Multidimensional Scaling and Isomap

Department of Electrical & Computer Engineering,
University of Waterloo, ON, Canada

Data and Knowledge Modeling and Analysis (ECE 657A)

Course Instructor: Prof. Mark Crowley
TA and Presenter of Slides: Benyamin Ghojogh

Dataset Notations

$$\text{training dataset: } \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n, \boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n} \tag{1}$$

$$\text{embedding of training data: } \{\boldsymbol{y}_i \in \mathbb{R}^p\}_{i=1}^n, \boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n] \in \mathbb{R}^{p \times n} \tag{2}$$

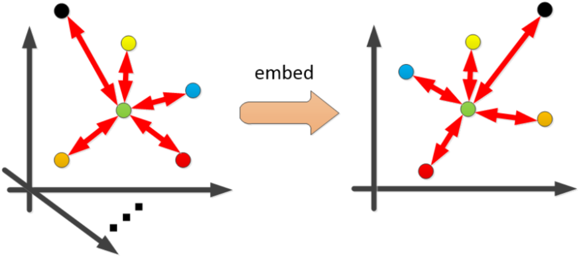
$$p \leq d, \quad \text{usually: } p \ll d \tag{3}$$

Lecture Outline

Classical MDS

- minimizing the differences of **distances** in the input space and the embedding space. It is equivalent to minimizing the differences of **similarities** in these spaces.
- If two points are similar/dissimilar, in the subspace, they should also be similar/dissimilar.

$$\underset{\boldsymbol{Y}}{\text{minimize}} \ ||D_X - D_Y||_F^2, \implies \underset{\boldsymbol{Y}}{\text{minimize}} \ ||\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}||_F^2 \qquad (4)$$



Classical MDS

$$||\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}||_F^2 = \textbf{tr}\big[(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y})^\top (\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y})\big]$$
$$= \textbf{tr}\big[(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y})(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y})\big] = \textbf{tr}\big[(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y})^2\big]$$

$$\boxed{\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V} \Delta \boldsymbol{V}^\top} \tag{5}$$

$$\boxed{\boldsymbol{Y}^\top \boldsymbol{Y} = \boldsymbol{Q} \Psi \boldsymbol{Q}^\top} \tag{6}$$

$$||\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}||_F^2 = \textbf{tr}\big[(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y})^2\big] = \textbf{tr}\big[(\boldsymbol{V} \Delta \boldsymbol{V}^\top - \boldsymbol{Q} \Psi \boldsymbol{Q}^\top)^2\big]$$
$$= \textbf{tr}\big[(\boldsymbol{V} \Delta \boldsymbol{V}^\top - \boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{Q} \Psi \boldsymbol{Q}^\top \boldsymbol{V}\boldsymbol{V}^\top)^2\big]$$
$$= \textbf{tr}\Big[\big(\boldsymbol{V}(\Delta - \boldsymbol{V}^\top \boldsymbol{Q} \Psi \boldsymbol{Q}^\top \boldsymbol{V})\boldsymbol{V}^\top\big)^2\Big] = \textbf{tr}\Big[\boldsymbol{V}^2(\Delta - \boldsymbol{V}^\top \boldsymbol{Q} \Psi \boldsymbol{Q}^\top \boldsymbol{V})^2(\boldsymbol{V}^\top)^2\Big]$$
$$= \textbf{tr}\Big[(\boldsymbol{V}^\top)^2 \boldsymbol{V}^2(\Delta - \boldsymbol{V}^\top \boldsymbol{Q} \Psi \boldsymbol{Q}^\top \boldsymbol{V})^2\Big] = \textbf{tr}\Big[(\underbrace{\boldsymbol{V}^\top \boldsymbol{V}}_{\boldsymbol{I}})^2(\Delta - \boldsymbol{V}^\top \boldsymbol{Q} \Psi \boldsymbol{Q}^\top \boldsymbol{V})^2\Big]$$
$$= \textbf{tr}\Big[(\Delta - \boldsymbol{V}^\top \boldsymbol{Q} \Psi \boldsymbol{Q}^\top \boldsymbol{V})^2\Big].$$

