机器学习引论

彭玺
pengxi@scu.edu.cn
www.pengxi.me

提纲

提纲

一．Review

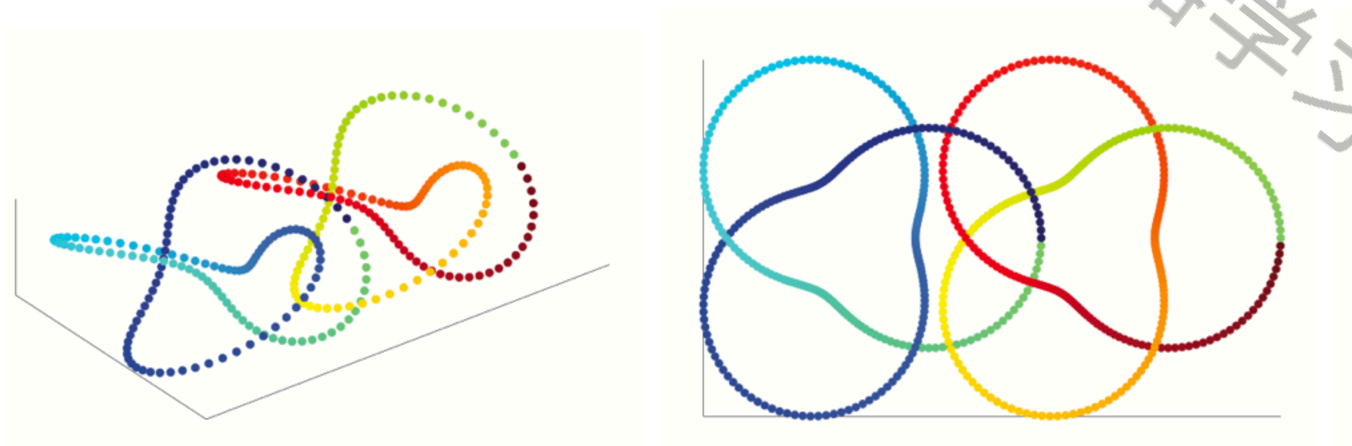
二．Neighborhood Preserving Embedding

三．Locality Preserving Projections

四．Summary of Dimension Reduction

一、Review

Nonlinear Dimensionality Reduction



(a) (b)

- Sam T. Roweis1 and Lawrence K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding, Science, 2000;
- Lawrence K. Saul and Sam T. Roweis1, Think Globally, Fit Locally- Unsupervised Learning of Low Dimensional Manifolds, JMLR2003.

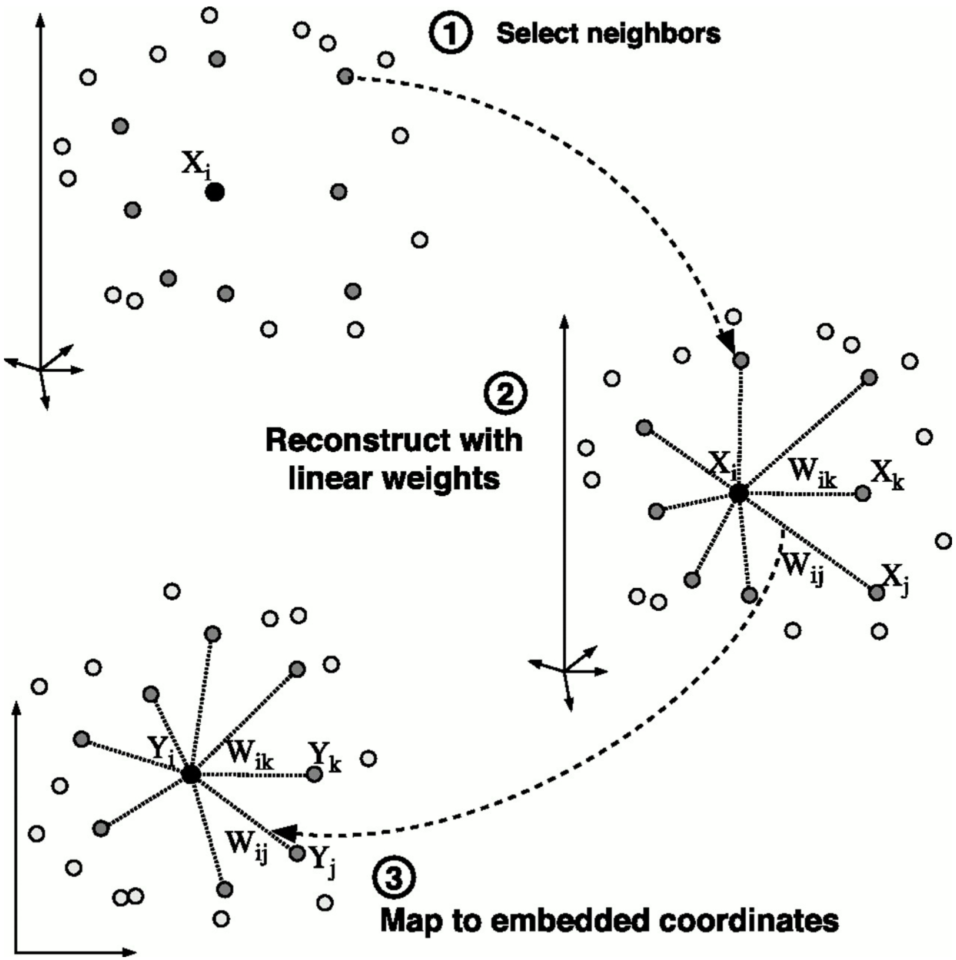
一、Review

- A manifold is a topological space which is locally Euclidean.
- Euclidean space is a simplest example of a manifold.
- The dimension of a manifold is the minimum integer number of co-ordinates necessary to identify each point in that manifold.

