Distance Metric Learning in Data Mining (Part II)

Fei Wang and Jimeng Sun

IBM TJ Watson Research Center

Outline

Part I - Applications

- Motivation and Introduction
- Patient similarity application

Part II - Methods

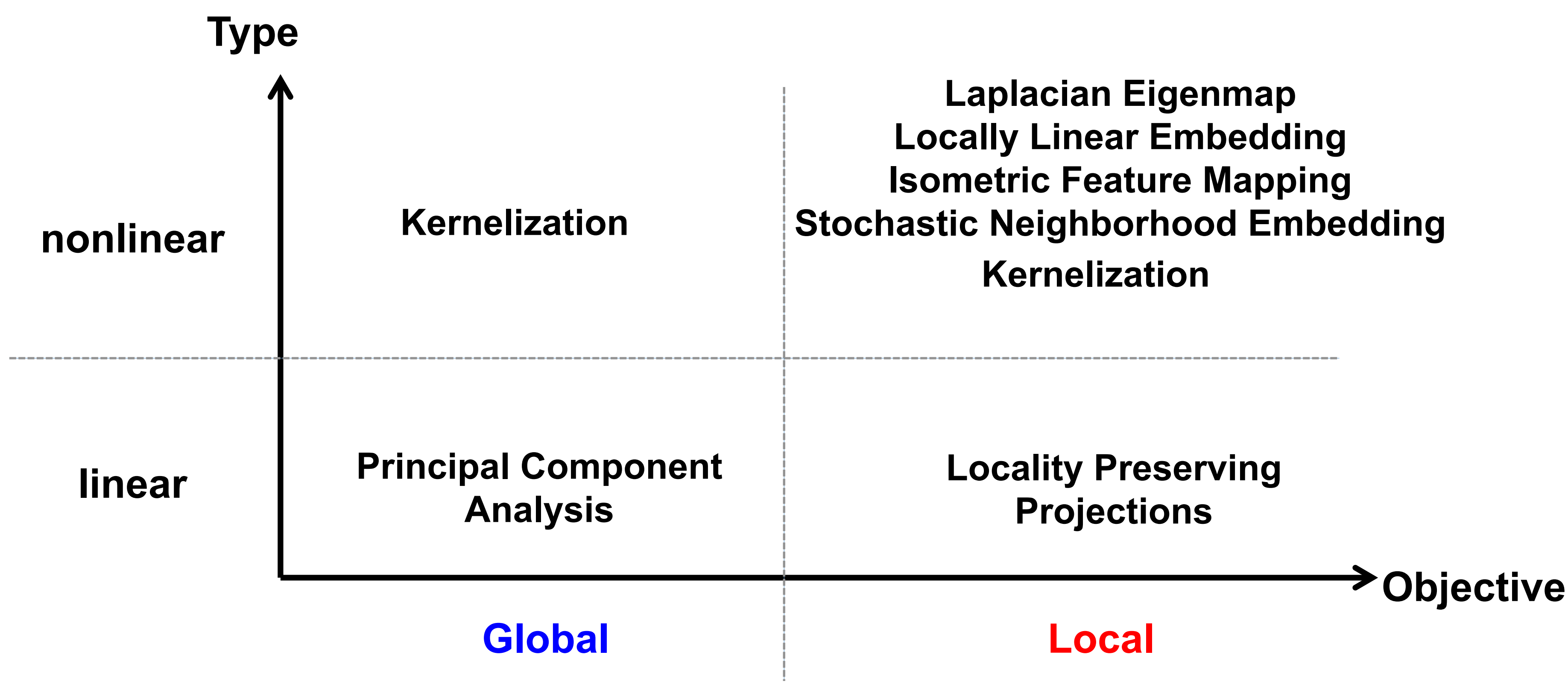
- Supervised Metric Learning
- Unsupervised Metric Learning
- Semi-supervised Metric Learning
- Challenges and Opportunities for Metric Learning

Metric Learning Meta-algorithm

- Embed the data in some space
  - Usually formulate as an optimization problem
    - Define objective function
      - Separation based
      - Geometry based
      - Information theoretic based
    - Solve optimization problem
      - Eigenvalue decomposition
      - Semi-definite programming
      - Gradient descent
      - Bregman projection
- Euclidean distance on the embedded space

Unsupervised metric learning

- Learning pairwise distance metric purely based on the data, i.e., without any supervision



Outline

Part I - Applications

- Motivation and Introduction
- Patient similarity application

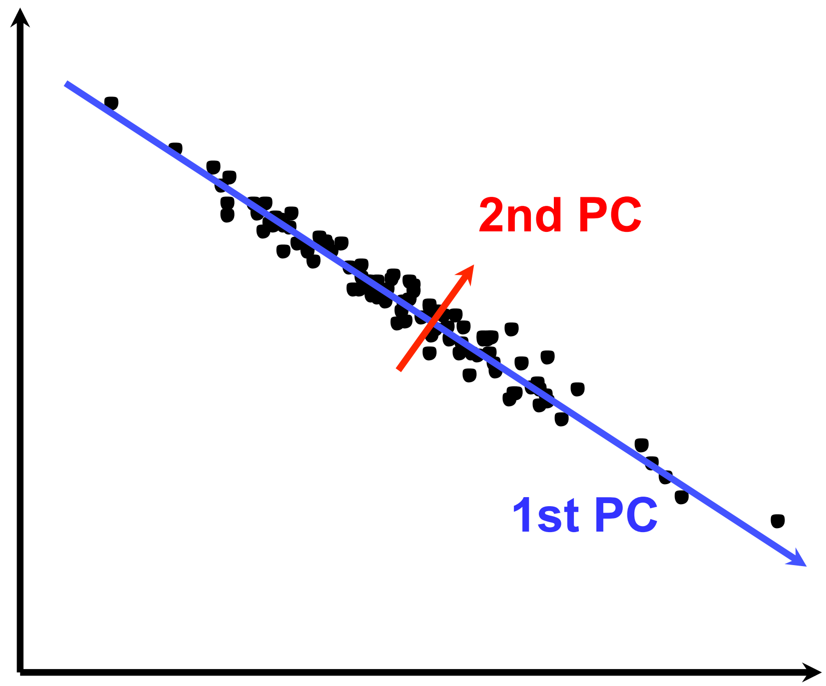
Part II - Methods

- <u>Unsupervised Metric Learning</u>
- Supervised Metric Learning
- Semi-supervised Metric Learning
- Challenges and Opportunities for Metric Learning

