Data Embedding: A Geometric Perspective

Instructor: Hao Su

Feb 7, 2019
Agenda

• General theories of embedding

• Algorithms of computing data embedding
Agenda

• General theories of embedding

• Algorithms of computing data embedding
Many Names

• Dimensionality reduction

• Embedding

• Multidimensional scaling

• Manifold learning

...
Basic Task

Given pairwise distances extract an embedding.

Is it always possible?
What dimensionality?
Metric Space

Ordered pair \((M, d)\) where \(M\) is a set and \(d\) satisfies

\[
\begin{align*}
d(x, y) & \geq 0 \\
d(x, y) &= 0 \iff x = y \\
d(x, y) &= d(y, x) \\
d(x, z) &\leq d(x, y) + d(y, z)
\end{align*}
\]

\(\forall x, y, z \in M\)
Many Examples of Metric Spaces

\[ \mathbb{R}^n, d(x, y) := \| x - y \|_p \]

\[ S \subset \mathbb{R}^3, d(x, y) := \text{geodesic} \]

\[ C^\infty(\mathbb{R}), d(f, g)^2 := \int_{\mathbb{R}} (f(x) - g(x))^2 \, dx \]
Isometry  [ahy-som-i-tree]:
A map between metric spaces that preserves pairwise distances.
Q: Can you *always* embed a metric space isometrically in?
Q:

Can you always embed a finite metric space isometrically in ?
Disappointing Example

\[ X := \{a, b, c, d\} \]

\[ d(a, d) = d(b, d) = 1 \]

\[ d(a, b) = d(a, c) = d(b, c) = 2 \]

\[ d(c, d) = 1.5 \]

Cannot be embedded in Euclidean space!
Embedding Manifold to Euclidean Space
A **Riemannian space** \((M, g)\) is a real, smooth manifold \(M\) equipped with an inner product \(g_p\) on the tangent space \(T_pM\) at each point \(p\) that varies smoothly from point to point.
Embedding of Riemannian Manifold

Strong Whitney embedding theorem:

Any smooth real $m$-manifold (required also to be Hausdorff and second-countable) can be smoothly embedded in the real $2m$-space ($\mathbb{R}^{2m}$), if $m > 0$.

- This is the best linear bound on the smallest-dimensional Euclidean space that all $m$-dimensional manifolds embed in.
- Although every $n$-manifold embeds in $\mathbb{R}^{2n}$, one can frequently do better.
Embedding of Riemannian Manifold

**Theorem.** Let \((M,g)\) be a Riemannian manifold and \(f: M^m \to \mathbb{R}^n\) a short \(C^\infty\)-embedding (or immersion) into Euclidean space \(\mathbb{R}^n\), where \(n \geq m+1\). Then for arbitrary \(\varepsilon > 0\) there is an embedding (or immersion) \(f_\varepsilon: M^m \to \mathbb{R}^n\) which is

i. in class \(C^1\),

ii. isometric: for any two vectors \(v, w \in T_x(M)\) in the tangent space at \(x \in M\),

\[
g(v, w) = \langle df_\varepsilon(v), df_\varepsilon(w) \rangle,
\]

iii. \(\varepsilon\)-close to \(f\):

\[
\|f(x) - f_\varepsilon(x)\| < \varepsilon \quad \forall x \in M.
\]

In particular, any \(m\)-dimensional Riemannian manifold admits an isometric \(C^1\)-embedding into an *arbitrarily small neighborhood* in \(2m\)-dimensional Euclidean space.

For example, it follows that any closed oriented Riemannian surface can be \(C^1\) isometrically embedded into an arbitrarily small \(\varepsilon\)-ball in Euclidean 3-space (for small \(\varepsilon\) there is no such \(C^2\)-embedding since from the formula for the Gauss curvature an extremal point of such an embedding would have curvature \(\geq \varepsilon^{-2}\)).
Embedding in $\ell_p$
Approximate Embedding

\[
\begin{align*}
\text{expansion}(f) & := \max_{x,y} \frac{\mu(f(x), f(y))}{\rho(x, y)} \\
\text{contraction}(f) & := \max_{x,y} \frac{\rho(x, y)}{\mu(f(x), f(y))} \\
\text{distortion}(f) & := \text{expansion}(f) \times \text{contraction}(f)
\end{align*}
\]
Well-Known Result

**Theorem (Bourgain, 1985).**
Let \((X, d)\) be a metric space on \(n\) points. Then,

\[
(X, d) \xrightarrow{O(\log n)} \ell_p O(\log^2 n)
\]