Classical MDS

$$\mathbb{R}^{n\times n} \ni \boldsymbol{M} := \boldsymbol{V}^\top \boldsymbol{Q} \tag{7}$$

$$||\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}||_F^2 = \textbf{tr}\Big[(\Delta - \boldsymbol{M}\Psi\boldsymbol{M}^\top)^2\Big] \tag{8}$$

$$\underset{\boldsymbol{Y}}{\text{minimize}} \;\; ||\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}||_F^2 \equiv \underset{\boldsymbol{M},\Psi}{\text{minimize}} \;\; \textbf{tr}\Big[(\Delta - \boldsymbol{M}\Psi\boldsymbol{M}^\top)^2\Big]. \tag{9}$$

$$\begin{aligned} c_1 &= \textbf{tr}\Big[(\Delta - \boldsymbol{M}\Psi\boldsymbol{M}^\top)^2\Big] = \textbf{tr}(\Delta^2 + (\boldsymbol{M}\Psi\boldsymbol{M}^\top)^2 - 2\Delta\boldsymbol{M}\Psi\boldsymbol{M}^\top) \\ &= \textbf{tr}(\Delta^2) + \textbf{tr}((\boldsymbol{M}\Psi\boldsymbol{M}^\top)^2) - 2\,\textbf{tr}(\Delta\boldsymbol{M}\Psi\boldsymbol{M}^\top). \end{aligned}$$

$$\begin{aligned} &\mathbb{R}^{n\times n} \ni \frac{\partial c_1}{\partial \boldsymbol{M}} = 2(\boldsymbol{M}\Psi\boldsymbol{M}^\top)\boldsymbol{M}\Psi - 2\Delta\boldsymbol{M}\Psi \overset{\text{set}}{=} 0 \\ &\implies (\boldsymbol{M}\Psi\boldsymbol{M}^\top)(\boldsymbol{M}\Psi) = (\Delta)(\boldsymbol{M}\Psi) \implies \boxed{\boldsymbol{M}\Psi\boldsymbol{M}^\top = \Delta} \end{aligned} \tag{10}$$

Classical MDS

$$c_1 = \mathbf{tr}(\Delta^2) + \mathbf{tr}((\boldsymbol{M}\Psi\boldsymbol{M}^\top)^2) - 2\,\mathbf{tr}(\Delta\boldsymbol{M}\Psi\boldsymbol{M}^\top).$$

$$\mathbb{R}^{n\times n} \ni \frac{\partial c_1}{\partial \Psi} = 2\boldsymbol{M}^\top(\boldsymbol{M}\Psi\boldsymbol{M}^\top)\boldsymbol{M} - 2\boldsymbol{M}^\top\Delta\boldsymbol{M} \overset{\text{set}}{=} 0$$

$$\implies \boldsymbol{M}^\top(\boldsymbol{M}\Psi\boldsymbol{M}^\top)\boldsymbol{M} = \boldsymbol{M}^\top(\Delta)\boldsymbol{M} \implies \boxed{\boldsymbol{M}\Psi\boldsymbol{M}^\top = \Delta} \tag{11}$$

$$\boxed{\boldsymbol{M} = \boldsymbol{I}} \tag{12}$$

$$\boxed{\Psi = \Delta} \tag{13}$$

$$\text{We had: } \boldsymbol{M} = \boldsymbol{V}^\top\boldsymbol{Q} \implies \boldsymbol{V}^\top\boldsymbol{Q} = \boldsymbol{I} \implies \boxed{\boldsymbol{Q} = \boldsymbol{V}} \tag{14}$$

$$\text{We had: } \boldsymbol{Y}^\top\boldsymbol{Y} = \boldsymbol{Q}\Psi\boldsymbol{Q}^\top = \boldsymbol{Q}\Psi^{\frac{1}{2}}\Psi^{\frac{1}{2}}\boldsymbol{Q}^\top \implies \boldsymbol{Y} = \Psi^{\frac{1}{2}}\boldsymbol{Q}^\top$$

$$\implies \boxed{\boldsymbol{Y} = \Delta^{\frac{1}{2}}\boldsymbol{V}^\top} \tag{15}$$

