LAIS, Lecture #19

Principal Component Analysis (PCA)

$X \in \mathbb{R}^{D \times N}$ data matrix

rows: features

columns: samples of some population

Ex 1 recommendation systems

columns: users

rows: products

X_{ij}: rating of product i by user j

Ex 2 biomedical

rows: genes

columns: patients

X_{ij}: expression level of gene i on patient j.

Rem 3 A plethora of such instances in science and engineering.

Rem 4 In many cases D or N or both can be very large. The need arises to reduce the size of the data without losing much information

Dfn 5

"dimensionality reduction"
Replace $\mathcal{X} \in \mathbb{R}^{D \times N}$ by $\mathcal{Y} \in \mathbb{R}^{d \times n}$ with $d \ll D$, $n \ll N$ s.t. $\mathcal{Y}$ contains almost same information as $\mathcal{X}$. $\boxdot$

Rem 6 Dimensionality reduction by linear projection. Let $\mathcal{S} \subseteq \mathbb{R}^D$ be a linear subspace of dimension $r < D$. Let $\pi_{\mathcal{S}} : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ be the orthogonal projection onto $\mathcal{S}$

$\pi_{\mathcal{S}} = \tau_{U_{\mathcal{S}} U_{\mathcal{S}}^T}$, $U_{\mathcal{S}}$: orthonormal basis of $\mathcal{S}$.

$\mathcal{X} = [x_1 \cdots x_N]$, $x_j \in \mathbb{R}^D$

$\mathcal{X}_{\mathcal{S}} = [\pi_{\mathcal{S}}(x_1) \cdots \pi_{\mathcal{S}}(x_N)] \in \mathbb{R}^{D \times N}$

$$\begin{tikzcd}
\mathbb{R}^D \arrow[r, "\pi_{\mathcal{S}}"] \arrow[dr, "\pi_{\mathcal{S}}", "U_{\mathcal{S}} U_{\mathcal{S}}^T"'] & \mathbb{R}^D \\
 & \mathcal{S} \arrow[u, hook] \arrow[r, "\sim", "U_{\mathcal{S}}^T"'] & \mathbb{R}^r
\end{tikzcd}$$

$$\mathcal{X} \longmapsto \underbrace{U_{\mathcal{S}} U_{\mathcal{S}}^T \mathcal{X}}_{\mathcal{X}_{\mathcal{S}}} \longmapsto U_{\mathcal{S}}^T \mathcal{X} = \mathcal{Y} \in \mathbb{R}^{r \times N}$$

error: $\|\mathcal{X} - \mathcal{X}_{\mathcal{S}}\|_F$

Dfn 7 [subspace version of PCA]

$$\min_{\substack{S \subseteq \mathbb{R}^D \\ \dim S \le r}} \sum_{j \in [N]} \| x_j - \Pi_S(x_j) \|_2^2 \quad (\alpha)$$

$$\Longleftrightarrow$$

$$\min_{\substack{S \subseteq \mathbb{R}^D \\ \dim S \le r}} \| \mathcal{X} - \mathcal{X}_S \|_F^2 \quad \boxed{\equiv}$$

Rem 8 In practice dimensionality reduction is meaningful because there are many patterns in the data $\Rightarrow \mathcal{X}$ is approximately low-rank $\boxed{\equiv}$

Dfn 9 [matrix version of PCA]

$$\min_{\substack{A \in \mathbb{R}^{D \times N} \\ \mathrm{rank}(A) \le r}} \| \mathcal{X} - A \|_F^2 \quad (\beta)$$

$$\Longleftrightarrow$$

$$\min_{\substack{B \in \mathbb{R}^{D \times r} \\ G \in \mathbb{R}^{r \times N}}} \| \mathcal{X} - BG \|_F^2 \quad \boxed{\equiv}$$

Prp 10 $(\alpha) \Longleftrightarrow (\beta)$ $\boxed{\equiv}$

Rem 11 $\mathcal{X}_S = \underbrace{U_S}_{B} \underbrace{U_S^T \mathcal{X}}_{G}$ $\boxed{\equiv}$

$BG = Q\underbrace{RG}_{G'}$

$BG = \underline{U_B} \underbrace{\Sigma_B V_B^T G}_{G'}$

Lem 12

U: or
$\in \mathbb{R}^{D \times r}$
projec
Then
orthog
Prf
is a
i) "e
bas
an
[U
[U

$\|_2^2$

of PCA]

$$\min_{\substack{A\in\mathbb{R}^{D\times N}\\ \operatorname{rank}(A)\le r}} \|X-A\|_F^2 \quad (b)$$

$$\Longleftrightarrow$$

$$\min_{\substack{B\in\mathbb{R}^{D\times r}\\ G\in\mathbb{R}^{r\times N}}} \|X-BG\|_F^2$$

$\blacksquare$

Prp 10 $(a)\Longleftrightarrow(b)$ $\blacksquare$

Rem 11 $X_S = \underbrace{U_S}_{B}\underbrace{U_S^T X}_{G}$ $\blacksquare$

$BG = Q\underbrace{RG}$

$BG = \underline{U_B}\,\Sigma_B\underbrace{V_B^T G}_{G'}$

Lem 12 $S\subseteq\mathbb{R}^D$, $\dim S = r$

U: orthonormal basis of S $\in\mathbb{R}^{D\times r}$ UU^T: orthogonal projection onto S.

Then $I-UU^T$ is the orthogonal projection onto $S^\perp$.

