(1) Note Title

Multidimensional Scaling (MDS)

Spring 2014 11/2/2010

MDS is a linear projection method. It is related to PCA. MDS and PCA can be used for non-linear projection – next lecture.

<u>Key Idea of MDS</u>: project to preserve the distances $|\underline{x}_i - \underline{x}_j|$ between the datapoints $\{\underline{x}_i : i=1 \text{ to } N\}$,

i.e. $\underline{x}_i \to \underline{y}_i$ such that $|\underline{x}_i - \underline{x}_j| \approx |\underline{y}_i - \underline{y}_j|$,

but the $\underline{y}$'s have lower dimension.

This projection constraint is imposed on the dot products $\underline{x}_i \cdot \underline{x}_j \approx \underline{y}_i \cdot \underline{y}_j$

this will imply ~~that~~ $|\underline{x}_i - \underline{x}_j| \approx |\underline{y}_i - \underline{y}_j|$.

<u>Important Property</u>: we only need to know $|\underline{x}_i - \underline{x}_j| \triangleq \Delta_{ij}$ in order to calculate $\underline{x}_i \cdot \underline{x}_j$. This will be useful for non-linear applications later. Also, sometimes only $|\underline{x}_i - \underline{x}_j|$ is specified.

<u>Result</u>: $$\underline{x}_i \cdot \underline{x}_j = \frac{1}{2N} \sum_k \Delta_{ik}^2 + \frac{1}{2N} \sum_l \Delta_{lj}^2 - \frac{1}{2N^2} \sum_{lk} \Delta_{lk}^2 - \frac{1}{2} \Delta_{ij}^2$$

provided $\sum_i \underline{x}_i = 0$ (**Subtract $\frac{1}{N} \sum_{i=1}^{N} \underline{x}_i$ from data to ensure this**)

<u>Proof</u>: $\Delta_{ij}^2 = |\underline{x}_i|^2 + |\underline{x}_j|^2 - 2 \underline{x}_i \cdot \underline{x}_j$

Let $T = \sum_i |\underline{x}_i|^2$, Note that $\sum_i \underline{x}_i \cdot \underline{x}_j = 0 = \sum_j \underline{x}_i \cdot \underline{x}_j$.

Then $\sum_k \Delta_{ik}^2 = N|\underline{x}_i|^2 + T$, $\sum_l \Delta_{lj}^2 = N|\underline{x}_j|^2 + T$, $\sum_{lk} \Delta_{lk}^2 = 2NT$.

the result follows by substitution.

Define the <u>Gram matrix</u>

$$G_{ij} = \underline{x}_i \cdot \underline{x}_j = \sum_k \Delta_{ik}^2 + \frac{1}{2N} \sum_l \Delta_{lj}^2 - \frac{1}{2N^2} \sum_{lk} \Delta_{lk}^2 - \frac{1}{2} \Delta_{ij}^2 .$$

Define an error function

$$err(\underline{y}) = \sum_{ij} (G_{ij} - \underline{y}_i \cdot \underline{y}_j)^2$$

$\underline{x}_i$ lies in D-dim space, G_{ij} is $N \times N$ matrix

$\underline{y}_i$ is a vector in d-dim space $\quad d << D$, $d < N$

(2) Minimize $\mathrm{err}(\underline{y}) = \sum_{ij} (G_{ij} - \underline{y}_i \cdot \underline{y}_j)^2$.

Do spectral decomposition:

$$\underline{\underline{G}} = \sum_{\alpha=1}^{N} \lambda_\alpha \underline{v}_\alpha \underline{v}_\alpha^T$$

$\lambda_1 \geq \ldots \geq \lambda_n \geq 0$ eigenvalues of $\underline{\underline{G}}$

$\underline{v}_\alpha \cdot \underline{v}_\beta = \delta_{\alpha\beta}$ eigenvectors

Claim: optimal minimization is $y^i_\alpha = \sqrt{\lambda_\alpha}\, v^\alpha_i$

i.e. $\underline{y}_i = (\sqrt{\lambda_1}\, v^1_i, \sqrt{\lambda_2}\, v^2_i, \ldots, \sqrt{\lambda_n}\, v^N_i)$

N-dimension

Proof. $\underline{y}_i \cdot \underline{y}_j = \sum_\alpha \sqrt{\lambda_\alpha}\, v^\alpha_i \sqrt{\lambda_\alpha}\, v^\alpha_j = \sum_\alpha \lambda_\alpha \underline{v}_\alpha \underline{v}_\alpha^T = G_{ij}$

This gives $\mathrm{err} = 0$.