Any \(n\)-point metric space \((X, D)\) can be embedded in \(\ell_2\) (in fact, in every \(\ell_p\)) with distortion \(O(\log n)\).
metric, the simplest in many respects, and the most restricted. Every finite \( \ell_2 \) metric embeds isometrically in \( \ell_p \) for all \( p \). More generally, we have the following

**THEOREM 8.1.1**  \textit{Dvoretzky’s theorem (a finite quantitative version)}

For every \( d \) and every \( \varepsilon > 0 \) there exists \( n = n(d, \varepsilon) \leq 2^{O(d/\varepsilon^2)} \) such that \( \ell^d_2 \) can be \((1+\varepsilon)\)-embedded in every \( n \)-dimensional normed space.
THE EUCLIDEAN SPACES $\ell_2^d$

- Dimension reduction:

**THEOREM 8.2.3** Johnson and Lindenstrauss [JL84]

For every $\varepsilon > 0$, any $n$-point $\ell_2$ metric can be $(1+\varepsilon)$-embedded in $\ell_2^{O(\log n/\varepsilon^2)}$. 
The spaces $\ell^d_\infty$ are the richest (and thus generally the most difficult to deal with); every $n$-point metric space $(X, D)$ embeds isometrically in $\ell^n_\infty$. To see this, write $X = \{x_1, x_2, \ldots, x_n\}$ and define $f: X \to \ell^n_\infty$ by $f(x_i)_j = D(x_i, x_j)$.

**THEOREM 8.2.2**

For an integer $b > 0$ set $c = 2b - 1$. Then any $n$-point metric space can be embedded in $\ell^d_\infty$ with distortion $c$, where $d = O(bn^{1/b} \log n)$. 
Graph Embedding
THEOREM 8.3.2  Rao [Rao99]
Any n-point planar-graph metric can be embedded in $\ell_2$ with distortion $O(\sqrt{\log n})$. 
<table>
<thead>
<tr>
<th>FROM</th>
<th>TO</th>
<th>DISTORTION</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>any constant-degree expander</td>
<td>( \ell_p, 1 \leq p &lt; \infty )</td>
<td>( O(\log n) )</td>
<td>[Bou85]</td>
</tr>
<tr>
<td>( k )-reg. graph, ( k \geq 3 ), girth ( g )</td>
<td>( \ell_p, p &lt; \infty ) fixed ( \ell_2 )</td>
<td>( \Omega(\log n) )</td>
<td>[LLR95]</td>
</tr>
<tr>
<td>any</td>
<td>( \ell_\infty^{O(\log n)} )</td>
<td>( \Omega(\sqrt{g}) )</td>
<td>[LMN02]</td>
</tr>
<tr>
<td>some</td>
<td>( \Omega(n^{1/b}) )-dim. normed space</td>
<td>( 2b - 1, b = 1, 2, \ldots )</td>
<td>[Mat96]</td>
</tr>
<tr>
<td>any</td>
<td>( \ell_1^1 )</td>
<td>( 2b - 1, b = 1, 2, \ldots ) (Erdős's conj.)</td>
<td>[Mat96]</td>
</tr>
<tr>
<td>any</td>
<td>( \ell_p^d, d ) fixed</td>
<td>( \Theta(n) )</td>
<td>[Mat90]</td>
</tr>
<tr>
<td>( \ell_2 ) metric</td>
<td>( \ell_2^{O(\log n/\varepsilon^2)} )</td>
<td>( O(n^{2/d \log^{3/2} n}), \Omega(n^{1/[(d+1)/2]} )</td>
<td>[Mat90]</td>
</tr>
<tr>
<td>( \ell_1 ) metric</td>
<td>( \ell_1^{n^\alpha}, 0 &lt; \alpha &lt; 1 )</td>
<td>( 1 + \varepsilon )</td>
<td>[JL84]</td>
</tr>
<tr>
<td>planar or forbidden minor</td>
<td>( \ell_2 )</td>
<td>( \Omega(\sqrt{\log n}) )</td>
<td>[BC03]</td>
</tr>
<tr>
<td>series-parallel</td>
<td>( \ell_2 )</td>
<td>( \Omega(\sqrt{\log n}) )</td>
<td>[Rao99]</td>
</tr>
<tr>
<td>planar</td>
<td>( \ell_\infty^{O(\log^2 n)} )</td>
<td>( O(1) )</td>
<td>[NR02]</td>
</tr>
<tr>
<td>outerplanar or series-parallel</td>
<td>( \ell_1 )</td>
<td>( O(1) )</td>
<td>implicit in [Rao99]</td>
</tr>
<tr>
<td>tree</td>
<td>( \ell_1 )</td>
<td>( O(1) )</td>
<td>[GNRS99]</td>
</tr>
<tr>
<td></td>
<td>( \ell_\infty^{O(\log n)} )</td>
<td>( 1 )</td>
<td>(folklore)</td>
</tr>
<tr>
<td></td>
<td>( \ell_2 )</td>
<td>( 1 )</td>
<td>[LLR95]</td>
</tr>
<tr>
<td></td>
<td>( \ell_2^d )</td>
<td>( \Theta((\log \log n)^{1/2}) )</td>
<td>[Bou86, Mat99]</td>
</tr>
<tr>
<td></td>
<td>( \ell_2^2 )</td>
<td>( O(n^{1/(d-1)}) )</td>
<td>[Gup00]</td>
</tr>
<tr>
<td></td>
<td>tree, unit edges</td>
<td>( \Theta(\sqrt{n}) )</td>
<td>[BMMV02]</td>
</tr>
<tr>
<td>Hausdorff metric over ((X, D))</td>
<td>( \ell_\infty^{\lceil X \rceil} )</td>
<td>( 1 )</td>
<td>[FCI99]</td>
</tr>
<tr>
<td>Hausd. over ( s )-subsets of ((X, D))</td>
<td>( \ell_\infty^{O(1)\lceil X \rceil^\alpha \log \Delta} )</td>
<td>( c(\alpha) )</td>
<td>[FCI99]</td>
</tr>
<tr>
<td>Hausd. over ( s )-subsets of ( \ell_p^k )</td>
<td>( \ell_\infty^{O(1/\varepsilon)O(k) \log \Delta} )</td>
<td>( 1 + \varepsilon )</td>
<td>[FCI99]</td>
</tr>
<tr>
<td>EMD over ((X, D))</td>
<td>( \ell_1 )</td>
<td>( O(\log</td>
<td>X</td>
</tr>
<tr>
<td>Levenshtein metric</td>
<td>( \ell_1 )</td>
<td>( O(\log d \cdot \log^* d) )</td>
<td>[ADG+03]</td>
</tr>
<tr>
<td>block-edit metric over ( \Sigma^d )</td>
<td>( \ell_\infty^{O(B \log n)} )</td>
<td>( 1 )</td>
<td>[MS00, CM02]</td>
</tr>
<tr>
<td>(1,2)-B metric</td>
<td>convex comb. of dom. trees (HSTs)</td>
<td>( O(\log n) )</td>
<td>[FRT03]</td>
</tr>
<tr>
<td></td>
<td>convex comb. of spanning trees</td>
<td>( 2^{O(\sqrt{\log n \log \log n})} )</td>
<td>[AKPW95]</td>
</tr>
</tbody>
</table>
Agenda

