Comparison of Dimensionality Reduction techniques for Dataset X

Ellango Jothimurugesan, Yilun Zhou, Roger Zou

April 25, 2015

Abstract

We will introduce some popular dimensionality reduction methods: Principal Component Analysis (PCA), Locally Linear Embedding (LLE), and Isomap. This survey assumes familiarity with elementary linear algebra. Some preliminary concepts will be given without proof.

Preliminaries

Eigenvalue Decomposition (EVD)

We will only consider the EVD for symmetric matrices, and so will only review properties applied to matrices of this type:

Definition *(Symmetric Matrix)* Let A be a $n \times n$ matrix. Then A is *symmetric* if $A = A^T$.

We can consider the eigenvectors and eigenvalues of symmetric matrices (and square matrices in general):

Definition *(Eigenvectors and Eigenvalues)* Let A be a $n \times n$ real matrix. A non-zero vector $\mathbf{v}$ is an *eigenvector* if and only if

$$A\mathbf{v} = \lambda\mathbf{v}$$

where λ is the corresponding (scalar) *eigenvalue.*

Intuitively, if we consider A to be a linear map $A : \mathbf{R}^n \to \mathbf{R}^n$, then an eigenvector $\mathbf{v}$ is a vector that has its direction preserved and scaled by λ under A.

An important result of linear algebra is the spectral theorem, which formally states that:

Theorem 0.1. (Spectral Theorem and EVD) *Let A be a $n \times n$ real, symmetric matrix. Then there exists exactly n eigenvalues (not necessarily distinct) $\lambda_1, \ldots, \lambda_n$, with corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ that form an orthonormal basis. Furthermore, there exists the decomposition:*

$$A = Q\Lambda Q^T$$

where Q is an orthogonal matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and Λ is a diagonal matrix with $\lambda_1, \ldots, \lambda_n$ along the diagonal.

Singular Value Decomposition (SVD)

We can perform a related, extremely useful, factorization to any real $m \times n$ matrix:

Theorem 0.2. (Existence of SVD) *Let A be a real $m \times n$ matrix. Then there exists orthogonal matrices*

$$U = \begin{bmatrix} \mathbf{u}_1 & \ldots & \mathbf{u}_m \end{bmatrix} \qquad V = \begin{bmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_n \end{bmatrix}$$

with $\mathbf{u}_i \in \mathbb{R}^m$ and $\mathbf{v}_j \in \mathbb{R}^n$, s.t.

$$A = U\Sigma V^T$$

where Σ is a diagonal matrix with singular values $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$ along the diagonal, $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are the left singular vectors, and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the right singular vectors.

This is closely related to the EVD. Indeed, it is useful to observe that: Given $A = U\Sigma V^T$, we have that:

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma^2 U^T$$
$$A^T A = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^2 V^T$$

Therefore, we can see that

1. The square of the singular values of A are the eigenvalues of the symmetric $n \times n$ matrix $A^T A$ or AA^T.

2. The left singular vectors of A are the eigenvectors of AA^T.

3. The right singular vectors of A are the eigenvectors of A^TA.

Dimensionality reduction for linear manifolds

We are now ready to introduce the first technique that can be used for dimensionality reduction. We assume that the data lies in some k-dimensional approximately linear manifold in a larger m-dimensional vector space $\mathbb{R}^m$. The goal of dimensionality reduction is to re-express the data in k necessary dimensions, rather than the much larger m original dimensions. Two fundamentally related methods fall under this category: PCA and MDS. Both methods are very efficient (requiring only matrix operations and factorizations) because they exploit the linearity assumption.

Principal Component Analysis (PCA)

Let X be a zero-meaned $m \times n$ data matrix, where m is the dimension of the data, and n the number of samples.

Definition *(Covariance matrix)* We define the covariance matrix of X to be

$$C_X = \frac{1}{n-1} XX^T$$

to be a symmetric, $m \times m$ matrix that quantifies the pairwise correlations between all data dimensions.