Principal Component Analysis (PCA)

- Find successive projection directions which maximize the data variance

$$\max_{\mathbf{W}} \; tr\left(\mathbf{W}^\top \bar{\mathbf{X}} \bar{\mathbf{X}}^\top \mathbf{W}\right)$$
$$s.t. \quad \mathbf{W}^\top \mathbf{W} = \mathbf{I}$$



Geometry based objective
Eigenvalue decomposition
Linear mapping
Global method

The pairwise Euclidean distances will not change after PCA

Kernel Mapping

Map the data into some high-dimensional feature space, such that the nonlinear problem in the original space becomes the linear problem in the high-dimensional feature space. We do not need to know the explicit form of the mapping, but we need to define a kernel function.

$$\phi : (x_1, x_2) \longrightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

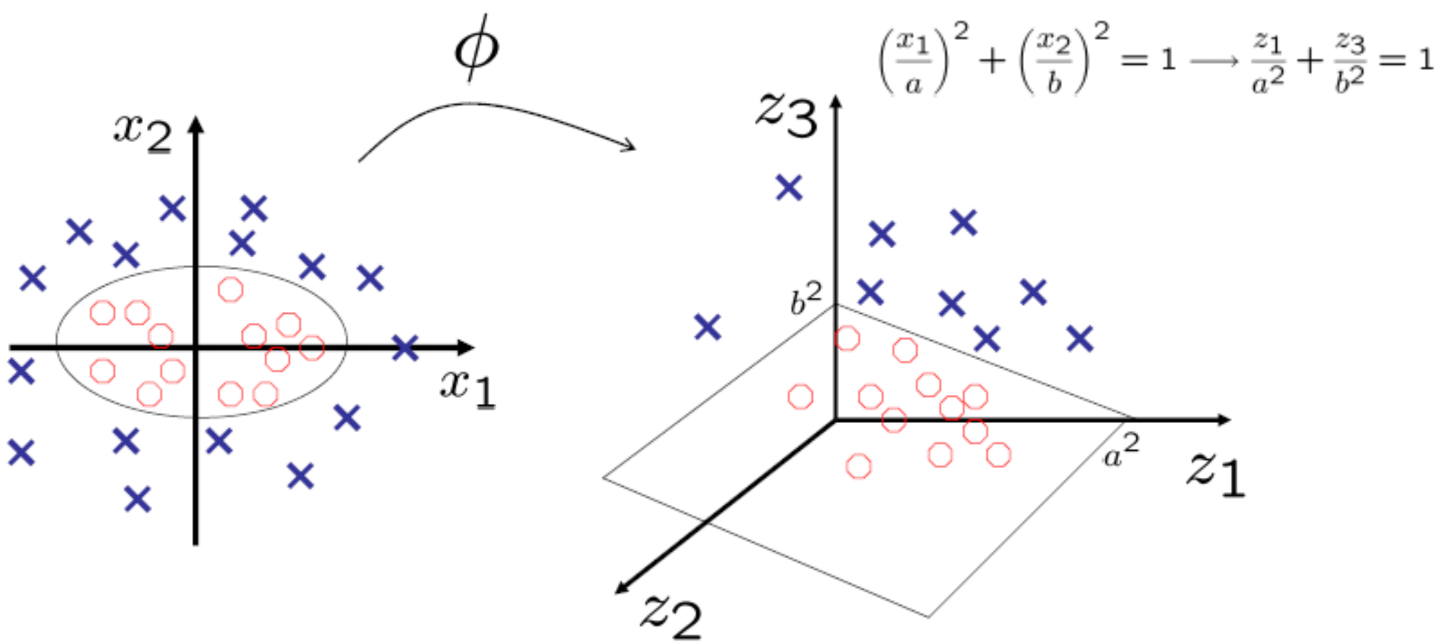


Figure from http://omega.albany.edu:8008/machine-learning-dir/notes-dir/ker1/ker1.pdf

Kernel Principal Component Analysis (KPCA)

Perform PCA in the feature space

$$\boxed{\mathbf{C}\mathbf{v} = \frac{1}{n-1}\bar{\mathbf{\Phi}}\bar{\mathbf{\Phi}}^\top\mathbf{v} = \lambda\mathbf{v}}$$

$$\bar{\mathbf{\Phi}} = \left[\phi(\mathbf{x}_1) - \bar{\phi}(\mathbf{x}), \phi(\mathbf{x}_2) - \bar{\phi}(\mathbf{x}), \cdots, \phi(\mathbf{x}_n) - \bar{\phi}(\mathbf{x})\right]$$

With the representer theorem $\mathbf{v} = \bar{\mathbf{\Phi}}\boldsymbol{\alpha}$

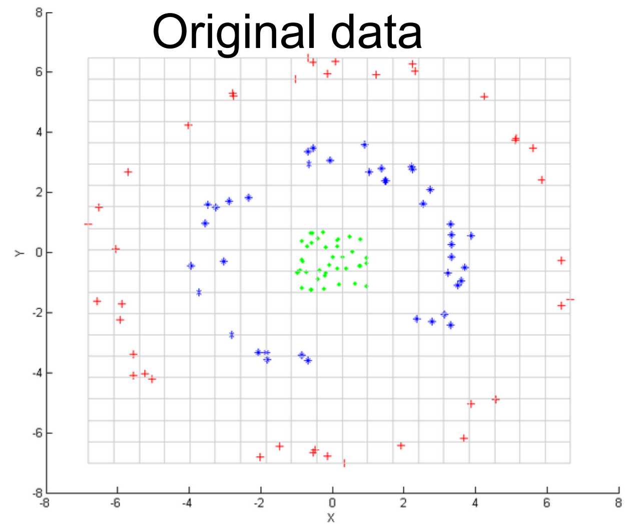
$$\boxed{\frac{1}{n-1}\bar{\mathbf{K}}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}}$$

$$\bar{\mathbf{K}}_{ij} = \langle\phi(\mathbf{x}_i) - \bar{\phi}(\mathbf{x}), \phi(\mathbf{x}_j) - \bar{\phi}(\mathbf{x})\rangle = \langle\bar{\phi}(\mathbf{x}_i), \bar{\phi}(\mathbf{x}_j)\rangle$$

$$\bar{\phi}(\mathbf{x}_i)^\top\mathbf{v} = \sum_{u=1}^{n}\alpha_u\langle\bar{\phi}(\mathbf{x}_i), \bar{\phi}(\mathbf{x}_u)\rangle = \bar{\mathbf{K}}_{i\cdot}\boldsymbol{\alpha}$$

