Visualizing Data using t-SNE

Taran Lynn, Xiaoli Yang, Xiaoxing Chen

October 21, 2020

Isometric Mapping (Isomap)

a non-linear dimensionality reduction method which tries to preserve the geodesic distances in the lower dimension

1. Nearest neighbor search

   Isomap starts by creating a neighborhood network.

2. Shortest-path graph search

   Isomap uses graph distance to the approximate geodesic distance between all pairs of points.

3. Partial eigenvalue decomposition

   And then, through eigenvalue decomposition of the geodesic distance matrix, it finds the low dimensional embedding of the dataset.

Isometric Mapping (Isomap)

Complexity

$$\underbrace{O[D\log(k)N\log(N)]}_{\text{nearest neighbors search}} + \underbrace{O[N^2(k+\log(N))]}_{\text{shortest-path graph search}} + \underbrace{O[dN^2]}_{\text{partial eigenvalue decomposition}}$$

- N: number of training data points
- D: input dimension
- k: number of nearest neighbors
- d: output dimension

Locally Linear Embedding (LLE)

A topology preserving manifold learning method

Assumptions:

- Data is well sampled i.e. density of the dataset is high.
- Dataset lies on a smooth manifold.

1. Nearest neighbor search

A distance metric is needed to measure the distance between the two points and classify them as neighbors. For example Euclidean, Mahalanobis, hamming and cosine. Either e-neighborhood or K-nearest neighbors will be used to create a neighborhood matrix.

2. Weight Matrix Construction

Each point of the dataset is reconstructed as a linear weighted sum of its neighbors.

3. Partial Eigenvalue Decomposition

Create each point in lower dimension using its neighbors and local W matrix. The neighborhood graph and the local Weight matrix capture the topology of the manifold.

Locally Linear Embedding (LLE)

A topology preserving manifold learning method

Complexity

$$\underbrace{O[D\log(k)N\log(N)]}_{\text{nearest neighbors search}} + \underbrace{O[DNk^3]}_{\text{weight matrix construction}} + \underbrace{O[dN^2]}_{\text{partial eigenvalue decomposition}}$$

- N: number of training data points
- D: input dimension
- k: number of nearest neighbors
- d: output dimension

Weakness: Sensitive to outliers and noise

Datasets have a varying density and it is not always possible to have a smooth manifold.

Sammon Mapping

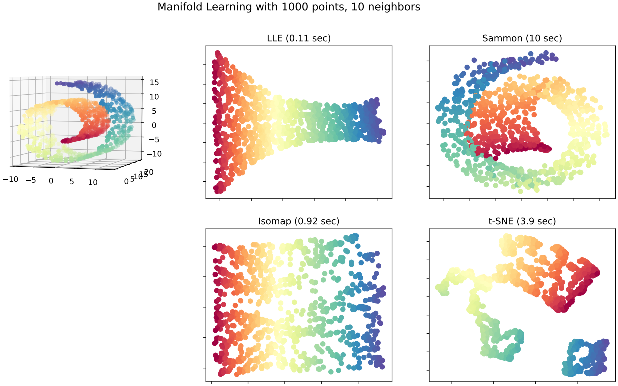
Cost Function

$$E_s = \frac{1}{\sum_i \sum_{j>i} d_{ij}} \sum_i \sum_{j>i} d_{ij} \frac{(d_{ij} - \|r_i - r_j\|^2)^2}{d_{ij}}$$

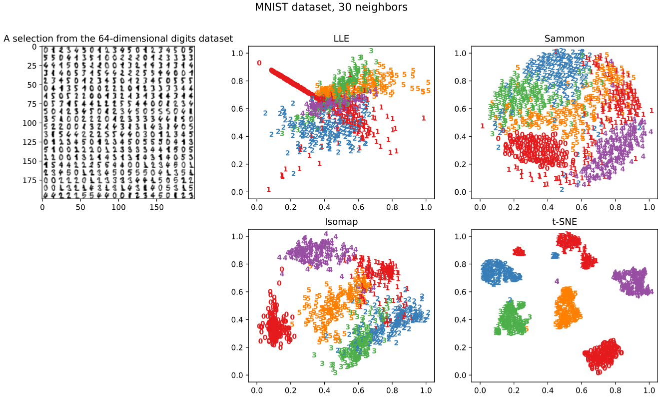
Steps

- Distance calculation
- Distance matrix construction
- Minimizing the projection error

Algorithm Comparison



Algorithm Comparison



Weakness of t-SNE

- Dimensionality reduction for other purposes
- Curse of intrinsic dimensionality
- Non-convexity of the t-SNE cost function

Conclusion

Computational complexity: $O(n^2)$
Memory complexity: $O(n^2)$