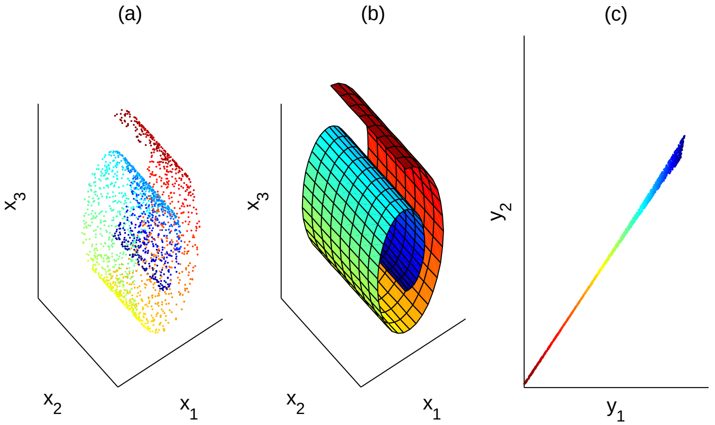
Metody redukcji wymiarowości danych Locally Linear Embedding

Piotr Lipiński

wykład z eksploracji danych

Introduction



Overview of LLE

- for each data point, determine its K nearest neighbors in the data sample
- for each data point, approximate it by a linear combination of its K nearest neighbors
- find an embedding of the data sample in a lower dimension search space such that the mapping of each data point is a linear combination of mappings of its K nearest neighbors with the same linear coefficients as in the original search space

Step 1

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbb{R}^d$.

For each $\mathbf{x}_i$, $i = 1, 2, \ldots, N$, let

$$n_1^{(i)}, n_2^{(i)}, \ldots, n_K^{(i)} \in \{1, 2, \ldots, N\} \setminus \{i\}$$

denote the indices of its K nearest neighbors in the data sample.

Step 2

Each data point $\mathbf{x}_i \in \mathcal{D}$ is approximated by a linear combination of its K nearest neighbors, i.e. the goal is to find linear coefficients

$$w_1^{(i)}, w_2^{(i)}, \ldots, w_K^{(i)} \in \mathbb{R}$$

minimizing the error function

$$||\mathbf{x}_i - \tilde{\mathbf{x}}_i||^2, \quad \text{where} \quad \tilde{\mathbf{x}}_i = \sum_{k=1}^{K} w_k^{(i)} \mathbf{x}_{n_k^{(i)}}, \tag{1}$$

under the constraint $\sum_{k=1}^{K} w_k^{(i)} = 1$.

Step 3

A mapping from the original search space to the reduced search space is constructed so that the mapping of each data point was approximated by the linear combination of mappings of its K nearest neighbors with the same linear coefficients as in the original search space, ie. the goal is to find a reduced data sample $\mathcal{R} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\} \subset \mathbb{R}^l$, where each reduced data point $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{il})^T \in \mathbb{R}^l$, for $i = 1, 2, \ldots, N$, is a data point in the reduced search space $\mathbb{R}^l$, minimizing the error function

$$\sum_{i=1}^{N} ||\mathbf{y}_i - \tilde{\mathbf{y}}_i||^2, \quad \text{where} \quad \tilde{\mathbf{y}}_i = \sum_{k=1}^{K} w_k^{(i)} \mathbf{y}_{n_k^{(i)}}, \tag{2}$$

under the constraint $d^{-1}\mathbf{Y}^T\mathbf{Y} = \mathbf{I}$, where $\mathbf{Y} \in \mathbb{R}^{l \times N}$ is the matrix with columns $\mathbf{y}_i$ and $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix (i.e. the covariance matrix of the reduced data sample is the identity matrix).

Step 1 - implementation

For each data point $\mathbf{x}_i \in \mathcal{D}$, its K nearest neighbors in the current data sample $\mathcal{D}$ are determined. Let $n_1^{(i)}, n_2^{(i)}, \ldots, n_K^{(i)}$ be the indices of the successive nearest neighbors of the data point $\mathbf{x}_i$.