Geometry based objective
Eigenvalue decomposition
Nonlinear mapping
Global method

Original data



Embedded data after KPCA

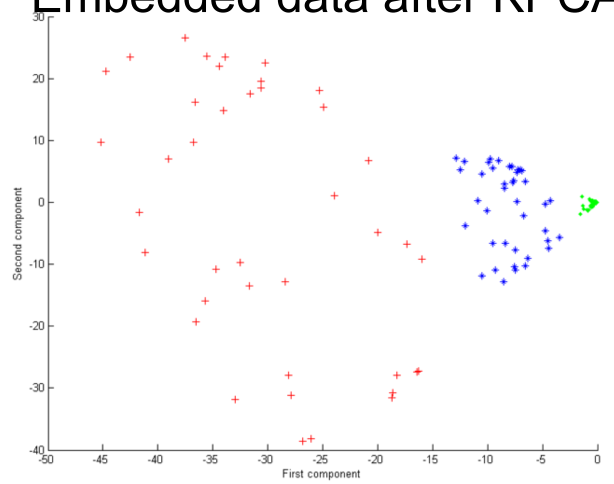


Figure from http://en.wikipedia.org/wiki/Kernel_principal_component_analysis

Locality Preserving Projections (LPP)

Find linear projections that can preserve the localities of the data set

$$\boxed{\omega_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)}$$

$$\min_{\mathbf{W}^\top \mathbf{X}\mathbf{D}\mathbf{X}^\top \mathbf{W} = \mathbf{I}} \boxed{\sum_{ij} \|\mathbf{W}^\top \mathbf{x}_i - \mathbf{W}^\top \mathbf{x}_j\|^2 \omega_{ij}}$$

$$= tr\left(\mathbf{W}^\top \mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{W}\right)$$

Degree Matrix

$$\mathbf{D}_{ii} = \sum_j \omega_{ij}$$

Laplacian Matrix

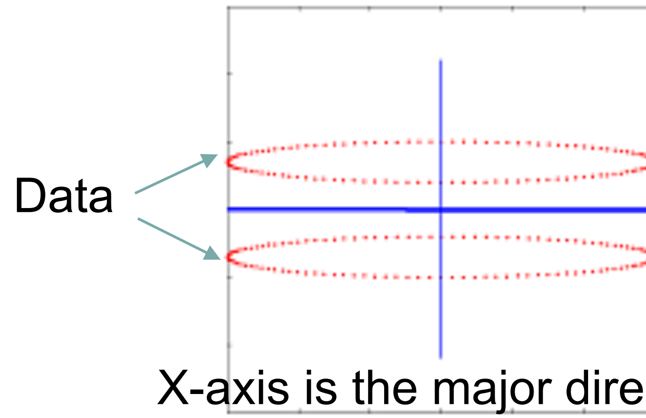
$$\mathbf{L} = \mathbf{D} - \mathbf{\Omega}$$

$$\mathbf{\Omega}_{ij} = \omega_{ij}$$

$$\boxed{\mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{w} = \lambda \mathbf{X}\mathbf{D}\mathbf{X}^\top \mathbf{w}}$$

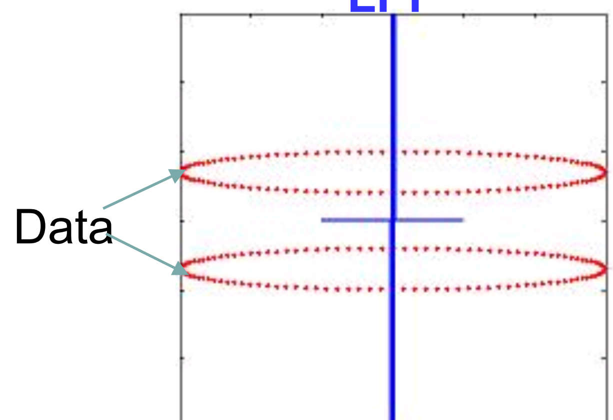
Geometry based objective
Generalized eigenvalue decomposition
Linear mapping
Local method

PCA



X-axis is the major directi

LPP



Y-axis is the major projection direction

X. He and P. Niyogi. Locality Preserving Projections. NIPS 2003.

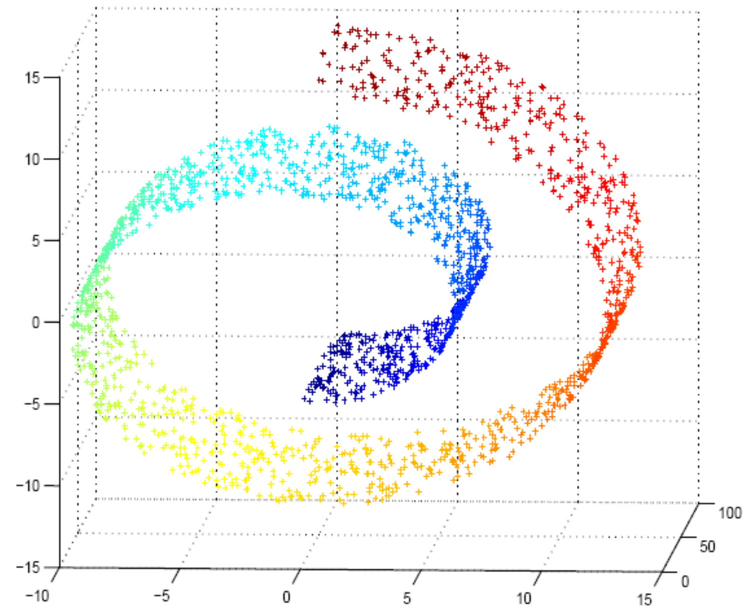
Laplacian Embedding (LE)

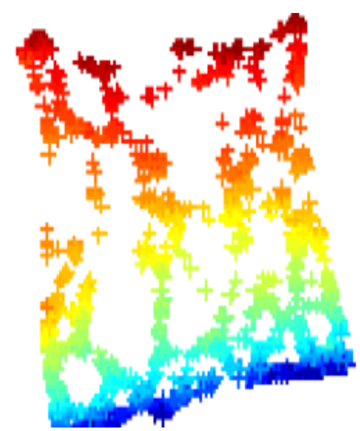
LE: Find embeddings that can preserve the localities of the data set

$$\min_{\mathbf{y}^\top \mathbf{D}\mathbf{y}=1} \sum_{i=1}^{n} (y_i - y_j)^2 \omega_{ij} = \mathbf{y}^\top \mathbf{L}\mathbf{y}$$

