Video

Bookmark this page

How Does Isomap Work?

Video

Download video file

Transcripts

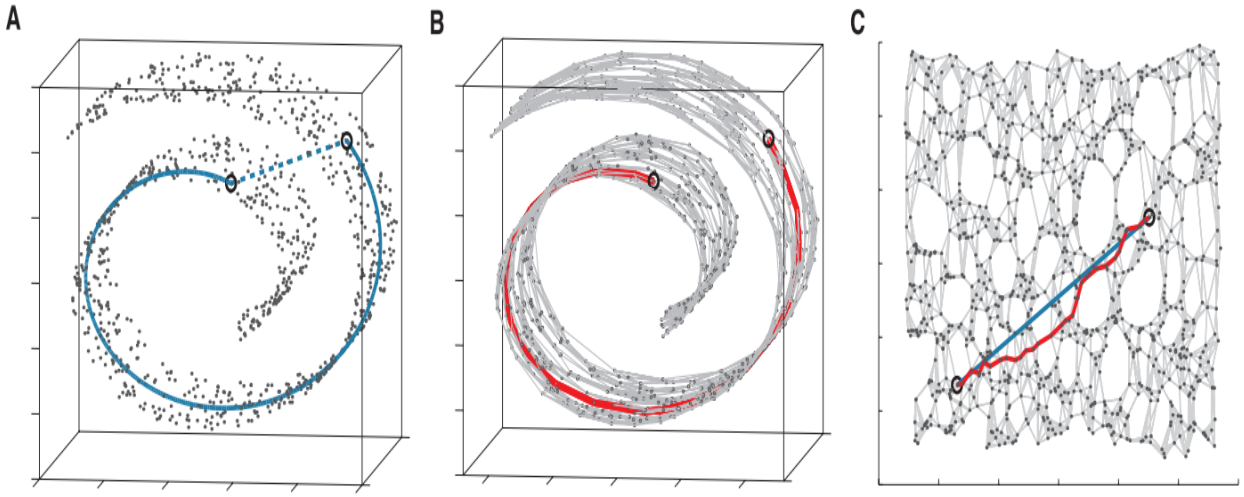
Download SubRip (.srt) file

Download Text (.txt) file

Isomap operates by first computing each record's nearest neighbors. This is done by comparing each sample to every other sample in the dataset. Only a sample's K-nearest samples qualify for being included in its nearest-neighborhood samples list. A

neighborhood graph is then constructed by linking each sample to its K-nearest neighbors. The result is similar to map of roads that is traversed in order to move from point to point.

Just as you would drive waypoint to waypoint in order to navigate to a final destination, so too does Isomap travel from sample to sample, taking the shortest neighborhood paths between any two distant samples in your dataset. The straightforward, e.g. high-dimensional, direct Euclidean distance between any two records fails to properly account for any intrinsic, nonlinear geometry present within your dataset's features. Isomap is able to intelligently recover and estimate a lower-dimensional embedding, also known as a manifold, by traversing the shortest distances between samples, hopping along through the calculated neighborhood map. It is with this map that Isomap calculates a projection, or reduced dimensional embedding that represents your dataset through multi-dimensional scaling.



A comparison by Josh Tenenbaum of the direct Euclidean vs manifold neighborhood path distances of the classic 'swiss-roll', as appears in Science 290 (5500), December 2000.

You've already encountered one *form* of multi-dimensional scaling. Do you recall what it was? It was actually none other than PCA itself! Multi dimensional scaling is a process of taking samples in *N*-dimensional space, and representing them on some other *M*-dimensional space, while attempting to preserve the inter-sample distances as much as possible. If you're going from a lower to higher dimensionality, then the distances can be preserved perfectly. But when you reduce dimensions, as you've seen some information is lost.

Fun fact, the actual distance formula used to calculate the K-nearest samples doesn't even have to be 100% precise. As long as it works reasonably well for 'close by' samples. The truth of the matter is that with Isomap, samples far away from one another aren't included in each other's neighborhood map, and therefore do not affect the final simplified manifold produced by multi dimensional scaling. Due to this, some very nice speed optimizations can be enacted with the distance calculations.

English