$\mathbf{x}_i \in \mathcal{R}^D, \quad \mathbf{D}_i \in \mathcal{R}^{D\times k}, \quad \mathbf{W}_{ij} \in \mathcal{R}^1, \quad \mathbf{W}_i \in \mathcal{R}^k$

$\mathbf{y}_i \in \mathcal{R}^d, \quad \hat{\mathbf{D}}_i \in \mathcal{R}^{d\times k}, \quad \mathbf{D}_{ij} \in \mathcal{R}^D, \quad \hat{\mathbf{D}}_{ij} \in \mathcal{R}^d$

一、Review



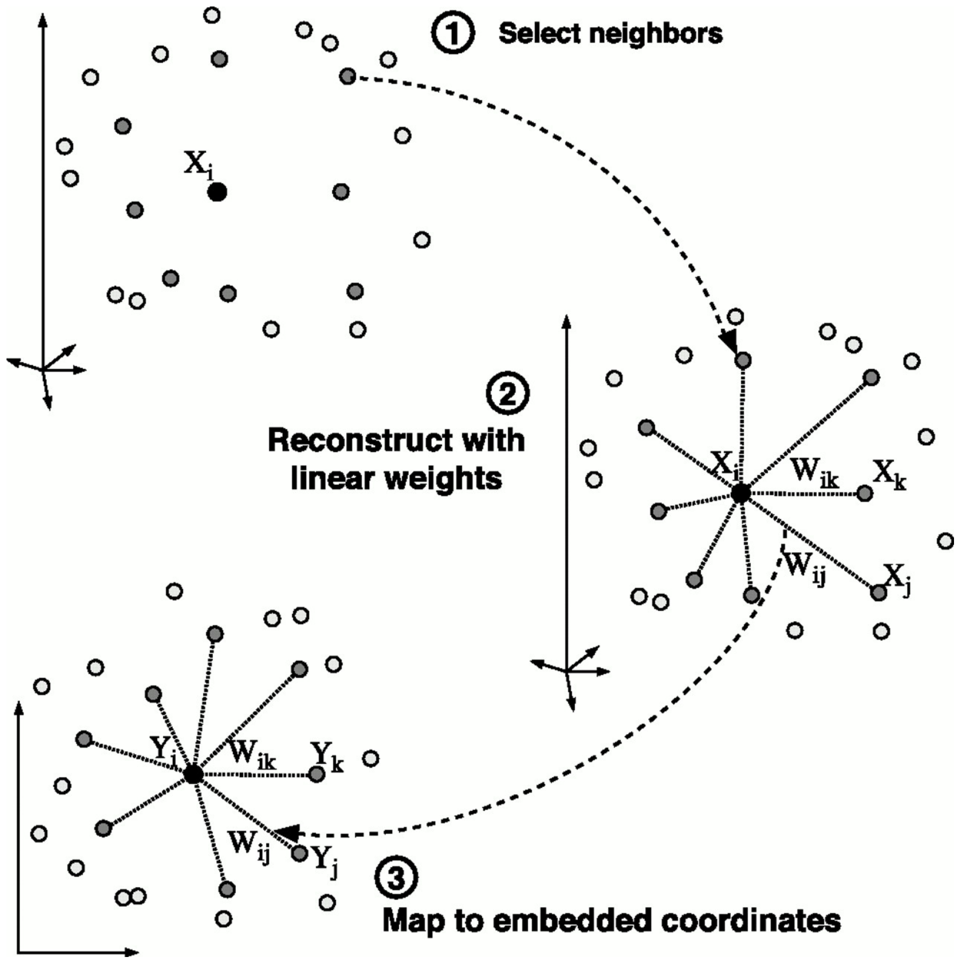
- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$

To find a set of Euclidean space because a manifold is a topological space which is locally Euclidean.

$\mathbf{x}_i \in \mathcal{R}^D, \quad \mathbf{D}_i \in \mathcal{R}^{D\times k}, \quad \mathbf{W}_{ij} \in \mathcal{R}^1, \quad \mathbf{W}_i \in \mathcal{R}^k$

$\mathbf{y}_i \in \mathcal{R}^d, \quad \hat{\mathbf{D}}_i \in \mathcal{R}^{d\times k}, \quad \mathbf{D}_{ij} \in \mathcal{R}^D, \quad \hat{\mathbf{D}}_{ij} \in \mathcal{R}^d$

一、Review



- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$
- Linear: compute the linear reconstruction coefficient $\mathbf{W}_{ij}$ w.r.t. $\mathbf{D}_i$ via

$$\min_{\mathbf{W}_{ij}} \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2$$

$$\text{s.t.} \sum_{j=1}^{k} \mathbf{W}_{ij} = 1$$

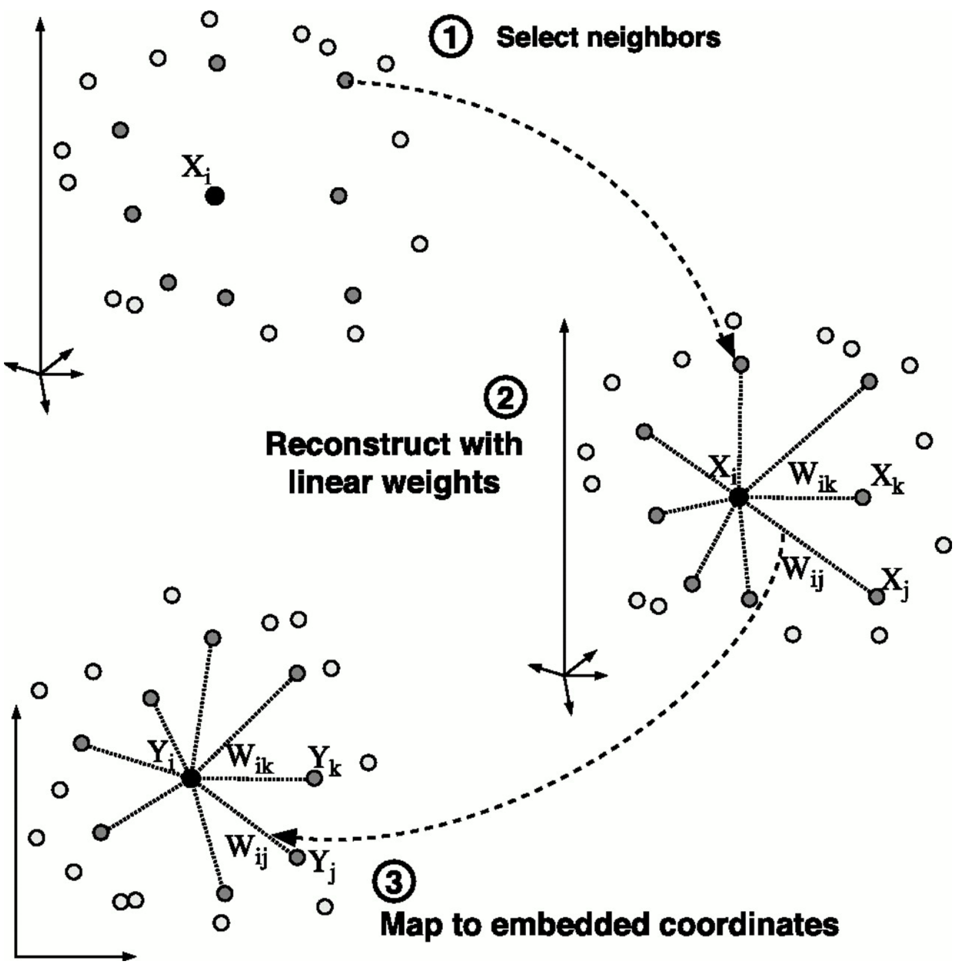
As $\mathbf{x}_i$ and $\mathbf{D}_i$ lies on the Euclidean space, there could be linearly represented (by $\mathbf{W}_{ij}$) each other thanks to the property of linear space.

