Dimensionality Reduction

Jianlin Cheng

Outline

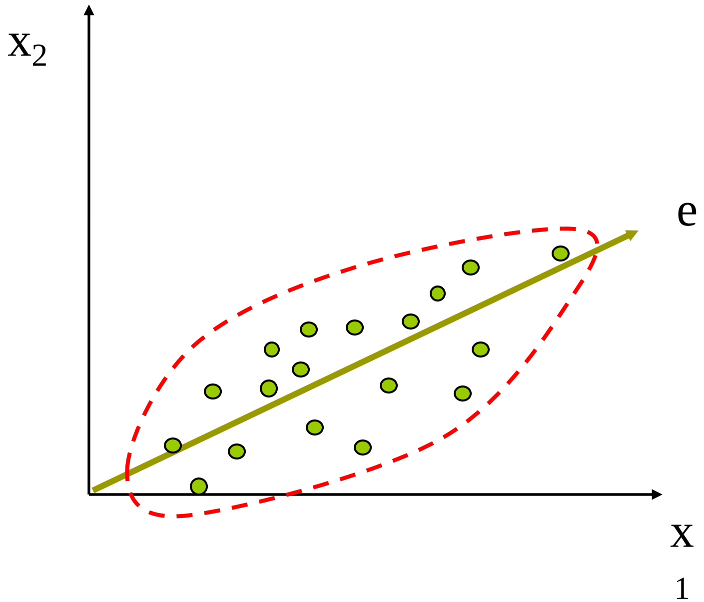
- Principal Component Analysis (PCA)
- Multidimensional Scaling (MDS)
- Isomap
- Kernel PCA
- Locally Linear Embedding (LLE)
- Laplacian Eigenmaps (LEM)
- Sammon Mapping
- Mutilayer Autoencoder
- Other Nonconvex methods

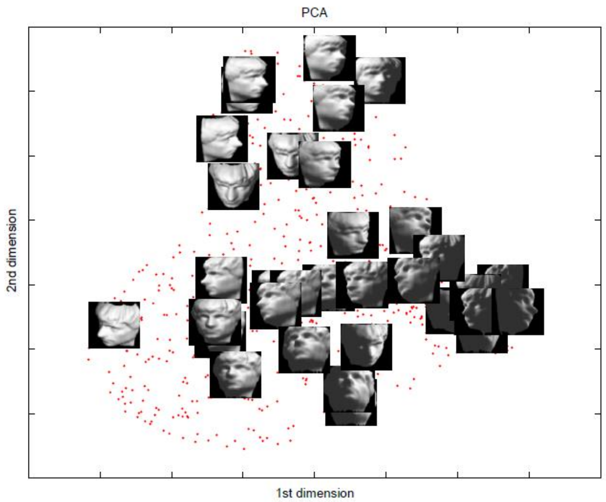
Models

- Linear methods
  - Principal component analysis (PCA)
  - Multidimensional scaling (MDS)
  - Independent component analysis (ICA)
- Nonlinear methods
  - Kernel PCA
  - Locally linear embedding (LLE)
  - Laplacian eigenmaps (LEM)

Principal Component Analysis (PCA)

