER 15, 2023

RESEARCH ON DIMENSIONALITY REDUCTION

UMAP PROJECTION BASED

JOSEPH JESÚS AGUILAR RODRÍGUEZ
MANUEL NICOLÁS NAHIB FRANCO VALENCIA
MARCO ALEJANDRO GONZÁLEZ BARBUDO

UMAP Projection Based:

UMAP is a dimensionality reduction technique that aims to map high-dimensional data to a lower-dimensional space while preserving the underlying structure of the data. It does this by constructing a fuzzy topological representation of the data in the high-dimensional space and then projecting this representation into a lower-dimensional space.

Here's how UMAP works, which involves a form of projection:

1. High-Dimensional Space: UMAP starts with your dataset in a high-dimensional space. This could be data with many features or dimensions.
2. **Neighborhood Graph Construction:** UMAP starts by constructing a neighborhood graph from the high-dimensional data. This graph represents the relationships between data points. It connects nearby points and is based on a measure of similarity, often using distance metrics like Euclidean distance or cosine similarity.
3. Constructing a Fuzzy Representation: UMAP constructs a fuzzy simplicial set, which is a mathematical structure that captures the relationships and similarities between data points. In this step, UMAP calculates pairwise similarities between data points, typically using a Gaussian kernel or another similarity metric. These similarities represent how close or similar two data points are in the high-dimensional space.
4. Creating a Lower-Dimensional Projection: UMAP then aims to project this fuzzy topological representation into a lower-dimensional space. This is where the term "projection" comes into play. The goal is to find a lower-dimensional representation of the data that preserves the topological structure and pairwise similarities as much as possible.
5. Optimization: UMAP uses optimization techniques, including stochastic gradient descent, to find the optimal lower-dimensional representation. The optimization process aims to minimize a cost function that measures the discrepancy between the fuzzy representation in the high-dimensional space and the lower-dimensional projection.
6. Result: The final output of UMAP is a lower-dimensional representation of your data, where each data point is mapped to a point in this space. This lower-dimensional space typically has a much lower dimensionality than the original space, making it suitable for visualization and further analysis.

The key advantages of UMAP compared to other dimensionality reduction techniques like t-SNE are its scalability and speed. UMAP tends to be faster and can handle larger datasets more efficiently, while still providing meaningful representations of the data. It also offers better preservation of global structure,

making it useful for various applications, including clustering, visualization, and exploration of high-dimensional data.

UMAP is commonly used in fields like data analysis, bioinformatics, natural language processing, and computer vision to gain insights into complex datasets and visualize them in a way that can help researchers and analysts better understand their underlying structure and patterns.

Installation

Conda install, via the excellent work of the conda-forge team:

conda install -c conda-forge umap-learn

The conda-forge packages are available for linux, OS X, and Windows 64 bit.

PyPI install, presuming you have numba and sklearn and all its requirements (numpy and scipy) installed:

pip install umap-learn

References:

- Coenen, A., & Pearce, A. (n.d.). Understanding umap. PAIR Page Redirection. https://pair-code.github.io/understanding-umap/#:~:text=UMAP%2C%20at%20its%20core%2C%20works,as%20structurally%20similar%20as%20possible
- McInnes, L., Healy, J., & Melville, J. (2020, September 18). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Tutte Institute for Mathematics and Computing.
- Uniform manifold approximation and projection for dimension reduction. UMAP. (n.d.). https://umap-learn.readthedocs.io/en/latest/