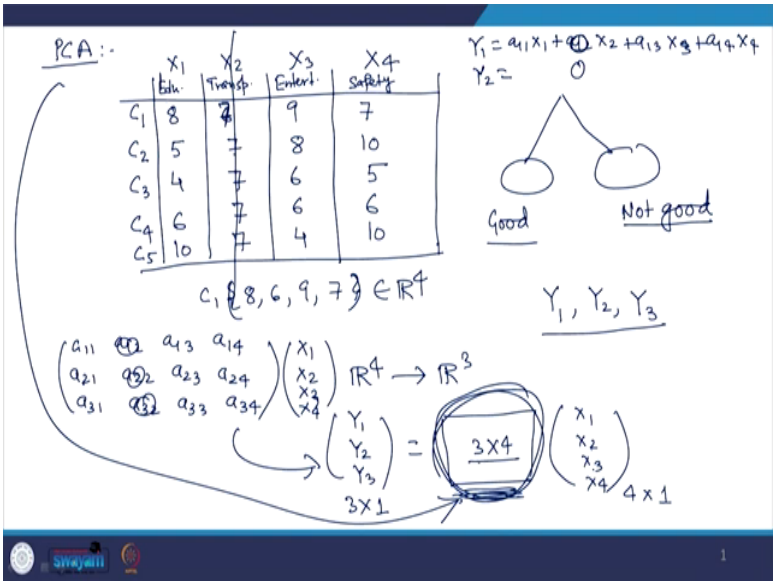
Essential Mathematics for Machine Learning

Prof. Sanjeev Kumar
Department of Mathematics
Indian Institute of Technology, Roorkee

Lecture – 16
Principal Component Analysis

Hello friends. So, welcome to the module 16 and this is the first lecture of week number 4 of this course Essential Mathematics for Machine Learning. And in this lecture, we will talk about a very popular and very important concept in machine learning for reducing the dimension of the data that is called Principal Component Analysis. So, it is a manifold learning technique and really very very applicable among machine learning research here and today, we will explore that what is this basically; how it will work on the data dimension reduction.

(Refer Slide Time: 01:11)



So, in sort I will write it PCA. So, here P stands for principal, C stands for component and A stands for analysis. Now, basically what is this PCA? So, take a very simple example. Suppose, I am having some data of four city. So, cities are C 1, C 2, C 3 and C 4 and then, what I am having? One parameter is education, transport, entertainment and the last one more parameter is safety that is law and order. So, suppose, I am having this kind of that for four cities or let us say for five cities.

So, let us I have graded all those parameters education, transport facility, entertainment opportunity and safety on a 10 pointy scale and let us say see for city C 1 it is 8, 6, 9 and 7. Like for C 2 city 2, all those parameters is let us say 5 7 8 10. Let us say for third city it is 4 7 6 and let us say 5. Then, for C 4, it is let us say something like some random value, let me take 6 7 6 6 and for C 5, let us say 10 7 4 and let us say 10.

So, now I have to classify all these 5 cities into two classes; one is good city for leaving, another one is not good based on these parameters. So, basically what I am having? I am having 5 cities and I am having the vector feature vector corresponding to each city, like the for city C 1 my vector is 8 6 9 7. So, it is a four-dimensional vector belongs to R4. It is a vector in R4. Similarly, for C 2, C 3, C 4 and C 5.

Now, I have to classify all those city based on these parameters. But what is happening here, suppose I do not want to classify with all four features. First I want to reduce this data into three features. So, instead of these four features, four attributes, education, transport, entertainment and safety, I want three features so that I can plot a three-dimensional plot, I can have for these all five city data and then, I can classify them by using some hyper plane in R3.

So, what I need? If I am saying this let us say F1, F2, F3 or let a better to write X1, X2, X3, X4. So, what I want? I want to go from R 4 to R 3 means I want Y1, Y2, Y3 and I want to apply it on X1, X2, X3 and X4. So, X1, X2, X3 and X4. So, what will be here? So, what you have to do? It is a 3 by 1, it is a 4 by 1. So, what you need here to get a 3 by 1 vector? So, you