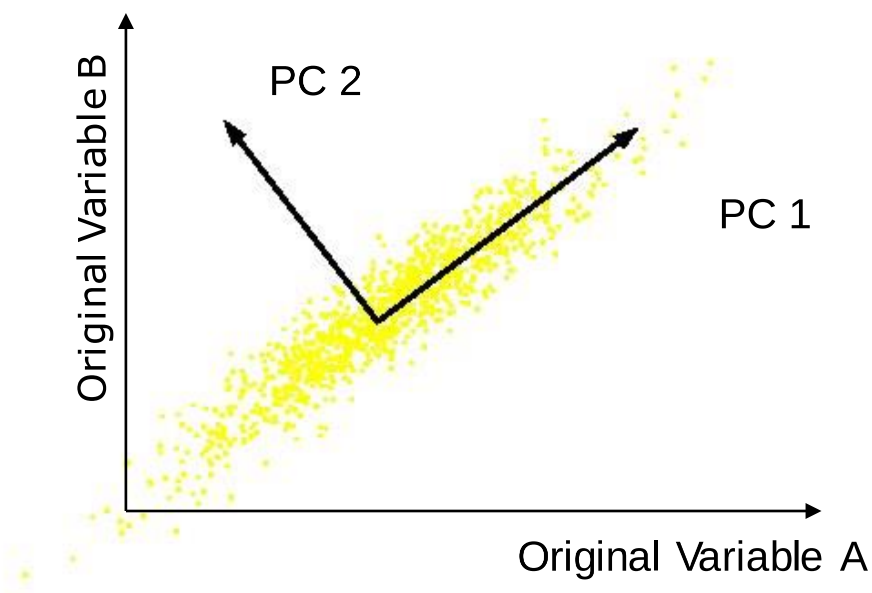
- History: Karl Pearson, 1901
- Find projections that capture the largest amounts of variation in data
- Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space





PCA

- Definition: Given a set of data $X \in R^{d \times N}$, find the principal axes are those orthonormal axes onto which the variance retained under projection is maximal



Formulation

- Variance on the first dimension
  - $\mathrm{var}(\boldsymbol{U}_1)=\mathrm{var}(\mathbf{w}^T X)=\mathbf{w}^T\mathbf{S}\mathbf{w}$
  - $\mathbf{S}$: covariance matrix of X
- Objective: the variance retains the maximal
- Formulation

$$\max_{\mathbf{w}} \quad \mathbf{w}^T\mathbf{S}\mathbf{w}$$
$$\text{s.t.} \quad \mathbf{w}^T\mathbf{w}=1$$

- Solving procedure
  - Construct Langrangian $L(\mathbf{w},\lambda_1)=\mathbf{w}^T\mathbf{S}\mathbf{w}-\lambda_1(\mathbf{w}^T\mathbf{w}-1)$
  - Set the partial derivative on to zero

$$\frac{\partial L}{\partial \mathbf{w}}=0 \Rightarrow \mathbf{S}\mathbf{w}=\lambda_1\mathbf{w}$$

  - As **w ≠ 0** then **w** must be an eigenvector of **S** with eigenvalue λ_1

$$\mathbf{w}^T\mathbf{S}\mathbf{w}=\lambda_1\mathbf{w}^T\mathbf{w}=\lambda_1$$

How to choose k

Choosing k (number of principal components)

Average squared projection error: $\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)} - x_{approx}^{(i)}\|^2$

Total variation in the data: $\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2$

Typically, choose k to be smallest value so that

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2} \leq 0.01 \qquad (1\%)$$

"99% of variance is retained"

Multidimensional Scaling (MDS)

- History: T. Cox and M. Cox, 2001
- Attempts to preserve pairwise distances
- Different formulation of PCA, but yields similar result form

$$\min_{Y} \sum_{i=1}^{N} \sum_{j=1}^{N} (d_{ij}^{(X)} - d_{ij}^{(Y)})^2$$

where $d_{ij}^{(X)} = \|x_i - x_j\|^2$ and $d_{ij}^{(Y)} = \|y_i - y_j\|^2$.

- Transformation

$$\min_{Y} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i^\top x_j - y_i^\top y_j)^2$$

Isomap

- History: J. Tenenbaum et al, Science 1998
- A nonlinear generalization of classical MDS
- Perform MDS, not in the original space, but in the geodesic space
- Procedure-similar to LLE
  1. Find neighbors of each data point
  2. Compute geodesic pairwise distances (e.g., shortest path) between all points
  3. Embed the data via MDS

Kernel PCA

- History: S. Mika et al, NIPS, 1999
- Data may lie on or near a nonlinear manifold, not a linear subspace
- Find principal components that are nonlinearly to the input space via nonlinear mapping

$$\Phi : x \to \mathcal{H} \qquad x \mapsto \Phi(x)$$

- Objective $\min_{U_k} \sum_{i=1}^{N} \left\| \Phi(\mathbf{x}_i) - U_k U_k^T \Phi(\mathbf{x}_i) \right\|^2$

