Assignment 8

Research Article Review

Description

The goal of the assignment is to review two research articles that are focused on spectral methods for dimensionality reduction by Tenenbaum et al. and Roweis and Saul.

Background

Nonlinearities in data lead to unreliable results when using methods such as simple PCA for dimensionality reduction. This is because simpler methods are unable to realize similarities between points in higher dimensional space when those points lie on a manifold. The papers being reviewed discuss methods of dimensionality reduction that effectively determine and represent these relationships.

Reviews

A Global Geometric Framework for Nonlinear Dimensionality Reduction uses the domain of visual perception to demonstrate the use of the Isomap algorithm. The algorithm has three steps:

1. Construct neighborhood graph – Find the nearest neighbors for each point.
2. Compute shortest paths – Compute the shortest paths between all pairs of points.
3. Construct d-dimensional embedding – Find point location in lower dimensional space.

There were two very interesting portions in this paper to me. The first was the graph of the headshots with different poses and lighting conditions, as well as the graph of twos written in different ways. Both graphs are great visualizations of the way that we are able to extract relevant data from higher dimensional spaces, and how the isometric mappings we produce can represent transitions that are interpretable. I was especially struck by Fig. 4, as the transitions were so clear. I think that it's rare that we get the opportunity to so easily visualize important algorithms.

Nonlinear Dimensionality Reduction by Locally Linear Embedding briefly covers another method of dimensionality reduction that seeks to eliminate the need to compute pair-wise distances of all datapoints in the space. The steps of the algorithm are similar to those of Isomap[1]:

1. Construct neighborhood graph – Find the nearest neighbors for each point.
2. Compute weighted points – Use nearest neighbors to compute a point that is a combination of their linear weights
3. Map to d-dimensional embedding – Map points to embedding space

[1] https://medium.com/analytics-vidhya/locally-linear-embedding-lle-data-mining-b956616d24e9

There are some upsides to this algorithm. It is less computationally expensive than Isomap, and it has the ability to find features that may not be accurately represented in the reduced space of the Isomap algorithm.

The Isomap Algorithm and Topological Stability is a rebuttal to alleged major errors with Isomap. Namely that if the neighborhood size is defined as too large relative to the manifold, we may include points that should not have been considered local, thereby seriously altering the neighborhood map and the geodesic distances, and shortest paths. This means that we must have a good idea of what neighborhood size to use, which is inconvenient and can be challenging in high dimensional spaces. The rebuttal makes the following points. Isomap is able to take datapoints with an unknown manifold structure and find the correct 2-d embedding. LLE also suffers in the presence of excessive noise. Isomap is not 'topologically unstable' if it is considered across even a small range of neighborhood sizes.

My Thoughts

As I mentioned, I believe that the interpretability and the ability to visualize an algorithm are important factors in our being able to effectively communicate the results of it. The Isomap and LLE algorithms provide both of these features, and are viable options for dimensionality reduction in the presence of higher dimensional manifolds. They perform extremely well on the Swiss roll dataset, and the results look nice. It does seem to me that the LLE approach is slightly superior to Isomap because it is more efficient. It also makes sense to me to use the weighted average of the neighbors as the new mapped point so long as the data is sufficiently dense.

I don't have a lot of exposure to high dimensional data with obvious manifolds such as the Swiss-roll, so while I do see that the algorithms are effective, I can't think of any datasets personally that would fall into this category. It is a really interesting approach though.

I think that the future of this research should probably be in actually recognizing manifolds in higher dimensional data. When we don't have a nice 3-d Swiss roll to look at, we cannot simply visualize the manifold, we have to have some indicators that there are manifolds present.

I could see the inverse of this method being used for interpolating between image frames to create very realistic videos using a few still images. If the image traits are isometric features, then we could request an update along various features dimensions, and the result would be a new image with the same core structure but an update to the chosen feature. This is similar to the way that we already use encoder / decoder systems to add features to images, but this would be a transformation (not simply an interpolation) between like images.

Conclusion

Isomap and LLE are proven ways to find manifolds in higher dimensional data, and reduce dimensions while maintaining the relationship among those points. There are potential flaws with each, but these can be avoided through trial and error, or a knowledge of the structure of the data. I really enjoyed this course, thank you!