UNIVERSITAT DE BARCELONA

Course. Introduction to Machine Learning
Work 2.
Dimensionality Reduction with PCA and truncatedSVD and Visualization using PCA and ISOMAP

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona

Contents

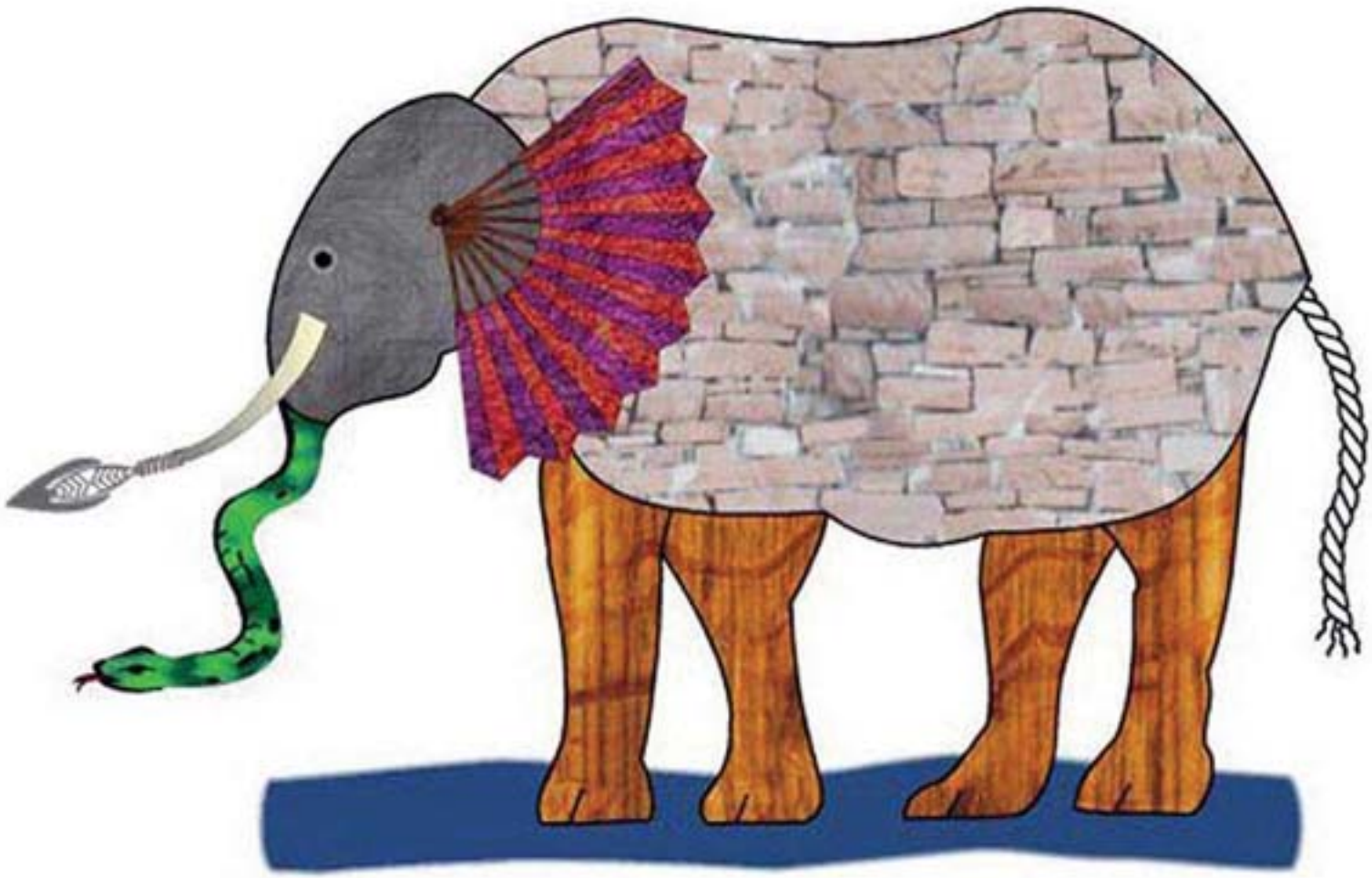
Introduction

Introduction

Unsupervised learning is a class of machine learning algorithms which involves modelling the underlying structure or distribution of the ***"unlabeled"*** data.