Lecture Outline

Generalized Classical MDS

In classical MDS, we had:

$$\underset{\boldsymbol{Y}}{\text{minimize}} \; ||\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}||_F^2 \tag{16}$$

- Solution: eigenvalue decomposition:
$\boxed{\text{linear kernel} = \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\Delta\boldsymbol{V}^\top} \implies \boxed{\boldsymbol{Y} = \Delta^{(1/2)}\boldsymbol{V}^\top}$
- In generalized classical MDS, we use a general kernel. The **kernel** can be written in terms of a valid metric **distance matrix** $\boldsymbol{D}$:

$$\boxed{\boldsymbol{K} = -\frac{1}{2}\boldsymbol{H}\boldsymbol{D}\boldsymbol{H}}$$

where $\boldsymbol{H}$ is the centring matrix: $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}11^\top$. Proof: Next slides.

Generalized Classical MDS

The squared Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, we have:

$$\begin{aligned} d_{ij}^2 &= ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &= \boldsymbol{x}_i^\top \boldsymbol{x}_i - \boldsymbol{x}_i^\top \boldsymbol{x}_j - \boldsymbol{x}_j^\top \boldsymbol{x}_i + \boldsymbol{x}_j^\top \boldsymbol{x}_j \\ &= \boldsymbol{x}_i^\top \boldsymbol{x}_i - 2\boldsymbol{x}_i^\top \boldsymbol{x}_j + \boldsymbol{x}_j^\top \boldsymbol{x}_j = \boldsymbol{G}_{ii} - 2\boldsymbol{G}_{ij} + \boldsymbol{G}_{jj}, \end{aligned}$$

where $\mathbb{R}^{n\times n} \ni \boldsymbol{G} := \boldsymbol{X}^\top \boldsymbol{X}$ is the Gram matrix. If $\mathbb{R}^n \ni \boldsymbol{g} := [\boldsymbol{g}_1, \dots, \boldsymbol{g}_n] = [\boldsymbol{G}_{11}, \dots, \boldsymbol{G}_{nn}] = \mathbf{diag}(\boldsymbol{G})$, we have:

$$d_{ij}^2 = \boldsymbol{g}_i - 2\boldsymbol{G}_{ij} + \boldsymbol{g}_j$$

$$\boxed{\boldsymbol{D} = \boldsymbol{g}1^\top - 2\boldsymbol{G} + 1\boldsymbol{g}^\top = 1\boldsymbol{g}^\top - 2\boldsymbol{G} + \boldsymbol{g}1^\top}$$

Generalized Classical MDS

$$
\begin{aligned}
\boldsymbol{HDH} &= (\boldsymbol{I} - \frac{1}{n}11^\top)\boldsymbol{D}(\boldsymbol{I} - \frac{1}{n}11^\top) \\
&= (\boldsymbol{I} - \frac{1}{n}11^\top)(1\boldsymbol{g}^\top - 2\boldsymbol{G} + \boldsymbol{g}1^\top)(\boldsymbol{I} - \frac{1}{n}11^\top) \\
&= \Big[\underbrace{(\boldsymbol{I} - \frac{1}{n}11^\top)1}_{=0}\boldsymbol{g}^\top - 2(\boldsymbol{I} - \frac{1}{n}11^\top)\boldsymbol{G} + (\boldsymbol{I} - \frac{1}{n}11^\top)\boldsymbol{g}1^\top\Big](\boldsymbol{I} - \frac{1}{n}11^\top) \\
&= -2(\boldsymbol{I} - \frac{1}{n}11^\top)\boldsymbol{G}(\boldsymbol{I} - \frac{1}{n}11^\top) + (\boldsymbol{I} - \frac{1}{n}11^\top)\boldsymbol{g}\underbrace{1^\top(\boldsymbol{I} - \frac{1}{n}11^\top)}_{=0} \\
&= -2(\boldsymbol{I} - \frac{1}{n}11^\top)\boldsymbol{G}(\boldsymbol{I} - \frac{1}{n}11^\top) = -2\,\boldsymbol{HGH}
\end{aligned}
$$

$$
\boldsymbol{HGH} = \boldsymbol{HX}^\top\boldsymbol{XH} = -\frac{1}{2}\boldsymbol{HDH} \overset{centered}{\Longrightarrow} \boldsymbol{X}^\top\boldsymbol{X} = -\frac{1}{2}\boldsymbol{HDH} \tag{17}
$$

$$
\boxed{\mathbb{R}^{n\times n} \ni \boldsymbol{K} = -\frac{1}{2}\boldsymbol{HDH}} \tag{18}
$$