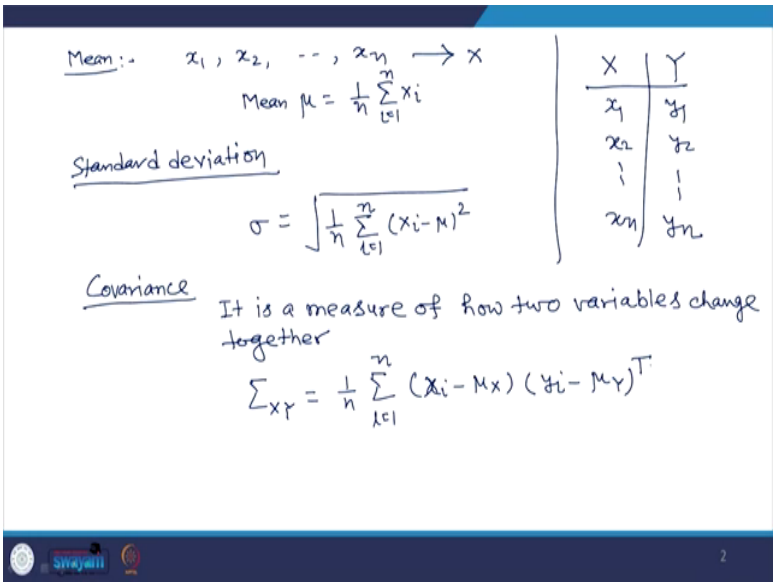
need a 3 by 4 matrix that is a matrix like a 11 a 12 a 13 a 14, a 21 a 22 a 23 a 24, a 31 a 32 a 33 a 34. So, that if I multiply it on to X1, X2, X3 and X4, I got this vector Y1, Y2, Y3.

So, what I want to say here? Y1 is nothing just linear combination of all these four feature. Similarly, Y2 will be a 21 into X1 plus a 22 into X2 plus a 23 into X3 plus a 24 into X4 and similarly, for Y3. So, how to find out this matrix? Because if I am having this matrix, then what will happen? I can transform my four-dimensional data set to a three-dimensional data set or suppose, I want to go to two-dimensional data set it will become a 2 by 4 matrix and what should be there? That the maximum information of the data should be preserve even though I transform to a lower dimensional space.

For example, here if I say for classification which of the feature vector is not having much information? The vector or the column in which I am having minimum variation. So, what is that column? This one. So, if it is also 7 here, then if you remove this column, then it will not make any because all the cities are having same value. So, it will not make any difference in the classification. So, that is the way of doing it. So, but here I want a linear combination. For example, here X2 this component will become 0 0. So, this column will become 0 0 0 here, in in that way I want.

So, I want to preserve maximum information; the same time, I want to reduce the dimension of the data. So, PCA principal component analysis is a tool for doing this kind of dimension reduction. Means, what is the objective of PCA? To find out this particular matrix, this 3 by 4 or whatever transformation matrix. So, that is the overall idea of PCA.

(Refer Slide Time: 08:26)



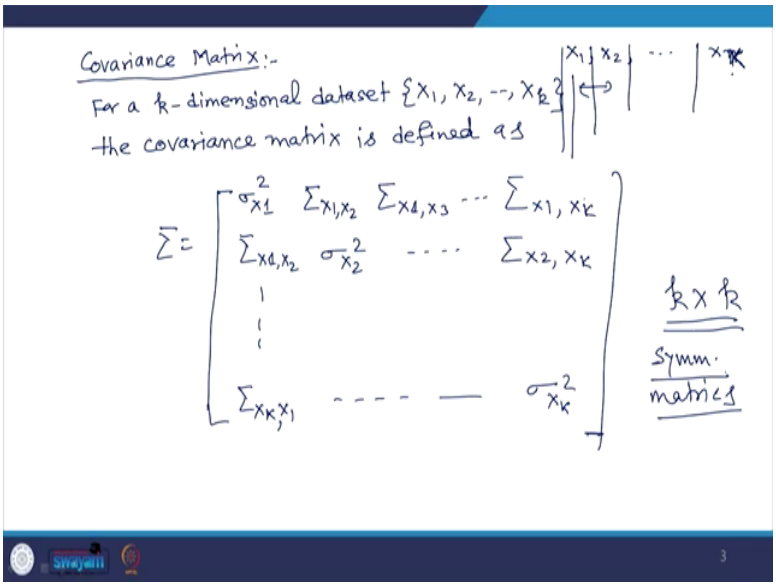
Now, some definitions which we need in PCA, the first definition is mean. So, as you know if you are having samples, let us say x1, x2 xn, then mean is a measure of central tendency and it is defined by mu and mean mu will be given as 1 by n, number of samples and then sum of all samples. Another important thing is standard deviation. So, again how to measure standard deviation? So, it will be sigma and sigma is given by square root 1 by n summation i equals to 1 to n xi minus mu square and what it will give? It will give that the measure of variability about the mean. Means how? My data is deviated about the mean.

The next one is covariance. So, this I am taking only one variable that is X; suppose, I am having two variables one is X, another one is Y. So, it is having value x1 x2 xn and it is having y 1 y 2 yn. So, it is a measure of how two variables change together and it is defined as sigma

XY like for this X and Y 1 by n because n number of samples i equals to 1 to n and then xi minus mu x yi minus mu y transpose.