The embedded Xi

The relationship between Xi and Xj





Geometry based objective
Generalized eigenvalue decomposition
Nonlinear mapping
Local method

M. Belkin, P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation, June 2003; 15 (6):1373-1396

Locally Linear Embedding (LLE)

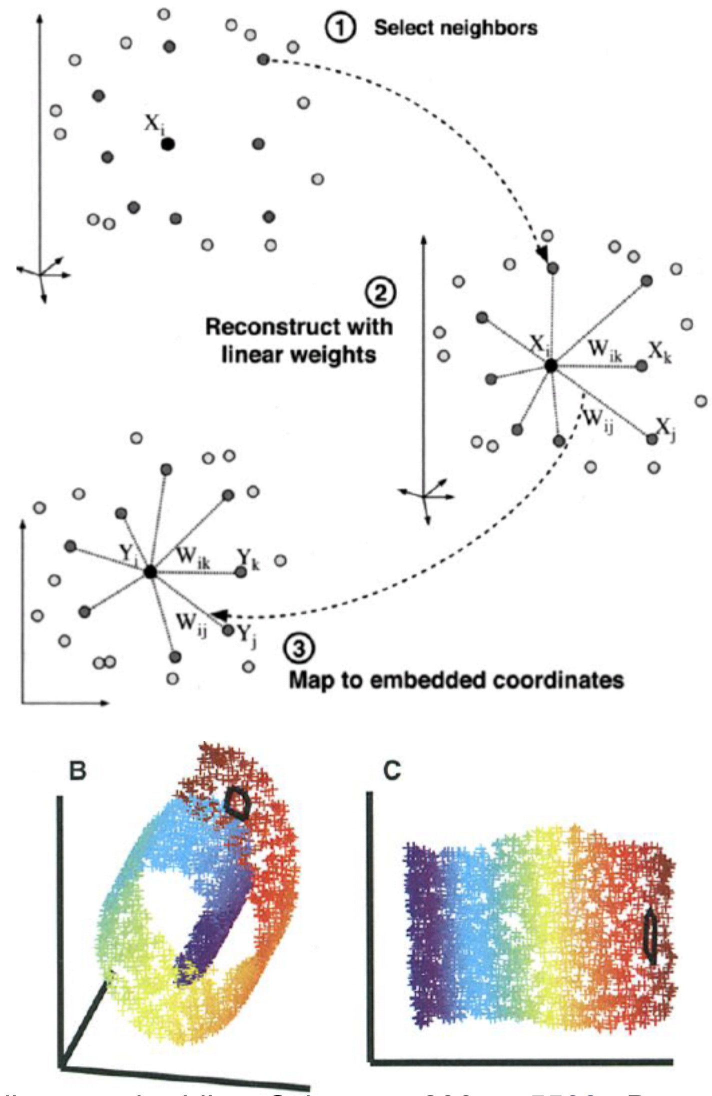
Obtain data relationships

$$\min_{\omega_{ij}} \sum_i \left\| \mathbf{x}_i - \sum_{\mathbf{x}_j \in \mathcal{N}_i} \omega_{ij} \mathbf{x}_j \right\|^2$$

Obtain data embeddings

$$\min_{\{\mathbf{y}_i\}_{i=1}^n} \sum_i \left\| \mathbf{y}_i - \sum_{\mathbf{x}_j \in \mathcal{N}_i} \omega_{ij} \mathbf{y}_j \right\|^2$$

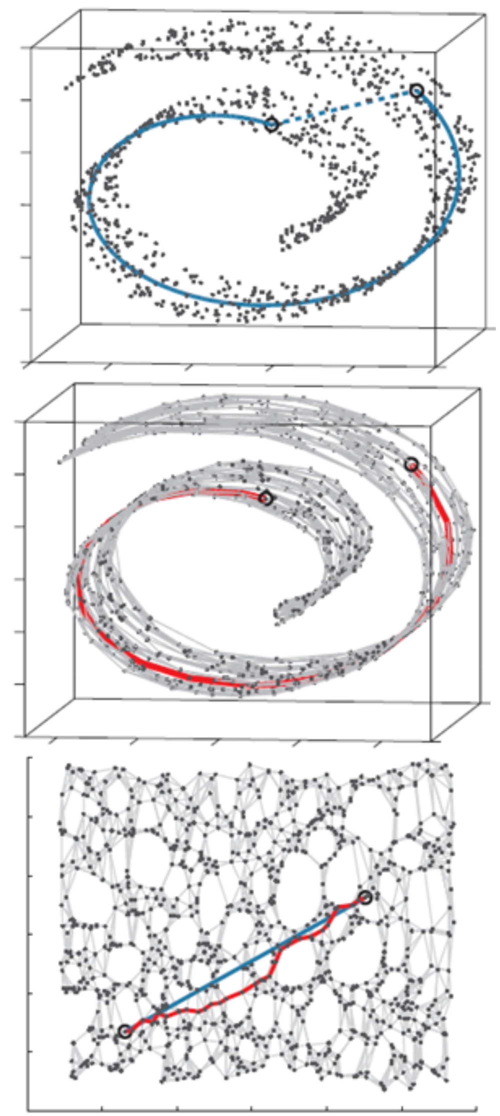
Geometry based objective
Generalized eigenvalue decomposition
Nonlinear mapping
Local method



Sam Roweis & Lawrence Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, v.290 no.5500 , Dec. 22, 2000. pp.2323—2326.