Generalized Classical MDS

$$\boxed{\text{a valid kernel} = \boldsymbol{K} = -\frac{1}{2}\boldsymbol{H}\boldsymbol{D}\boldsymbol{H} = \boldsymbol{V}\Delta\boldsymbol{V}^\top} \tag{19}$$

$$\boldsymbol{Y}^\top\boldsymbol{Y} = \boldsymbol{Q}\Psi\boldsymbol{Q}^\top \tag{20}$$

$$\boxed{\boldsymbol{Y} = \Delta^{(1/2)}\boldsymbol{V}^\top} \tag{21}$$

Classical MDS with Euclidean distance is equivalent to Principal Component Analysis (PCA) [1]. Moreover, the generalized classical MDS is equivalent to kernel PCA.
Proof: See [2].

Lecture Outline

Metric MDS

$$\underset{\{\boldsymbol{y}_i\}_{i=1}^n}{\text{minimize}} \quad c_2 := \left(\frac{\sum_{i=1}^n \sum_{j=1,j<i}^n \big(d_x(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_y(\boldsymbol{y}_i, \boldsymbol{y}_j)\big)^2}{\sum_{i=1}^n \sum_{j=1,j<i}^n d_x^2(\boldsymbol{x}_i, \boldsymbol{x}_j)} \right)^{\frac{1}{2}} \tag{22}$$

or, without the normalization factor:

$$\underset{\{\boldsymbol{y}_i\}_{i=1}^n}{\text{minimize}} \quad c_2 := \left(\sum_{i=1}^n \sum_{j=1,j<i}^n \big(d_x(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_y(\boldsymbol{y}_i, \boldsymbol{y}_j)\big)^2 \right)^{\frac{1}{2}} \tag{23}$$

The classical MDS is a linear method and has a closed-form solution; however, the metric and non-metric MDS methods are **nonlinear** but do *not have closed-form solutions* and should be solved **iteratively**.

$$y_{i,k}^{(\nu+1)} := y_{i,k}^{(\nu)} - \eta \left| \frac{\partial^2 c_2}{\partial y_{i,k}^2} \right|^{-1} \frac{\partial c_2}{\partial y_{i,k}} \tag{24}$$

$$y_{i,k}^{(\nu+1)} := y_{i,k}^{(\nu)} - \eta \, \frac{\partial c_2}{\partial y_{i,k}} \tag{25}$$

Sammon mapping [3] is a special case of metric MDS.

Lecture Outline

Non-Metric MDS

In non-metric MDS, rather than using a distance metric, $d_y(\boldsymbol{x}_i, \boldsymbol{x}_j)$, for the distances between points in the embedding space, we use $f(d_y(\boldsymbol{x}_i, \boldsymbol{x}_j))$ where $f(.)$ is a non-parametric monotonic function. Only the order of dissimilarities is important rather than the amount of dissimilarities:

$$d_y(\boldsymbol{y}_i, \boldsymbol{y}_j) \leq d_y(\boldsymbol{y}_k, \boldsymbol{y}_\ell) \iff f(d_y(\boldsymbol{y}_i, \boldsymbol{y}_j)) \leq f(d_y(\boldsymbol{x}_k, \boldsymbol{y}_\ell)) \tag{26}$$