• General theories of embedding

• Algorithms of computing data embedding
Euclidean Case

\[ D_{ij} = \left\| x_i - x_j \right\|_2^2, \quad D \in \mathbb{R}^{n \times n} \]

**Proposition.**

Proof:

\[ D = -2X^\top X + \text{diag}(X^\top X)1^\top + 1\text{diag}(X^\top X)^\top \]

Embedding via eigenvalue problem (take \( x_1 = 0 \):

\[
\left\| x_i - x_j \right\|_2^2 = \left\| x_i \right\|_2^2 + \left\| x_j \right\|_2^2 - 2x_i \cdot x_j
\]

\[ x_i \cdot x_j = \frac{1}{2} \left[ \left\| x_i \right\|_2^2 + \left\| x_j \right\|_2^2 - \left\| x_i - x_j \right\|_2^2 \right] \]
Gram Matrix \([\text{gram mëy-triks}]\): A matrix of inner products

\[ X^T X \]
Classical Multidimensional Scaling

1. Double centering: \( B := -\frac{1}{2} J D J \)
   Centering matrix \( J := I - \frac{1}{n} 11^\top \)

2. Find \( m \) largest eigenvalues/eigenvectors

3. \( X = E_m \Lambda_m^{1/2} \)

"MDS"

Stress Majorization

\[ \min_X \sum_{ij} \left( d_{ij}^0 - \| x_i - x_j \|_2^2 \right)^2 \]

Nonconvex!