The intuition behind PCA is to find some orthonormal basis $\mathbf{p}_1, \dots, \mathbf{p}_m$ in $\mathbb{R}^m$ that transforms the data in the standard basis with coefficients in X to this special basis represented by coefficients in Y such that the covariance matrix of Y, C_Y is diagonalized.

In other words, we wish to find some matrix P where

$$Y = PX$$

such that the covariance matrix of Y,

$$C_Y = \frac{1}{n-1} YY^T$$

is diagonalized. Furthermore, the *rows* $\mathbf{p}_1, \dots, \mathbf{p}_m$ in $\mathbb{R}^m$ of P are exactly the basis vectors we're looking for. This can be seen easily by considering

$$y_i = \sum_{j=1}^{m} \mathbf{p}_j^T \mathbf{x}_i$$

Therefore, the goal of PCA is to find P.

Theorem 0.3. (PCA) *Let X and Y be a $m \times n$ matrix, where C_Y is diagonalized, and let $X = U\Sigma V^T$ be the singular value decomposition of X. Then the matrix P s.t. $Y = PX$ is*

$$P = U^T$$

Proof. Let $C_Y = \frac{1}{n-1} YY^T$ be the covariance matrix of Y. We wish to find the P s.t. C_Y is diagonal.

$$\begin{aligned} C_Y &= \frac{1}{n-1} YY^T \\ &= \frac{1}{n-1} (PX)(PX)^T \\ &= \frac{1}{n-1} P(XX^T)P^T \end{aligned}$$

Taking the SVD of X, we have

$$\begin{aligned} &= \frac{1}{n-1} P(U\Sigma V^T)(U\Sigma V^T)^T P^T \\ &= \frac{1}{n-1} P(U\Sigma^2 U^T)P^T \end{aligned}$$

Here we make the observation that if $P = U^T$, we have by substituting that

$$\begin{aligned} C_Y &= \frac{1}{n-1} U^T U \Sigma^2 U^T U \\ &= \frac{1}{n-1} \Sigma^2 \\ &= \frac{1}{n-1} \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \end{bmatrix} \end{aligned}$$

This completes the proof. □

We now also derived a simple algorithm to compute PCA of the matrix X:

1. Take the SVD of $X = U\Sigma V^T$;

2. return $Y = U^T X$.

This, Y is a $m \times n$ matrix of the transformed data into a more "natural" basis (i.e. C_Y is diagonalized).

PCA for dimensionality reduction

A consequence of the eigenvalue decomposition above is that there is a natural ordering to the singular values.

$$\sigma_1 \geq \ldots \sigma_r \geq \sigma_{r+1} = \ldots = \sigma_m$$

Since they correspond to variances in each principal direction $\mathbf{p}_i$, if we wish to find the first three principal components (i.e. to have data in $\mathbb{R}^3$), let

$$P_3 = \begin{bmatrix} \mathbf{p}_1^T \\ \mathbf{p}_2^T \\ \mathbf{p}_3^T \end{bmatrix}$$

be a $3 \times m$ matrix. Then

$$Y_3 = P_3 X$$

returns a $3 \times n$ data matrix Y_3, with n samples in the 3 principal dimensions that account for the most variance.

Classical Multidimensional Scaling (CMDS)

Given some distance matrix D, where d_{ij} measures the dissimilarity between elements i and j, MDS attempts to find a specified low-dimensional representation that preserves distances as much as possible. Suppose X is some (possibly unknown) $m \times n$ data matrix of m dimensions and n samples that generated D. Furthermore, there is a k-dimensional manifold embedded in X, which we wish to represent with Y. For simplicity we will measure distances with the Euclidean Metric.