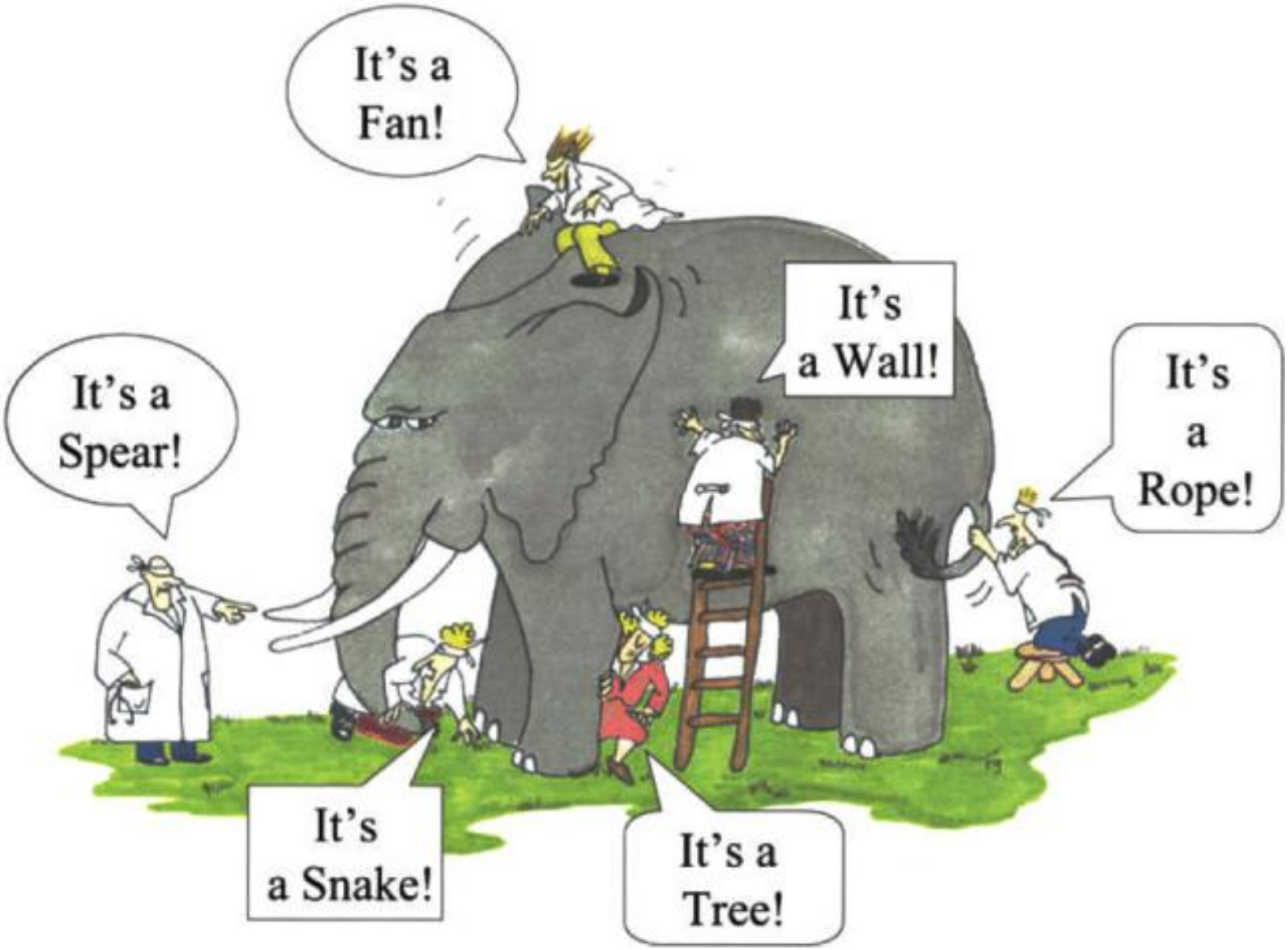
Unlabeled data means the classification or categorization is not available in the observations.

Introduction



Parable: *The blind men and an elephant*

Introduction

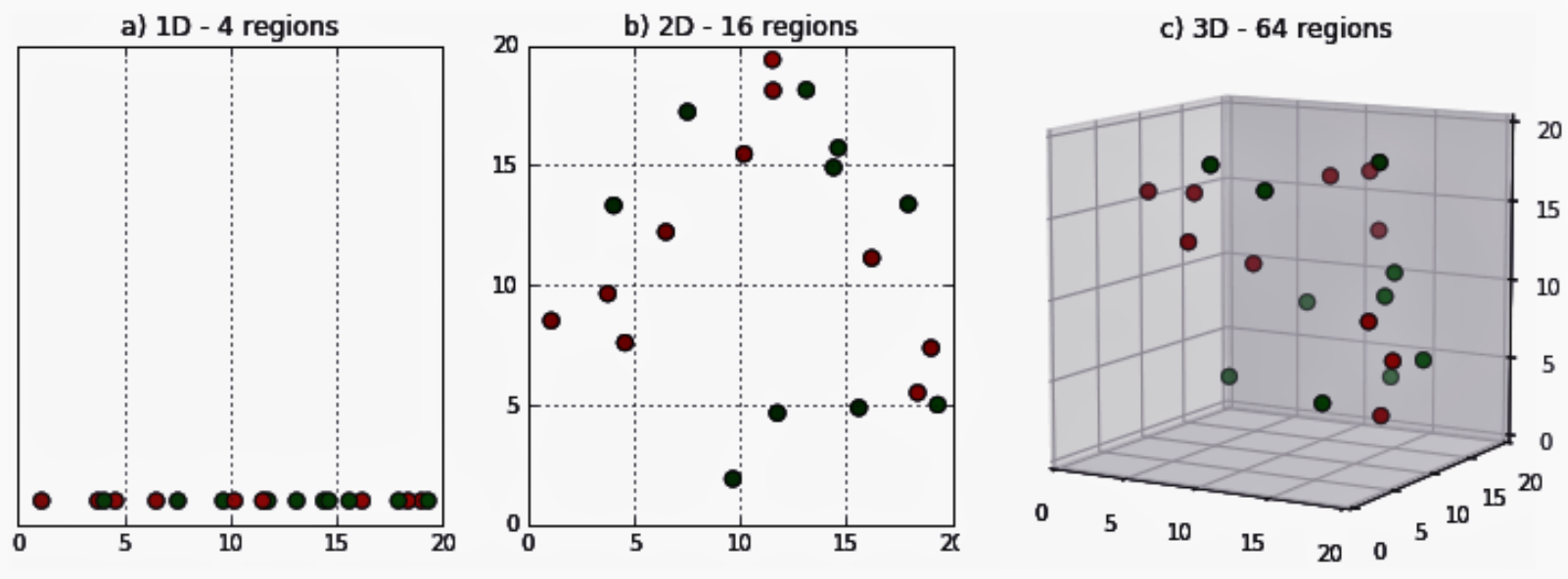


Introduction

The **goal** of Work 2 is to…

1. Reduce dimensionality with PCA (your own code) and truncatedSVD from sklearn library
2. Analyse K-Means (your own code) and Birch (from sklearn) with and without dimensionality reduction
3. Analyze different low-dimensional visualization algorithms
4. Compare visualizations with PCA and ISOMAP