We can reduce the dimension by

truncating $\underline{y}_i$ to $\underline{y}_i = (\sqrt{\lambda_1}\, v^1_i, \ldots, \sqrt{\lambda_d}\, v^d_i)$ for $d < N$

In this case $G_{ij} \neq \underline{y}_i \cdot \underline{y}_j$

$$\underline{y}_i \underline{y}_j = \sum_{\alpha=1}^{d} \lambda_\alpha v^\alpha_i v^\alpha_j, \quad G_{ij} = \sum_{\alpha=1}^{N} \lambda_\alpha v^\alpha_i v^\alpha_j$$

Hence $$G_{ij} - \underline{y}_i \cdot \underline{y}_j = \sum_{\alpha=d+1}^{N} \lambda_\alpha v^\alpha_i v^\alpha_j$$

Claim $$\sum_{ij} (G_{ij} - \underline{y}_i \cdot \underline{y}_j)^2 = \sum_{\alpha=d+1}^{N} \lambda_\alpha^2$$

Proof. $$\sum_{ij} \left\{ \sum_{\alpha=d+1}^{N} \sum_{\beta=d+1}^{N} \lambda_\alpha \lambda_\beta v^\alpha_i v^\alpha_j v^\beta_i v^\beta_j \right\},$$

$$\sum_i v^\alpha_i v^\beta_i = \delta^{\alpha\beta}, \quad \sum_j v^\alpha_j v^\beta_j = \delta^{\alpha\beta}$$

$$\sum_{\alpha=d+1}^{N} \sum_{\beta=d+1}^{N} \lambda_\alpha \lambda_\beta \delta^{\alpha\beta} \delta^{\alpha\beta} = \sum_{\alpha=d+1}^{N} \lambda_\alpha^2.$$

Hence we project to d-dimension

provided $\sum_{\alpha=d+1}^{N} \lambda_\alpha^2$ is small, or $\dfrac{\sum_{\alpha=d+1}^{N} \lambda_\alpha^2}{\sum_{\alpha=1}^{N} \lambda_\alpha^2}$ is small.

$$\underline{y}_i = (\sqrt{\lambda_1}\, v^1_i, \ldots, \sqrt{\lambda_d}\, v^d_i)$$

(3) Relation between MDS & PCA?

Both linear. Both depend on eigenvectors/eigenvalue.

Recall PCA (subtract mean to ensure $\sum_i \underline{x}_i = 0$)

$$\underline{x}_p = (\underline{x}.\underline{e}_1, \ldots, \underline{x}.\underline{e}_d)$$

the $\underline{e}$'s are eigenvectors of $K_{ab} = \frac{1}{N}\sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T$

MDS $\underline{y}_i = (\sqrt{\lambda_1} v_i^1, \ldots, \sqrt{\lambda_d} v_i^d)$

the $\underline{v}$'s are eigenvectors of $G_{ij} = \sum_{i=1}^{N} \underline{x}_i . \underline{x}_j$

Claim: the eigenvalues of $\underline{\underline{G}}$ and $\underline{\underline{K}}$ are the same. The eigenvectors are closely related.

Proof Let $\underline{\underline{X}}$ be an $N \times D$ matrix $i = 1$ to N no. of points, $a = 1$ to D space dimension

with elements X_{ia} a^{th} component of i^{th} datapoint

Consider $\underline{\underline{X}}\,\underline{\underline{X}}^T$ $N \times N$ matrix, $(\underline{\underline{X}}\,\underline{\underline{X}}^T)_{ij} = \sum_a X_{ia} X_{ja}$

$\underline{\underline{X}}^T \underline{\underline{X}}$ $D \times D$ matrix, $(\underline{\underline{X}}^T \underline{\underline{X}})_{ab} = \sum_i X_{ia} X_{ib}$

Both are square matrices and both are positive definite, so they have positive eigenvectors.

