MATH466/MATH766
Math of machine learning
03/17 Lecture 15 PCA MDS

References:

- Ch14.5, 14.8 of The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman

Todays contents:

- Principal component analysis (PCA)
- Multidimensional scaling (MDS)

Important concepts:

- principal components
- singular value decomposition
- kernel

Recommend reading:

-

Warm up

1. Let I_n be $n\times n$ identity matrix, 1_n be $n\times 1$ vector with all entries 1, $J := I_n - \frac{1}{n} 1 1^T$

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_n^T \end{bmatrix} \in \mathbb{R}^{n\times m} \quad \text{then} \quad JA = \_\_\_\_\_$$

$$B = [b_1\ b_2\ \cdots\ b_n] \in \mathbb{R}^{m\times n} \qquad BJ = \_\_\_\_\_$$

2. Let $\{v_1, v_2, \cdots, v_d\} \subseteq \mathbb{R}^D$ be an orthonormal basis, $x \in \mathbb{R}^D$

the projection of x to span $\{v_1, v_2, \cdots, v_d\}$ is _____

$\tilde{x} = \sum_{i=1}^{d} y_i v_i$ $(y_i \in \mathbb{R})$. then $\|\tilde{x}\|_2^2 =$ _____

3. $A \in \mathbb{R}^{m\times n}$, the SVD of A is _____

the eigen-decomposition of A^TA is _____

of AA^T is _____

Multidimensional Scaling (MDS)

"Preserve" pairwise dissimilarity

Data $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^D$

dissimilarity $\delta_{ij} = \text{dist}(x_i, x_j)$ e.g. $\|x_i - x_j\|_p$

Let the d-dim representation be $\{y_i\}_{i=1}^n$, $y_i \in \mathbb{R}^d$

transformed distance $s_{ij} = \|y_i - y_j\|_2$

Want: $s_{ij} \approx \delta_{ij}$

only need to know the pairwise distances

- metric MDS

$$\min_{\{y_i\}_i} \text{stress}^2(\{y_i\}_i) := \sum_{i<j} (s_{ij} - \delta_{ij})^2$$

- non-metric MDS

$$\min_{\{y_i\}_i} \text{stress}^2 := \frac{\sum_{i<j} (f(\delta_{ij}) - s_{ij})^2}{\sum_{i<j} s_{ij}^2}$$

for some monotonically increasing f

$f(\delta_{ij})$ is interpreted as the transformed dissimilarity.

that preserves the order / ranking of the dissimilarities

- stress is non-convex in $\{y_i\}_i$ and does not have explicit solution in general
- in practice, optimized by an algorithm known as "stress majorization"

- classical MDS "preserve" pairwise inner product

Let $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$ be centered

i.e. $\sum_{i=1}^n x_i = 0$, $\sum_{i=1}^n y_i = 0$

$\left(\frac{X = [x_1, \cdots, x_n]}{Y = [y_1, \cdots, y_n]} \right)$ ↓ different from notation in supervised learning

$\|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2 x_i^T x_j$

Denote $A \in \mathbb{R}^{n \times n}$, $A_{ij} = \|x_i - x_j\|^2$

$G \in \mathbb{R}^{n \times n}$, $G_{ij} = x_i^T x_j$ $(G = X^T X)$

$a \in \mathbb{R}^{n \times 1}$, $a_i = \|x_i\|^2$

then $A = a 1^T + 1 a^T - 2G$

$JAJ = J(a1^T + 1a^T - 2G)J$

$= 0 + 0 - 2 JX^T XJ = -2G$ (b/c $\{x_i\}$ are centered)

$\Rightarrow G = -\frac{1}{2} JAJ$

When both $\{x_i\}_i$, $\{y_i\}_i$ are centered

$\|y_i - y_j\| \approx \|x_i - x_j\| \Rightarrow y_i^T y_j \approx x_i^T x_j$

"preserve" pairwise inner product well

is "necessary" to "preserving" pairwise distance well

Classical MDS: $\min_{\{y_i\}_{i=1}^n} \sum_{i=1}^n \sum_{j=1}^n (y_i^T y_j - x_i^T x_j)^2$

$\min_Y \|Y^T Y - X^T X\|_{Fro}^2$

Denote the eigen-decomposition of $X^T X$ be $V \Lambda V^T$

the best $Y = \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \sqrt{\lambda_2} & \\ & & \ddots \\ & & & \sqrt{\lambda_d} \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_d^T \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} v_1^T \\ \sqrt{\lambda_2} v_2^T \\ \vdots \\ \sqrt{\lambda_d} v_d^T \end{bmatrix}$, $y_i = \begin{bmatrix} \sqrt{\lambda_1} v_{1i} \\ \sqrt{\lambda_2} v_{2i} \\ \vdots \\ \sqrt{\lambda_d} v_{di} \end{bmatrix}$

Principal Component Analysis (PCA)

$\{x_i\}_{i=1}^n \subseteq \mathbb{R}^D$, want a low-dim representation $\{y_i\}_{i=1}^n \subseteq \mathbb{R}^d$ $(d < D)$

1. minimize reconstruction error

Want to find a d-dimensional space

$$\mu + W_d \sim \mu + \text{span}\{w_1, \dots, w_d\}$$ o.n basis

that minimize the distance b/t x_i and its projection to $\mu + W_d$

$$\min_{\substack{\mu, \{w_1,\dots,w_d\}\text{ o.n.} \\ \{y_{ij}\}_{i=1,j=1}^{n,d}}} \sum_{i=1}^n \| x_i - (\mu + \sum_{j=1}^d y_{ij} w_j) \|^2$$

