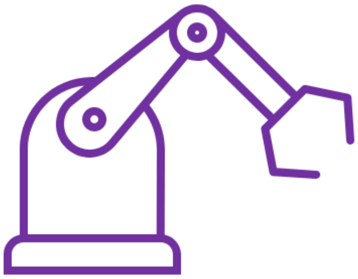
UNIVERSIDAD POLITÉCNICA DE YUCATÁN

BIS Universities

UMAP

Team 6



Teacher: Victor Ortiz

University: Polytechnical University of Yucatán

Date: September 15, 2023

Group: 9°A robotics

- Eduardo Antonio Flores Arellano
- Axel Esdras Flores Islas
- Edwin Antonio Can Pinto
- Mauricio Daniel Zaldivar Medina
- Angel Iván Mayo Carrillo

UMAP

UMAP, developed by McInnes, is an innovative technique that offers several advantages over t-SNE, most notably increased speed and better preservation of the data's global structure. In this article, we will delve into the theory behind UMAP to gain a deeper understanding of how the algorithm operates, how to utilize it effectively, and how its performance compares to that of t-SNE.

Dimensionality reduction is a powerful tool for machine learning practitioners to visualize and understand large, high dimensional datasets.

UMAP is an incredibly potent tool in the data scientist's toolkit, offering several advantages over t-SNE. While both UMAP and t-SNE produce somewhat similar results, the increased speed, superior preservation of global structure, and more interpretable parameters make UMAP a more efficient instrument for visualizing high-dimensional data. It's crucial to bear in mind that the no dimensionality reduction technique is flawless – by necessity, we're distorting the data to fit it into lower dimensions – and UMAP is no exception. However, by cultivating an intuitive understanding of how the algorithm operates and grasping how to fine-tune its parameters, we can harness this powerful tool more effectively to visualize and comprehend extensive, high-dimensional datasets.

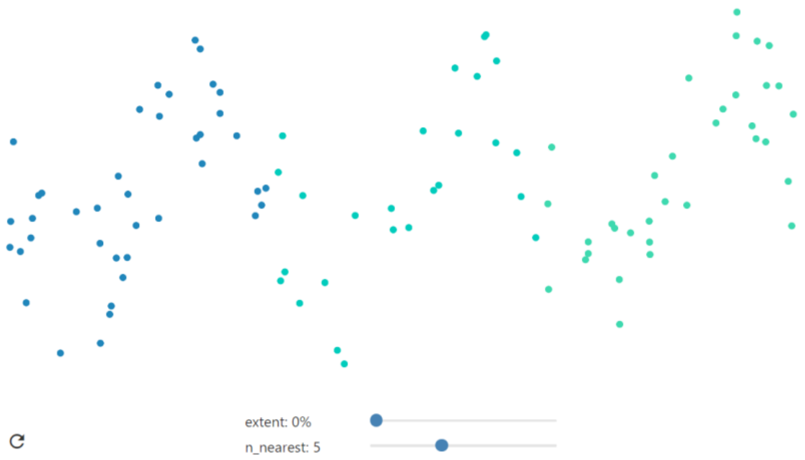
Underlying Principles of UMAP

Uniform Manifold Approximation and Projection (UMAP) is grounded in several fundamental mathematical principles to achieve its goal of dimensionality reduction while preserving the inherent data structure. The primary underlying mathematical principles of UMAP include:

Topology and Geometry

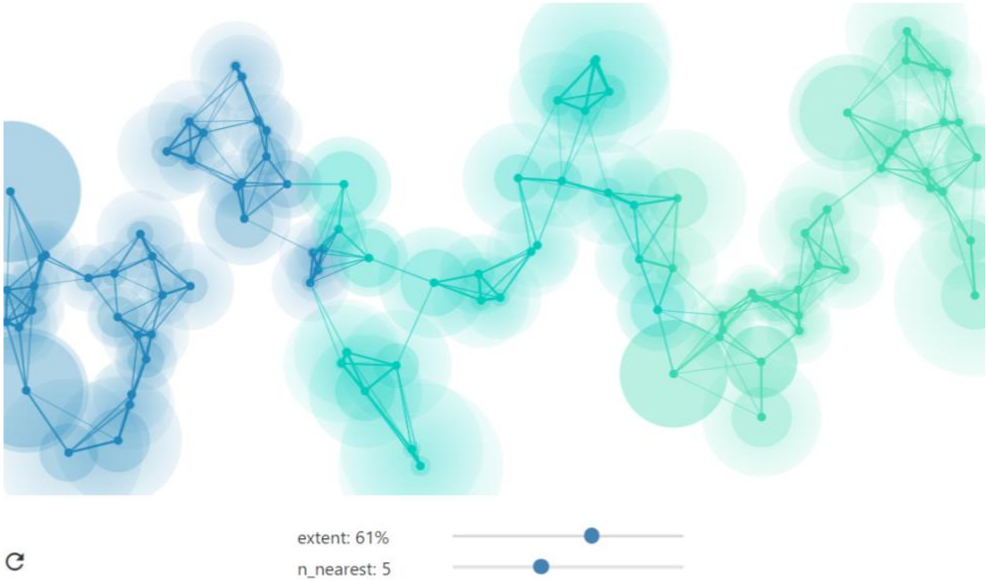
The algorithm relies on a number of insights from algebraic topology and Riemannian geometry. UMAP essentially constructs a weighted graph from the high dimensional data, with edge strength representing how “close” a given point is to another, then projects this graph down to a lower dimensionality.