Here, the neighborhood size k should be larger than the intrinsic dimension of manifold.

$\mathbf{x}_i \in \mathcal{R}^D$, $\mathbf{D}_i \in \mathcal{R}^{D\times k}$, $\mathbf{W}_{ij} \in \mathcal{R}^1$, $\mathbf{W}_i \in \mathcal{R}^k$

$\mathbf{y}_i \in \mathcal{R}^d$, $\hat{\mathbf{D}}_i \in \mathcal{R}^{d\times k}$, $\mathbf{D}_{ij} \in \mathcal{R}^D$, $\hat{\mathbf{D}}_{ij} \in \mathcal{R}^d$

一、Review



- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$
- Linear: compute the linear reconstruction coefficient $\mathbf{W}_{ij}$ w.r.t. $\mathbf{D}_i$ via

$$\min_{\mathbf{W}_{ij}} \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2$$

$$\text{s.t.} \sum_{j=1}^{k} \mathbf{W}_{ij} = 1$$

- Embedding: using $\mathbf{W}$ as an invariance for DR by embedding it into the manifold via:

$$\min_{\mathbf{Y}} \|\mathbf{y}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\hat{\mathbf{D}}_{ij}\|_2^2$$

$$\text{s.t.} \mathbf{y}_i^T\mathbf{y}_i = 1$$

$\hat{\mathbf{D}}_i$ denotes the neighbors of $\mathbf{x}_i$ in the projection space.

一、Review

- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$
- Linear: compute the linear reconstruction coefficient $\mathbf{W}_{ij}$ w.r.t. $\mathbf{D}_i$ via

$$\min_{W} \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 \qquad \text{s.t.} \sum_{j=1}^{k} \mathbf{W}_{ij} = 1$$

with following optimization process:

$$\mathcal{L} = \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij}) = \boxed{\|\sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij})}$$

$$= \|\sum_{j=1}^{k}(\mathbf{x}_i - \mathbf{D}_{ij})\mathbf{W}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij}) = \mathbf{w}^T(\mathbf{X}_i - \mathbf{D}_i)^T(\mathbf{X}_i - \mathbf{D}_i)\mathbf{w} + \lambda(1 - \mathbf{1}^T\mathbf{w})$$

一、Review

- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$
- Linear: compute the linear reconstruction coefficient $\mathbf{W}_{ij}$ w.r.t. $\mathbf{D}_i$ via

$$\min_W \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 \qquad \text{s.t.} \sum_{j=1}^{k} \mathbf{W}_{ij} = 1$$

with following optimization process:

$$\mathcal{L} = \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij}) = \|\sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij})$$

$$= \boxed{\|\sum_{j=1}^{k}(\mathbf{x}_i - \mathbf{D}_{ij})\mathbf{W}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij})} = \mathbf{w}^T(\mathbf{X}_i - \mathbf{D}_i)^T(\mathbf{X}_i - \mathbf{D}_i)\mathbf{w} + \lambda(1 - \mathbf{1}^T\mathbf{w})$$

一、Review

- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$
- Linear: compute the linear reconstruction coefficient $\mathbf{W}_{ij}$ w.r.t. $\mathbf{D}_i$ via

$$\min_{W} \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 \qquad \text{s.t.} \sum_{j=1}^{k} \mathbf{W}_{ij} = 1$$

with following optimization process:

$$\mathcal{L} = \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij}) = \|\sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij})$$

$$= \|\sum_{j=1}^{k}(\mathbf{x}_i - \mathbf{D}_{ij})\mathbf{W}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij}) = \boxed{\mathbf{w}^T(\mathbf{X}_i - \mathbf{D}_i)^T(\mathbf{X}_i - \mathbf{D}_i)\mathbf{w} + \lambda(1 - \mathbf{1}^T\mathbf{w})}$$

w is a vector whose elements are $\mathbf{W}_{ij}$

$\mathbf{X}_i$ is a matrix whose column is $\mathbf{x}_i$

一、Review

$$\mathbf{G}_i = (\mathbf{X}_i - \mathbf{D}_i)^T(\mathbf{X}_i - \mathbf{D}_i)$$

- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$
- Linear: compute the linear reconstruction coefficient $\mathbf{W}_{ij}$ w.r.t. $\mathbf{D}_i$ via

$$\min_W \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 \qquad \text{s.t.} \sum_{j=1}^{k} \mathbf{W}_{ij} = 1$$

with following optimization process:

$$\mathcal{L} = \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij}) = \|\sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij})$$

$$= \|\sum_{j=1}^{k}(\mathbf{x}_i - \mathbf{D}_{ij})\mathbf{W}_{ij}\|_2^2 + \lambda(1 - \sum_{j=1}^{k} \mathbf{W}_{ij}) = \boxed{\mathbf{w}^T(\mathbf{X}_i - \mathbf{D}_i)^T(\mathbf{X}_i - \mathbf{D}_i)\mathbf{w} + \lambda(1 - \mathbf{1}^T\mathbf{w})}$$

Let $\frac{\partial L}{\partial \mathbf{W}} = 0$, we have $(\mathbf{X}_i - \mathbf{D}_i)^T(\mathbf{X}_i - \mathbf{D}_i)\mathbf{w} = \lambda\mathbf{1}$

$$\rightarrow \mathbf{w} = \frac{\mathbf{G}_i^{\dagger}\mathbf{1}}{\mathbf{1}^T\mathbf{G}_i^{\dagger}\mathbf{1}}$$

λ is a constant which is used to achieve the constraint.

Note that, a small number will be added onto the main diagonal entries of $\mathbf{G}_i^{\dagger}$ for nonsingularity.