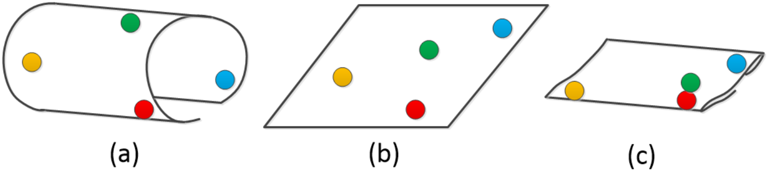
$$\underset{\{\boldsymbol{y}_i\}_{i=1}^n}{\text{minimize}} \quad c_3 := \left(\frac{\sum_{i=1}^n \sum_{j=1, j<i}^n \big(d_x(\boldsymbol{x}_i, \boldsymbol{x}_j) - f(d_y(\boldsymbol{y}_i, \boldsymbol{y}_j))\big)^2}{\sum_{i=1}^n \sum_{j=1, j<i}^n d_x^2(\boldsymbol{x}_i, \boldsymbol{x}_j)} \right)^{\frac{1}{2}} \tag{27}$$

$$y_{i,k}^{(\nu+1)} := y_{i,k}^{(\nu)} - \eta \left| \frac{\partial^2 c_3}{\partial y_{i,k}^2} \right|^{-1} \frac{\partial c_3}{\partial y_{i,k}} \tag{28}$$

$$y_{i,k}^{(\nu+1)} := y_{i,k}^{(\nu)} - \eta \, \frac{\partial c_3}{\partial y_{i,k}} \tag{29}$$

Lecture Outline

Why Need Nonlinear Method?

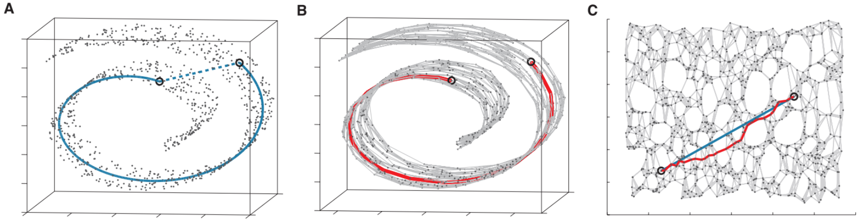


(a) A 2D nonlinear manifold where the data exist on in the 3D original space. As the manifold is nonlinear, the geodesic distances of points on the manifold are different from their Euclidean distances. (b) The correct unfolded manifold where the geodesic distances of points on the manifold have been preserved. (c) Applying the linear PCA, which takes Euclidean distances into account, on the nonlinear data where the found subspace has ruined the manifold so the far away red and green points have fallen next to each other.

The credit of this example is for Prof. Ali Ghodsi.

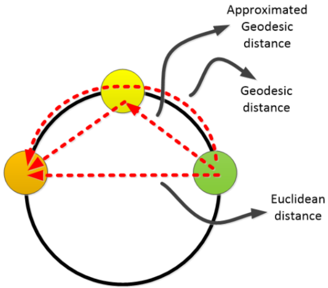
Isomap

- Isomap was first proposed in [4].
- The subspace (manifold) of data may not be linear (for example, Swiss roll).
- The dotted line is the Euclidean distance. The blue line is the geodesic distance. The red line is an approximation for the geodesic distance.



Isomap

Calculating the geodesic distances are hard. so, we use k-nearest neighbor (kNN) graph and by piece-wise Euclidean distances, we approximate it. The Dijkstra or Floyd-Warshal algorithm is used to find the shortest paths.

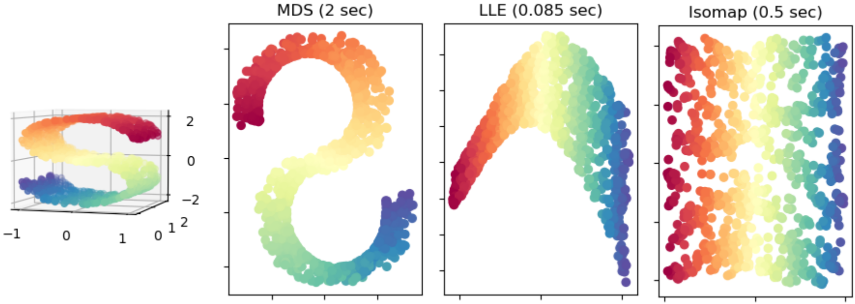


We use the geodesic distance matrix in $\boldsymbol{K} = -\frac{1}{2}\boldsymbol{HDH}$ and then apply MDS.

$$\boxed{\boldsymbol{K} = -\frac{1}{2}\boldsymbol{HDH} = \boldsymbol{V}\Delta\boldsymbol{V}^\top} \quad \boxed{\boldsymbol{Y} = \Delta^{(1/2)}\boldsymbol{V}^\top}$$