Isometric Feature Mapping (ISOMAP)

1. **Construct the neighborhood graph**
2. **Compute the shortest path length (geodesic distance) between pairwise data points**
3. **Recover the low-dimensional embeddings of the data by Multi-Dimensional Scaling (MDS) with preserving those geodesic distances**



Geometry based objective
Eigenvalue decomposition
Nonlinear mapping
Local method

J. B. Tenenbaum, V. de Silva and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 2000.

Stochastic Neighbor Embedding (SNE)

The probability that i picks j as its neighbor

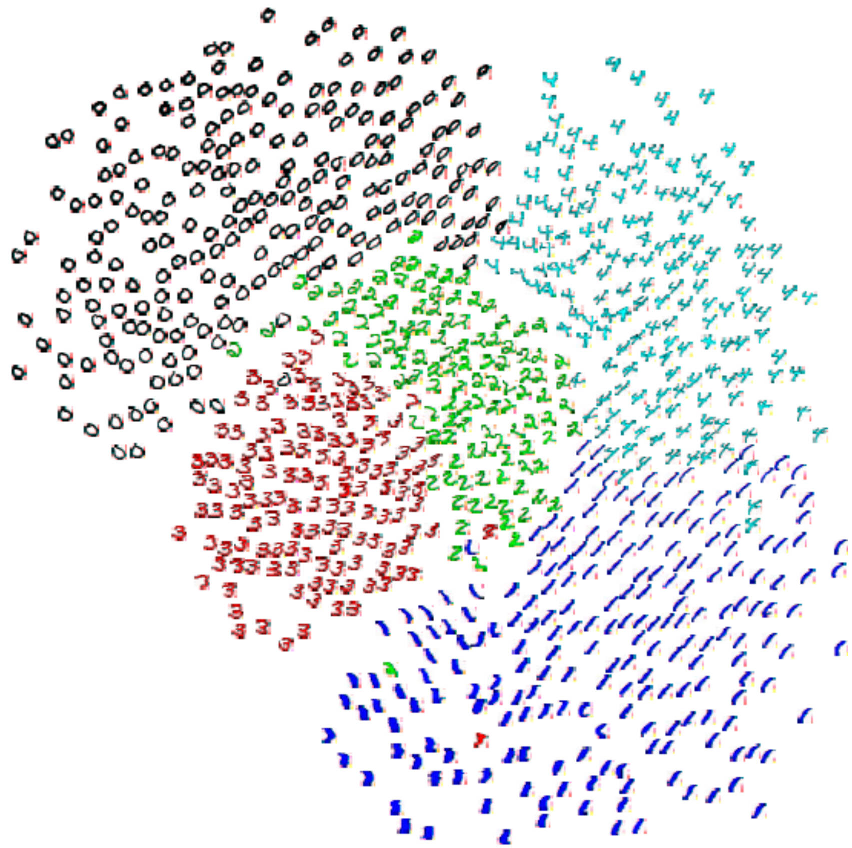
$$p_{ij} = \frac{\exp\left(-d_{ij}^2\right)}{\sum_{k \neq i} \exp\left(-d_{ik}^2\right)}$$

Neighborhood distribution in the embedded space

$$q_{ij} = \frac{\exp\left(-\left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\mathbf{y}_i - \mathbf{y}_k\right\|^2\right)}$$

Neighborhood distribution preservation

$$\mathcal{J} = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}} = \sum_i KL(P_i \| Q_i)$$



Information theoretic objective
Gradient descent
Nonlinear mapping
Local method

Hinton, G. E. and Roweis, S. Stochastic Neighbor Embedding. NIPS 2003.

Summary: Unsupervised Distance Metric Learning

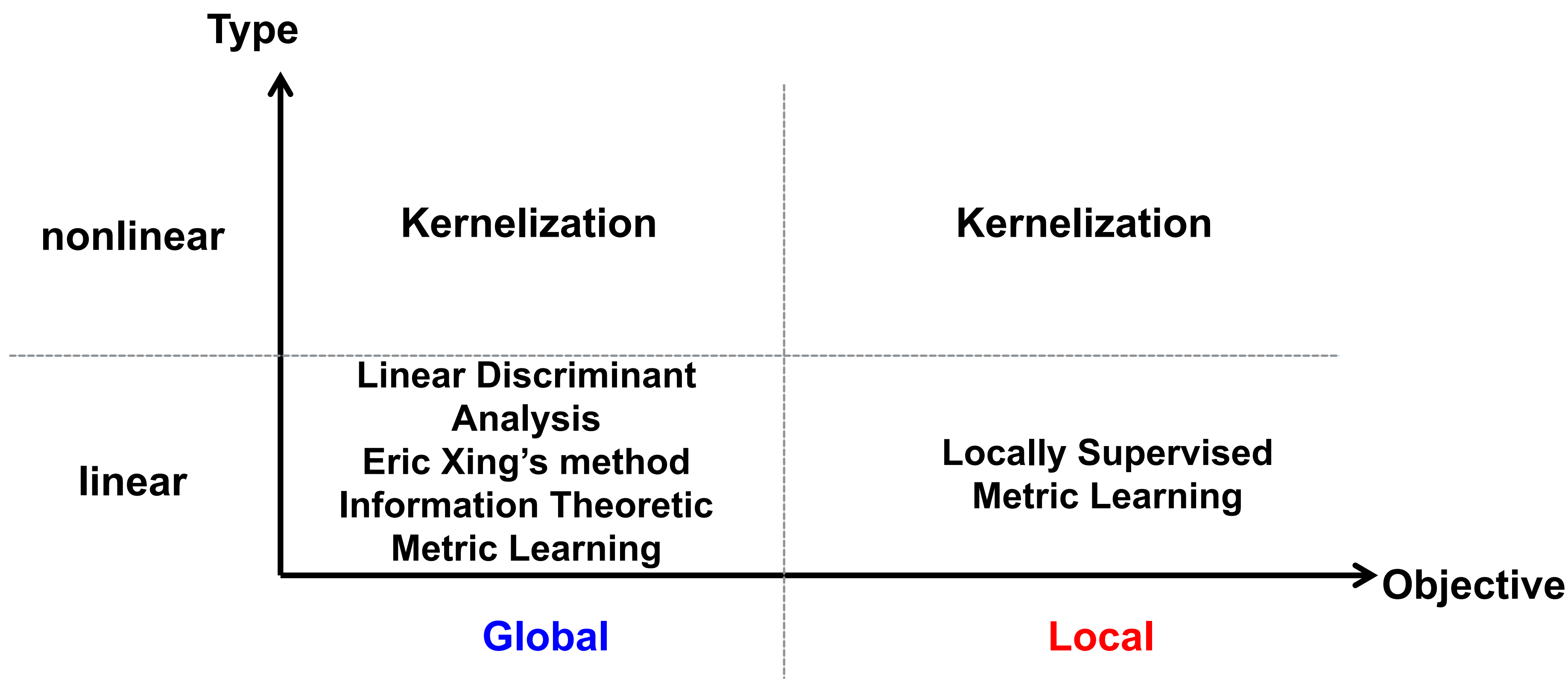
	PCA	LPP	LLE	Isomap	SNE
Local		√	√	√	√
Global	√				
Linear	√	√			
Nonlinear			√	√	√
Separation					
Geometry	√	√	√	√	
Information theoretic					√
Extensibility	√	√			