- Solution found by SVD: $\Phi(X) = U \Sigma V^T$

  U contains the eigenvectors of $\Phi(X)\Phi(X)^T$

Locally Linear Embedding (LLE)

- History: S. Roweis and L. Saul, Science, 2000
- Procedure
  1. Identify the neighbors of each data point
  2. Compute weights that best linearly reconstruct the point from its neighbors

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \|\mathbf{x}_i - \sum_{j=1}^{k} w_{ij} \mathbf{x}_{N_i(j)}\|^2$$

  3. Find the low-dimensional embedding vector which is best reconstructed by the weights determined in Step 2

$$\min_{Y} \sum_{i=1}^{N} \|\mathbf{y}_i - \sum_{j=1}^{k} w_{ij} \mathbf{y}_{N_i(j)}\|^2$$

Laplacian Eigenmaps (LEM)

- History: M. Belkin and P. Niyogi, 2003
- Similar to locally linear embedding
- Different in weights setting and objective function
  - Weights $$W_{ij} = \begin{cases} 1 & i, j \text{ are connected} \\ \exp\left(\frac{-\|x_i - x_j\|^2}{s}\right) & \text{otherwise} \end{cases}$$
  - Objective

$$\min_{Y} \sum_{i=1}^{N} \sum_{j=1}^{N} (\mathbf{y}_i - \mathbf{y}_j) W_{ij}$$

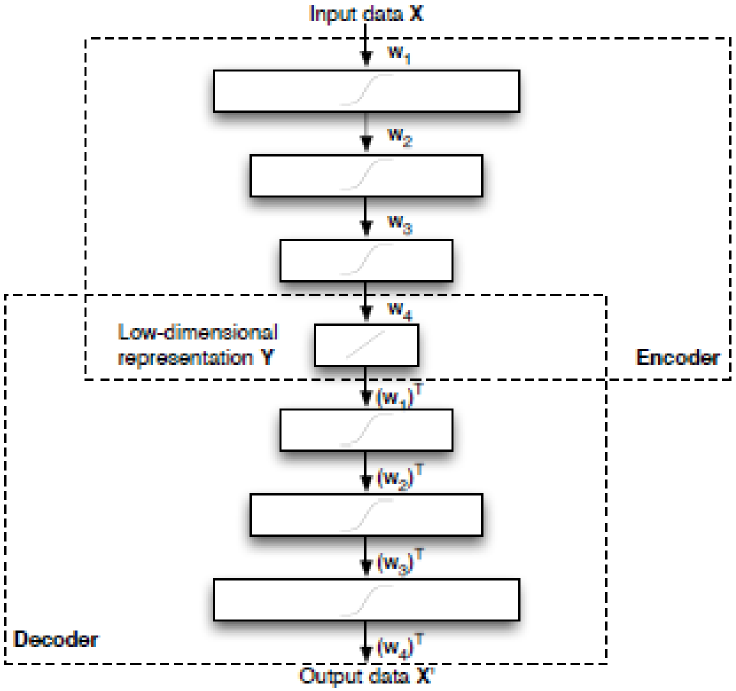
Sammon's Mapping

- A variation on MDS
- Preserve the distances between nearby points than between those which are far a part

- formulation

$$\phi(\mathbf{Y}) = \frac{1}{\sum_{ij} d_{ij}} \sum_{i \neq j} \frac{(d_{ij} - \|\mathbf{y}_i - \mathbf{y}_j\|)^2}{d_{ij}}$$

Multilayer Autoencoders



Summary

--nonlinear techniques for dimensionality reduction are, despite their large variance, often not capable of outperforming traditional linear techniques such as PCA

--new nonlinear techniques for dimensionality reduction required

(i) do not suffer from the presence of trivial optimal solutions

(ii) may be based on non-convex objective functions

(iii) do not rely on neighbourhood graphs to model the local structure of the data manifold