

**Deep Learning**  
**Prof. Mitesh M. Khapra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Madras**

**Module – 7.2**

**Lecture – 07**

**Link between PCA and Auto encoders**

So, we will move to the next module where I would like to show you a Link between PCA and Auto encoders.

(Refer Slide Time: 00:20)

PCA

$P^T X^T X P = D$

- We will now see that the encoder part of an autoencoder is equivalent to PCA if we
  - use a linear encoder ✓
  - use a linear decoder ✓
  - use squared error loss function ✓
  - normalize the inputs to

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left( x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$

NPTEL Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 7

So, this is what I am trying to show you that under certain conditions, PCA is or rather an auto encoder is equivalent to a PCA. And the conditions are; if you use a linear encoder, if you use a linear decoder, if you use a squared error loss function and if you normalize the inputs to this. So, for the time being, just ignore the last bullet. Let us look at the other 3 bullets using squared error loss functions.

So, remember I gave you different choices right? You could have used the cross entropy or the squared error loss. But I am going to prove this equivalence only under the condition when we have the squared error loss. What do I mean the encoder is a linear encoder?  $g$  is a linear function. We are not using a sigmoid or any logistic or anything like that and linear decoder. Again the same thing; we are not using the sigmoid or soft

max or anything at the output; it is a linear function. Under these conditions, I will show that or I will try to show you that PCAs equal auto encoders equal to PCA.

What does this mean actually? Ok, now what do I mean by it is equivalent? What do I have to show you actually? How many of you understand what I am trying to prove? How many of you can mathematically define it? Ok, so, we will try to make this clear over the next 15 minutes.

(Refer Slide Time: 01:43)

• First let us consider the implication of normalizing the inputs to

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left( x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$

• The operation in the bracket ensures that the data now has 0 mean along each dimension  $j$  (we are subtracting the mean)

• Let  $X'$  be this zero mean data matrix then what the above normalization gives us is  $X = \frac{1}{\sqrt{m}} X'$

• Now  $(X)^T X = \frac{1}{m} (X')^T X'$  is the covariance matrix (recall that covariance matrix plays an important role in PCA)

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 7 17/55

First, let us look at the last condition right; which I ignored., I always anticipate all this right. So, I have full faith in you guys ok. What is this mean? Now, what I am doing? Centering the data and I am also doing 1 by square root of m; why? Mean, as the standard deviation [FL] fine.

So, the operation in the bracket ensures that your data now has become 0 centered right. It is a 0 mean. And now let  $X$  dash be this matrix this one right such that, all its elements are 0, mean is this still a flicker again alright.

So, let I am calling  $X$  dash as this matrix ok. So, this matrix, where I also have 1 by square root of m, I can write it as everyone gets this is simple. Now do you see where this is headed? What would  $X$  transpose  $X$  be? Covariance matrix; So, I needed that 1 by m right at the out.

So, now this is the co-variance matrix. So, if I do this normalization to the original data and then if I take let  $x$  dash be that quantity and then if I take  $x$  transpose  $x$  then I will get the co variance matrix everyone gets this that I did this to get the co variance matrix. So, that I mean I did this. So, that when I take  $x$  transpose  $x$  I get the co variance matrix after this normalization only it will be the covariance matrix.

(Refer Slide Time: 03:04)

• First we will show that if we use linear decoder and a squared error loss function then

• The optimal solution to the following objective function

$$\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \hat{x}_{ij})^2$$

is obtained when we use a linear encoder.

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 7

So, first we will show that, if we use the linear encoder, decoder and a squared error loss function, then the optimal solution to the following objective function. What does this objective function?

Student: Squared error.

Squared error loss is obtained when we use a linear encoder. Do you understand the implication of this? What does being stated here? Ok, So, I have fixed the decoder. I have said that the decoder is going to be a decoder. I have fixed the encoder or I have fixed the loss function. This is going to be a squared error loss function. This is given to me. Now under these conditions, I am trying to minimize this loss function ok.

Then I am telling you that the only solution to this is that the function dash should be a linear function which function? The function  $g$  should be a linear function you cannot choose sigmoid or logistic or anything else right? The optimal solution will occur when  $g$  is a linear function everyone gets what is being stated here ok?