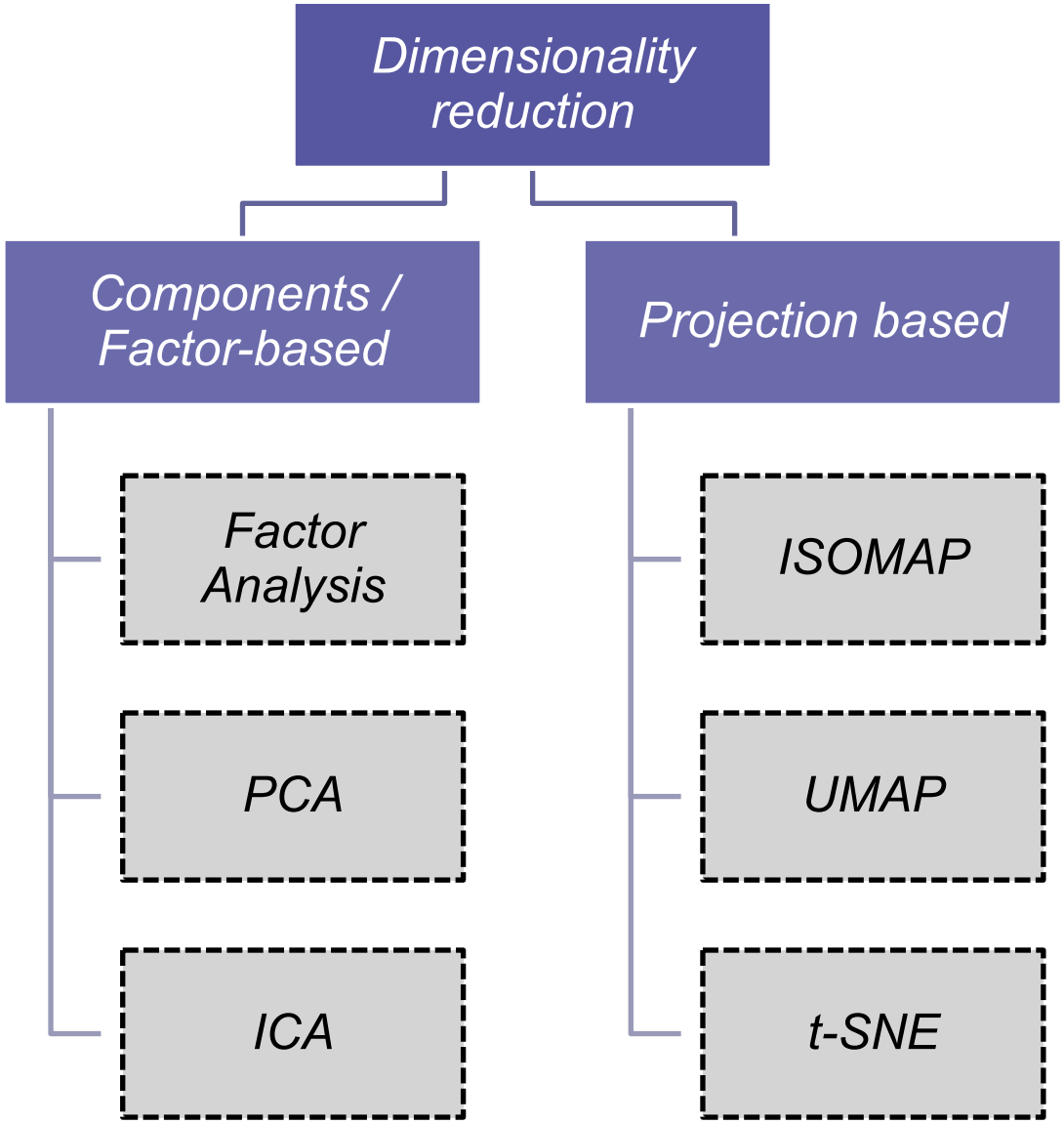
Curse of dimensionality



One approach to mitigate it is to reduce dimensionality

Dimensionality reduction simply refers to the process of reducing the number of attributes in a dataset while keeping as much of the variation in the original dataset as possible.

Approaches for dimensionality reduction



Note: Incomplete list, just with a few examples of each category

Main approaches for Dimensionality Reduction

The two main approaches:

- **Projection**: This technique deals with projecting every data point which is in high dimension, onto a subspace suitable lower-dimensional space in a way which approximately preserves the distances between the points
- **Manifold Learning**: Many dimensionality reduction algorithms work by modelling the manifold on which the training instance lie; this is called *Manifold learning*.
  - It relies on the manifold hypothesis or assumption, which *holds that most real-world high-dimensional datasets lie close to a much lower-dimensional manifold,* this assumption in most of the cases is based on observation or experience rather than theory or pure logic

Dimensionality reduction

- The goal
  - The meaningful low-dimensional structures hidden in their high-dimensional observations
- Linear Method
  - **PCA** (Principal Component Analysis)
    - preserves the variance
  - **MDS** (Multidimensional Scaling)
    - preserves inter-point distance
- Non-Linear Method
  - **ISOMAP** (Isometric Feature Mapping)
    - preserves the intrinsic geometry of the data
  - **LLE** (Locally Linear Embedding)
  - …

Principal Component Analysis

PCA

- **Principal Component Analysis (PCA)** is a **dimension-reduction** tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.