一、Review

- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$
- Linear: compute the linear reconstruction coefficient $\mathbf{W}_{ij}$ w.r.t. $\mathbf{D}_i$ via

$$\min_{W} \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 \qquad \text{s.t.} \sum_{j=1}^{k} \mathbf{W}_{ij} = 1$$

- Embedding: using $\mathbf{W}$ as an invariance for DR by embedding it into a low dimensional space via:

$$\min_{\mathbf{y}_i} \sum_{i=1}^{i} \|\mathbf{y}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\hat{\mathbf{D}}_{ij}\|_2^2 \quad \text{s.t. } \mathbf{y}_i^T\mathbf{y}_i = 1$$

$$\rightarrow \min_{\mathbf{Y}} \|\mathbf{Y} - \mathbf{Y}\mathbf{W}\|_F^2 \quad \text{s.t. } tr(\mathbf{Y}^T\mathbf{Y}) = 1$$

一、Review

- Locally: for each data point $\mathbf{x}_i$, finding its k nearest neighbors $\mathbf{D}_i$
- Linear: compute the linear reconstruction coefficient $\mathbf{W}_{ij}$ w.r.t. $\mathbf{D}_i$ via

$$\min_{W} \|\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{D}_{ij}\|_2^2 \qquad \text{s.t.} \sum_{j=1}^{k} \mathbf{W}_{ij} = 1$$

- Embedding: using **W** as an invariance for DR by embedding it into a low dimensional space via:

$$\min_{\mathbf{y}_i} \sum_{i=1}^{i} \|\mathbf{y}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\hat{\mathbf{D}}_{ij}\|_2^2 \quad \text{s.t. } \mathbf{y}_i^T\mathbf{y}_i = 1$$

$$\rightarrow \min_{\mathbf{Y}} \|\mathbf{Y} - \mathbf{Y}\mathbf{W}\|_F^2 \quad \text{s.t. } tr(\mathbf{Y}^T\mathbf{Y}) = 1$$

Let $\mathcal{L} = \|\mathbf{Y} - \mathbf{Y}\mathbf{W}\|_F^2 + \lambda trace(\mathbf{I} - \mathbf{Y}^T\mathbf{Y})$ and its derivative w.r.t. $\mathbf{Y}$ be zero, then we have

$$2(\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^T\mathbf{Y}^T = 2\lambda\mathbf{Y}^T$$

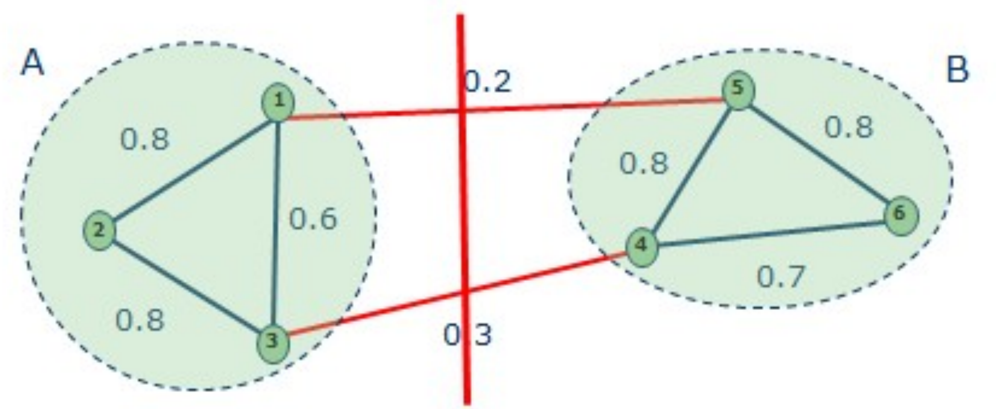
Clearly, the optimal Y consists of *d* eigenvectors corresponding to d smallest nonzero eigenvalue of $(\mathbf{I}\text{-}\mathbf{W})(\mathbf{I}\text{-}\mathbf{W})^T$

一、Review

$$\mathbf{W}_{ij} \in \mathcal{R}^1$$
$$\mathbf{x}_i, \mathbf{x}_j \in \mathcal{R}^D$$
$$\epsilon, t > 0$$

- Step 1: find k nearest neighbors for each data point
- Step 2: obtain a local invariance by constructing a similarity graph via

$$\mathbf{W}_{ij} = \begin{cases} \exp^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{t}} & if \|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon \text{ or they are knn} \\ 0 & \text{otherwise} \end{cases}$$



$\mathbf{W}_{ij} \in \mathcal{R}^1$

$\mathbf{x}_i, \mathbf{x}_j \in \mathcal{R}^D$

$\epsilon, t > 0$

一、Review

- Step 1: find k nearest neighbors for each data point
- Step 2: obtain a local invariance by constructing a similarity graph via

$$\mathbf{W}_{ij} = \begin{cases} \exp^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{t}} & if \|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon \text{ or they are knn} \\ 0 & \text{otherwise} \end{cases}$$

- Step 3: embed **W** into a low-dimensional space by

$$\min_{\mathbf{Y}} \sum_i \sum_j \mathbf{W}_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|_2^2$$

$$\text{s.t. } \mathbf{Y}\mathbf{D}\mathbf{Y}^T = \mathbf{I}$$

一、Review

$$\mathbf{W}_{ij} \in \mathcal{R}^1$$
$$\mathbf{x}_i, \mathbf{x}_j \in \mathcal{R}^D$$
$$\epsilon, t > 0$$

- Step 1: find k nearest neighbors for each data point
- Step 2: obtain a local invariance by constructing a similarity graph via

$$\mathbf{W}_{ij} = \begin{cases} \exp^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{t}} & if \|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon \text{ or they are knn} \\ 0 & \text{otherwise} \end{cases}$$