(Refer Slide Time: 04:02)

So, this summation that I have written right or. In fact, this the entire objective that I have written is actually equivalent to this objective. Is this fine with everyone? Even though I have not defined what H is just fine with everyone. So, we had this X it was X 1 to X m ok. I had picked one of these Xs. What is the dimension of this?

Student: 1 cross m.

1 cross m and then I had multiplied it by a weight matrix W. Not W star; remember that what do the dimension of W.

Student: n n.

N cross k and what will I get as the output.

Student: (Refer Time: 04:58).

I got an H which was 1 cross k, what did I do this?

Student: Multiply it by

Multiply it by.

Student: W star.

W star which was k cross m and what did I get as the output?

Student:  $\hat{x}$ .

$\hat{x}$  which was  $1 \times n$  right. So, what I am telling you is that, I could do this together for all these  $X$ 's. I could do this operation at one go and I can call this as  $X$  matrix and what will I get here?  $h_1$  to  $h_2$  to  $h_m$ . And I can call it as the  $H$  matrix and I multiply it by  $W^*$  and what do I get?  $\hat{X}$  ok. It that fine ok, but without defining these things also it was fine. So, it does not matter ok.

(Refer Slide Time: 05:51)

• This is equivalent to

$$\min_{\hat{x}} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \hat{x}_{ij})^2 \quad (1)$$
$$\min_{W, H} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - HW^*)^2$$
$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$$
$$\hat{X} = U \Sigma V^T$$
$$\begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix} = \begin{bmatrix} \hat{x}_{11} & \dots & \hat{x}_{1n} \\ \vdots & \ddots & \vdots \\ \hat{x}_{m1} & \dots & \hat{x}_{mn} \end{bmatrix}$$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 7

So, now how many of you get that this quantity is the same as this quantity? Oh you get it? Ok fine, I thought you are answering why now what you it just obvious now how do I explain this was the Frobenius norm of a matrix? Some of the squares of the elements.

Now, what is the matrix  $X$ ? It is the  $X_{11}$  up to  $X_{1n}$  and  $X_{m1}$  up to  $X_{mn}$  and all elements in between right? What is the matrix  $H W^*$ ; we just did that the same thing expect that it is  $\hat{X}$ .

Student: (Refer Time: 06:04).

I take the difference between these 2 what do I get? Every element of that matrix is equal to this quantity that I have underlined right. So, I get a new matrix such that every element of that matrix is equal to this quantity. Is that fine? Now, I am taking the square of every element of that matrix and adding them up what is that equal to?

Student: (Refer Time: 06:15).

A Frobenius norm how many of you get that now? Almost everyone, ok. So, this is equivalent to the Frobenius norm ok. Now, where have you seen the Frobenius norm before what did we show in the SVD theorem?

Let us try to connect things right if you do not learn how to connect things it is going to be very difficult. What is this  $\hat{X}$ ? It is a dash of  $X$ .

Student: Reconstruction.

Reconstruction it is a dash of  $X$  approximation. What is the solution to this optimization problem? What is the solution to this optimization problem? I shall started off with the answer that we saw this in the SVD theorem and then I asked you a question what 30 hours 32 hours; not even 32 hours are passed since we did this. Come on, what is the solution to this? No, no that is fine.

But what is the solution  $\hat{X}$  is equal to what? The best approximation to  $X$  is given by what? Is it fine yeah, yeah. So, some  $k$  yeah, but it is going to come from the SVD theorem, right is that fine? It depends on what rank approximation you want, but it the best approximation to this is going to be given by the SVD of  $X$ , is it ok? Everyone gets that yes forgot about it, but now do you remember it all those extra lectures 8' O clock in the morning.

