Kristine Wiggins, Austin Lackey, and Niko DeNiro
Alex Elchesen
DSCI 475 Topological Data Analysis
3 May 2023

Exploring Grouping Techniques using Isomap Dimensionality Reduction

Introduction

In recent years, image classification of high dimensional data has been an essential task of computer vision and machine learning. Our group decided to use the technique of Isomap Nonlinear Dimensionality Reduction in order to visualize high dimensional data in a low-dimensional space that captures the underlying structure. Isomapping is also useful for identifying patterns within the data and compressing it for efficient processing. Real-world examples where this technique is used include analyzing brain activity, image, genetic, audio, and stock market data. This paper covers the methodology and results of our project while demonstrating the effectiveness of Isomap using real world data.

Opening with what Isomapping (short for Isometric Mapping) is, it's important to know what geodesics are. The goal of this mapping is to maintain a geodesic distance between two points. Geodesic is more formally defined as the shortest path on the surface itself.

By understanding the pair-wise geodesic distances, Isomap aims to approximate the geometry of the data before projecting it down into the specified dimension. There are three main steps to isomapping. The first is with k-Nearest Neighbors(k-NN). This method identifies the k nearest neighbors to each data point in the high-dimensional space, and uses these neighbors to construct a lower-dimensional representation of the data, based on the distances between its k nearest neighbors. Using k-NN, a kernel matrix is constructed that represents the pairwise similarities between data points. Here is an example of an equation for computing the Euclidean distances:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

After this, the Dijkstra or Floyd-Warshall algorithm is used. Dijkstra's algorithm, also called the Single Source Shortest Path(SSSP), finds the shortest path from a single source node to a given destination node. The equation for this is:

$$distance[v] = min(distance[v], distance[u] + w(u, v))$$

Where distance[v] is the tentative distance to node v, distance[u] is the tentative distance to the current node u, and w(u,v) is the weight of the edge between nodes u and v.

Floyd-Warshall's algorithm, or All-Pairs Shortest Paths(APSP), finds the shortest paths between all pairs of nodes in the graph. A matrix is made up of the distances between all pairs of nodes and iterates through and updates each distance over time. The equation for this is:

$$d[i, j] = min(d[i, j], d[i, k] + d[k, j])$$

Where d[i,j] is the distance between nodes i and j, and d[i,k] and d[k,j] are the distances between nodes i and k and between nodes k and j, respectively.

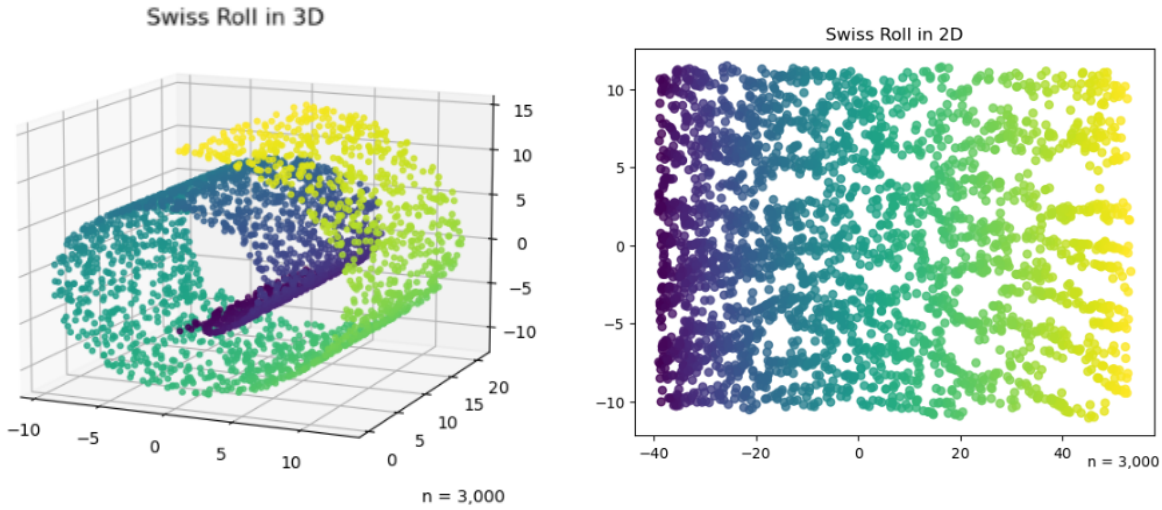
Using the pairwise distances from either one of these algorithms, one can construct a kernel matrix which captures the pairwise similarities between data points. After this, one can perform the partial eigenvalue decomposition of the kernel matrix to find the low-dimensional embedding. Once this method is executed, the data can be visualized using scatter plots, heat maps, or other data analysis techniques.

Pictures like facial imaging can be used with isomapping to extract certain features that correlate to each other by computing the geodesic distances between these features. There are many other applications like robotics, computer vision and machine learning that can use isomapping to help visualize complex high-dimensional datasets.