- Step 3: embed **W** into a low-dimensional space by

$$\min_{\mathbf{Y}} \sum_i \sum_j \mathbf{W}_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \qquad \mathbf{W}_{ij}\uparrow \longrightarrow (\mathbf{y}_i\text{-}\mathbf{y}_j)\downarrow$$

$$\text{s.t. } \mathbf{YDY}^T = \mathbf{I}$$

一、Review

$$
\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=1}^{k}\|\mathbf{y}_i-\mathbf{y}_j\|^2 W_{ij}\\
&=\sum_{i=1}^{k}\sum_{j=1}^{k}(\mathbf{y}_i^{\intercal}\mathbf{y}_i-2\mathbf{y}_i^{\intercal}\mathbf{y}_j+\mathbf{y}_j^{\intercal}\mathbf{y}_j)W_{ij}\\
&=\sum_{i=1}^{k}(\sum_{j=1}^{k}W_{ij})\mathbf{y}_i^{\intercal}\mathbf{y}_i+\sum_{j=1}^{k}(\sum_{i=1}^{k}W_{ij})\mathbf{y}_j^{\intercal}\mathbf{y}_j\\
&\quad-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^{\intercal}\mathbf{y}_j W_{ij}\\
&=2\sum_{i=1}^{k}D_{ij}\mathbf{y}_i^{\intercal}\mathbf{y}_i-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^{\intercal}\mathbf{y}_j W_{ij}\\
&=2\sum_{i=1}^{k}(\sqrt{D_{ij}}\mathbf{y}_i)^{\intercal}(\sqrt{D_{ij}}\mathbf{y}_i)-2\sum_{i=1}^{k}\mathbf{y}_i^{\intercal}(\sum_{j=1}^{k}\mathbf{y}_j W_{ij})\\
&=2Tr[(\mathbf{Y}\sqrt{\mathbf{D}})(\mathbf{Y}\sqrt{\mathbf{D}})^{\intercal}]-2\sum_{i=1}^{k}\mathbf{y}_i^{\intercal}(\mathbf{Y}\mathbf{W}^{\intercal})_i\\
&=2Tr[\mathbf{Y}(\mathbf{D}-\mathbf{W})\mathbf{Y}^{\intercal}]=2Tr[\mathbf{Y}\mathbf{L}\mathbf{Y}^{\intercal}]
\end{aligned}
$$

一、Review

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\|\mathbf{y}_i-\mathbf{y}_j\|^2 W_{ij}$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{k}(\mathbf{y}_i^\intercal\mathbf{y}_i-2\mathbf{y}_i^\intercal\mathbf{y}_j+\mathbf{y}_j^\intercal\mathbf{y}_j)W_{ij}$$

$$=\sum_{i=1}^{k}(\sum_{j=1}^{k}W_{ij})\mathbf{y}_i^\intercal\mathbf{y}_i+\sum_{j=1}^{k}(\sum_{i=1}^{k}W_{ij})\mathbf{y}_j^\intercal\mathbf{y}_j$$

$$-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^\intercal\mathbf{y}_j W_{ij}$$

$$=2\sum_{i=1}^{k}D_{ij}\mathbf{y}_i^\intercal\mathbf{y}_i-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^\intercal\mathbf{y}_j W_{ij}$$

$$=2\sum_{i=1}^{k}(\sqrt{D_{ij}}\mathbf{y}_i)^\intercal(\sqrt{D_{ij}}\mathbf{y}_i)-2\sum_{i=1}^{k}\mathbf{y}_i^\intercal(\sum_{j=1}^{k}\mathbf{y}_j W_{ij})$$

$$=2Tr[(\mathbf{Y}\sqrt{\mathbf{D}})(\mathbf{Y}\sqrt{\mathbf{D}})^\intercal]-2\sum_{i=1}^{k}\mathbf{y}_i^\intercal(\mathbf{Y}\mathbf{W}^\intercal)_i$$

$$=2Tr[\mathbf{Y}(\mathbf{D}-\mathbf{W})\mathbf{Y}^\intercal]=2Tr[\mathbf{Y}\mathbf{L}\mathbf{Y}^\intercal]$$

一、Review

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\|\mathbf{y}_i-\mathbf{y}_j\|^2 W_{ij}$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{k}(\mathbf{y}_i^\intercal\mathbf{y}_i-2\mathbf{y}_i^\intercal\mathbf{y}_j+\mathbf{y}_j^\intercal\mathbf{y}_j)W_{ij}$$

$$=\sum_{i=1}^{k}(\sum_{j=1}^{k}W_{ij})\mathbf{y}_i^\intercal\mathbf{y}_i+\sum_{j=1}^{k}(\sum_{i=1}^{k}W_{ij})\mathbf{y}_j^\intercal\mathbf{y}_j$$

$$-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^\intercal\mathbf{y}_j W_{ij}$$

$$=2\sum_{i=1}^{k}D_{ij}\mathbf{y}_i^\intercal\mathbf{y}_i-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^\intercal\mathbf{y}_j W_{ij}$$

$$=2\sum_{i=1}^{k}(\sqrt{D_{ij}}\mathbf{y}_i)^\intercal(\sqrt{D_{ij}}\mathbf{y}_i)-2\sum_{i=1}^{k}\mathbf{y}_i^\intercal(\sum_{j=1}^{k}\mathbf{y}_j W_{ij})$$

$$=2Tr[(\mathbf{Y}\sqrt{\mathbf{D}})(\mathbf{Y}\sqrt{\mathbf{D}})^\intercal]-2\sum_{i=1}^{k}\mathbf{y}_i^\intercal(\mathbf{Y}\mathbf{W}^\intercal)_i$$

$$=2Tr[\mathbf{Y}(\mathbf{D}-\mathbf{W})\mathbf{Y}^\intercal]=2Tr[\mathbf{Y}\mathbf{L}\mathbf{Y}^\intercal]$$

一、Review

$$
\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=1}^{k}\|\mathbf{y}_i-\mathbf{y}_j\|^2 W_{ij}\\
&=\sum_{i=1}^{k}\sum_{j=1}^{k}(\mathbf{y}_i^\intercal\mathbf{y}_i-2\mathbf{y}_i^\intercal\mathbf{y}_j+\mathbf{y}_j^\intercal\mathbf{y}_j)W_{ij}\\
&=\sum_{i=1}^{k}(\sum_{j=1}^{k}W_{ij})\mathbf{y}_i^\intercal\mathbf{y}_i+\sum_{j=1}^{k}(\sum_{i=1}^{k}W_{ij})\mathbf{y}_j^\intercal\mathbf{y}_j\\
&\quad-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^\intercal\mathbf{y}_j W_{ij}\\
&=2\sum_{i=1}^{k}D_{ij}\mathbf{y}_i^\intercal\mathbf{y}_i-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^\intercal\mathbf{y}_j W_{ij}\\
&=2\sum_{i=1}^{k}(\sqrt{D_{ij}}\mathbf{y}_i)^\intercal(\sqrt{D_{ij}}\mathbf{y}_i)-2\sum_{i=1}^{k}\mathbf{y}_i^\intercal(\sum_{j=1}^{k}\mathbf{y}_j W_{ij})\\
&=2Tr[(\mathbf{Y}\sqrt{\mathbf{D}})(\mathbf{Y}\sqrt{\mathbf{D}})^\intercal]-2\sum_{i=1}^{k}\mathbf{y}_i^\intercal(\mathbf{Y}\mathbf{W}^\intercal)_i\\
&=2Tr[\mathbf{Y}(\mathbf{D}-\mathbf{W})\mathbf{Y}^\intercal]=2Tr[\mathbf{Y}\mathbf{L}\mathbf{Y}^\intercal]
\end{aligned}
$$