Step 2 - implementation

For each data point $\mathbf{x}_i \in \mathcal{D}$, linear coefficients $w_1^{(i)}, w_2^{(i)}, \ldots, w_K^{(i)}$ are determined by minimizing the error function (1), which may be transformed, taking into consideration the constraint $\sum_{k=1}^{K} w_k^{(i)} = 1$, to

$$||\mathbf{x}_i - \tilde{\mathbf{x}}_i||^2 = ||\sum_{k=1}^{K} w_k^{(i)} \mathbf{x}_i - \sum_{k=1}^{K} w_k^{(i)} \mathbf{x}_{n_k^{(i)}}||^2 = \tag{3}$$

$$= ||\sum_{k=1}^{K} w_k^{(i)} (\mathbf{x}_i - \mathbf{x}_{n_k^{(i)}})||^2 = ||\sum_{k=1}^{K} w_k^{(i)} \mathbf{z}_k||^2, \tag{4}$$

where $\mathbf{z}_k = \mathbf{x}_{n_k^{(i)}} - \mathbf{x}_i$.

Step 2 - implementation, continued

Defining the matrix $\mathbf{Z} \in \mathbb{R}^{d \times K}$ as the matrix with columns $\mathbf{z}_k$ and the vector $\mathbf{w} \in \mathbb{R}^K$ as the vector with coordinates $w_k^{(i)}$ (certainly, the matrix $\mathbf{Z}$ and the vector $\mathbf{w}$ depends on i, but we omit it here for the sake of simplicity of the notation), the error function (1) becomes

$$||\sum_{k=1}^{K} w_k^{(i)} \mathbf{z}_k||^2 = ||\mathbf{Z}\mathbf{w}||^2 = (\mathbf{Z}\mathbf{w})^T(\mathbf{Z}\mathbf{w}) = \mathbf{w}^T\mathbf{Z}^T\mathbf{Z}\mathbf{w} = \mathbf{w}^T\mathbf{V}\mathbf{w}, \quad (5)$$

where $\mathbf{V} = \mathbf{Z}^T\mathbf{Z} \in \mathbb{R}^{d \times d}$. Moreover, the constraint $\sum_{k=1}^{K} w_k^{(i)} = 1$ may be transformed to $\mathbf{1}^T\mathbf{w} = 1$, where $\mathbf{1} \in \mathbb{R}^K$ is the vector of ones.

Step 2 - implementation, continued

Therefore, in order to minimize the error function (1), the Lagrange multiplier method may be used, i.e. the following equation system, with the Lagrange multiplier λ, must be solved:

$$\frac{\partial \mathbf{w}^T \mathbf{V} \mathbf{w}}{\partial w_i} = \lambda \frac{\partial \mathbf{1}^T \mathbf{w}}{\partial w_i}, \tag{6}$$

for each $i = 1, 2, \ldots, K$ with the constraint $\mathbf{1}^T \mathbf{w} = 1$.
Since

$$\nabla \mathbf{w}^T \mathbf{V} \mathbf{w} = 2\mathbf{V}\mathbf{w}, \quad \text{and} \quad \nabla \mathbf{1}^T \mathbf{w} = \mathbf{1}^T, \tag{7}$$

the equation system (6) is equivalent to the matrix equation

$$2\mathbf{V}\mathbf{w} = \lambda \mathbf{1}^T, \tag{8}$$

thus, if $\mathbf{V}$ is invertible,

$$\mathbf{w} = \frac{\lambda}{2} \mathbf{V}^{-1} \mathbf{1}^T, \tag{9}$$

and λ must be adjusted to fulfil the constraint $\mathbf{1}^T \mathbf{w} = 1$. If $\mathbf{V}$ is not invertible, the error function should be modified by some regularization component and minimized in a similar way.

Step 3 - implementation

Reduced data points $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N \in \mathbb{R}^l$ are determined by minimizing the error function (2).

Assume at the beginning that $l = 1$ and then each $\mathbf{y}_i$ is just a real number. Thus, the error function (2) may be transformed to

$$\sum_{i=1}^{N} ||\mathbf{y}_i - \tilde{\mathbf{y}}_i||^2 = \sum_{i=1}^{N} ||\mathbf{y}_i - \sum_{k=1}^{K} w_k^{(i)} \mathbf{y}_{n_k^{(i)}}||^2 =$$