Outline

- Unsupervised Metric Learning
- <u>Supervised Metric Learning</u>
- Semi-supervised Metric Learning
- Challenges and Opportunities for Metric Learning

Supervised Metric Learning

- Learning pairwise distance metric with data and their supervision, i.e., data labels and pairwise constraints (must-links and cannot-links)



Linear Discriminant Analysis (LDA)

Suppose the data are from C different classes

$$\Sigma_{\mathcal{C}} = \frac{1}{C}\sum_{c}\frac{1}{n_c}\sum_{\mathbf{x}_i\in c}(\mathbf{x}_i-\bar{\mathbf{x}}_c)(\mathbf{x}_i-\bar{\mathbf{x}}_c)^\top$$

$$\Sigma_{\mathcal{S}} = \frac{1}{C}\sum_{c}(\bar{\mathbf{x}}_c-\bar{\mathbf{x}})(\bar{\mathbf{x}}_c-\bar{\mathbf{x}})^\top$$

$$\min_{\mathbf{W}^\top\mathbf{W}=\mathbf{I}} \frac{tr(\mathbf{W}^\top\Sigma_{\mathcal{C}}\mathbf{W})}{tr(\mathbf{W}^\top\Sigma_{\mathcal{S}}\mathbf{W})}$$

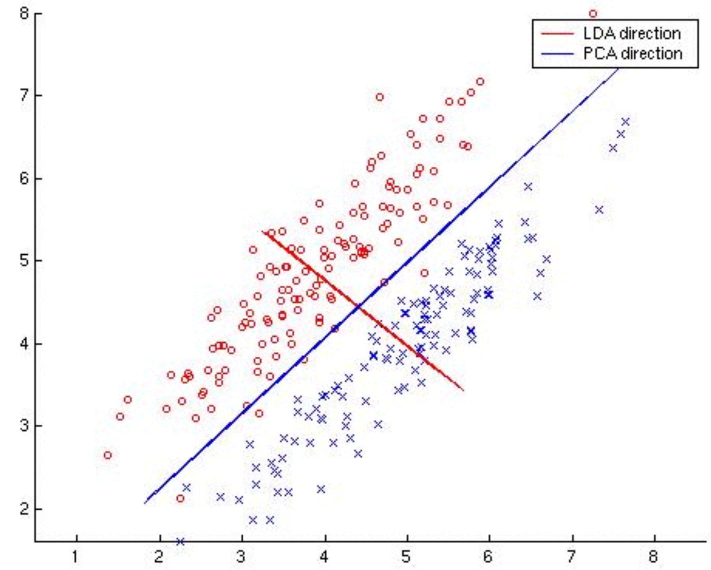


Figure from http://cmp.felk.cvut.cz/cmp/software/stprtool/examples/ldapca_example1.gif

Separation based objective
Eigenvalue decomposition
Linear mapping
Global method

Keinosuke Fukunaga. Introduction to statistical pattern recognition. Academic Press Professional, Inc. 1990 .
Y. Jia, F. Nie, C. Zhang. Trace Ratio Problem Revisited. IEEE Trans. Neural Networks, 20:4, 2009.

Distance Metric Learning with Side Information (Eric Xing's method)

Must-link constraint set: $\mathcal{M}$
each element is a pair of data that come from the same class or cluster

Cannot-link constraint set: $\mathcal{C}$
each element is a pair of data that come from different classes or clusters

$$\max_{\mathrm{M}} \sum_{(\mathrm{x}_i,\mathrm{x}_j)\in\mathcal{C}} (\mathrm{x}_i - \mathrm{x}_j)^\top \mathrm{M}(\mathrm{x}_i - \mathrm{x}_j)$$

$$s.t. \quad \sum_{(\mathrm{x}_u,\mathrm{x}_v)\in\mathcal{M}} (\mathrm{x}_u - \mathrm{x}_v)^\top \mathrm{M}(\mathrm{x}_u - \mathrm{x}_v) \leqslant 1$$

$$\mathrm{M} \succeq 0$$

Separation based objective
Semi-definite programming
No direct mapping
Global method

E.P. Xing, A.Y. Ng, M.I. Jordan and S. Russell,
Distance Metric Learning, with application to Clustering with side-information. NIPS 16. 2002.

Information Theoretic Metric Learning (ITML)

Suppose we have an initial Mahalanobis distance parameterized by M_0 a set $\mathcal{M}$ of must-link constraints and a set $\mathcal{C}$ of cannot-link constraints. ITML solves the following optimization problem

$$\mathcal{D}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^\top \mathbf{M}(\mathbf{x}-\mathbf{y})}$$

$$\begin{aligned} \min_{\mathcal{M}} \quad & d_{logdet}(\mathrm{M}, \mathrm{M}_0) \\ s.t. \quad & (\mathrm{x}_i - \mathrm{x}_j)^\top \mathrm{M}(\mathrm{x}_i - \mathrm{x}_j) \geqslant l, \ (\mathrm{x}_i, \mathrm{x}_j) \in \mathcal{C} \\ & (\mathrm{x}_u - \mathrm{x}_v)^\top \mathrm{M}(\mathrm{x}_u - \mathrm{x}_v) \leqslant u, \ (\mathrm{x}_u, \mathrm{x}_v) \in \mathcal{M} \end{aligned}$$

Information theoretic objective
Bregman projection
No direct mapping
Global method

$$d_{logdet}(\mathrm{M}, \mathrm{M}_0) = tr(\mathrm{M}\mathrm{M}_0^{-1}) - \log det(\mathrm{M}\mathrm{M}_0^{-1}) - n$$

Scalable and efficient

Jason Davis, Brian Kulis, Prateek Jain, Suvrit Sra, & Inderjit Dhillon. Information-Theoretic Metric Learning. ICML 2007