UMAP relies on constructing what’s known as a Čech complex, which is a way of representing a topology combinatorially. To get there, a basic building block called a simplex is used. Geometrically, a simplex is a k-dimensional object formed by connecting k + 1 points - for example, **0-simplex** is a point, a **1-simplex** is a line, and a **2-simplex** is a triangle.



extent: 0%

n_nearest: 5



extent: 61%

n_nearest: 5

Diffusion

Is a process in which particles or information spreads out from a concentrated area to less concentrated areas. UMAP uses diffusion to model the spread of information from high-dimensional space to low-dimensional space. It helps to capture the local relationships between data points, ensuring that similar points remain close in the low-dimensional representation.

Cross-Entropy Minimization

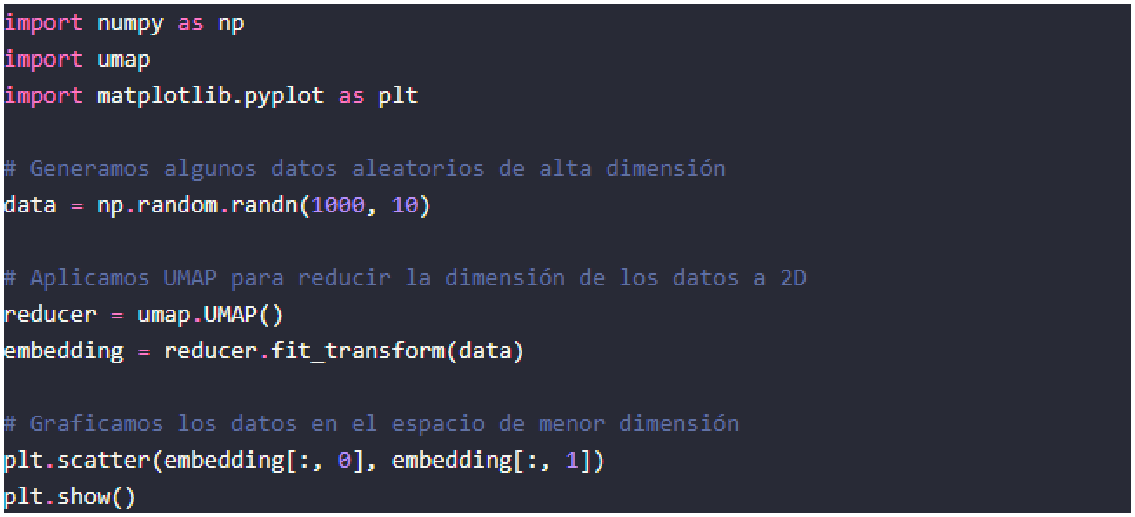
Cross-entropy is a measure of the difference between two probability distributions. In UMAP, cross-entropy minimization is used to optimize the mapping from the high-dimensional space to the low-dimensional space.

Stochastic Optimization

UMAP executes by combining topological, geometric, and probabilistic techniques to effectively achieve dimensionality reduction of data while preserving the underlying structure. Thanks to these mathematical principles, UMAP has become a powerful tool in data visualization and analysis. Preserves data structures by constructing a weighted graph that captures local relationships among points in the high-dimensional space and by employing a diffusion process to propagate this information into the low-dimensional space. This approach ensures that close relationships between points are maintained in the final projection, making UMAP an effective technique for dimensionality reduction and data structure preservation.

How Does UMAP Work?