So, it may be positive, it may be negative or it may be 0. So, if it is positive means the two data are changing in the same direction. If one is increasing, another one is also increasing. If one is decreasing, another one is also decreasing. If it is negative that is a covariance between two data, then the direction of the change are opposite to each other. Means, one is increasing, another one is decreasing and vice versa. If it is 0 means the two data are just independent of each other, you cannot comment anything about their behavior together.

(Refer Slide Time: 11:37)



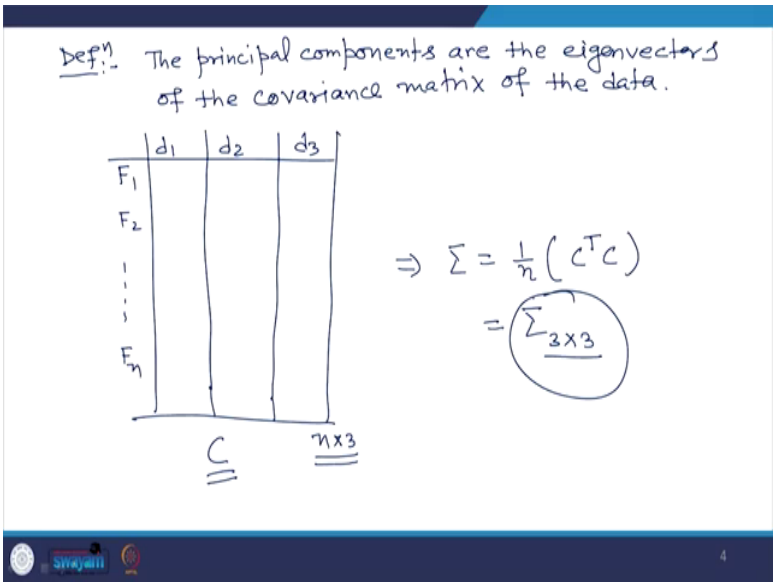
And the next one is very important that is covariance matrix. So, this particular matrix tells you, if you are having n-dimensional data in which direction it is having maximum variation. So, let us say you are having like this X1, X2, X n. So, n-dimensional data or let us say

k-dimensional data. So, for a k-dimensional data, data sets or columns are let us say X1, X2 X k. The covariance matrix is defined as sigma equals to the variance of data X1, means variance of this column..

Then, covariance between first and second column; then covariance between first and third column and then, in that way covariance between first and kth column. Then, this will be again covariance between second and first column which will be same because it is a scalar quantity, then variance of second column and that way finally, what I will be having summation X k, X1 means covariance between kth and first column which is similar to this one.

So, it will be a symmetric matrix because the covariance within first and second column equals to covariance between second and first column and similarly, for any two columns and finally, here you will be having variance of last column that is kth column. So, it will be having a k by k matrix and it is symmetric matrix. And if it is symmetric matrix, it will be having real eigenvalues and you will be having always orthogonal decomposition, means you can have orthogonal eigenvectors of this matrix. So, this is the covariance matrix.

(Refer Slide Time: 14:46)

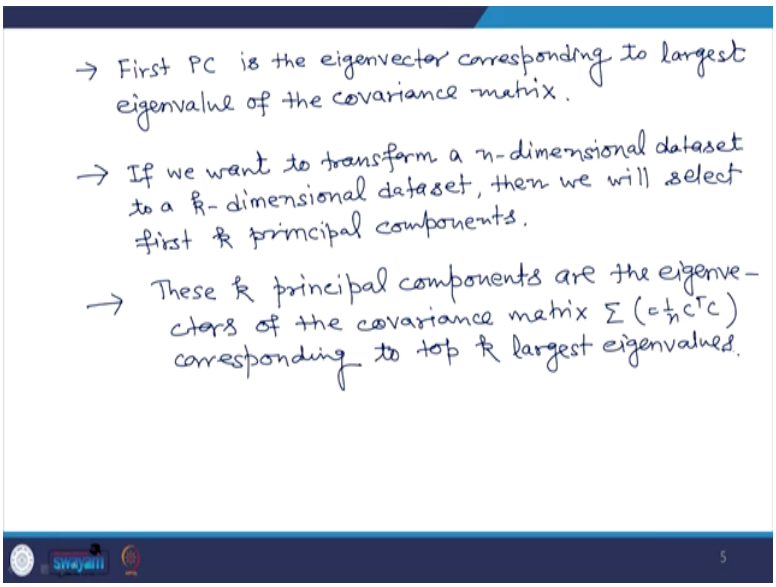


Now, another definition. The principal components are the eigenvectors of the covariance matrix of the data. So, for example, if you are having let us say data which is having for example, F1, F2, F n feature vectors. So, n feature vectors and each feature vectors is having 3 attributes; d 1, d 2, d 3. So, it is a three-dimensional data of n samples, where is sample is having three attributes. So, it is sample is a a feature vector is in R 3.