Best paper award

Summary: Supervised Metric Learning

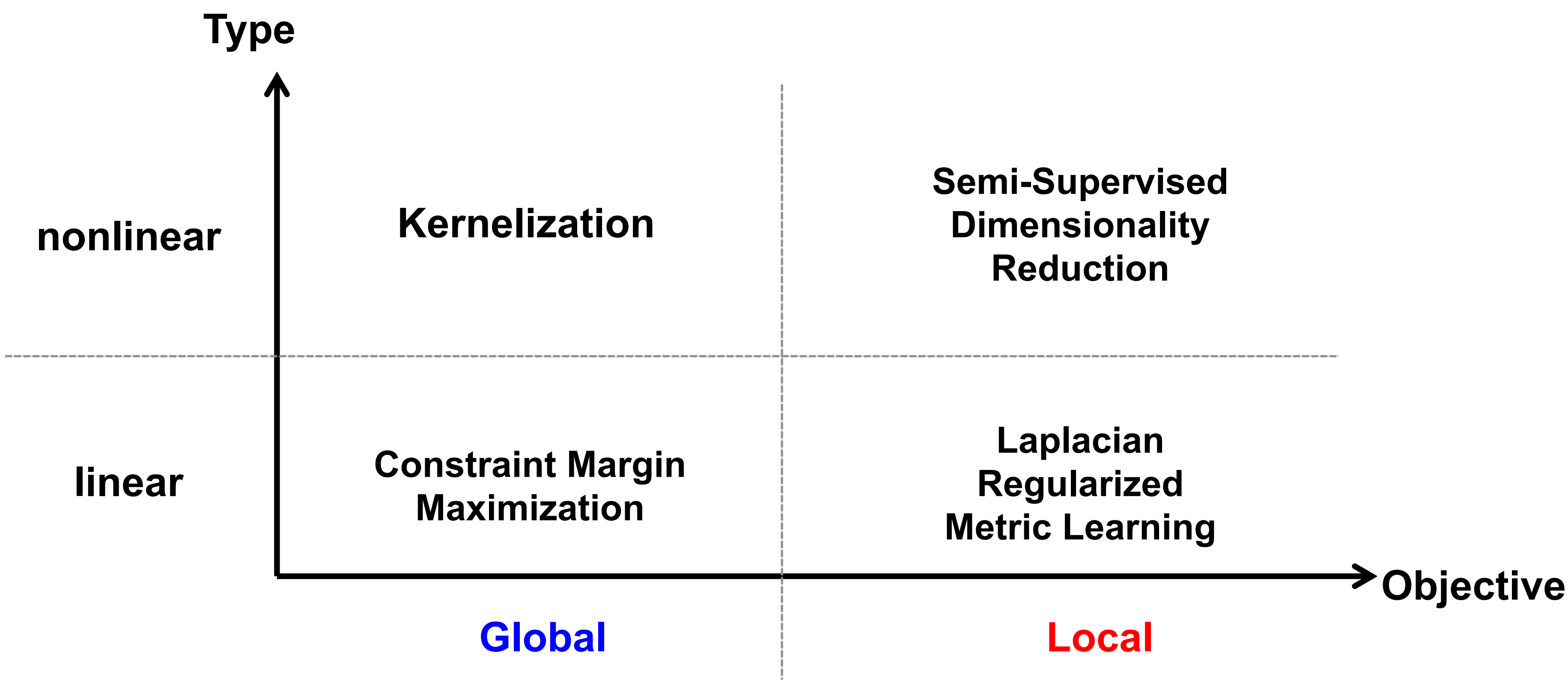
	LDA	Xing	ITML	LSML
Label	√			√
Pairwise constraints		√	√	√
Local				√
Global	√	√	√	
Linear	√			√
Nonlinear				
Separation	√	√	√	√
Geometry				
Information theoretic			√	
Extensibility	√			√

Outline

- Unsupervised Metric Learning
- Supervised Metric Learning
- Semi-supervised Metric Learning
- Challenges and Opportunities for Metric Learning

Semi-Supervised Metric Learning

- Learning pairwise distance metric using data with and without supervision



Constraint Margin Maximization (CMM)

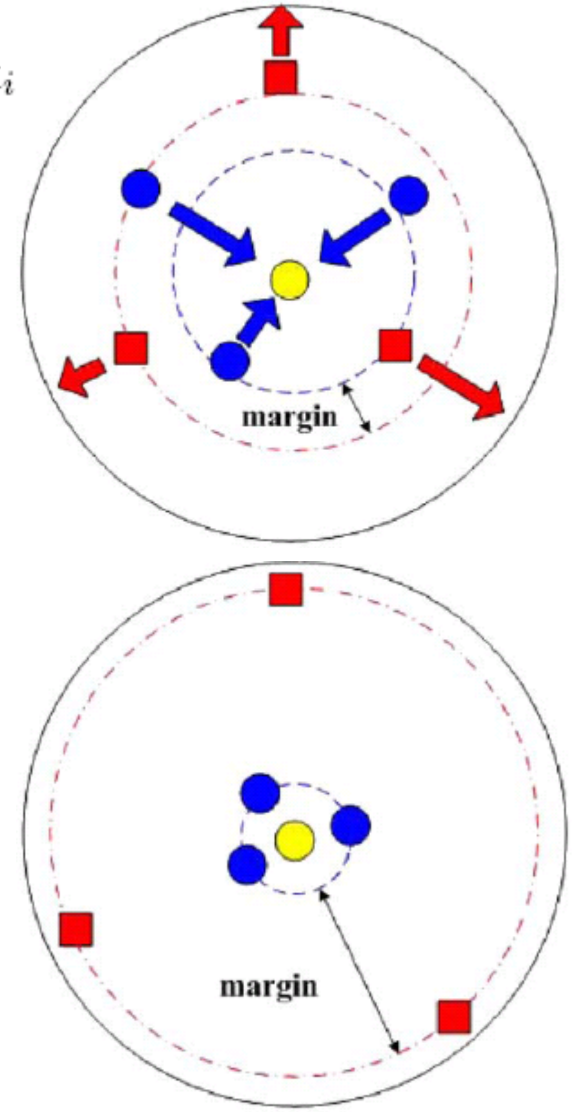
$$\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$$

Scatterness **Compactness**

$$\delta = \boxed{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} \frac{(\mathbf{y}_i-\mathbf{y}_j)^T(\mathbf{y}_i-\mathbf{y}_j)}{N_{\mathcal{C}}}} - \boxed{\sum_{(\mathbf{x}_k,\mathbf{x}_l)\in\mathcal{M}} \frac{(\mathbf{y}_k-\mathbf{y}_l)^T(\mathbf{y}_k-\mathbf{y}_l)}{N_{\mathcal{M}}}}$$

$$\mathcal{J} = \frac{1}{N}\sum_{i,j=1}^{N} \boxed{(\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j)^T(\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j)} + \beta\delta$$

