Mathematical techniques in data science

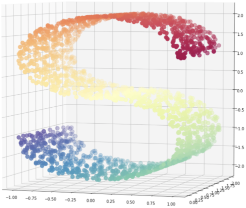
Lecture 23: Manifold learning

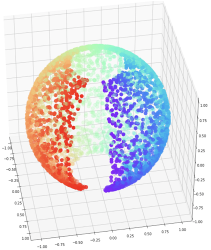
April 17th, 2019

Schedule

Week	Chapter
1	Chapter 2: Intro to statistical learning
3	Chapter 4: Classification
4	Chapter 9: Support vector machine and kernels
5, 6	Chapter 3: Linear regression
7	Chapter 8: Tree-based methods + Random forest
8	
9	Neural networks
12	**PCA → Manifold learning**
11	Clustering: K-means → Spectral Clustering
10	Bayesian methods + UQ
13	Reinforcement learning/Online learning/Active learning
14	Project presentation

Manifold learning





- high-dimensional data often has a low-rank structure
- question: how can we discover low dimensional structures in data?

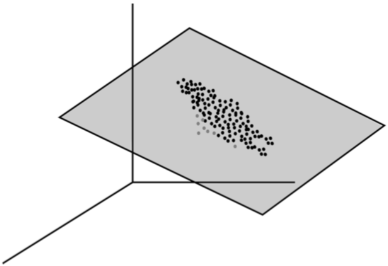
Some definitions

- Metric space: a space on which one can compute the distance between any two points
- Manifold: every point has a neighborhood that is homeomorphic to an open subset of an Euclidean space
- One may say that a manifold is locally Euclidean while globally its structure is more complex
- The dimension of a manifold is equal to the dimension of this Euclidean space

Topics

- Linear methods
    - *Principal component analysis*
    - **Multi-dimensional scaling (MDS)**
- Non linear methods
    - **Isomap**
    - Spectral embedding
    - Locally linear embedding (LLE)
    - t-distributed Stochastic Neighbor Embedding (t-SNE)

Principal component analysis



Problem: How can we discover low dimensional structures in data?

- Principal components analysis: construct projections of the data that capture most of the *variability* in the data.
- Provides a low-rank approximation to the data.
- Can lead to a significant dimensionality reduction.

Multidimensional scaling

Multidimensional scaling (MDS)

- is a means of visualizing the level of similarity of individuals of a dataset
- seeks a low-dimensional representation of the data that respects the distances in the original high-dimensional space
- the goal of an MDS analysis is to find a spatial configuration of objects when all that is known is some measure of their general (dis)similarity

Problem settings

- The data to be analyzed is a collection of n objects on which a distance function is defined: d_{ij} is the distance between objects i and object j
- Given d_{ij}, MDS want to finds vector $z_1, z_2, \ldots, z_n \in \mathbb{R}^d$ such that

$$d_{ij} \approx \|z_i - z_j\|$$

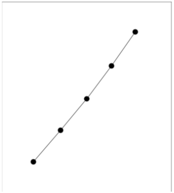
- MDS is formulated as an optimization problem

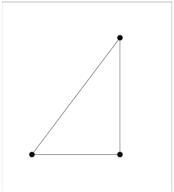
$$\min_{x_1,\ldots,x_n} \sum_{i<j} (d_{ij} - \|x_i - x_j\|)^2$$

Problem settings

$$D = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 0 & 3 & 4 \\ 3 & 0 & 5 \\ 4 & 5 & 0 \end{bmatrix}$$





MDS is formulated as an optimization problem

$$\min_{x_1,\ldots,x_n} \sum_{i<j} (d_{ij} - \|x_i - x_j\|)^2$$

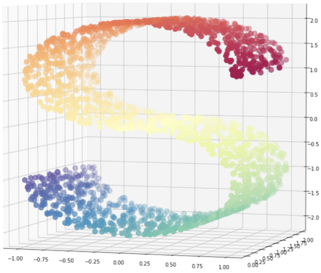
MDS

- MDS is formulated as an optimization problem

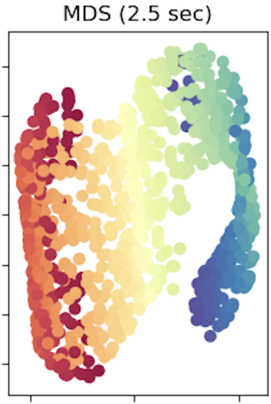
$$\min_{x_1,\ldots,x_n} \sum_{i<j} (d_{ij} - \|x_i - x_j\|)^2$$