- PCA works by identifying the hyperplane which lies closest to the data and then projects the data on that hyperplane while retaining most of the variation in the data set
  - PCA is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*.
  - The first principal component accounts for as much as of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible

PCA

- Traditionally, PCA is performed on a **square symmetric** matrix. It can be:
  - A **SSCP** matrix (pure sums of squares and cross products),
  - A **Covariance** matrix (scaled sums of squares and cross products), or,
  - A **Correlation** matrix (sums of squares and cross products from standardized data).
- The analysis results for objects of types SSCP and Covariance do not differ, since these objects only differ in a global scaling factor.
- A correlation matrix is used if the variances of individual variates differ much, or if the units of measurement of the individual variates differ.

PCA

- In this work, you have to:
  - Implement your own code of PCA, using **a covariance** matrix
  - Compare and analyze your results to the ones obtained using:

- `sklearn.decomposition.PCA` (https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html) and,

- `sklearn.decomposition.IncrementalPCA` (https://scikit-learn.org/stable/auto_examples/decomposition/plot_incremental_pca.html)

TruncatedSVD

truncated SVD

- **Singular-value decomposition** (SVD) is a dimension reduction technique for matrices that reduces the matrix into its component to simplify the calculation.
  - It works better with sparse data (i.e., data with many zero values)
  - https://personal.math.vt.edu/embree/cmda3606/chapter6.pdf (optional reading)
- **Truncated SVD** works by performing SVD on the data matrix and then truncating the resulting matrices to a lower rank. It factorizes data matrix where the number of columns is equal to the truncation.

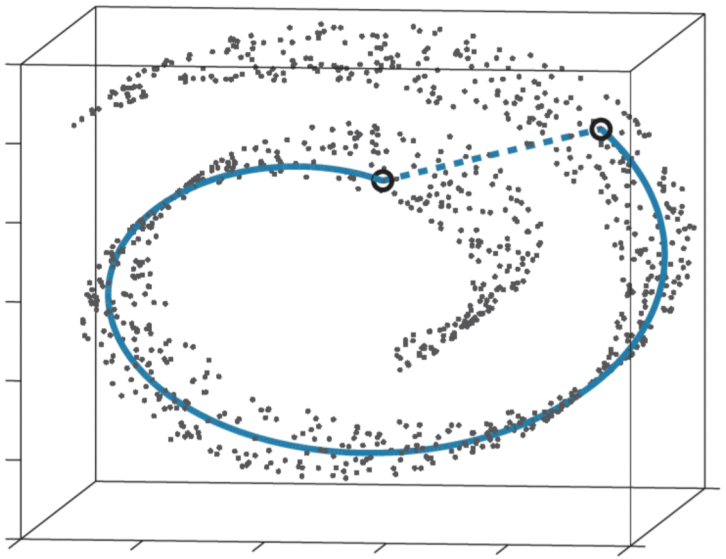
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

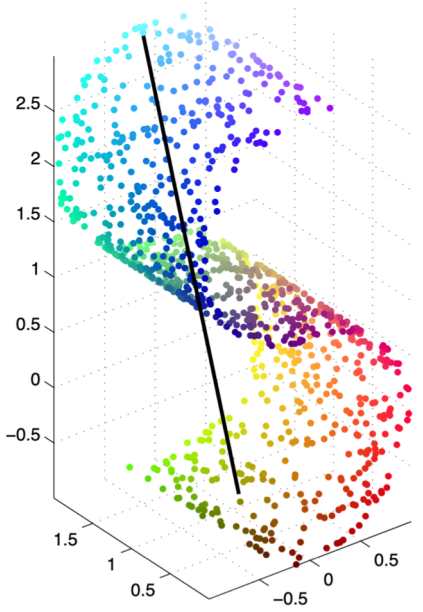
ISOMAP

Isometric Mapping

ISOMAP

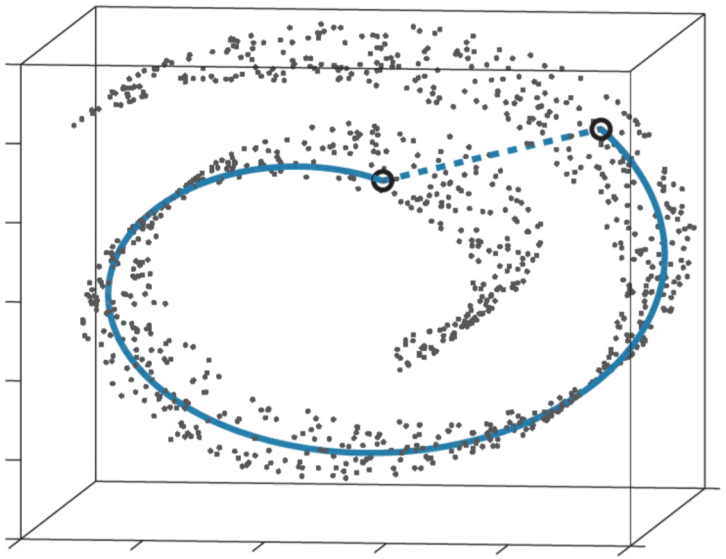
- It is a nonlinear dimensionality reduction technique
- It is a combination of the Floyd-Warshall algorithm with classic Multidimensional Scaling (MDS)