$$= \sum_{i=1}^{N} (\mathbf{y}_i^2 - \mathbf{y}_i \sum_{k=1}^{K} w_k^{(i)} \mathbf{y}_{n_k^{(i)}} - \sum_{k=1}^{K} w_k^{(i)} \mathbf{y}_{n_k^{(i)}} \mathbf{y}_i + (\sum_{k=1}^{K} w_k^{(i)} \mathbf{y}_{n_k^{(i)}})^2) =$$

$$= \sum_{i=1}^{N} (\mathbf{y}_i^2) - \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{y}_i w_k^{(i)} \mathbf{y}_{n_k^{(i)}} - \sum_{i=1}^{N} \sum_{k=1}^{K} w_k^{(i)} \mathbf{y}_{n_k^{(i)}} \mathbf{y}_i + \sum_{i=1}^{N} (\sum_{k=1}^{K} w_k^{(i)} \mathbf{y}_{n_k^{(i)}})^2) =$$

Step 3 - implementation, continued

$$= \sum_{i=1}^{N}(\mathbf{y}_i^2) - \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{y}_i w_k^{(i)}\mathbf{y}_{n_k^{(i)}} - \sum_{i=1}^{N}\sum_{k=1}^{K} w_k^{(i)}\mathbf{y}_{n_k^{(i)}}\mathbf{y}_i + \sum_{i=1}^{N}(\sum_{k=1}^{K} w_k^{(i)}\mathbf{y}_{n_k^{(i)}})^2) =$$

$$= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T(\mathbf{WY}) - (\mathbf{WY})^T\mathbf{Y} + (\mathbf{WY})^T(\mathbf{WY}) =$$

$$= \mathbf{Y}^T(\mathbf{I}-\mathbf{W})\mathbf{Y} - (\mathbf{WY})^T(\mathbf{I}-\mathbf{W})\mathbf{Y} =$$

$$= (\mathbf{Y}^T - (\mathbf{WY})^T)(\mathbf{I}-\mathbf{W})\mathbf{Y} = \mathbf{Y}^T(\mathbf{I}-\mathbf{W})^T(\mathbf{I}-\mathbf{W})\mathbf{Y} = \mathbf{Y}^T\mathbf{MY},$$

where $\mathbf{Y} \in \mathbf{R}^{l\times N}$ is the matrix with columns $\mathbf{y}_i$, $\mathbf{W} \in \mathbf{R}^{N\times N}$ is the matrix with elements $w_{ij} = w_k^{(i)}$ if $\mathbf{x}_j$ is the k-th nearest neighbor of $\mathbf{x}_i$ and $w_{ij} = 0$ otherwise, $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix, and $\mathbf{M} = (\mathbf{I}-\mathbf{w})^T(\mathbf{I}-\mathbf{w}) \in \mathbb{R}^{N\times N}$.

Step 3 - implementation, continued

Therefore, in order to minimize the error function(2) under the constraint $d^{-1}\mathbf{Y}^T\mathbf{Y} = \mathbf{I}$, the Lagrange multiplier method may be used, i.e. the following equation system, with the Lagrange multiplier λ, must be solved:

$$\frac{\partial \mathbf{Y}^T\mathbf{M}\mathbf{Y}}{\partial \mathbf{y}_i} = \lambda \frac{\partial d^{-1}\mathbf{Y}^T\mathbf{Y}}{\partial \mathbf{y}_i}, \tag{10}$$

for each $i = 1, 2, \ldots, N$ with the constraint $d^{-1}\mathbf{Y}^T\mathbf{Y} = \mathbf{I}$.
Since

$$\nabla \mathbf{Y}^T\mathbf{M}\mathbf{Y} = 2\mathbf{M}\mathbf{Y} \quad \text{and} \quad \nabla d^{-1}\mathbf{Y}^T\mathbf{Y} = 2d^{-1}\mathbf{Y}, \tag{11}$$

the equation system (10) is equivalent to the matrix equation

$$\mathbf{M}\mathbf{Y} = \lambda d^{-1}\mathbf{Y}, \tag{12}$$

thus, $\mathbf{Y}$ is an eigenvector of the matrix $\mathbf{M}$. As the error function (2) is being minimized, the eigenvector $\mathbf{Y}$ should correspond to the smallest non-zero eigenvalue of the matrix $\mathbf{M}$. In order to generalize the calculation for $l > 1$, the successive eigenvectors of the matrix $\mathbf{M}$ should be taken to determine the successive coordinates of the mappings $\mathbf{y}_i$.