$\underline{\underline{X}}\,\underline{\underline{X}}^T$ is used for MDS, $\underline{\underline{X}}^T \underline{\underline{X}}$ is used for PCA

Suppose $\underline{e}, \lambda$ are an eigenvector, eigenvalue of $\underline{\underline{X}}^T \underline{\underline{X}}$

$$\underline{\underline{X}}^T \underline{\underline{X}}\, \underline{e} = \lambda \underline{e}$$

So $$\underline{\underline{X}}\, \underline{\underline{X}}^T \underline{\underline{X}}\, \underline{e} = \lambda \underline{\underline{X}}\, \underline{e}$$

$$(\underline{\underline{X}}\,\underline{\underline{X}}^T)(\underline{\underline{X}}\,\underline{e}) = \lambda (\underline{\underline{X}}\,\underline{e})$$

So $(\underline{\underline{X}}\,\underline{e})$ is an eigenvector (un-normalized) of $\underline{\underline{X}}\,\underline{\underline{X}}^T$ with eigenvalue λ.

Similarly if $\underline{\underline{X}}\,\underline{\underline{X}}^T \underline{v} = \lambda \underline{v}$ then $(\underline{\underline{X}}^T \underline{v})$ is an eigenvector (un-normalized) of $\underline{\underline{X}}^T \underline{\underline{X}}$ with eigenvalue λ.

Conclusion → the two matrices have the same eigenvalues and related eigenvectors.

(4)

Result → the truncation conditions for MDS and PCA are similar $\frac{\sum_{i=1}^{\Delta} \lambda_i^2}{\sum_{i=1}^{1} \lambda_i^2} > \text{Threshold}$.

MDS projects $\underline{x}_i$ to $\underline{y}_i = (\sqrt{\lambda_1}\, v_i^1, \ldots, \sqrt{\lambda_d}\, v_i^d)$

PCA projects $\underline{x}$ to $\underline{y} = (\underline{x}\cdot\underline{e}_1, \ldots \underline{x}\cdot\underline{e}_d)$
where the $\underline{e}$'s and the $\underline{v}$'s are related (see previous page).

Note: The equivalence between the eigenvalues of $\underline{\underline{X}}\,\underline{\underline{X}}^T$ and $\underline{\underline{X}}^T\underline{\underline{X}}$ has computational importance
If $N \ll D$, faster to compute eigenvalues/vectors for $(\underline{\underline{X}}\,\underline{\underline{X}}^T)$, then convert to eigenvalues/vectors of $\underline{\underline{X}}^T\underline{\underline{X}}$

Note: to do PCA we have to know the data $\{\underline{x}_i\}$, but to do MDS we only need to know Δ_{ij}.
In some applications it is possible to specify Δ_{ij} but not the $\{\underline{x}_i\}$.

Deeper Understanding: This relationship between $\underline{\underline{X}}\,\underline{\underline{X}}^T$ and $\underline{\underline{X}}^T\underline{\underline{X}}$ can be used to proove SVD:

$\underline{\underline{X}} = \underline{\underline{F}}\,\underline{\underline{D}}\,\underline{\underline{E}}$ where $\underline{\underline{F}}\,\underline{\underline{F}} = \underline{\underline{I}}$ ← Identity.
$\underline{\underline{D}} = \begin{pmatrix} d_1 & 0 \\ 0 & \ddots\, d_D \end{pmatrix}$ $\underline{\underline{E}}\,\underline{\underline{E}}^T = \underline{\underline{I}}$

This is a generalization of the spectral decomposition
$\underline{G} = \sum \lambda_a \underline{v}_a \underline{v}_a^T$ $\underline{\underline{X}}$ is $N \times D$ $N \neq D$
to any matrix $\underline{\underline{X}}$ → so $\underline{\underline{X}}$ is not square.

It follows that $\underline{\underline{X}}^T\underline{\underline{X}} = (\underline{\underline{E}}^T\underline{\underline{D}}\,\underline{\underline{F}}^T)(\underline{\underline{F}}\,\underline{\underline{D}}\,\underline{\underline{E}})$
$= \underline{\underline{E}}^T\underline{\underline{D}}^2\underline{\underline{E}}$

spectral decomposition → with $\lambda_1 = d_1^2, \lambda_2 = d_2^2 \ldots$

and $\underline{\underline{X}}\,\underline{\underline{X}}^T = \underline{\underline{F}}\,\underline{\underline{D}}\,\underline{\underline{E}}\,\underline{\underline{E}}^T\underline{\underline{D}}\,\underline{\underline{F}}^T = \underline{\underline{F}}\,\underline{\underline{D}}^2\underline{\underline{F}}^T$

So SVD is like the square root **of spectral decomposition.!**