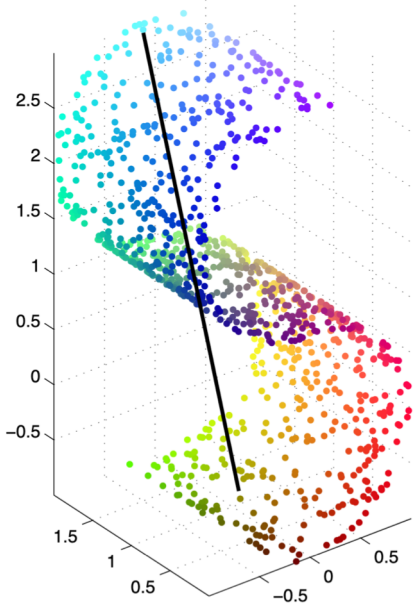




ISOMAP

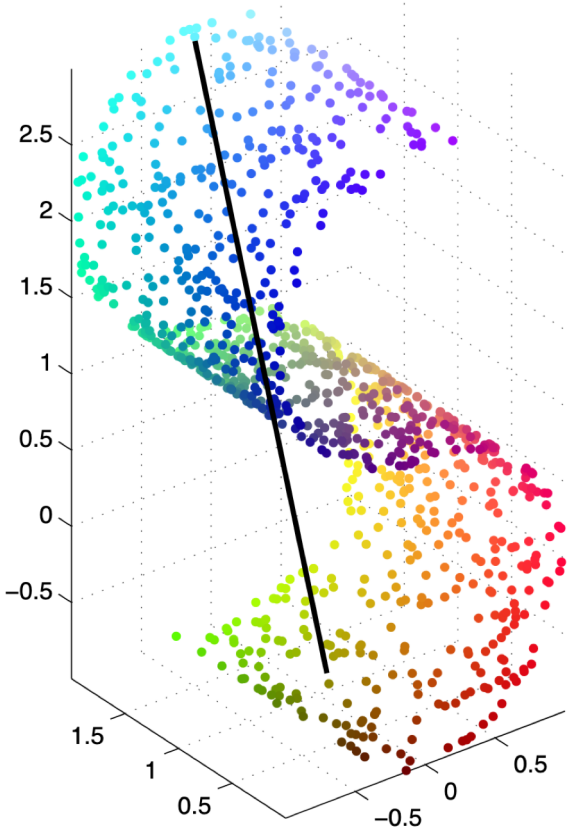
- It is a nonlinear dimensionality reduction technique
  - Many datasets contain essential nonlinear structures that are invisible to PCA and MDS
  - MDS preserve all squared distances, but sacrifice preservation of large distances in favor of preserving small ones.
  - For example de swiss roll, or the s-curve





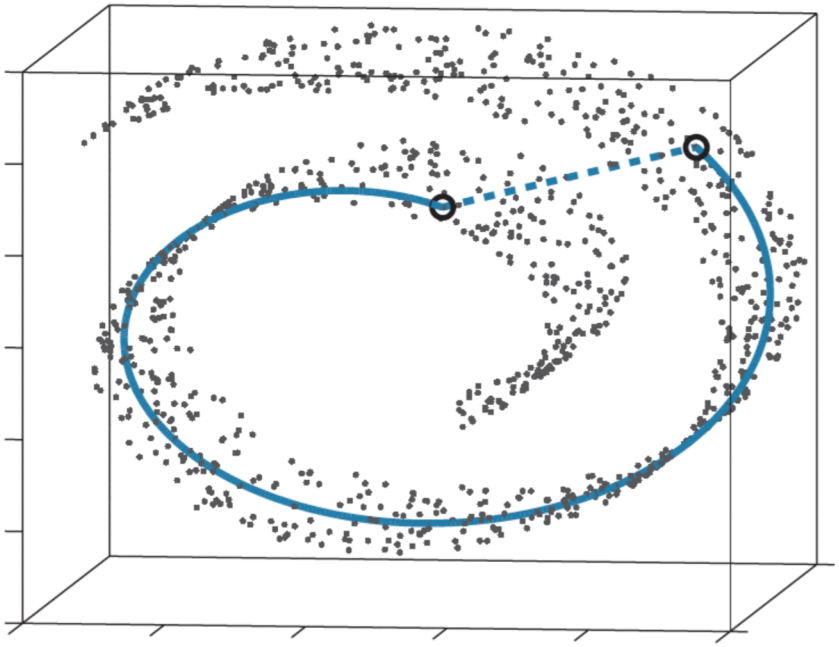
ISOMAP

- If the data lies along a manifold embedded in a high-dimensional space, long Euclidean distances might not follow the manifold
- Therefore directly preserving long Euclidean distances would not "unfold" the manifold



Euclidean distance between two points corresponds to the length of the line connecting the points. The line usually does not follow the manifold of the data

ISOMAP



Euclidean distance can "jump" across the manifold, while the "ideal" distance goes along the manifold

- Euclidean distance in the original space might not be appropriate
  - Replace by **geodesic distance** [Tenenbaum et al., 2000 "*A Global Geometric Framework for Nonlinear Dimensionality Reduction*"]
  - Approximate geodesic distance as shortest distance along a *neighborhood graph* of the data

ISOMAP

- Preserves the intrinsic geometry of the data
- Uses the **geodesic manifold distances** between all pairs
- Uses multi-dimensional scaling to embed corresponding points into a new space
- Advantage over Euclidean distance: points that are somewhat near in Euclidean space, but are far away in **geodesic distance** are considered far away from one-another

ISOMAP algorithm

Step 1. Construct neighborhood graph, G

- Compute matrix $D_G = \{d_x(i,j)\}$
- $d_x(i,j)$= Euclidean distance between neighbors

Step 2. Compute shortest paths

- Compute matrix $D_G = \{d_G(i,j)\}$
- $d_G(i,j)$= approx. geodesic distance

Step 3. Construct *d*-dimensional embedding

- Apply MDS to D_G instead of D_X

ISOMAP algorithm

Step 1. Construct neighborhood graph

- Determine which points are neighbors on the manifold M.
- Two simple methods are to connect each point to all points within a fixed radius ϵ, or to all of its *k* nearest neighbors
- These neighborhood relations are represented as a weighted graph G over the data points, with edges of weight $d_x(i,j)$ between neighboring points.