Now, this I can read as n by 3 matrix and if it is having k attributes, it will become a n by k matrix. So, now, how to find out covariance matrix of this? The covariance matrix will be this. So, you just need or use this table as a matrix ok. Let us say matrix C and then, the covariance matrix will become 1 by n into C transpose into C where, C is this matrix n by 3 matrix.

So, what it will be? It will be a 3 by 3 matrix and if you are having k features or k attributes, then it will be a k by k matrix and in that way, the principal components are the eigenvectors of this matrix, that is your covariance matrix of the data.

(Refer Slide Time: 17:23)



So, then, in continuation first. So, in sort for principal component, I am writing PC. So, first Principal Component is the eigenvector corresponding to largest eigenvalue of the covariance matrix that is the covariance matrix will be having eigenvalues and whatever will be the largest eigenvalue, the eigenvector corresponding to that eigenvalue will be the first principal component. And what is the meaning of first principal component, that in the direction of that vector the dataset will be having the maximum variability maximum variation.

So, using this fact, if we want to transform a n-dimensional dataset to a k-dimensional set dataset, then we will select first k principal components and what are these k principal

components? These k principal components are the eigenvectors of the covariance matrix sigma which is nothing just 1 by n C transpose into C, where C is the data matrix corresponding to top k largest eigenvalues.

So, what you have to do, if you are having a n dimensional data means your data is in Rn and you want to reduce it in let us say in R k, you want to reduce it in k-dimensions, where k is less than n? Using the original n-dimensional data, first what you will do? You will find out the covariance matrix of the data.

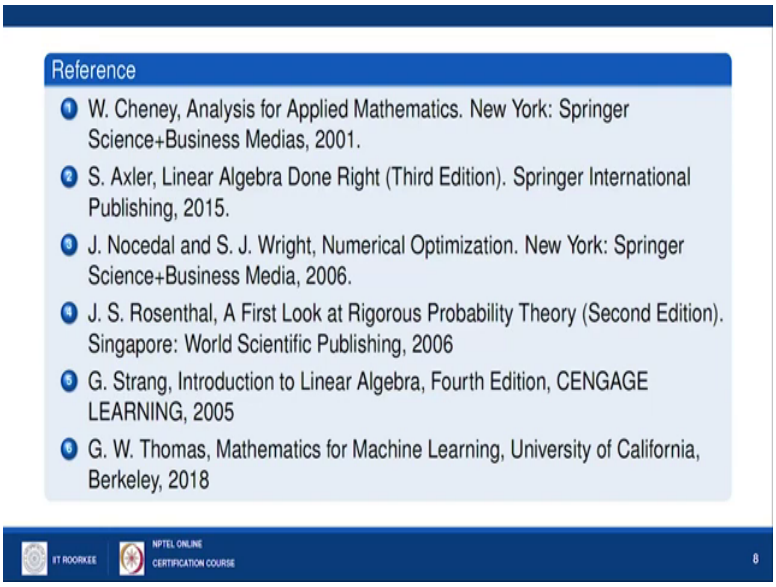
Once you are having covariance matrix of that data that will be obviously a n by n symmetric matrix. Then, you will calculate the eigenvalues. So, you will be having n eigenvalues of that matrix. Now, you select top k eigenvalues largest. Once you are having those top k largest eigenvalues, what you do? You can pick the vectors corresponding to those k eigenvalues.

So, you will get k eigenvectors, those are orthogonal because you are covariance matrix is a symmetric matrix. Now, what you are having? You are having those k orthogonal eigenvectors of the covariance matrix.

Those k orthogonal eigenvectors will expand a k-dimensional space. You project your n-dimensional data to that k-dimensional space and then, your data will become k-dimensional. So, this is the overall idea of principal component analysis. This is having a close lesson with singular value decomposition also and that we will explore. There is a question that why to choose the top eigenvectors corresponding to top eigenvalues.

So, in the next lecture, we will prove it that why we will preserve the maximum information about the data although by projecting it to lower dimensional space, by using the eigenvectors corresponding to largest eigenvalues. That we will prove, then we will take couple of example, we will see how we can utilize those examples and finally, we will we will see the link of this principal component analysis with singular value decompositions because it is basically very easy if you see it in terms of singular value decomposition.

(Refer Slide Time: 23:33)



So, these are the references.

Thank you very much.