- the idea is simple, but is easily generalizable

MDS



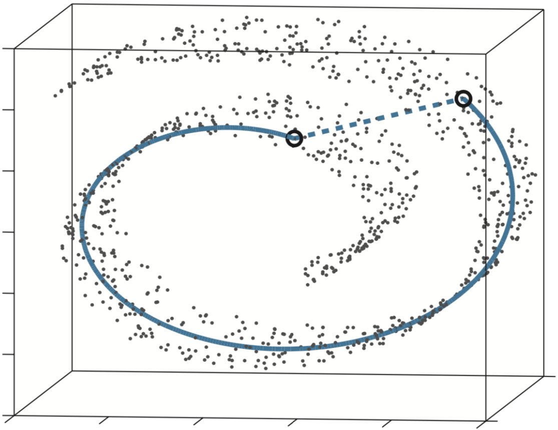
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00



MDS (2.5 sec)

Isometric feature mapping (Isomap)

Distance on a manifold



Isomap

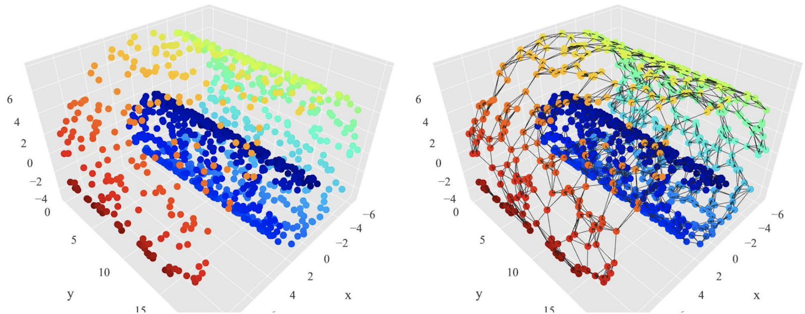
Isomap differs from MDS in one vital way - the construction of the distance matrix.

- In MDS, the distance between two points is just the euclidean distance
- In Isomap, the distances between points are the weight of the shortest path in a point-graph

Isomap: neighbor graph

- For each point, determine either
  - K nearest neighbors
  - all points in a fixed radius
- Construct a neighborhood graph.
  - each point is connected to other if it is a K nearest neighbor.
  - edge length equal to Euclidean distance between the points

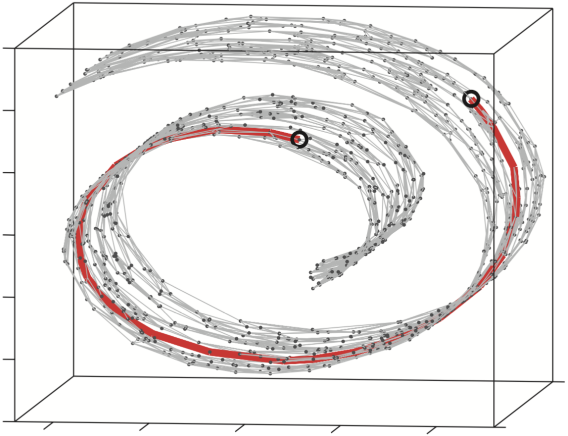
Neighbor graph



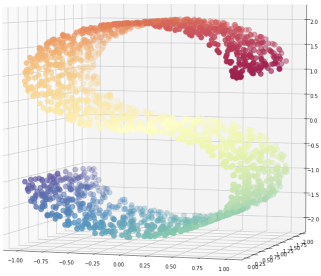
Isomap: compute intrinsic distance

- Compute shortest path between two nodes
  - Dijkstra's algorithm
  - Floyd–Warshall algorithm
- Compute lower-dimensional embedding using MDS
- The graph distance is non-Euclidean, so when embedded back into Euclidean space, some distortion occur

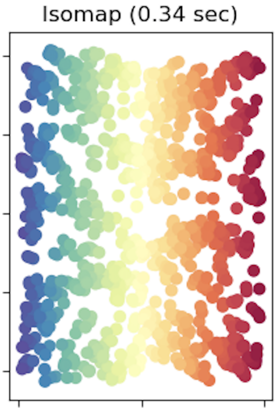
Intrinsic distance



Isomap



2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00



Isomap (0.34 sec)

Locally linear embedding

Locally linear embedding

- A manifold is locally Euclidean while globally its structure is more complex
- Locally, the relation between data points in a neighborhood is linear/affine
- Idea: try to preserve this linear structure

Locally linear embedding

1. For each data point x_i in p dimensions, we find its K-nearest neighbors $\mathcal{N}(i)$ in Euclidean distance.

2. We approximate each point by an affine mixture of the points in its neighborhood:
$$\min_{W_{ik}} ||x_i - \sum_{k \in \mathcal{N}(i)} w_{ik} x_k||^2 \tag{14.102}$$
over weights w_{ik} satisfying $w_{ik} = 0,\ k \notin \mathcal{N}(i),\ \sum_{k=1}^{N} w_{ik} = 1$. w_{ik} is the contribution of point k to the reconstruction of point i. Note that for a hope of a unique solution, we must have $K < p$.

3. Finally, we find points y_i in a space of dimension $d < p$ to minimize
$$\sum_{i=1}^{N} ||y_i - \sum_{k=1}^{N} w_{ik} y_k||^2 \tag{14.103}$$
with w_{ik} fixed.