**SMACOF:**
Scaling by Majorizing a Complicated Function

SMACOF Potential Terms

\[ \sum_{ij} (d^0_{ij})^2 = \text{const.} \]

\[ \sum_{ij} \| x_i - x_j \|_2^2 = \text{tr}(XVX^\top), \quad \text{where} \quad V = 2nI - 211^\top \]

\[ \sum_{ij} d^0_{ij} \| x_i - x_j \|_2 = \text{tr}(XB(X)X^\top) \]

where \( b_{ij}(X) := \begin{cases} 
- \frac{2d^0_{ij}}{\| x_i - x_j \|_2} & \text{if } x_i \neq x_j, i \neq j \\
0 & \text{if } x_i = x_j, i \neq j \\
- \sum_{j \neq i} b_{ij} & \text{if } i = j 
\end{cases} \]
**SMACOF Lemma**

\[
\sum_{ij} (d_{ij}^0)^2 = \text{const.}
\]

\[
\sum_{ij} \|x_i - x_j\|_2^2 = \text{tr}(XVX^\top)
\]

\[
\sum_{ij} d_{ij}^0 \|x_i - x_j\|_2 = \text{tr}(XB(X)X^\top)
\]

where \( b_{ij}(X) := \begin{cases} 
- \frac{2d_{ij}^0}{\|x_i - x_j\|_2} & \text{if } x_i \neq x_j, i \neq j \\
0 & \text{if } x_i = x_j, i \neq j \\
- \sum_{j \neq i} b_{ij} & \text{if } i = j 
\end{cases} \)

**Lemma.** Define

\[\tau(X, Z) := \text{const.} + \text{tr}(XVX^\top) - 2\text{tr}(XB(Z)Z^\top)\]

Then,

\[\tau(X, X) \leq \tau(X, Z) \quad \forall Z\]

with equality exactly when \( X = Z \).
**SMACOF: Single Step**

\[ X^{k+1} \leftarrow \min_X \tau(X, X^k) \]

\[ \tau(X, Z) := \text{const.} + \text{tr}(XVX^\top) - 2\text{tr}(XB(Z)Z^\top) \]

\[ \implies 0 = \nabla_X [\tau(X, X^k)] \]

\[ = 2XV - 2X^kB(X^k) \]

\[ \implies X^{k+1} = X^kB(X^k)V^+ \]

\[ V^+ = (2nI - 211^\top)^+ \]

\[ = \frac{1}{2n} \left( I - \frac{11^\top}{n} \right)^+ \]

\[ = \frac{1}{2n} \left( I - \frac{11^\top}{n} \right) \]

**Majorization-Minimization algorithm**

Objective convergence:

\[ \tau(X^{k+1}, X^{k+1}) \leq \tau(X^k, X^k) \]
More General Metric MDS

• In general, we minimize directly the square loss on distances

\[
\text{stress} = \mathcal{L}(\hat{d}_{ij}) = \left( \frac{\sum_{i<j}(\hat{d}_{ij} - f(d_{ij}))^2}{\sum d_{ij}^2} \right)^{\frac{1}{2}}
\]

• Sammon mapping

\[
\text{Sammon's stress}(\hat{d}_{ij}) = \frac{1}{\sum_{\ell<k} d_{\ell k}} \sum_{i<j} \frac{(\hat{d}_{ij} - d_{ij})^2}{d_{ij}}
\]

• This weighting system normalizes the squared-errors in pairwise distances by using the distance in the original space. As a result, Sammon mapping preserves the small $d_{ij}$, giving them a greater degree of importance in the fitting procedure than for larger values of $d_{ij}$

Generally solved by gradient descent
Non-Linear Dimensionality Reduction
Non-Linear Dimensionality Reduction

- Many data sets contain essential nonlinear structures that “invisible” to PCA and MDS
- Must resort to some nonlinear dimensionality reduction approaches

Data sampled from a non-linear manifold
The Choice of Distance

- We can try to capture the manifold structure through the right notion of distance directly on the manifold (geodesic distance)
The Challenge of NLDR

• An unsupervised learning algorithm must discover the global internal coordinates of the manifold without external signals that suggest how the data should be embedded in low dimensions.
Isomap
ISOMAP
(J. B. Tenenbaum, V. de Silva and J. C. Langford)

• Example of non-linear structure (Swiss roll)
  • Only the geodesic distances reflect the true low-dimensional geometry of the manifold

• ISOMAP (Isometric Feature Mapping)
  • Preserves the intrinsic geometry of the data
  • Uses the geodesic manifold distances between all pairs
ISOMAP (Algorithm Description)

- **Step 1**
  - Form a near-neighbor graph $G$ on the original data points, weighing the edges based on their original distances $d_x(i, j)$.

- **Step 2**
  - Estimate the geodesic distances $d_G(i, j)$ between all pairs of points on the sampled manifold by computing their shortest path distances in the graph $G$.