(Refer Slide Time: 08:26)

$$\min_{\hat{X}} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \hat{x}_{ij})^2 \quad (1)$$

- This is equivalent to

$$\min_{W \cdot H} (\|X - HW^*\|_F)^2 \quad \|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$$

(just writing the expression (1) in matrix form and using the definition of  $\|A\|_F$  (we are ignoring the biases)

- From SVD we know that optimal solution to the above problem is given by

$$HW^* = U_{:, \leq k} \Sigma_{k,k} V_{:, \leq k}^T$$

- By matching variables one possible solution is

$$\begin{aligned} H &= U_{:, \leq k} \Sigma_{k,k} \\ W^* &= V_{:, \leq k}^T \end{aligned}$$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 7

So, that means,  $H W^*$  should be equivalent to this that we know from the SVD theorem that, the optimal solution is going to be given by SVD. So, if I just compare terms ok, then I could write that one solution is this that  $H$ .  $H$  is equal to  $U$  into  $\sigma$  and  $W^*$  is equal to  $V^T$ . I could have chosen the other solution also where  $H$  is equal to  $V$  or sorry  $U$  and  $W^*$  is equal to  $\sigma V^T$ , ok. But I will work with this particular solution. You see, this I am just matching variables right? It is said that,  $A B$  is equal to  $C D E$ . So, I am saying that  $A$  is equal to  $C D$  and  $B$  is equal to  $E$ , is that fine? Ok.

Now, we will work with this. So, and we will try to show something; so, let us see what we are trying to show.

(Refer Slide Time: 09:14)

We will now show that  $H$  is a linear encoding and find an expression for the encoder weights  $W$

$$H = WX + \sigma(WX)$$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 7

Now, first thing that we will show is that  $H$  is actually a linear encoding. So, what does this mean? You first always understand what has been tried to prove right? I am saying that, I am going to show that  $H$  is a linear encoding of  $X$ , then what is it that I am trying to show?

I am trying to show that  $H$  is equal to a linear encoding of  $X$  when  $H$  is of the form  $W X$  and not something of the form  $W \text{ sigmoid of } W X$  or something like that or any other non-linearity for that matter. Is the statement clear? That is what I am trying to show. When I say  $H$  is a linear encoding, I mean that  $H$  is obtained by a linear transformation of  $X$ .

(Refer Slide Time: 09:52)

We will now show that  $H$  is a linear encoding and find an expression for the encoder weights  $W$

$$H = U_{:, \leq k} \Sigma_{k,k}$$

$$H = WX$$

Now,  $H$  as we defined on the previous slide is equal to this. Now, if I already had an  $X$  here, then I was done, but I do not have any  $X$  there yet. So, I want to get to a form where I can show that  $H$  is equal to  $W$  in to  $X$ . So, I will just do some simple trickery and arrive try to do arrive at that form.

(Refer Slide Time: 10:13)

We will now show that  $H$  is a linear encoding and find an expression for the encoder weights  $W$

$$H = U_{:, \leq k} \Sigma_{k,k}$$

$$= (XX^T)(XX^T)^{-1} U_{:, \leq k} \Sigma_{k,k} \quad (\text{pre-multiplying } (XX^T)(XX^T)^{-1} = I)$$

$$= (XV \Sigma^T U^T)(U \Sigma V^T V \Sigma^T U^T)^{-1} U_{:, \leq k} \Sigma_{k,k} \quad (\text{using } X = U \Sigma V^T)$$

$$= XV \Sigma^T U^T (U \Sigma \Sigma^T U^T)^{-1} U_{:, \leq k} \Sigma_{k,k} \quad (V^T V = I)$$

$$= XV \Sigma^T U^T U (\Sigma \Sigma^T)^{-1} U^T U_{:, \leq k} \Sigma_{k,k} \quad ((ABC)^{-1} = C^{-1} B^{-1} A^{-1})$$

$$= XV \Sigma^T (\Sigma \Sigma^T)^{-1} U^T U_{:, \leq k} \Sigma_{k,k} \quad (U^T U = I)$$

$$= XV \Sigma^T \Sigma^{T-1} \Sigma^{-1} U^T U_{:, \leq k} \Sigma_{k,k} \quad ((AB)^{-1} = B^{-1} A^{-1})$$

$$= XV \Sigma^{-1} I_{:, \leq k} \Sigma_{k,k} \quad (U^T U_{:, \leq k} = I_{:, \leq k})$$

$$= X V I_{:, \leq k}$$

$$H = X V_{:, \leq k}$$

Thus  $H$  is a linear transformation of  $X$  and  $W = V_{:, \leq k}$

So, the first thing I am going to do is pre multiplying pre multiply by this quantity and this is fair because this is just equal to  $I$  what next I will write these 3  $X$ s as  $U \Sigma V^T$  transpose and I will leave one  $X$  as it is that ok.