Theorem 0.4. (CMDS) *Let D be a real, symmetric $n \times n$ dissimilarity/distance matrix generated with a Euclidean metric from X, an unknown $m \times n$ data matrix of n samples and m dimensions. If $D = V\Sigma^2 V^T$ is the eigenvalue decomposition of D and Σ_m is the first m rows of Σ, then*

$$X = \Sigma_m V^T$$

not necessarily unique.

Proof. Let x_i be the i-th m-dimensional element of X. Then the euclidean distance between x_i and x_j is:

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$$

Writing out the terms, we have that

$$d_{ij} = \|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_2^2 - 2\mathbf{x}_i^T \mathbf{x}_j$$

For convenience, we can center and rescale d_{ij}. These are acceptable, linear operations on the space that imply the non-uniqueness of X.

$$\tilde{d}_{ij} = -\frac{1}{2}\left(d_{ij} - \|\mathbf{x}_i\|_2^2 - \|\mathbf{x}_j\|_2^2\right) = \mathbf{x}_i^T \mathbf{x}_j$$

Let $\tilde{D}$ be the centered and rescaled distance matrix. Then we can represent in matrix notation:

$$\tilde{D} = X^T X$$

Since $\tilde{D}$ is symmetric, we can take its eigenvalue decomposition to get:

$$\begin{aligned} \tilde{D} &= V \Lambda V^T \\ &= V \Sigma^2 V^T \\ &= (V\Sigma)(\Sigma V^T) \\ &= (\Sigma V^T)^T (\Sigma V^T) \end{aligned}$$

Therefore,

$$X = \Sigma V^T$$

But since X is assumed to be embedded in m dimensional space, we can select the first m rows of Σ (Σ_m). So, instead

$$X = \Sigma_m V^T$$

The non-uniqueness of X can be explicitly showed by the fact that for some orthogonal matrix Q,

$$\hat{X} = QX$$

also satisfies the necessary and sufficient condition

$$D = X^T X = (QX)^T (QX) = X^T Q^T Q X = \hat{X}^T \hat{X}$$

This completes the proof. □

CMDS for dimensionality reduction

But assuming there is a k-dimensional approximately linear manifold in X, we can reconstruct it by taking the first k rows of Σ instead of the first m. We now have a simple procedure to compute the CMDS of D for $k << m$ dimensions:

1. Compute the centered, rescaled $\tilde{D}$ from the original D.

2. Take the EVD of $D = V\Sigma^2 V^T$

3. Select the first k singular values in Σ, i.e. in MATLAB notation....

$$\tilde{\Sigma} = \Sigma(1:k,:)$$

4. return $Y = \tilde{\Sigma}V^T$.

Note Y is a $k \times n$ reconstructed data matrix of n samples in k dimensions, as desired.

Dimension reduction for non-linear manifolds

PCA and MDS are simple and efficient methods of dimensionality reduction. They are guaranteed to find data structure that lie on a linear subspace of the input data that lie in the originally high-dimensional vector space. However, these methods fail when the structure takes the form of a nonlinear manifold (generalization of a surface to higher dimensions). One popular toy example is the "swiss roll". Here we introduce two techniques that attempts to address these issues: Isomap and Locally Linear Embedding (LLE).

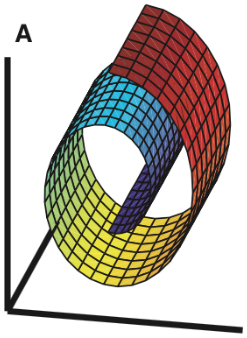

Figure 1: (From Roweis, 2000) The "swiss roll".

Isomap

Recall that Classical MDS (CMDS) finds an embedding that preserves the pairwise distances between data points, and only require a similarity "metric" matrix D as input. Isomap extends CMDS by producing D that accurately represents the metric on the possibly nonlinear manifold.

