An Introduction to Graph-based Dimensionality Reduction

Extended Abstract

Christoph Luhter

December 2020

Extended Abstract

The abundance of high-dimensional data in fields like language processing or bioinformatics often makes the use of dimensionality reduction techniques inevitable. The goal of such methods is to find a meaningful low-dimensional representation of the input data. This can mitigate the problem of sparse data in a high-dimensional space. It also helps to visualize the data in an accessible dimension and thus, for example, facilitates exploratory analysis. Classical methods, like Principal Component Analysis (cf. [2], chapter 1) and Multidimensional Scaling (MDS, cf. [2], chapter 1) find a linear mapping from the input space to the embedding space, while graph-based methods are able to introduce non-linearity to such mappings. Two graph-based approaches are Isomap [4] and Laplacian Eigenmaps (LEM, [1]). Before explaining the principle of these methods, however, it is imperative to introduce graphs as a mean to represent data.

Assume, as starting point, a set $\mathbf{X} = \{x_1, ..., x_n\}$, $x_i \in \mathbb{R}^D$ for all $i = 1, ..., n$, of vector valued data for a high dimension D. The first step in both Isomap and LEM is to create a graph that represents the data and is required for further processing. To this end, you draw a vertex v_i, $i = 1, ..., n$, for every data point and add undirected edges to nodes representing points that are close in the original data. There are to ways to determine, which nodes are connected via an edge. You can either connect a node v_i to the k nearest neighbours of x_i in the original data or to each node representing data points in an ϵ-radius around x_i. This procedure creates a so called neighbouring (or adjacency) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the set of all vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$. To include more information in these graphs, each edge is assigned a weight which describes the (dis-)similarity of data points - you can think of the Euclidean distance, which serves as the weight in Isomap.

The latter is the first graph-based method to be presented. The key assumption of Isomap, which is also made in LEM, is that the high-dimensional data can be characterized by points that lie in a submanifold with lower dimension. After construction of the required graph, Isomap approximates the geodesic of each pair of observations for the underlying submanifold making use of the weighted edges in the graph. This yields a matrix of pairwise geodesics (or *dissimilarities*) of size $n \times n$. The latter then is processed as in classical MDS which provides lower dimensional vectors as embeddings of the original data. Note that MDS is a stand-alone *linear* dimension reduction technique. Due to its integral role in the Isomap algorithm, it is also part of the introduced approaches. In short, MDS requires a matrix of pairwise dissimilarities and tries to find low-dimensional vectors, whose pairwise Euclidean distances approximate the input dissimilarity values. While the MDS algorithm requires a matrix of distances or more generally any proximity measure, there is no restriction on how this matrix has come about. We already know, how it is built in the Isomap approach, but for mere linear dimension reduction, these distances can simply be the Euclidean distances from the input data.

The second graph-based approach to be presented is Laplacian Eigenmaps. In short, LEM minimizes the following criterion with respect to $\mathbf{Y} = \{y_1, ..., y_n\}$, $y_i \in \mathbb{R}^d$, $d << D$:

$$E_{LEM} = \sum_{i,j=1}^{n} ||y_i - y_j||_2^2 \cdot w_{i,j}$$

where $w_{i,j}$ are the weights of the edges in the graph, but for LEM these weights have to go towards zero for more distant points unlike a metric. An optimal vector y_i then is the embedding of x_i. At this point, it might be worth to distinguish *local* and *global* methods of dimensionality reduction. LEM is a local method, which means that it prioritizes preserving the local geometry of the data. This becomes clear when looking at E_{LEM}. The closer two data points are, the higher $w_{i,j}$ and subsequently the penalty in the minimization problem. Isomap, on the contrary, is a global method, which additionally tries to map distant input vectors to faraway embeddings. Again, looking at E_{LEM}, the crux of LEM, is to rewrite the minimization problem to an eigenvalue problem for the so called graph Laplacian. For now, it shall suffice that the graph Laplacian is a matrix which contains information about the neighbouring structure of the graph and especially, how connected it is. For LEM, it is deduced from a graph constructed as explained above. In brief, the eigenvectors associated with the d smallest non-zero eigenvalues of a normalized version of the Laplacian solve the minimization problem from above and serve as embeddings for our data.

In addition to explainig the three methods, MDS, Isomap and LEM, in greater detail, they will also be applied to three toy data sets to compare their performance. Furthermore, the paper will discuss extensions to Isomap and LEM. While Isomap can be speeded up using so called landmark points [3], global LEM (cf. [2], chapter 2) adds the use of global information of the geometry to standard Laplacian Eigenmaps. Furthermore, LEM has a natural link to spectral clustering, which will also be covered albeit in lesser detail. Altogether, the focus of the seminar paper will be on Isomap, LEM, a comparison of both and their applications.

References

[1] M. Belkin and P. Niyogi. “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation”. In: *Neural Computation* 15.6 (2003), pp. 1373–1396. DOI: 10.1162/089976603321780317.

[2] Yunqian Ma and Yun Fu. *Manifold Learning Theory and Applications*. CRC Press, 2012.

[3] Vin Silva and Joshua Tenenbaum. “Global Versus Local Methods in Nonlinear Dimensionality Reduction”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press, 2003, pp. 721–728. URL: https://proceedings.neurips.cc/paper/2002/file/5d6646aad9bcc0be55b2c82f69750387-Paper.pdf.

[4] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. “A Global Geometric Framework for Nonlinear Dimensionality Reduction”. In: *Science* 290.5500 (2000), pp. 2319–2323. ISSN: 0036-8075. DOI: 10.1126/science.290.5500.2319. eprint: https://science.sciencemag.org/content/290/5500/2319.full.pdf. URL: https://science.sciencemag.org/content/290/5500/2319.