ISOMAP algorithm

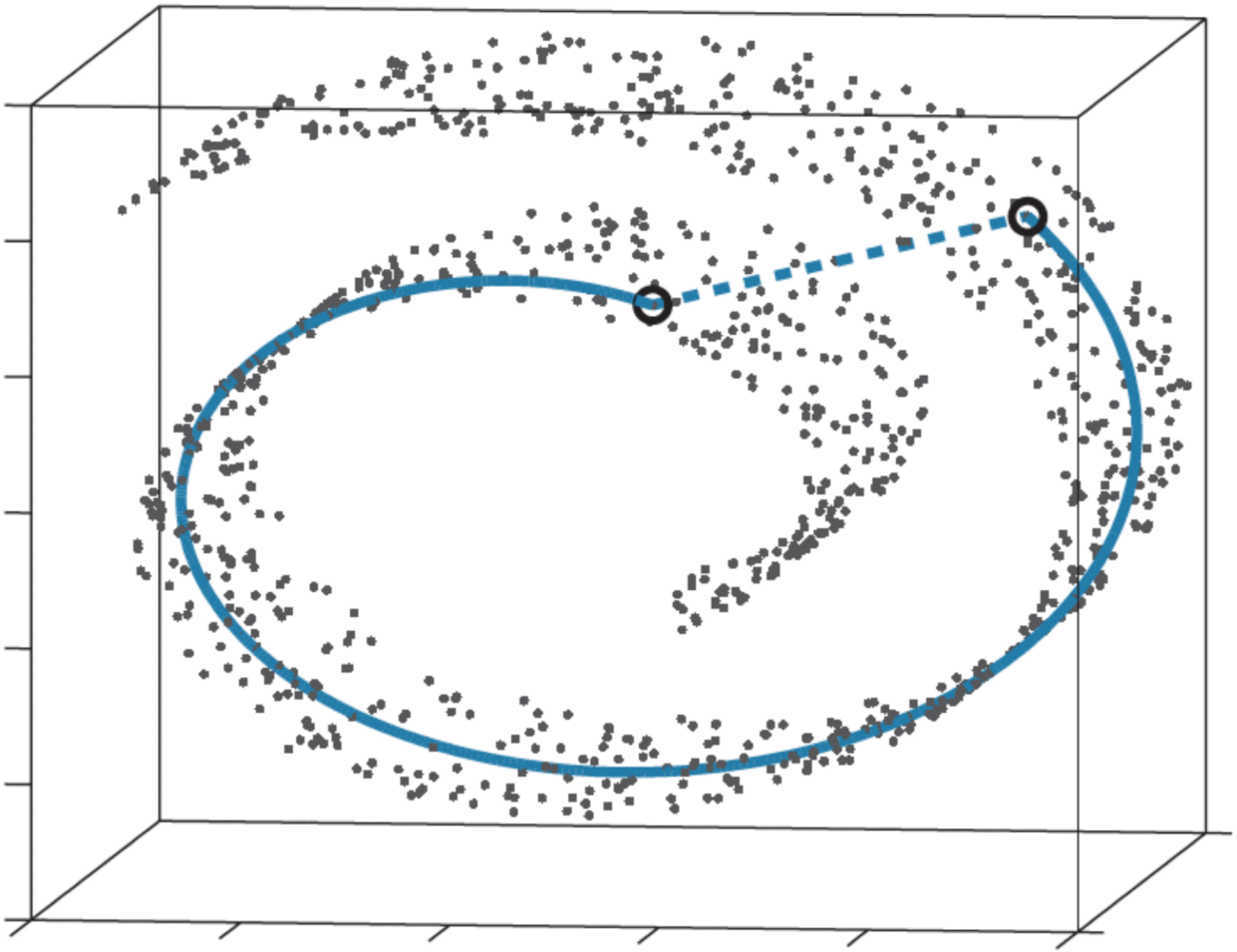
Step 2. Compute shortest paths

- Isomap estimates the geodesic distances $d_M(i,j)$ between all pair of points on the manifold *M* by computing their shortest path distance $d_G(i,j)$ in the graph *G*
- Can be done using Floyd's algorithm or Dijkstra's algorithm