一、Review

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\|\mathbf{y}_i-\mathbf{y}_j\|^2 W_{ij}$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{k}(\mathbf{y}_i^\intercal\mathbf{y}_i-2\mathbf{y}_i^\intercal\mathbf{y}_j+\mathbf{y}_j^\intercal\mathbf{y}_j)W_{ij}$$

$$=\sum_{i=1}^{k}(\sum_{j=1}^{k}W_{ij})\mathbf{y}_i^\intercal\mathbf{y}_i+\sum_{j=1}^{k}(\sum_{i=1}^{k}W_{ij})\mathbf{y}_j^\intercal\mathbf{y}_j$$

$$-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^\intercal\mathbf{y}_j W_{ij}$$

$$=2\sum_{i=1}^{k}D_{ij}\mathbf{y}_i^\intercal\mathbf{y}_i-2\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{y}_i^\intercal\mathbf{y}_j W_{ij}$$

$$=2\sum_{i=1}^{k}(\sqrt{D_{ij}}\mathbf{y}_i)^\intercal(\sqrt{D_{ij}}\mathbf{y}_i)-2\sum_{i=1}^{k}\mathbf{y}_i^\intercal(\sum_{j=1}^{k}\mathbf{y}_j W_{ij})$$

$$=2Tr[(\mathbf{Y}\sqrt{\mathbf{D}})(\mathbf{Y}\sqrt{\mathbf{D}})^\intercal]-2\sum_{i=1}^{k}\mathbf{y}_i^\intercal(\mathbf{Y}\mathbf{W}^\intercal)_i$$

$$=2Tr[\mathbf{Y}(\mathbf{D}-\mathbf{W})\mathbf{Y}^\intercal]=2Tr[\mathbf{Y}\mathbf{L}\mathbf{Y}^\intercal]$$

Then, the loss is as blow

$$\min_{\mathbf{Y}} \quad Tr(\mathbf{YLY}^T)$$
$$\text{s.t.} \quad \mathbf{YDY}^T=\mathbf{I}$$

一、Review

Let $\mathcal{L} = Tr(\mathbf{YLY}^T + \mathbf{\Lambda}(\mathbf{I} - \mathbf{YDY}^T))$,

where $\mathbf{\Lambda}$ is a diagonal matrix whose entries are Lagrange multipliers.

We compute the derivative of ζ with respect to $\mathbf{Y}$ as

$$\frac{\partial \zeta}{\partial \mathbf{Y}} = \mathbf{LY} - \mathbf{DY\Lambda}$$

The optimal $\mathbf{Y}$ satisfies

$$\mathbf{LY} - \mathbf{DY\Lambda} = \mathbf{0} \tag{5}$$

which is a generalized eigenvalue problem, we turn Equa. (6) into a simple eigenvalue problem by post-multiplying $\mathbf{D}^{-1}$, The optimal $\mathbf{Y}$ satisfies

$$\mathbf{D}^{-1}\mathbf{LY} = \mathbf{Y\Lambda} \tag{6}$$

Note that: **L** is a symmetric matrix

提纲

一．Review

二．Neighborhood Preserving Embedding

三．Locality Preserving Projections

四．Summary of Dimension Reduction

二、Neighborhood Preserving Embedding

Two limitations suffered by LLE and LE

LLE

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

LE

$$\mathbf{D}^{-1}\mathbf{L}\mathbf{Y} = \lambda\mathbf{Y}$$

- Scalability issue: the complexity of them is proportional to $O(mn^2)$.

二、Neighborhood Preserving Embedding

Two limitations suffered by LLE and LE

LLE

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

LE

$$\mathbf{D}^{-1}\mathbf{L}\mathbf{Y}=\lambda\mathbf{Y}$$

- Scalability issue: the complexity of them is proportional to O(mn^2).
- Out-of-sample issue: **L** (**W**) depends on the whole data set, thus making impossibility in handling new coming data points.

二、Neighborhood Preserving Embedding

Step 1&2 are the same with LLE

Step 3: Let **y**=**A**ᵀ**x**, we have

$$\sum_{i=1}^{n} \|\mathbf{y}_i - \sum_{j=1}^{k} \mathbf{W}_{ij}\mathbf{y}_j\|_2^2 = \|\mathbf{Y} - \mathbf{Y}\mathbf{W}\|_F^2$$

$$= \|\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{W}\|_F^2$$

$$= Tr(\mathbf{A}^T\mathbf{X}(\mathbf{I}-\mathbf{W})(\mathbf{I}-\mathbf{W})^T\mathbf{X}^T\mathbf{A})$$

$$= Tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A})$$

$$\mathbf{y}^T\mathbf{y} = 1 \rightarrow \mathbf{A}^T\mathbf{x}\mathbf{x}^T\mathbf{A} = 1$$

二、Neighborhood Preserving Embedding

$$\mathcal{L} = Tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A} + \lambda(\mathbf{I} - \mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A})$$

Let $\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 0$, then

$$\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{A}$$

二、Neighborhood Preserving Embedding

$$\mathcal{L} = Tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A} + \lambda(\mathbf{I} - \mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A})$$

Let $\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 0$, then

$$\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{A}$$

The optimal **A** consists of eigenvectors corresponding to d smallest nonzero eigenvalues of

$$(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{X}\mathbf{M}\mathbf{X}^T$$

二、Neighborhood Preserving Embedding

$$\mathcal{L} = Tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A} + \lambda(\mathbf{I} - \mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A})$$

Let $\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 0$, then

$$\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{A}$$

The optimal **A** consists of eigenvectors corresponding to d smallest nonzero eigenvalues of

$$(\mathbf{X}\mathbf{X}^T)^{\dagger}\mathbf{X}\mathbf{M}\mathbf{X}^T$$

Then, for any new coming data point **z**, one could obtain its low-dimensional features via $\mathbf{W}^T\mathbf{z}$.