Now, just can you just try to see what the next step would be this  $V$  transpose  $V$  will disappear because it is equal to  $I$ . Now what happened here? I actually just expanded this inverse. So, I will think of this as  $A B C$ . So,  $A B C$  inverse is equal to  $C$  inverse  $B$  inverse  $A$  inverse.

So, I have just applied that it just that my inverse is a very straight forward matrices here they are just the transform of the original matrices. Everyone gets this step. Well you can stare at for a for a few more seconds if you want. How many of you do not get this? How many of you get this? Ok, now what is next this  $U$  transpose?  $U$  disappears.

Student: (Refer Time: 10:59).

This also disappears.

Student: (Refer Time: 11:03).

No.

Student: (Refer Time: 11:06).

It is this  $U$  is only. The first  $k$  columns of  $U$  right, this is not the entire  $U$ . This is just the first  $k$  columns of  $U$ , fine. Now what next  $A$  into  $B$  inverse is?

Student:  $B$  inverse.

$B$  inverse  $A$  inverse what will happen? Now that quantity will disappear. So, what do you have left now ok. So, this is something ok. So, now, let us look at this is let us say this is  $n$  cross  $n$  and this is  $n$  cross  $k$  what is the output going to be.

Student:  $n$  cross  $k$ .

$n$  cross  $k$  and what is the output going to look like is the first  $k$  columns of.

Student: Identity.

The identity matrix everyone gets that if you do not you can just work it out with the small matrix after going home and you will get it right if; so if I done the full multiplication, I would have got the identity matrix. But I am just talking the first  $k$

columns. So, I will get the first  $k$  columns of the identity matrix. Do not feed too much. If you are not getting this, you can just work it out on paper and you will get it.

So, I get the first  $k$  columns of the identity matrix and this inverse disappears this sigma transpose into sigma transpose (Refer Time: 12:19) now what next what is this product going to be the first  $k$  elements of.

Student: Sigma inverse.

Sigma inverse and that is going to get multiplied by sigma  $k$  cross  $k$ . So, that will give me the first  $k$  elements of.

Student: Identity.

Matrix there is some very simple matrix operations where you are just taking some columns right. So, if you do not understand this right. Now do not worry. You can work it out. Everyone is confident, they can do this, please raise your hands if you are confident. And now, what do I finally, get this multiplication will give me.

Student: The first  $k$  columns.

The first  $k$  columns of  $V$  ok; so, have we come to the desired form what I have shown. Now  $H$  is a dash of  $X$  or linear transformation of  $X$ ; that means, my optimal encoder was a linear encoder and what was the optimal weight matrix  $w$  the first  $k$  columns of  $V$  yeah I someone pointed it last time also I could not. I ignored it. I will just pretend I understood.

But I get it I know that there is a simpler solution. I do not know why do it this way, but there is a simpler solution. I just like making life miserable for you guys, but, but the point is, you can figure it out, that it is a it is a linear transformation of  $X$  now.

(Refer Slide Time: 14:00)

• We have encoder  $W = V_{:, \leq k}$

• From SVD, we know that  $V$  is the matrix of eigen vectors of  $X^T X$

• From PCA, we know that  $P$  is the matrix of the eigen vectors of the covariance matrix

• We saw earlier that, if entries of  $X$  are normalized by

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left( x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$

then  $X^T X$  is indeed the covariance matrix

• Thus, the encoder matrix for linear autoencoder ( $W$ ) and the projection matrix ( $P$ ) for PCA could indeed be the same. Hence proved

*Handwritten notes:*  $H = \frac{W}{\sqrt{m}} X$

*Handwritten scribble:*  ~~$H = \frac{W}{\sqrt{m}} X$~~

NPTEL | Mitesh M. Khapra | CS7015 (Deep Learning) : Lecture 7

We have that the encoder is equal to the first k columns of V ok. What is V eigenvectors of X transpose X ok?