UMAP works by using a cost function that balances the preservation of local distances, global distances, and the preservation of the topological structure. This cost function is optimized using gradient descent. The result is a low-dimensional representation of the data that can be easily visualized and analyzed.



```
import numpy as np
import umap
import matplotlib.pyplot as plt

# Generamos algunos datos aleatorios de alta dimensión
data = np.random.randn(1000, 10)

# Aplicamos UMAP para reducir la dimensión de los datos a 2D
reducer = umap.UMAP()
embedding = reducer.fit_transform(data)

# Graficamos los datos en el espacio de menor dimensión
plt.scatter(embedding[:, 0], embedding[:, 1])
plt.show()
```

Advantages and disadvantages

Advantages

1. UMAP is capable of preserving the global structure of the data, including topological structure and distances between points, making it a good choice for visualization and exploration.
2. UMAP is computationally efficient and scalable, making it suitable for large datasets.
3. UMAP can handle high-dimensional data, unlike some other dimensionality reduction methods that are limited to low-dimensional data.

Disadvantages

4. UMAP is a relatively new method and has not been as thoroughly tested and validated as some other dimensionality reduction methods.
5. UMAP may not perform well on data with complex non-linear structure or on data with non-uniform density.

Use Cases and Applications of UMAP

Uniform Manifold Approximation and Projection (UMAP) is a versatile dimensionality reduction technique that has found applications across various domains due to its ability to effectively preserve the global structure of high-dimensional data. Understanding its practical use cases can provide valuable insights into how UMAP can benefit data analysis and visualization efforts.

- Bioinformatics: UMAP has gained significant popularity in bioinformatics for its capacity to unravel complex biological datasets. It has been applied to tasks such as single-cell RNA sequencing analysis, where it aids in identifying distinct cell types, uncovering gene expression patterns, and exploring the cellular landscape. UMAP's ability to retain both local and global relationships in data makes it particularly well-suited for dissecting intricate biological systems.

- Natural Language Processing (NLP): UMAP has made inroads into NLP by assisting in the visualization of high-dimensional word embeddings and document representations. Researchers and practitioners use UMAP to create visually interpretable embeddings of text data, enabling semantic analysis, topic modeling, and document clustering. This is especially useful for understanding relationships between words or documents in large text corpora.

- Image Analysis: UMAP has proven valuable in image analysis tasks, such as computer vision and image classification. It can reduce the dimensionality of image feature representations while preserving the essential characteristics of the images. This makes it easier to visualize and explore image datasets, aiding in tasks like object recognition, image retrieval, and content-based image retrieval.

- Recommendation Systems: UMAP can enhance recommendation systems by reducing the dimensionality of user-item interaction data. It helps uncover latent patterns and similarities in user behavior, which can lead to more accurate and personalized recommendations. Online platforms, streaming services, and e-commerce platforms can leverage UMAP to improve user experiences and engagement.

- Anomaly Detection: UMAP can be used in anomaly detection applications to identify outliers or unusual patterns in high-dimensional data. By visualizing data in lower dimensions, anomalies become more apparent, allowing for quicker detection and response to potentially critical issues in various domains, including cybersecurity and quality control.
- Materials Science and Chemistry: Researchers in materials science and chemistry employ UMAP to analyze complex datasets related to material properties, molecular structures, and chemical interactions. UMAP's ability to capture the underlying structure of data facilitates the discovery of novel materials and compounds.

- Social Network Analysis: UMAP can assist in the exploration of social networks by reducing the dimensionality of user connections or interactions. It enables the visualization of community structures, the detection of influential nodes, and the analysis of user behavior patterns within online social platforms.

- Healthcare and Medical Research: UMAP aids in the analysis of healthcare data, such as patient records and medical images. It enables healthcare professionals and researchers to better understand patient populations, disease progression, and treatment responses.

Comparative Analysis with Other Dimensionality Reduction Techniques

To appreciate the strengths and versatility of Uniform Manifold Approximation and Projection (UMAP), it is instructive to contextualize it within the landscape of dimensionality reduction techniques and understand how it compares to other commonly used methods. While no single technique is universally superior, UMAP offers distinct advantages and characteristics when contrasted with alternative approaches.

1. Principal Component Analysis (PCA):

- PCA is a linear dimensionality reduction method that seeks to find orthogonal axes that capture the maximum variance in the data.
- UMAP vs. PCA: UMAP can capture nonlinear relationships in data, making it more suitable for complex, nonlinear datasets, while PCA assumes linearity.

2. t-Distributed Stochastic Neighbor Embedding (t-SNE):

- t-SNE is known for its effectiveness in preserving local structures and revealing clusters in data.
- UMAP vs. t-SNE: UMAP often exhibits superior preservation of global structures, making it advantageous for visualizing large datasets with both local and global patterns.

3. Multidimensional Scaling (MDS):

- MDS aims to represent the pairwise distances between data points in lower-dimensional space while preserving these distances as faithfully as possible.
- UMAP vs. MDS: UMAP incorporates topological and geometric information, offering a more robust approach for preserving the intrinsic structure of data, especially when dealing with nonlinear data.