提纲

一．Review

二．Neighborhood Preserving Embedding

三．Locality Preserving Projections

四．Summary of Dimension Reduction

三、Locality Preserving Projections

Step 1&2 are the same with LE

Step 3: Let $\mathbf{y}=\mathbf{A}^{T}\mathbf{x}$, we have

$$\min_{\mathbf{Y}} \sum_i \sum_j \mathbf{W}_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \quad \text{s.t. } tr(\mathbf{Y}\mathbf{D}\mathbf{Y}^T) = 1 \quad \rightarrow \quad \min_{\mathbf{W}} Tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \quad \text{s.t. } Tr(\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}) = 1$$

Using Lagrange Multipliers method, we have

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W} = \Lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W}$$

Namely, the optimal **W** consists of d smallest nonzero eigenvectors of

$$(\mathbf{X}\mathbf{D}\mathbf{X}^T)^{\dagger}\mathbf{X}\mathbf{L}\mathbf{X}^T$$

提纲

一．Review

二．Neighborhood Preserving Embedding

三．Locality Preserving Projections

四．Summary of Dimension Reduction

四、Summary

DR

PCA	Unsupervised Single Subspace
CCA	Weakly Supervised Multiple Subspace
LDA	Fully Supervised Multiple Subspace
LLE	embed the locally linear reconstruction coefficient from input to feature space
LE	embed the pairwise distance with heat kernel from input to feature space
NPE	out-of-sample extension of LLE
LPP	out-of-sample extension of LE

$C_{ij}=X_iX_j^T$ is the covariance matrix between these two views

四、Summary

PCA

$$\max_{\mathbf{W}} \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}$$
$$\text{s.t.}\mathbf{W}^T\mathbf{W}=\mathbf{I}$$

CCA

$$\max_{\mathbf{w}_1,\cdots,\mathbf{w}_k} \sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{W}_i^T\mathbf{C}_{ij}\mathbf{W}_j$$
$$\text{s.t.} \sum_{i=1}^{k} Tr(\mathbf{W}_i^T\mathbf{C}_{ii}\mathbf{W}_{ii})=1$$

LDA

$$\max_{\mathbf{W}} \frac{\mathbf{W}^T\mathbf{S}_B\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}$$

LLE

$$\min_{\mathbf{Y}} \ Tr(\mathbf{Y}\mathbf{M}\mathbf{Y}^{\mathbf{T}})$$
$$\text{s.t.} \ Tr(\mathbf{Y}\mathbf{Y}^T)=1$$

NPE

$$\min_{\mathbf{A}} \ Tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A})$$
$$\text{s.t.} \ Tr(\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A})=1$$

LE

$$\min_{\mathbf{Y}} \ Tr(\mathbf{Y}\mathbf{L}\mathbf{Y}^T)$$
$$\text{s.t.} \ Tr(\mathbf{Y}\mathbf{D}\mathbf{Y}^T)=1$$

LPP

$$\min_{\mathbf{A}} \ Tr(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A})$$
$$\text{s.t.} \ Tr(\mathbf{A}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{A})=1$$

$C_{ij}=X_iX_j^T$ is the covariance matrix between these two views

四、Summary

PCA

$$\max_{\mathbf{W}} \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}$$
$$\text{s.t.} \mathbf{W}^T\mathbf{W}=\mathbf{I}$$

CCA

$$\max_{\mathbf{W}_1,\cdots,\mathbf{W}_k} \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbf{W}_i^T\mathbf{C}_{ij}\mathbf{W}_j$$
$$\text{s.t.} \sum_{i=1}^{k} Tr(\mathbf{W}_i^T\mathbf{C}_{ii}\mathbf{W}_{ii})=1$$

LDA

$$\max_{\mathbf{W}} \frac{\mathbf{W}^T\mathbf{S}_B\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}$$

LLE

$$\min_{\mathbf{Y}} Tr(\mathbf{YMY}^\mathbf{T})$$
$$\text{s.t.} Tr(\mathbf{YY}^T)=1$$

NPE

$$\min_{\mathbf{A}} Tr(\mathbf{A}^T\mathbf{XMX}^T\mathbf{A})$$
$$\text{s.t.} Tr(\mathbf{A}^T\mathbf{XX}^T\mathbf{A})=1$$

LE

$$\min_{\mathbf{Y}} Tr(\mathbf{YLY}^T)$$
$$\text{s.t.} Tr(\mathbf{YDY}^T)=1$$

LPP

$$\min_{\mathbf{A}} Tr(\mathbf{A}^T\mathbf{XLX}^T\mathbf{A})$$
$$\text{s.t.} Tr(\mathbf{A}^T\mathbf{XDX}^T\mathbf{A})=1$$

$C_{ij}=X_iX_j^T$ is the covariance matrix between these two views

四、Summary

PCA

$$\max_{\mathbf{W}} \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}$$
$$\text{s.t.} \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

CCA

$$\max_{\mathbf{w}_1,\cdots,\mathbf{w}_k} \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbf{W}_i^T\mathbf{C}_{ij}\mathbf{W}_j$$
$$\text{s.t.} \sum_{i=1}^{k} Tr(\mathbf{W}_i^T\mathbf{C}_{ii}\mathbf{W}_{ii}) = 1$$

LDA

$$\max_{\mathbf{W}} \frac{\mathbf{W}^T\mathbf{S}_B\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}$$

LLE

$$\min_{\mathbf{Y}} \; Tr(\mathbf{Y}\mathbf{M}\mathbf{Y}^{\mathbf{T}})$$
$$\text{s.t.} \; Tr(\mathbf{Y}\mathbf{Y}^T) = 1$$

NPE

$$\min_{\mathbf{A}} \; Tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A})$$
$$\text{s.t.} \; Tr(\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}) = 1$$

LE

$$\min_{\mathbf{Y}} \; Tr(\mathbf{Y}\mathbf{L}\mathbf{Y}^T)$$
$$\text{s.t.} \; Tr(\mathbf{Y}\mathbf{D}\mathbf{Y}^T) = 1$$

LPP

$$\min_{\mathbf{A}} \; Tr(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A})$$
$$\text{s.t.} \; Tr(\mathbf{A}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{A}) = 1$$

$C_{ij}=X_iX_j^T$ is the covariance matrix between these two views

四、Summary

PCA

$$\max_{\mathbf{W}} \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}$$
$$\text{s.t.}\, \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

CCA

$$\max_{\mathbf{w}_1,\cdots,\mathbf{w}_k} \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbf{W}_i^T\mathbf{C}_{ij}\mathbf{W}_j$$
$$\text{s.t.} \sum_{i=1}^{k} Tr(\mathbf{W}_i^T\mathbf{C}_{ii}\mathbf{W}_{ii}) = 1$$