Lecture Outline

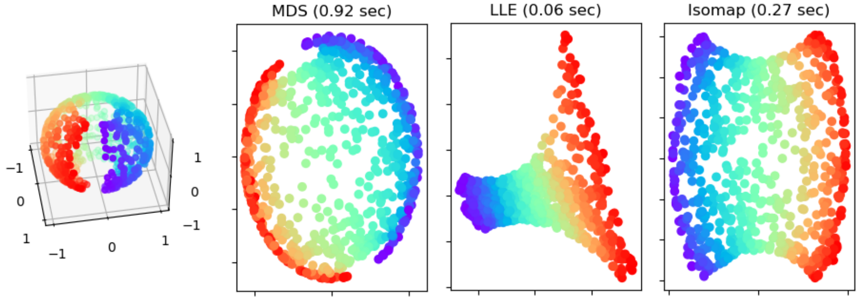
Example on Swiss Roll



Note that the MDS embedding captures the most variant direction, as in PCA.

The credit of image is for: `https://scikit-learn.org/stable/auto_examples/manifold/plot_compare_methods.html#sphx-glr-auto-examples-manifold-plot-compare-methods-py`

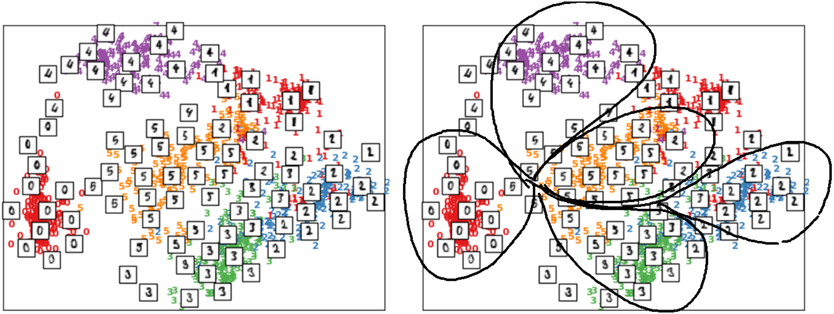
Example on Sphere



The credit of image is for: `https://scikit-learn.org/stable/auto_examples/manifold/plot_manifold_sphere.html#sphx-glr-auto-examples-manifold-plot-manifold-sphere-py`

Example on MNIST

The Isomap embedding:

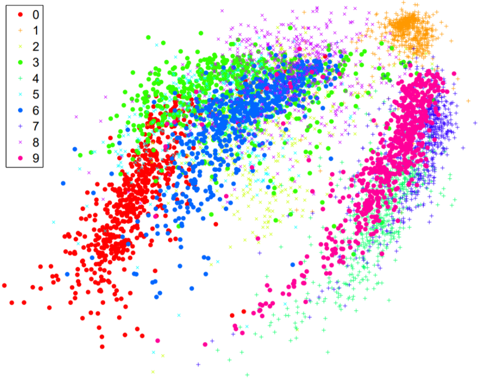


We have empirically seen that the embedding of Isomap is usually, but not always, like the legs of octopus.

The credit of image is for: `https://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html`

Example on MNIST

The Isomap embedding:



We have empirically seen that the embedding of Isomap is usually, but not always, like the legs of octopus.

The credit of image is for [5].

Useful Resources To Read

- Tutorial paper: "Multidimensional Scaling, Sammon Mapping, and Isomap: Tutorial and Survey" [2]
- Tutorial YouTube videos by Prof. Ali Ghodsi at University of Waterloo: [Click here]
- A book on MDS: [6]

References

[1] B. Ghojogh and M. Crowley, "Unsupervised and supervised principal component analysis: Tutorial," *arXiv preprint arXiv:1906.03148*, 2019.

[2] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Multidimensional scaling, Sammon mapping, and Isomap: Tutorial and survey," *arXiv preprint arXiv:2009.08136*, 2020.

[3] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on computers*, vol. 100, no. 5, pp. 401–409, 1969.

[4] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[5] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[6] M. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of data visualization*, pp. 315–347, Springer, 2008.