① the best $\mu = \bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$

WLOG, assume $\{x_i\}$ are centered, i.e. $\bar{x} = 0$

② the best $y_{ij} = x_i^T w_j$

③ the best w_j principal components

$$\min_{\{w_1,\dots,w_d\}\text{ o.n.}} \sum_{i=1}^n \| x_i - \sum_{j=1}^d (x_i^T w_j) w_j \|^2$$

Let $W := [w_1, w_2, \dots, w_d]$,

$X := [x_1, x_2, \dots, x_n]$ different from notation in supervised learning

$$\sum_{j=1}^d (x_i^T w_j) w_j = W \begin{bmatrix} w_1^T x_i \\ w_2^T x_i \\ \vdots \\ w_d^T x_i \end{bmatrix} = W W^T x_i$$

$$\min_{W \text{ o.n.}} \| X - \underbrace{WW^T X}_{\tilde{X}} \|_{Fro}^2 \qquad \text{rank}(\tilde{X}) = d$$

Let the SVD of X be $U \Sigma V^T$.

$U \in \mathbb{R}^{D \times D}$, $V \in \mathbb{R}^{n \times n}$

$\Sigma \in \mathbb{R}^{D \times n}$, $\Sigma = \begin{bmatrix} \sigma_1 & & & 0 \\ & \ddots & & \vdots \\ & & \sigma_D & 0 \end{bmatrix}$ $(D \le n)$ or $\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ \hline & 0 & \end{bmatrix}$ $(D > n)$

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_D \ge 0$$

$$X = \sum_{j=1}^{\min(D,n)} \sigma_j u_j v_j^T$$

A candidate of $\tilde{X}$: the best rank-d approximation $\sum_{j=1}^{d} \sigma_j u_j v_j^T$

this $\tilde{X}$ can be achieved by $W = [u_1, u_2, \cdots, u_d]$

$\Rightarrow$ Principal components $w_j = u_j$ $(j = 1, 2, \cdots, d)$

d-dim representation of x : $y = \begin{bmatrix} w_1^T x \\ w_2^T x \\ \vdots \\ w_d^T x \end{bmatrix} = W^T x$

of x_i : the i-th column of $W^T X = \begin{bmatrix} \sigma_1 v_1^T \\ \sigma_2 v_2^T \\ \vdots \\ \sigma_d v_d^T \end{bmatrix}$

$$y_i = \begin{bmatrix} \sigma_1 v_{1i} \\ \sigma_2 v_{2i} \\ \vdots \\ \sigma_d v_{di} \end{bmatrix}$$

PCA = (truncated) SVD

Q: What if the data is intrinsically nonlinear ?

2. Maximize projected norm

WLOG, assume $\{x_i\}_{i=1}^n$ is centered

If we pick o.n. basis $\{w_1, \cdots, w_d\}$

the projection of x_i to span $\{w_1, \cdots, w_d\}$ is

$$\tilde{x}_i := \sum_{j=1}^{d} (x_i^T w_j)\, w_j$$

$$\|\tilde{x}_i\|_2^2 = \sum_{j=1}^{d} (x_i^T w_j)^2$$

$$\max_{\{w_1,\cdots,w_d\}\ \text{o.n.}} \sum_{i=1}^{n} \sum_{j=1}^{d} (x_i^T w_j)^2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{d} w_j^T x_i\, x_i^T w_j$$

$$= \sum_{j=1}^{d} w_j^T \underbrace{\left(\sum_{i=1}^{n} x_i x_i^T \right)}_{XX^T} w_j$$

$$\max_{\{w_1,\cdots,w_d\}\ \text{o.n.}} \sum_{j=1}^{d} w_j^T \underbrace{XX^T}_{D\times D\ \text{matrix}} w_j$$

According to lecture 14,

w_j = eig-vec of XX^T corresponding to the j-th largest eig-val

(ex). check 1 $\Leftrightarrow$ 2.

3. Best low-rank estimation of the covariance matrix

Assume $x_i \sim_{iid} P(\mu, \Sigma)$

Covariance matrix : Σ

Sample covariance matrix : $S := \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$

WLOG, assume $\mu = 0$, $\sum_{i=1}^{n} x_i = 0$

$$S = \frac{1}{n} \sum_{i=1}^{n} XX^T = \frac{1}{n} \sum_{j=1}^{r(S)} \lambda_j(XX^T) \psi_j(XX^T) \psi_j^T(XX^T)$$

Projection of x_i : $\tilde{x}_i = \sum_{j=1}^{d} (x_i^T w_j) w_j$

Covariance matrix of projected samples :

$$S_d := \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} (x_i^T w_j)^2 w_j w_j^T$$

$$= \frac{1}{n} \sum_{j=1}^{d} \left(\sum_{i=1}^{n} (x_i^T w_j)^2 \right) w_j w_j^T$$

$\text{rank}(S_d) = d$

PCA solves $\min_{S_d} \| S_d - S \|_{Fro}^2$

> PCA gives the low-rank approximation of the sample covariance matrix.

4. In Euclidean space

Classical MDS = PCA (ex. check this)