Student: A.

What is the other thing that you know about the eigenvectors of X transpose X they are the solution for the.

Student: Eigen.

If you have given an matrix X then the PCA is the eigenvectors of the co variance matrix was the co variance matrix X transpose X what is are it is eigenvectors capital V right. So, what have we arrived at are we done with the proof. Yes, how many of you think that done with the proof? How many of you think that we are done now?

So, it is done right. So, we have proved what we wanted to prove right. So, what did we want to prove that you are doing auto encoders. You are trying to train an auto encoders and you are loss function is the squared error loss function. We saw a neat way of writing that squared error loss function as a matrix operation where X minus capital H into W.

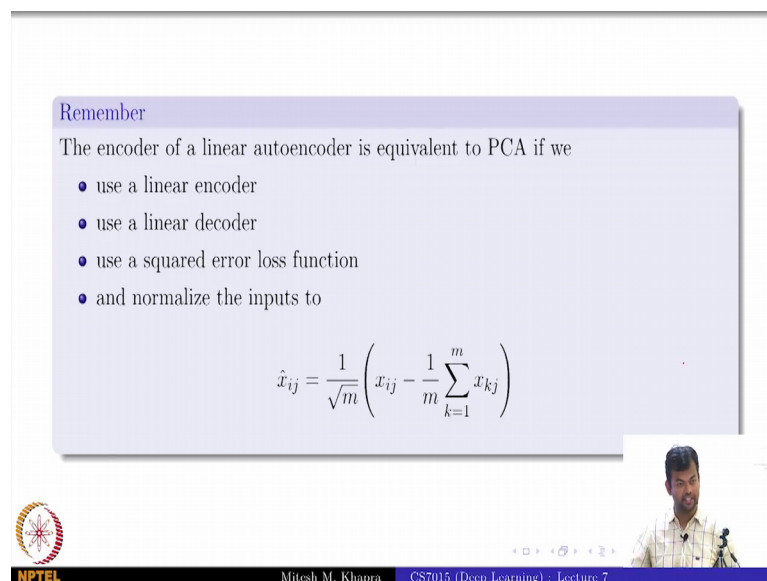
And then, we saw that these squared error loss function is nothing, but the Frobenius norm of this and we knew that the minima of this objective function the Frobenius norm of X minus H W would occur when S W is equal to SVD of X right? We started from

there and showed that H is actually a linear transformation of X and what was that linear transformation which matrix was used for the linear transformation V V. What is V? It is the eigenvectors of.

Student: X transpose.

X transpose X. So, what is happened in effect is that if I was trying to train my auto encoder with this objective function, the weights in my initial layer W would actually converge to v which are the eigenvectors of X transpose X; that means, the transformation that I have learnt this transformation which I have learnt is the same as a transformation that I have had learned using PCA. Because PCA would also have given me V into X where V was the eigenvectors of the co variance matrix and we just arrived at the same solution everyone gets it. Now we are done with the proof.

(Refer Slide Time: 16:12)



Remember

The encoder of a linear autoencoder is equivalent to PCA if we

- use a linear encoder
- use a linear decoder
- use a squared error loss function
- and normalize the inputs to

$$\hat{x}_{ij} = \frac{1}{\sqrt{m}} \left( x_{ij} - \frac{1}{m} \sum_{k=1}^m x_{kj} \right)$$

NPTEL

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 7

So, what we have proved is, under these specific conditions that the encoder of a linear auto encoder is linear auto encoder is equal to PCA if we use a linear decoder. If we use a squared error loss function and if we normalize the inputs to this and you understand why each of these steps was important why was the last step important.

Student: (Refer Time: 16:32).

Only then, we would have got the co variance matrix why was a step before that important because, only if it was the squared error loss we would have got that Frobenius

norm objective function right. And why was the linear decoder important again the same thing? Because  $X - HW$  we wanted it to be linear right is it fine.

So, you see why all these assumptions were important and under these conditions, we have proved that auto encoders E equivalent to PCA. How many of you are completely lost at this point? How many of you have followed 80 percent of what we have done? Ok.