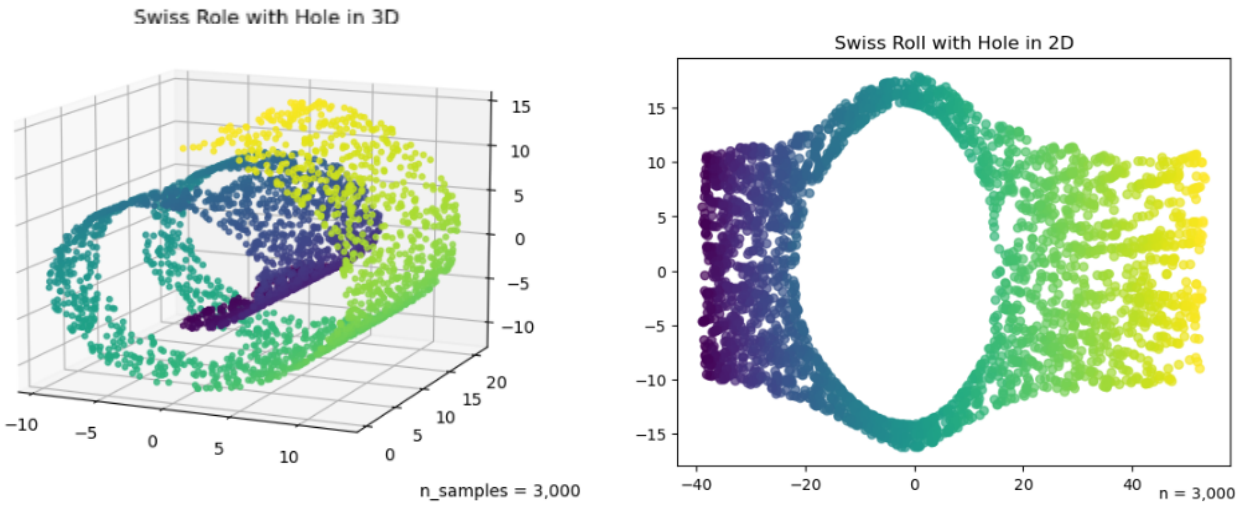
Methodology and Results

Isomap on Swiss Roll

Starting off, we explored the classic example of Isomap Dimensionality Reduction on a "Swiss roll". Using packages from scikit-learn, we created a synthetic dataset of 3,000 sample points making up a 3D Swiss roll. After creation of the role, we use compete dimensionality reduction from three dimensions down to two. We do this by using the Isomap mapping function from scikit-learn with 10 neighbors and plotting the results, as seen in the figures below.

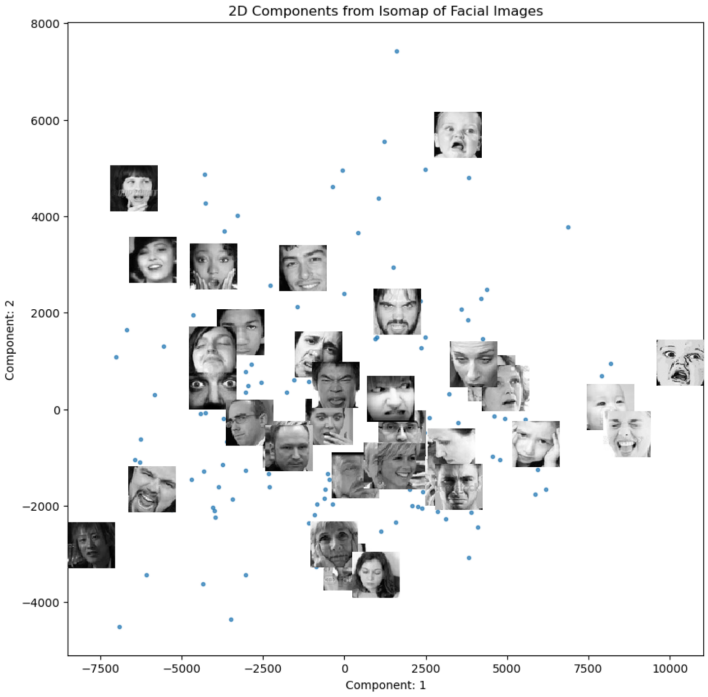


We can see how the roll was unrolled to preserve thegeodesic distances between colored points near eachother. Next, to further visualize how the Isometric Mapping algorithm is working, we completed another theoretical swiss roll example but with a hole.