- **Step 3**
  - Construct an embedding of the data in $d$-dimensional Euclidean space $Y$ that best preserves the distances (MDS).
Near Neighbor Graph

- **Step 1**
  - Determining neighboring points **within a fixed radius** based on the input space distance

\[ \# \ \epsilon\text{-radius} \quad \# \text{K-nearest neighbors} \]

\[ d_X(i, j) \]

- These neighborhood relations are represented as a weighted graph \( G \) over the data points.
Shortest Path Computation

- **Step 2**
  - Estimating the geodesic distances \(d_G(i, j)\) between all pairs of points on the manifold by computing their shortest path distances in the graph \(G\).
  - Can be done using Floyd/Warshall’s algorithm or Dijkstra’s algorithm

\[
d_G(i, j) = d_X(i, j) \quad \text{neighboring } \quad \text{otherwise}
\]

\[
d_G(i, j) = \infty \quad \text{otherwise}
\]

For \(k = 1, 2, \ldots, N\)

\[
d_G(i, j) = \min\{ d_X(i, j), d_X(i, k) + d_X(k, j) \}
\]
Euclidean Embedding

- **Step 3**
  - Constructing an embedding of the data in $d$-dimensional Euclidean space $Y$ that best preserves the inter-point distances
  
  - This is of course nothing but an MDS problem
Recovery Guarantees

- Isomap is guaranteed asymptotically to recover the true dimensionality and geometric structure of nonlinear manifolds.
- As the sample data points increases, the graph distances provide increasingly better approximations to the intrinsic geodesic distances.
ISOMAP Examples

# Face
: face pose and illumination

# Hand writing
: bottom loop and top arch

MDS: open triangles
Isomap: filled circles
Laplacian Eigenmaps
Laplacian Eigenmaps
(M. Belkin, P. Niyogi)

- Start same as Isomap, but use a spectral embedding in lieu of MDS

Hole distorts long geodesic distances
Locally Linear Embeddings
Locally Linear Embeddings (LLE)  
(S. T. Roweis and L. K. Saul)

- Define neighborhood relations between points (build NN graph)
  - $k$ nearest neighbors
  - $\varepsilon$-balls
- Find weights that reconstruct each data point from its neighbors:

$$\min \sum_{i} \sum_{j \in N(i)} w_{ij} x_{j} \quad \text{s.t.} \quad \sum_{i} w_{ij} = 1$$

- Find low-dimensional coordinates so that the same weights hold: $x_{1}^{'} , \ldots , x_{n}^{'} \in R^{d}$

$$\min \sum_{i} \sum_{j \in N(i)} w_{ij} x_{j}^{'} \quad \text{s.t.} \quad \sum_{i} w_{ij} = 1$$

$x_{i}^{'} = Y_{i}$
From Local to Global

- The weights \( w_{ij} \) capture the local shape
  - Invariant to translation, rotation and scale of the neighborhood
  - If the neighborhood lies on a manifold, the *local mapping* from the global coordinates \( (R^D) \) to the surface coordinates \( (R^d) \) is almost linear
  - Thus, the weights \( w_{ij} \) should hold also for manifold \( (R^d) \) coordinate system!

\[
\min \sum_{w_{ij} = 1} \left\| x_i - \sum_{j \in N(i)} w_{ij} x_j \right\|^2
\]

\[
\min_{x_1', \ldots, x_n'} \sum_i \left\| x_i' - \sum_{j \in N(i)} w_{ij} x_j' \right\|^2
\]
Solving the Minimizations

- Linear least squares (using Lagrange multipliers)

$$\min_{\sum_j w_{ij} = 1} \left\| x_i - \sum_{j \in N(i)} w_{ij} x_j \right\|^2$$

- To find $x_1', \ldots, x_n' \in \mathbb{R}^d$ that minimize,

$$\min_{x_1', \ldots, x_n'} \sum_{i} \left\| x_i' - \sum_{j \in N(i)} w_{ij} x_j' \right\|^2$$

a sparse eigenvalue problem is solved. Additional constraints are added for conditioning:

$$\sum_i x_i' = 0, \quad \frac{1}{n} \sum_i x_i' x_i'^T = I$$
## Comparison: ISOMAP vs. LLE

<table>
<thead>
<tr>
<th>ISOMAP</th>
<th>LLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global distances</td>
<td>Local averaging</td>
</tr>
<tr>
<td>$k$-NN graph distances</td>
<td>$k$-NN graph weighting</td>
</tr>
<tr>
<td>Largest eigenvectors</td>
<td>Smallest eigenvectors</td>
</tr>
<tr>
<td>Dense matrix</td>
<td>Sparse matrix</td>
</tr>
</tbody>
</table>

Many More Methods
Many NLDR Methods

From Wikipedia

Caltech 101

t-Distributed Stochastic Neighbor Embedding (t-SNE)