Prf i) "advanced" $UU^T+(I-UU^T)=I$ is an orthogonal resolution of the identity

ii) "elementary" Extend U to an orthonormal basis $[U\ V]\in\mathbb{R}^{D\times D}$ of $\mathbb{R}^D$. Then V is an orthonormal basis of $S^\perp$. Since $[U\ V]$ is orthonormal

$$[U\ V]\begin{bmatrix}U^T\\ V^T\end{bmatrix} = I_D.\ \blacksquare$$

LA IS, Lecture #19

Lem 13

$A \in \mathbb{R}^{D\times D}$, symmetric

$\lambda_1(A) \geq \cdots \geq \lambda_D(A)$

$$\sum_{j=r+1}^{D} \lambda_j(A) = \min_{\substack{V \in \mathbb{R}^{D\times(D-r)} \\ V^T V = I_{D-r}}} \mathrm{Trace}[V^T A V]$$

and the solution is given by $V^* = U_{:,(r+1:D)}$ and $A = U \Lambda U^T$ is the eigendecomposition of A

$\Lambda = \mathrm{diag}(\lambda_1(A), \ldots, \lambda_D(A))$. $\boxed{\Xi}$

Thm 14 [PCA]

Let $\mathcal{X} = U_{\mathcal{X}} \Sigma_{\mathcal{X}} V_{\mathcal{X}}^T$ be the thin SVD of $\mathcal{X}$.

Then the solution to (α) is $S^* = \mathcal{B}(U_{\mathcal{X},[:,(1:r)]})$

the leftmost $D\times r$ submatrix of $U_{\mathcal{X}}$

and the dimensionality reduction error is

$$\sum_{j=r+1}^{\min\{D,N\}} \sigma_{\mathcal{X},j}^2 = \|\mathcal{X} - \mathcal{X}_{S^*}\|_F^2$$

Prf $\|\mathcal{X}-\mathcal{X}_{S}\|_F^2 =$

$$= \|\mathcal{X} - U_S U_S^T \mathcal{X}\|_F^2$$

$$= \|(I - U_S U_S^T)\mathcal{X}\|_F^2$$

$$\overset{\text{Lem 12}}{=} \|U_{S^\perp} U_{S^\perp}^T \mathcal{X}\|_F^2$$

$$\overset{\|B\|_F^2 = \mathrm{Trace}[B^T B]}{=} \mathrm{Trace}[U_{S^\perp} U_{S^\perp}^T \mathcal{X}\mathcal{X}^T U_{S^\perp} U_{S^\perp}^T]$$

$$\overset{\mathrm{Trace}[AB] = \mathrm{Trace}[BA]}{=} \mathrm{Trace}[\underbrace{U_{S^\perp}^T}_{V^T} \underbrace{\mathcal{X}\mathcal{X}^T}_{A} \underbrace{U_{S^\perp}}_{V}]$$

$$\mathcal{X} = U_{\mathcal{X}} \Sigma_{\mathcal{X}} V_{\mathcal{X}}^T \Rightarrow$$

$$\mathcal{X}\mathcal{X}^T = U_{\mathcal{X}} \Sigma_{\mathcal{X}} \Sigma_{\mathcal{X}}^T U_{\mathcal{X}}^T$$

is an eigendecomposition of $\mathcal{X}\mathcal{X}^T$.

By Lem 13 the optimal solution to

$$\min_{\substack{S^\perp \subseteq \mathbb{R}^D \\ \dim S^\perp = D-r}} \mathrm{Trace}[U_{S^\perp}^T \mathcal{X}\mathcal{X}^T U_{S^\perp}]$$

is given by $U_{S^\perp} = U_{\mathcal{X},[:,(r+1:D)]}$

$$\Rightarrow U_{S^*} = U_{\mathcal{X},[:,(1:r)]}. \quad \square$$

Dfn 15 [Robust PCA] ← unsupervised machine learning

$$\underbrace{\mathcal{X}^*}_{\text{data model}} \in \mathbb{R}^{D\times N},\ \operatorname{rank}(\mathcal{X}^*) = r < \min\{D,N\}$$

given data $\tilde{\mathcal{X}}$: corrupted version of $\mathcal{X}^*$

goal: extract $\mathcal{X}^*$ or $\mathcal{S}^* = \mathcal{B}(\mathcal{X}^*)$ from $\tilde{\mathcal{X}}$.

Dfn 16 [Data corruptions]

i) additive noise $\tilde{\mathcal{X}} = \mathcal{X}^* + E$ ← dense matrix with small arbitrary entries

(just do PCA)

ii) sparse noise at unknown locations

$$\tilde{\mathcal{X}} = \mathcal{X}^* + E \leftarrow \text{sparse}$$

iii) outliers:

$$\tilde{\mathcal{X}} = [\mathcal{X}^*\ \mathcal{O}]\underbrace{\Pi}_{\text{unknown permutation}}$$

columns lie away from $\mathcal{S}^*$

iv) missing entries

$$\tilde{\mathcal{X}} = \mathcal{X}^* \odot \Omega$$

Ω: (0,1)-matrix "mask"

$\odot$: Hadamard product (entry by entry)

also known as "low-rank matrix completion"

LAIS, Lecture #19

"Low-rank matrices"

Dfn 17 $M(D\times N, r) = \{X \in \mathbb{R}^{D\times N} : \mathrm{rank}(X) \leq r\}$

$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

Thm 14

Let X

be the

Then t

is $S^* =$

and th

reduct

$\sum_{j=r+1}^{\min\{D,N\}} \sigma$