Maximum Variance
No label information



Geometry & Separation based objective
Eigenvalue decomposition
Linear mapping
Global method

F. Wang, S. Chen, T. Li, C. Zhang. Semi-Supervised Metric Learning by Maximizing Constraint Margin. CIKM08

Laplacian Regularized Metric Learning (LRML)

Smoothness term

$$t_1 = \sum_{i,j} \|\mathbf{W}^\top \mathbf{x}_i - \mathbf{W}^\top \mathbf{x}_j\|^2 \omega_{ij} = tr(\mathbf{W}^\top \mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{W}) = tr(\mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{M})$$

Compactness term

$$t_2 = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} \|\mathbf{W}^\top \mathbf{x}_i - \mathbf{W}^\top \mathbf{x}_j\|^2 = tr\left[\mathbf{M} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right]$$

Scatterness term

$$t_3 = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} \|\mathbf{W}^\top \mathbf{x}_i - \mathbf{W}^\top \mathbf{x}_j\|^2 = tr\left[\mathbf{M} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right]$$

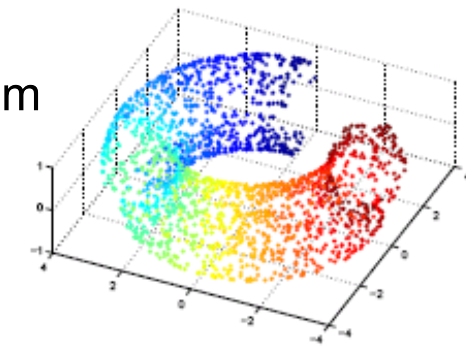
Geometry & Separation based objective
Semi-definite programming
No mapping
Local & Global method

$$\begin{aligned} \min_{\mathbf{M}} \quad & t + \gamma_1 t_2 + \gamma_2 t_3 \\ s.t. \quad & t_1 \leqslant t \\ & \mathbf{M} \succeq 0 \end{aligned}$$

S. Hoi, W. Liu, S. Chang. Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval. CVPR08

Semi-Supervised Dimensionality Reduction (SSDR)

Most of the nonlinear dimensionality reduction algorithms can finally be formalized as solving the following optimization problem



$$\min_{\mathbf{Y}^\top\mathbf{Y}=\mathbf{I}} tr(\mathbf{Y}^\top \mathbf{A} \mathbf{Y})$$

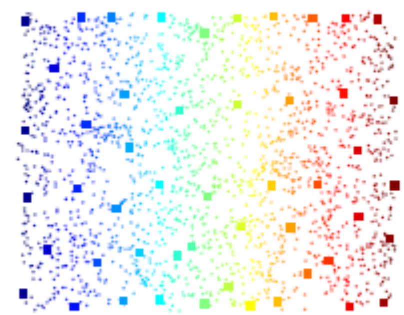
Low-dimensional embeddings

Algorithm specific symmetric matrix

$$[\mathbf{Y}_L^\top, \mathbf{Y}_U^\top] \begin{bmatrix} \mathbf{A}_{LL} & \mathbf{A}_{LU} \\ \mathbf{A}_{LU}^\top & \mathbf{A}_{UU} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_L \\ \mathbf{Y}_U \end{bmatrix}$$

$$= \cancel{\mathbf{Y}_L^\top \mathbf{A}_{LL} \mathbf{Y}_L} + \mathbf{Y}_U^\top \mathbf{A}_{LU}^\top \mathbf{Y}_L + \mathbf{Y}_L^\top \mathbf{A}_{LU} \mathbf{Y}_U + \mathbf{Y}_U^\top \mathbf{A}_{UU} \mathbf{Y}_U$$

$$\boxed{\mathcal{J}(\mathbf{Y}_U) = 2\mathbf{Y}_L^\top \mathbf{A}_{LU} \mathbf{Y}_U + \mathbf{Y}_U^\top \mathbf{A}_{UU} \mathbf{Y}_U}$$



Separation based objective
Eigenvalue decomposition
No mapping
Local method

X. Yang , H. Fu , H. Zha , J. Barlow. Semi-supervised nonlinear dimensionality reduction. ICML 2006.

Summary: Semi-Supervised Distance Metric Learning

	CMM	LRML	SSDM
Label	√	√	
Pairwise constraints	√	√	
Embedded coordinates			√
Local		√	√
Global	√		
Linear	√	√	
Nonlinear			√
Separation	√	√	
Locality-Preservation		√	√
Information theoretic			
Extensibility	√		

Outline

- Unsupervised Metric Learning
- Supervised Metric Learning
- Semi-supervised Metric Learning
- Challenges and Opportunities for Metric Learning

- Scalability
  - Online (Sequential) Learning (Shalev-Shwartz,Singer and Ng, ICML2004) (Davis et al. ICML2007)
  - Distributed Learning
- Efficiency
  - Labeling efficiency (Yang, Jin and Sukthankar, UAI2007)
  - Updating efficiency (Wang, Sun, Hu and Ebadollahi, SDM2011)
- Heterogeneity
  - Data heterogeneity (Wang, Sun and Ebadollahi, SDM2011)
  - Task heterogeneity (Zhang and Yeung, KDD2011)
- Evaluation

- Shai Shalev-Shwartz, Yoram Singer and Andrew Y. Ng. Online and Batch Learning of Pseudo-Metrics. *ICML 2004*
- J. Davis, B. Kulis, P. Jain, S. Sra, & I. Dhillon. Information-Theoretic Metric Learning. ICML 2007
- Liu Yang, Rong Jin and Rahul Sukthankar. Bayesian Active Distance Metric Learning. *UAI* 2007.
- Fei Wang, Jimeng Sun, Shahram Ebadollahi. Integrating Distance Metrics Learned from Multiple Experts and its Application in Inter-Patient Similarity Assessment. SDM2011.
- F. Wang, J. Sun, J. Hu, S. Ebadollahi. iMet: Interactive Metric Learning in Healthcare Applications. SDM 2011.
- Yu Zhang and Dit-Yan Yeung. Transfer Metric Learning by Learning Task Relationships. In: *Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1199-1208. Washington, DC, USA, 2010.

Thanks