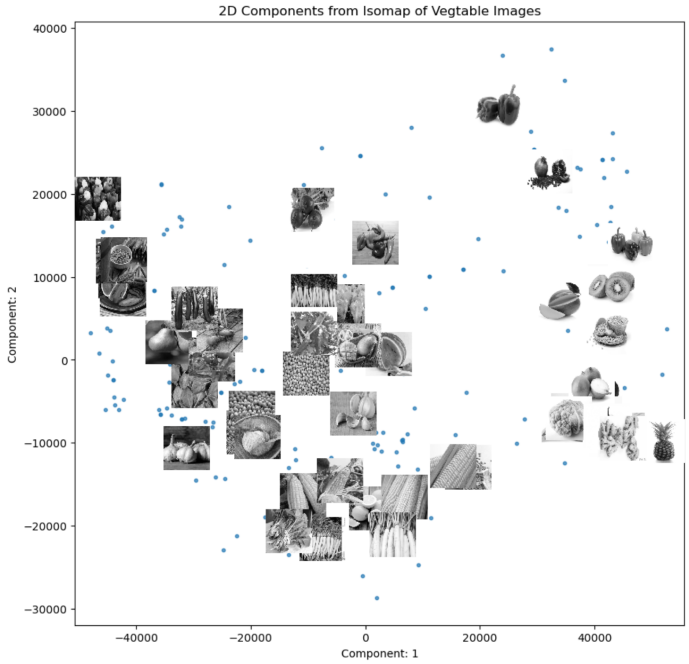
Here we can see the unravelling of the 3D spiral laid flat in a 2-dimensional space. The structure of the swiss roll with a hole in it is maintained in the reduction. The “bubble” shown in the reduced dimension plot can be explained by how the algorithm is attempting to maintain relative distances between the points at the edge of the hole.

Isomap on Facial Data



The first real world application we wanted to test Isomap on was a facial dataset that we found on kaggle. This consisted of images of people that were grouped into 7 different facial emotions. These images were already small and grayscale so we could immediately apply Isomap. On the image to the left we can immediately see that prominent features that were grouped was the lightness of the image. Images on the left had dark backgrounds, images on the button had lighter background with subjects that had darker attributes (i.e dark hair, shadows underneath eyes). Images on the top left had dark backgrounds with very light subjects, and images on the right had light backgrounds and light subjects. As far as emotions go, the groupings had nothing to do with emotions, there were a mix of emotions in every quadrant. This makes sense though, since each image was a unique person, the features that were the most prominent had to do with the brightness of the images. I think if we were to obtain facial data that kept the subject and lighting the same, we would start to see groupings of emotions.

Isomap on Vegetable Data



Since we came to the conclusion that the facial data was a bit messy, we decided to switch our focus to a dataset that was more consistent. We downloaded a set that consisted of images grouped into 36 different vegetables and fruits. When compared to the facial images, the fruit and vegetable images required more preprocessing techniques in order to properly perform Isomap. We needed to convert each image to grayscale, and size them down to 200x200 pixels since they were so large. As seen in the figure on the left, Isomap was able to group the images by a few different metrics. Images on the top-left were generally darker with black backgrounds. Images on the

bottom-right were often cleaner and had white backgrounds. We can see that the images on the bottom often had a linear pattern to them (i.e. vegetables that were long and skinny). Images on the right were often contained fruits that were cut open, and images on the top-right often contained vegetables that were in groups of 2 or 3. Also we can observe how images with backgrounds are grouped more to the left while groups with white/no backgrounds are grouped together on the right-most side.

Conclusion

In conclusion, we have found that using Isomap nonlinear dimensionality reduction techniques yields better results when using simulated data as compared to real-world data. We believe that this is due to the inconsistent noise that exists in real-world data which makes it much harder to visualize data in a reduced space. On the other hand, simulated data does an amazing job at showcasing how the Isomap algorithm works and it is much easier to interpret at a reduced scale. While Isomap is still very useful for real world applications, it requires a more extensive data cleaning process in order to account for the added noise and inconsistencies that are present.

References

Baeldung. “Dijkstra vs Floyd-Warshall Algorithms.” *Baeldung on Computer Science*, 13 Oct. 2021, www.baeldung.com/cs/dijkstra-vs-floyd-warshall.

Bueno, Luiz. “Rating Opencv Emotion Images.” Kaggle, 4 Dec. 2021, https://www.kaggle.com/datasets/juniorbueno/rating-opencv-emotion-images?resource=download.

“Isomap with Scikit-Learn.” Solem’s vision blog, 21 Mar. 2012, solem1.rssing.com/chan-7173164/article20.html.

Seth, Kritik. “Fruits and Vegetables Image Recognition Dataset.” Kaggle, 12 Feb. 2022, https://www.kaggle.com/datasets/kritikseth/fruit-and-vegetable-image-recognition?select=train.

“Sklearn.Datasets.Make_swiss_roll.” Scikit-Learn, scikit-learn.org/stable/modules/generated/sklearn.datasets.make_swiss_roll.html.

“Swiss Roll and Swiss-Hole Reduction.” Scikit-Learn, https://scikit-learn.org/stable/auto_examples/manifold/plot_swissroll.html

Tenenbaum, J B et al. “A global geometric framework for nonlinear dimensionality reduction.” *Science (New York, N.Y.)* vol. 290,5500 (2000): 2319-23. doi:10.1126/science.290.5500.2319