4. Autoencoders:

- Autoencoders are neural network-based techniques that can learn complex nonlinear representations of data.
- UMAP vs. Autoencoders: While autoencoders can be highly flexible, they may require more extensive hyperparameter tuning and training time compared to UMAP's efficient approach.

5. Isomap:

- Isomap constructs a graph representing the intrinsic geometric structure of the data and then projects it into a lower-dimensional space.
- UMAP vs. Isomap: UMAP often outperforms Isomap in terms of computational efficiency and the preservation of global structures.

6. Linear Discriminant Analysis (LDA):

- LDA is a supervised dimensionality reduction technique that seeks to maximize the separation between different classes or groups in the data.
- UMAP vs. LDA: UMAP is more versatile as it is not constrained by the need for class labels and can be applied to a broader range of unsupervised tasks.

7. Locally Linear Embedding (LLE):

- LLE aims to capture local relationships between data points by modeling them as linear combinations of their neighbors.
- UMAP vs. LLE: UMAP often exhibits better scalability and less sensitivity to the choice of parameters than LLE.

Best Practices for Using UMAP

Uniform Manifold Approximation and Projection (UMAP) is a powerful dimensionality reduction technique, but to harness its full potential and obtain meaningful results, it's important to follow best practices. Proper usage involves a combination of data preprocessing, parameter tuning, and result interpretation. Guidelines for effectively utilizing UMAP:

1. Data Preprocessing:

- Normalization: Ensure that your data is appropriately scaled or normalized, especially when working with features that have different units or scales. UMAP is sensitive to the magnitude of feature values.
- Outlier Handling: Address outliers in your data, as extreme values can disproportionately influence the UMAP projection. Consider using outlier detection techniques or robust data preprocessing methods.

2. Parameter Tuning:

- n_neighbors: This parameter controls the size of the local neighborhood considered for preserving local structures. Smaller values result in more emphasis on fine-grained details, while larger values focus on broader structures. Experiment with different values to find an appropriate balance for your dataset.
- min_dist: Adjust the "min_dist" parameter to control the minimum distance between points in the low-dimensional representation. Smaller values encourage more compact clusters, while larger values encourage more separation between clusters. Fine-tuning this parameter can significantly impact the results.
- Metric Choice: UMAP allows various distance metrics. Select a metric that aligns with the characteristics of your data. For example, you might use Euclidean distance for continuous data and cosine similarity for text data.

3. Visual Interpretation:

- Interpretability: UMAP offers a reduced-dimensional representation of your data, which is often easier to visualize and interpret. Pay attention to cluster structures, data separations, and distances between points to draw meaningful insights.
- Color Coding: Consider color-coding data points based on their original labels or attributes to further enhance interpretability. Visualizing clusters or categories can provide additional context.

4. Handling Large Datasets:

- UMAP is known for its computational efficiency, but with very large datasets, you may still encounter performance limitations. Consider subsampling or using approximate UMAP techniques for extremely large datasets.

5. Validation and Evaluation:

- Use domain-specific validation metrics or evaluation criteria to assess the quality of the UMAP projection. For example, silhouette score, Davies-Bouldin index, or other clustering quality measures can help gauge the separation and coherence of clusters in the reduced space.

6. Parameter Robustness Testing:

- Conduct robustness testing by evaluating the sensitivity of your UMAP results to changes in parameter values. This can provide insights into the stability of your projections.

7. Documentation and Reproducibility:

Keep detailed records of the parameters used, preprocessing steps, and the context of your analysis. This documentation is essential for reproducibility and troubleshooting.

8. Experimentation and Iteration:

Don't hesitate to experiment with different parameter settings and iterate on your UMAP analysis. The optimal parameters may vary depending on the specific dataset and objectives.

9. Resource Management:

Depending on the size and complexity of your dataset, UMAP may require substantial computational resources. Be mindful of memory and processing requirements when applying UMAP to large-scale projects.

10. Stay Informed:

Keep abreast of updates and developments in UMAP and the wider field of dimensionality reduction. New versions and improvements may impact the way UMAP performs or is applied.

References

Taniwa. (s. f.). *UMAP para descubrir tus datos*. https://taniwa.es/blog/umap/

Understanding UMAP. (n.d.). Github.Io. Retrieved September 16, 2023, from https://pair-code.github.io/understanding-umap/

UMAP — machine learning and data science compendium. (s/f). Lazyprogrammer.Me. Recuperado el 16 de septiembre de 2023, de https://lazyprogrammer.me/mlcompendium/dimension_reduction/umap.html