Step 3. Construct *d*-dimensional embedding

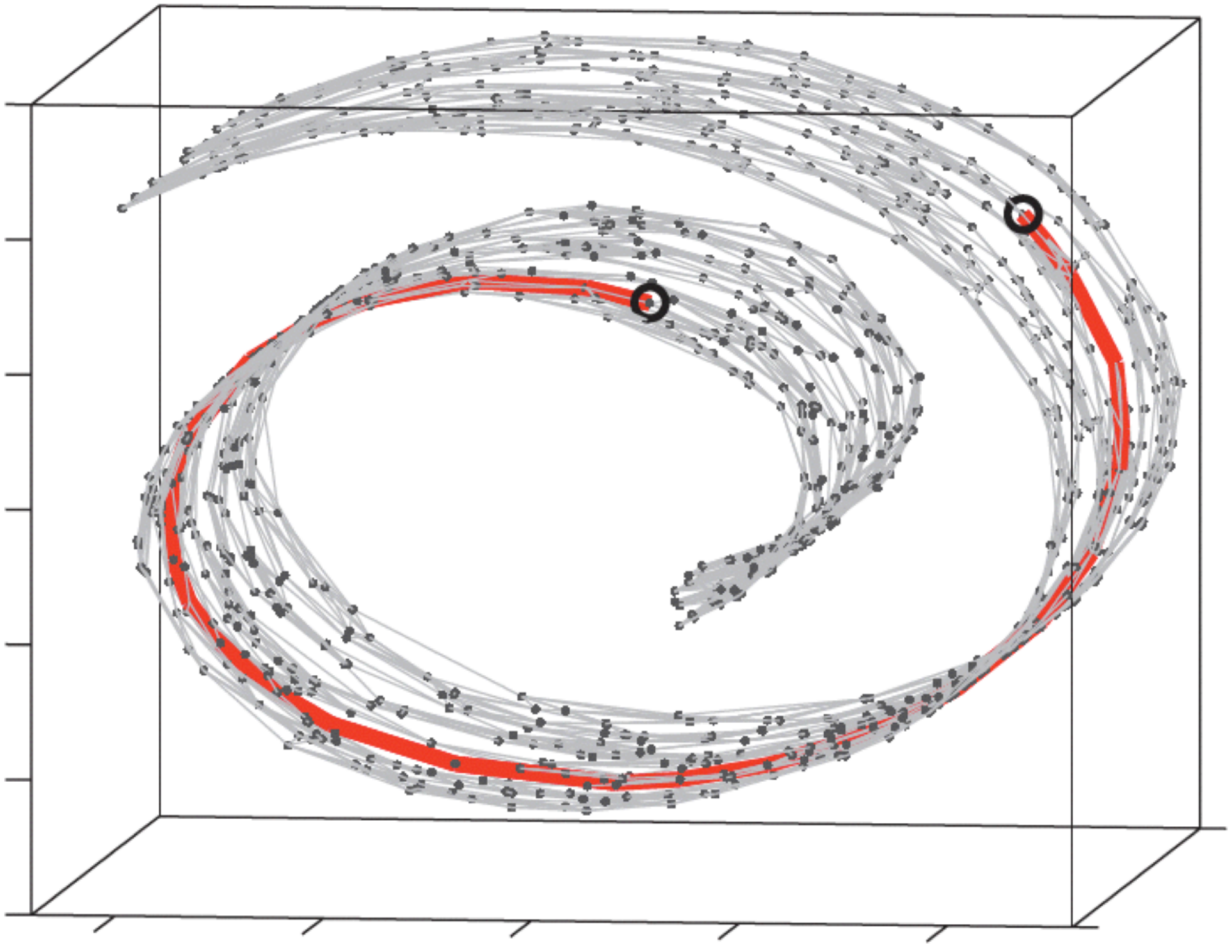
- Apply classical MDS to the matrix of graph distance $D_G = \{d_G(i,j)\}$, constructing an embedding of the data in a d-dimensional Euclidean space *Y* that best preserves the estimated intrinsic geometry of the manifold

ISOMAP: Swiss roll example



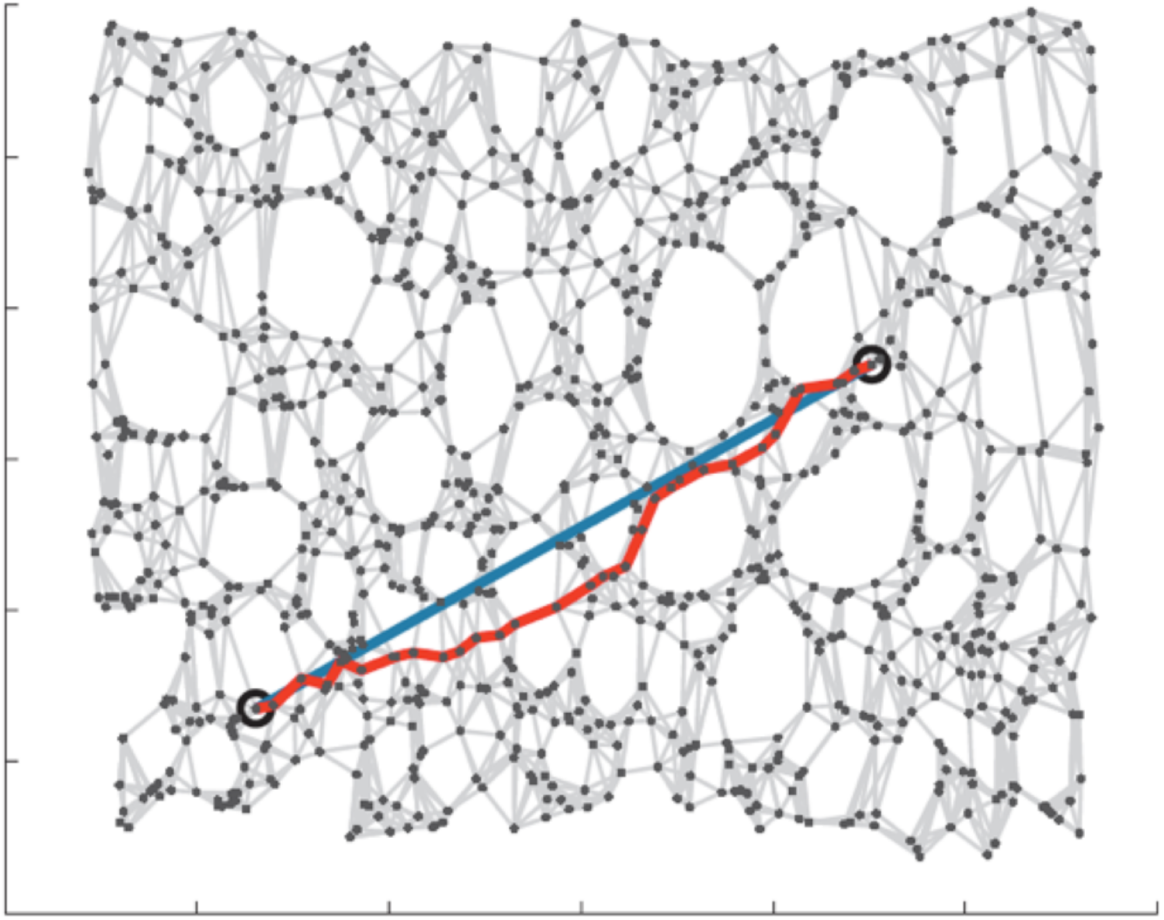
Euclidean distance can "jump" across the manifold, while the "ideal" distance goes along the manifold