LDA

$$\max_{\mathbf{W}} \frac{\mathbf{W}^T\mathbf{S}_B\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}$$

LLE

$$\min_{\mathbf{Y}} \; Tr(\mathbf{Y}\mathbf{M}\mathbf{Y}^{\mathbf{T}})$$
$$\text{s.t.} \; Tr(\mathbf{Y}\mathbf{Y}^T) = 1$$

NPE

$$\min_{\mathbf{A}} \; Tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A})$$
$$\text{s.t.} \; Tr(\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}) = 1$$

LE

$$\min_{\mathbf{Y}} \; Tr(\mathbf{Y}\mathbf{L}\mathbf{Y}^T)$$
$$\text{s.t.} \; Tr(\mathbf{Y}\mathbf{D}\mathbf{Y}^T) = 1$$

LPP

$$\min_{\mathbf{A}} \; Tr(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A})$$
$$\text{s.t.} \; Tr(\mathbf{A}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{A}) = 1$$

$C_{ij}=X_iX_j^T$ is the covariance matrix between these two views

四、Summary

PCA

$$\max_{\mathbf{W}} \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}$$
$$\text{s.t.} \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

CCA

$$\max_{\mathbf{W}_1,\cdots,\mathbf{W}_k} \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbf{W}_i^T\mathbf{C}_{ij}\mathbf{W}_j$$
$$\text{s.t.} \sum_{i=1}^{k} Tr(\mathbf{W}_i^T\mathbf{C}_{ii}\mathbf{W}_{ii}) = 1$$

LDA

$$\max_{\mathbf{W}} \frac{\mathbf{W}^T\mathbf{S}_B\mathbf{W}}{\mathbf{W}^T\mathbf{S}_W\mathbf{W}}$$

LLE

$$\min_{\mathbf{Y}} \ Tr(\mathbf{Y}\mathbf{M}\mathbf{Y}^{\mathbf{T}})$$
$$\text{s.t.} \ Tr(\mathbf{Y}\mathbf{Y}^T) = 1$$

NPE

$$\min_{\mathbf{A}} \ Tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A})$$
$$\text{s.t.} \ Tr(\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}) = 1$$

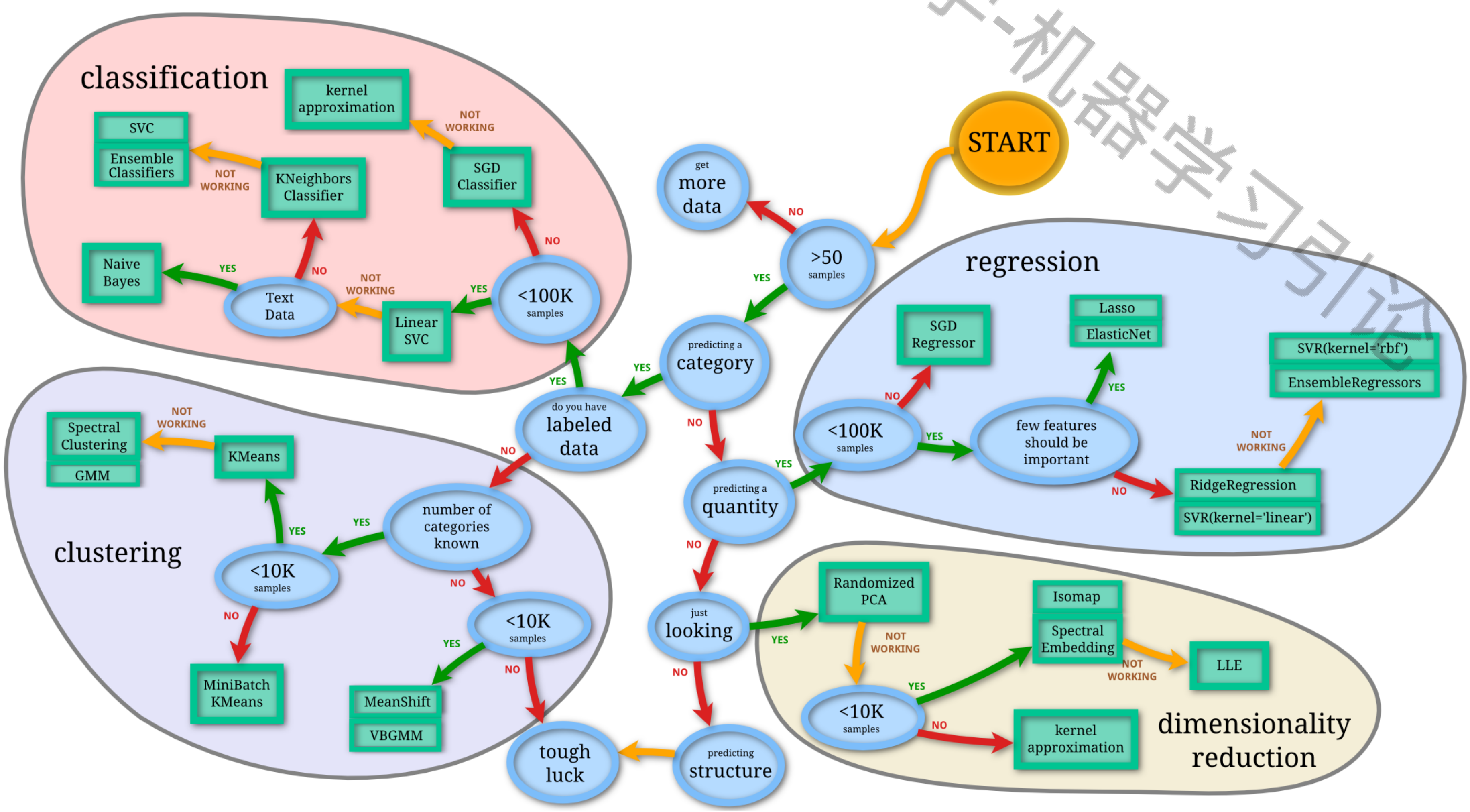
LE

$$\min_{\mathbf{Y}} \ Tr(\mathbf{Y}\mathbf{L}\mathbf{Y}^T)$$
$$\text{s.t.} \ Tr(\mathbf{Y}\mathbf{D}\mathbf{Y}^T) = 1$$

LPP

$$\min_{\mathbf{A}} \ Tr(\mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A})$$
$$\text{s.t.} \ Tr(\mathbf{A}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{A}) = 1$$

NEXT – Clustering



Q&A

THANKS!