**Diffusion Maps**

Patrick H. Rupp

E-mail: prupp1@jhu.edu

Abstract

Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd …

1. Background

Data is growing at an alarming rate as well as becoming more complex in nature. It is becoming common to have datasets with high number of dimensions. Analyzing this type of data is difficult and cumbersome often making it impossible to find the underlying correlations in the data.

Dimensionality reduction is a technique that attempts to identify dimensions which are correlated and reduce them to a smaller number of dimensions that can still represent the original shape of the data. There exists many different techniques each with their own advantages and limitations including both linear and non-linear approaches.

1.1 PCA

One of the most common dimensionality reduction technique is Principle Componenet Analysis (PCA). The PCA technique uses an approach that identifies features that are linearly correlated and attempts to reduce those dimensions by projecting the data in lower dimenions over new features axes. PCA is widely used within exploratory data anlaysis because visualizing correlations in high dimensional data is a difficult problem.

1.2 Diffusion Maps

Another technique that is becoming popular is called Diffusion Maps. This technique attempts to perform dimensionality reduction by attempting to find an underlying non-linear relationship in the data and projects the data to the lower dimension.

The special highlight with the diffusion map technique is that it is a non-linear approach unlike the PCA which is a linear approach. In practice, it is hard to find linearly correlated data which makes the diffusion maps a good option. It is able to perform the non-linear approach by assuming there is some type of lower dimension manifold tying the data together.

2. Dimension Reduction Experiment

The first experiement will be to test multiple differnet data sets to see how well the different dimensionality reduction techniques perform. The two main experiements will consist of using PCA as well as diffusion maps. In addition, the diffusion maps have three different implementations in this study to see how well they perform.

2.1 PCA

Tbd. Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd.

2.2 Diffusion Maps

Tbd. Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd.

*2.1.1 SKLearn*. Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd.

*2.1.2 KD Nuggets*. Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd.

*2.1.2 KD Nuggets Modified*. Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd.

3. Resource Usage Experiment

Tbd. Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd.

Tbd. Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd.

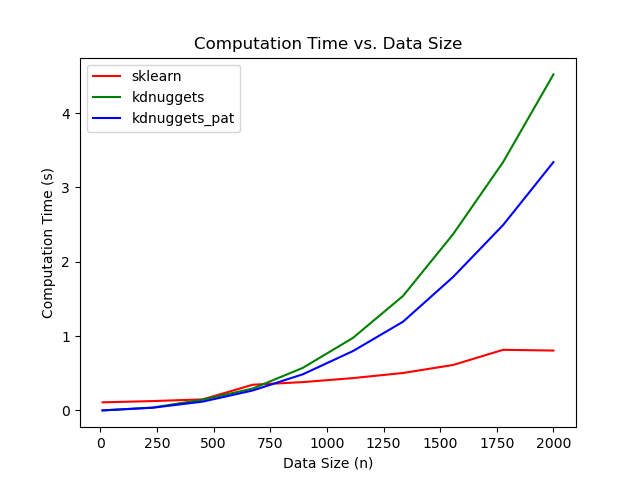


Figure I – Displays the computation time as a function of the size of the data for training the diffusion map. The knuggets has the fastest increase over time whereas the sklearn avoids the exponential growth due to not building the distance map between all points.

4. Conclusion

Tbd. Tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd tbd.

References

[1] Banisch, Ralf, Zofia Trstanova, Andreas Bittracher, Stefan Klus, and Peter Koltai. "Diffusion maps tailored to arbitrary non-degenerate Ito processes." arXiv preprint arXiv:1710.03484 (2017).

[2] Porte, Herbst, Hereman, Walt, An introduction to Diffusion Maps <https://inside.mines.edu/~whereman/talks/delaPorte-Herbst-Hereman-vanderWalt-DiffusionMaps-PRASA2008.pdf>

[3] Coifman, Lafon, Lee, Maggionni, Nadler, Warner, Zucker. “*Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*” <https://www.pnas.org/content/pnas/102/21/7426.full.pdf>

[4] Wolf, Guy. “Introduction to Data Mining: Diffusion Maps”. (2016). <http://cpsc445.guywolf.org/Slides/CPSC445%20-%20Topic%2010%20-%20Diffusion%20Maps.pdf>

[5] Raj, R., 2020. *Diffusion Map For Manifold Learning, Theory And Implementation - Kdnuggets*. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2020/03/diffusion-map-manifold-learning-theory-implementation.html> [Accessed 7 July 2020].