ISOMAP: Swiss roll example



Neighborhood graph (with N=1000 data points and K=7 neighbors connected to each point), geodesic approximated by shortest path along the graph

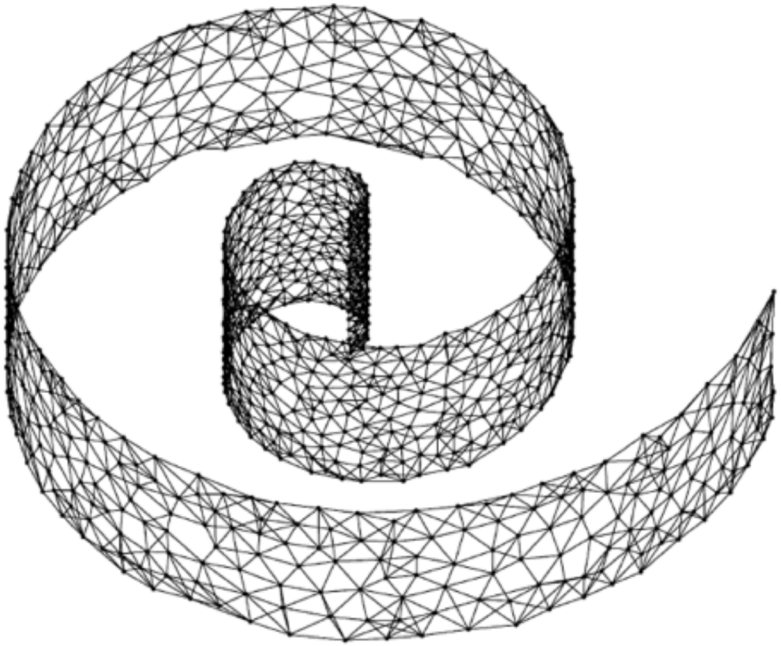
ISOMAP: Swiss roll example



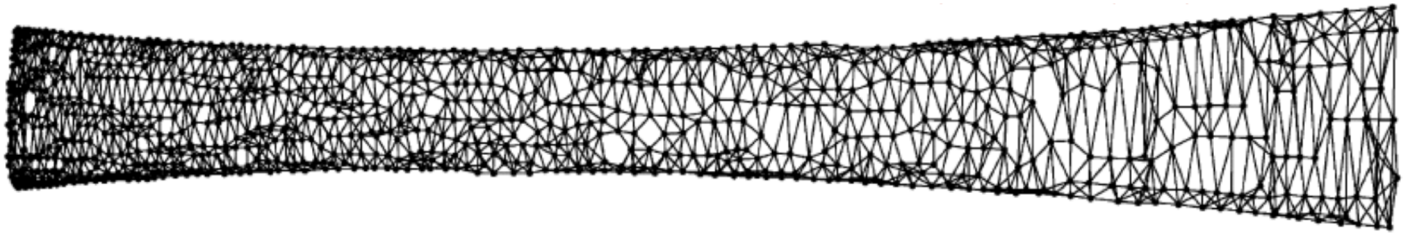
Two-dimensional embedding computed by MDS, to preserve the resulting approximate squared geodesic distances, as squared Euclidean distances on the display

ISOMAP

- Downsides of Isomap
  - Distances along the graph are only approximations of the true geodesic distances.
  - They overestimate the distances, especially large distances, when data is sparse (less “waypoints” to choose from → suboptimal path)



Overestimation of distances causes overstretching of faraway edges



ISOMAP

- Advantages of Isomap
  - Non-linear
  - Globally optimal
    - Still produces globally optimal low-dimensional Euclidean representation even though input space is highly folded, twisted, or curved.
  - It keeps the advantages of PCA and MDS
    - Non-iterative procedure
    - Polynomial procedure
    - Guaranteed convergence
  - Guarantee asymptotically to recover the true dimensionality
  - Represents the global structure of a dataset within a single coordinate system

ISOMAP

- Disadvantages of Isomap
  - Guarantee asymptotically to recover geometric structure of nonlinear manifolds
    - As N increases, pairwise distances provide better approximations to geodesics by “hugging surface” more closely
    - Graph discreteness overestimates $d_M(i,j)$
  - K must be high to avoid “linear shortcuts” near regions of high surface curvature
  - Mapping novel test images to manifold space

Links

- https://en.wikipedia.org/wiki/Distance_(graph_theory)

- **https://scikit-learn.org/stable/modules/generated/sklearn.manifold.lsomap.html#sklearn.manifold.lsomap**

- ORIGINAL PAPER (optional)
  https://wearables.cc.gatech.edu/paper_of_week/isomap.pdf

UNIVERSITAT DE BARCELONA

Course. Introduction to Machine Learning

Work 2

Dimensionality Reduction with PCA and truncatedSVD and Visualization using PCA and ISOMAP

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona