PCA algorithms proof

Friday, June 26, 2020 1:47 AM

So the problem ~~is~~ now

$$c^* = \arg\min_c \left(-2x^T D c + c^T c\right)$$

As our objective function is allways increasing then we can solve using derivatives that means

$$\nabla_c \left(-2x^T D c + c^T c\right)$$
$$= -2D^T x + 2c$$

as c^* is the value where our objective function reach the minimum then

$$-2D^T x + 2c^* = 0$$
$$2c^* = 2D^T x$$
$$c^* = D^T x$$

$\Rightarrow$ $f(x) = c = D^T x$ $\rightarrow$ coding function

$g(c) = Dc$ $\rightarrow$ decoding function

$r(x) = g(f(x)) = DD^T x$ $\rightarrow$ reconstruction function

Now we need to know what is the matrix $\boxed{D}$ knowing the matrix D we can map each $x^{(i)}$ to $c^{(i)}$, $\mathbb{R}^n \rightarrow \mathbb{R}^e$ with f and reconstruct the value again with an small error

As is the same Matrix D for all $x^{(i)}$ in $\{x^{(1)}, \ldots, x^{(n)}\}$ then we will no longer make the analysis with single $x^{(i)}$. Instead let's define

$$X = \begin{pmatrix} x^{(1)T} \\ \vdots \\ x^{(m)T} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

$X = (x^{(m)T}) \cdots$

$D \in \mathbb{R}^{n \times l}$

<u>before</u>

$$f(x) = D^T x \quad D \in \mathbb{R}^{n\times l},\ x \in \mathbb{R}^{n\times 1} \Bigg\} \begin{pmatrix} d^{(1)T} \\ \vdots \\ d^{(l)T} \end{pmatrix} x = \begin{pmatrix} d^{(1)T} x \\ d^{(2)T} x \\ \vdots \\ d^{(l)T} x \end{pmatrix}_{l \times 1}$$

$\Rightarrow d^{(1)T} x$ is a dot vector product

$d^{(1)} \in \mathbb{R}^{l \times 1}$

$x \in \mathbb{R}^{n \times 1}$

<u>Now</u>

$f(X) = XD \qquad X \in \mathbb{R}^{m\times n},\ D \in \mathbb{R}^{n \times l}$

$C = f(X) \in \mathbb{R}^{m \times l}$

$$\begin{pmatrix} x^{(1)T} \\ \vdots \\ x^{(m)T} \end{pmatrix}_{m\times n} \begin{pmatrix} d^{(1)} & \cdots & d^{(l)} \end{pmatrix}_{n \times l} = \begin{pmatrix} x^{(1)T} d^{(1)} & \cdots & x^{(1)T} d^{(l)} \\ x^{(2)T} d^{(1)} & \cdots & x^{(2)T} d^{(l)} \\ \vdots & & \vdots \\ x^{(m)T} d^{(1)} & \cdots & x^{(m)T} d^{(l)} \end{pmatrix}$$

$$= \begin{pmatrix} f(x_1)^T \\ \vdots \\ f(x^{(m)})^T \end{pmatrix} = \begin{pmatrix} x^{(1)T} D \\ \vdots \\ x^{(m)T} D \end{pmatrix}$$

We can see that each row is the corresponding mapping for $x^{(i)}\ \forall i \in \{1, \dots, m\}$
each row is $c^{(i)}\ \forall i \in \{1, \dots, m\}$.

$f(X) = XD = C$ where C is the reduce Matrix of X

Now the decode function for all C will be

$g(c^{(i)}) = Dc^{(i)} \qquad D \in \mathbb{R}^{n \times l} \quad c^{(i)} \in \mathbb{R}^{l \times 1}$

$= c^{(i)T} D^T$

$$\Rightarrow \begin{pmatrix} c^{(1)T} \\ \vdots \end{pmatrix} D^T \quad \begin{pmatrix} c^{(1)T} D^T \\ \vdots \end{pmatrix} = \begin{pmatrix} g(c^{(1)})^T \\ \vdots \end{pmatrix}$$

$$\begin{pmatrix} \vdots \\ c^{(m)T} \end{pmatrix} D^T = \begin{pmatrix} \vdots \\ c^{(m)T} D^T \end{pmatrix} \begin{pmatrix} \vdots \\ g(c^{(m)})^T \end{pmatrix}$$

$$\Rightarrow g(C) = CD^T \quad \Rightarrow r(X) = XDD^T$$

Now that we have our mapping configure to use over all of $x^{(i)}$ in the collections, to get the value of D we need to optimize the Frobenius Norm of the difference between X and $r(X)$

$$D^* = \underset{D}{\arg\min} \| X - XDD^T \|_F \qquad D^T D = I_l$$

$$\Rightarrow \quad = \underset{D}{\arg\min} \| X - XDD^T \|_F^2$$

$\|A\|_F^2 = Tr(A^T A) \Rightarrow$

$(ABC)^T = C^T B^T A^T$

$$= \underset{D}{\arg\min}\, Tr\Big(\underbrace{(X - XDD^T)^T (X - XDD^T)}_{①} \Big) \qquad ①$$

$$① \; (X - XDD^T)^T (X - XDD^T) = X^T X - X^T X DD^T - DD^T X^T X + DD^T X^T X DD^T$$

Reemplaze in ①

constant

$$\Rightarrow \underset{D}{\arg\min}\, \cancel{Tr(X^T X)} - Tr(X^T X DD^T) - Tr(DD^T X^T X) + Tr(DD^T X^T X DD^T)$$

$Tr(A+B)$
$= Tr(A) + Tr(B)$

if AB and BA are possibles
$\Rightarrow Tr(AB) = Tr(BA)$

$$\Rightarrow \underset{D}{\arg\min} - Tr(X^T X DD^T) - Tr(X^T X DD^T) + Tr(X^T X DD^T DD^T)$$

remember $D^T D = I_l$

$$\Rightarrow \underset{D}{\arg\min} - 2Tr(X^T X DD^T) + Tr(X^T X DD^T)$$

$$\Rightarrow \underset{D}{\arg\min} \left[- Tr(X^T X DD^T) \right]$$

one last rearragment.

remember that

$X \in \mathbb{R}^{m\times n}$

$D \in \mathbb{R}^{n\times e}$

$$\Rightarrow \underset{D}{\operatorname{argmax}} \operatorname{Tr}(\underbrace{D^T X^T X D}_{(2a)})$$

let Analyze (2a)

$D^T X^T X D$

$\Rightarrow$ $X^T X$ is a symmetric Matrix

$(X^T X)^T = X^T X^{T^T} = X^T X$ ✓

$\Rightarrow$ the eigenvalue decomposition is

$X^T X = U \Lambda U^T$ where U is a orthogonal matrix with each column is the eigenvector and Λ is a diagonal Matrix with the singular values

So if we create a function

$u(X^T X) = U^T X^T X U$

with U having the properties of D the maximum value for the trace correspond to the sum of maximum eigenvalues of the matrix $X^T X$, so and idea is that

$\underset{D}{\operatorname{Argmax}} \operatorname{Tr}(D^T X^T X D)$ proof by induction on e number of columns of D.

if $e = 1 \Rightarrow D = d \in \mathbb{R}^{n\times 1}$

$\underset{d}{\operatorname{Argmax}} \operatorname{Tr}(d^T X^T X d)$ $d^T d = 1_e$

$\Rightarrow$ the best value for d is the eigenvector of $X^T X$ corresponding to the largest eigenvalue of $X^T X$

Suppose for $e \Rightarrow D \in \mathbb{R}^{n\times e}$ (h.i)

$\operatorname{Argmax} [\operatorname{Tr}(D^T X^T X D)]$ $D^T D = 1_e$

D

it holds that the best value for D are the corresponding eigenvectors for the largest l eigenvalues of X^TX.

Proof for $l+1 \Rightarrow D \in \mathbb{R}^{n \times (l+1)}$

$\underset{D}{\text{Argmax}}\ Tr(D^TX^TXD)$

using hypothesis induction the first l columns need to be the first eigenvectors corresponding to the first eigenvalue

$$\begin{pmatrix} d^{(1)T} \\ \vdots \\ d^{(l+1)T} \end{pmatrix}_{(l+1)\times n} \begin{pmatrix} x^{(1)} & \dots & x^{(m)} \end{pmatrix}_{n \times m}$$

$$tr\left[\begin{pmatrix} d^{(1)T}x^{(1)} & \dots & d^{(1)T}x^{(m)} \\ \vdots & \ddots & \\ d^{(l+1)T}x^{(1)} & & d^{(l+1)T}x^{(m)} \end{pmatrix}_{(l+1)\times m} \begin{pmatrix} x^{(1)T}d^{(1)} & \dots & x^{(1)T}d^{(l+1)} \\ \vdots & \ddots & \vdots \\ x^{(m)T}d^{(1)} & & x^{(m)T}d^{(l+1)} \end{pmatrix}_{m \times (l+1)}\right]$$

$$= \Big(\underbrace{\overbrace{d^{(1)T}X^TXd^{(1)}}^{\lambda^{(1)}} + \dots}_{l} + d^{(l+1)T}X^TXd^{(l+1)}\Big) \quad \lambda^{l}\dots\lambda^{l}$$

where $\lambda^{(1)} > \lambda^2 > \lambda^3 > \dots > \lambda^{(l)} \rightarrow (h.1)$

= the first l is the corresponding Matrix $D' \in \mathbb{R}^{n \times l}$ where each $d^{(1)} \dots d^{(l)}$ are the eigenvectors corresponding to the first l largest eigenvalues then

$d^{(l+1)}$ must be the $l+1$ eigenvector corresponding to the $(l+1)$ eigenvalue of X^TX to maximize the trace.