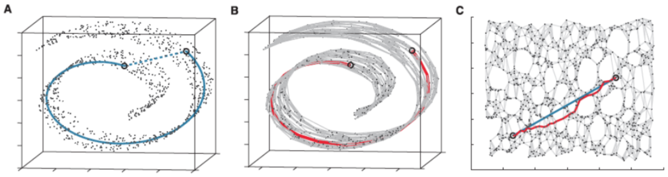

Figure 2: (From Tenenbaum, 2000) True geodesic distances between points on nonlinear manifold are approximated using Isomap.

To do so, for each point p, Isomap utilizes some neighborhood of points $\mathcal{N}(p)$, and connects p with $q \in \mathcal{N}(p)$ by an edge, with the edge cost represented by the distance between p and q, possibly by an euclidean metric. By performing this procedure for all data points, we construct a graph $G = (V, E)$. Each entry d_{ij} of D is now the length of the shortest path between vertex i and j. To compute d_{ij} for all i and j, there are efficient all-pairs shortest paths algorithms such as the Floyd-Warshall algorithm, $O(|V|^3)$. The dissimilarity/distance matrix D now becomes input to classical MDS. To summarize:

1. define some neighborhood measure $\mathcal{N}$: for each point/vertex p, for all $q \in \mathcal{N}(p)$, create edge $e(p,q)$. This forms undirected graph $G = (V, E)$.

2. compute all-pairs shortest paths on G. The output constructs the dissimilarity matrix D.

3. compute Classical MDS on D.

Locally Linear Embedding (LLE)

This method takes a different, but intuitive approach: Although the data, globally, may lie on some nonlinear manifold, a reasonable assumption is that the data, locally, is approximately linear.

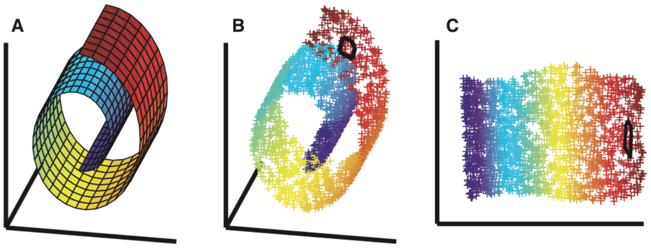

Figure 3: (From Roweis, 2000) The embedded lower-dimensional, non-linear manifold in high-dimensional space is detected with LLE.

Determining local weights

The first step of LLE is to solve the local problem. Given n data points in $\mathbb{R}^m$, let $\mathbf{x}_i$ be the i^{th} data point and $w_{j,i}$ be the scalar weight that represents the contribution of the j^{th} data point to the i^{th} reconstruction. Thus it has non-zero entries if some $\mathbf{x}_j$ is a neighbor of $\mathbf{x}_i$. Let $\mathcal{N}(i)$ be the set of neighbors to $\mathbf{x}_i$. To find the best weights, we wish to solve the following constrained, least squares problem:

$$\min_{w_{j,i}} \frac{1}{2}\|\mathbf{x}_i - \sum_{j\in\mathcal{N}(i)} w_{j,i}\mathbf{x}_j\|_2^2 \quad \text{s.t.} \quad \sum_{j\in\mathcal{N}(i)} w_{j,i} = 1$$

Global reconstruction

Given that weights are computed for each $\mathbf{x}_i$, we wish to find $\mathbf{y}_i$, the lower dimensional analog for each $\mathbf{x}_i$, by solving the following global minimization problem over all i.

$$\min_{\mathbf{y}_1,\cdots\mathbf{y}_n} \sum_{i=1}^{n} \|\mathbf{y}_i - \sum_{j\in\mathcal{N}(i)} w_{j,i}\mathbf{y}_j\|^2$$

References

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500), 2323-2326.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319-2323.

Tomasi, C. Accessed 2015. Orthogonal Matrices and the Singular Value Decomposition. https://www.cs.duke.edu/courses/fall13/compsci527/notes/svd.pdf

Accessed 2015. Other Dimension Reduction Techniques. http://www.stat.cmu.edu/ ryantibs/advmethods/notes/otherdr.pdf