Lecture – 12
Linear Algebra for Data science

This lecture is on linear algebra for data science. Linear algebra is a very fundamental part of data science and usually a typical one semester linear algebra course will run for about 36 hours. What we are trying to do here is we are trying to introduce the use of linear algebra in data science in a few hours. So, that necessarily means that we cannot cover the topic of linear algebra in all its detail; however, what we have attempted to do here is to identify the most im-portant concepts from linear algebra, that are useful in the eld of data science and in particular for the material that we are going to teach in this course. So, in that sense we have crystallized a few important concepts from linear algebra that the participants should learn and understand.

So, that is one thing that I would like to mention right at the beginning.The second thing that I would like to mention is the following; Linear algebra can be treated very theoretically very formally; however, in the short module on linear algebra which has relevance to data science ,what we have done is we have tried to explain the ideas in as simple fashion as possible without being too formal. However, we do not do any hand waving we teach linear algebra in a simple fashion. So, that is another thing that I would like you to remember as we go through this material. So, we first start by explaining what linear algebra is useful for.

(Refer Slide Time: 02:06)

Overview

Linear Algebra

Data Representation

Fundamental block in machine learning algorithms

Identifying linear relations
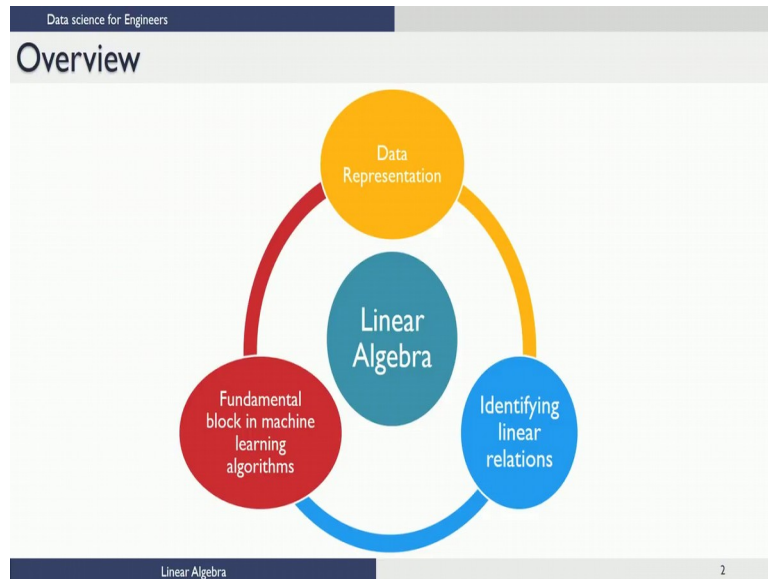
So, when one talks about data science, Data Representation becomes an important aspect of data science and data is represented usually in a matrix form and we are going to talk about this representation and concepts in matrices. The second important thing that one is interested from a data science perspective is, if this data contains several variables of interest, I would like to know how many of these variables are really important and if there are relationships between these variables and if there are these relationships, how does one un-cover these relationships?

So, that is also another interesting and important question that we need to answer from the viewpoint of understanding data. Linear algebraic tools allow us to understand this and that is something that we will teach in this course. The third block that we have basically says that the ideas from linear algebra become very very important in all kinds of machine learning algorithms.

So, one needs to have a good understanding of some of these concepts be-fore you can go and understand more complicated or more complex machine learning algorithms. So, in that sense also linear algebra is an important component of data science. So, we will start with matrices. Many of you would have seen matrices before. I am going to look at matrices and summarize the most important ideas that are relevant from a data science viewpoint. What is a matrix? Matrix is a form of organizing data into rows and columns. Now, there are many ways in which you can organize data. A matrix provides you a convenient way of organizing this data.

(Refer Slide Time: 01:48)

So, if you are an engineer and you are looking at data for multiple variables, at multiple times, how do you put this data together in a format that can be used later, is what a matrix is helpful for.

Now, matrices can be used to represent the data or in some cases matrices can also be used to represent equations and the matrix could have the coefficients in several equations as its component. Now, once we generate these matrices then you could use the linear algebra tools to understand and manipulate these matrices, so that you are able to derive useful information and useful knowledge from this data.

(Refer Slide Time: 04:44)

So, let us start and then try and understand how we can understand and study matrices.

(Refer Slide Time: 04:50)

- Usually matrices are used to store and represent the data on machines
- Matrix is a very natural approach for organizing data
- In general, data is organized in the following fashion
  - Rows represent samples
  - Columns represent the values of the variables (or attributes)
  - It is also possible to use rows for variables and columns for samples
  - However, we will stick to rows as samples and columns as variables in all of the material that will be presented

As I mentioned before matrices are usually used to store and represent data on machines and matrix is a very natural approach for organizing data. Typically when we have a matrix it is a rectangular structure with rows and columns. In general, we use the rows to represent samples and I will explain what I mean by this in subsequent slides and we use columns to represent the variables or attributes in the data.

Now, this is just one representation. It is possible that you might want to use rows to represent variables and columns to represent samples and there is nothing wrong with that; however, in this course and all the material that we present in this course we will stick to using rows to represent samples and columns to represent variables as far as this course is concerned.

Let me explain matrix using a real life example. Let us consider that you are an engineer and you are looking at a reactor which has multiple attributes and you are getting information from sensors such as pressure sensors, temperature sensors and density and so on.

Now let us assume that you have taken 1000 samples of these variables. Now you want to organize this data somehow. So that you can use it for purposes needed. One way to do this is to organize this in this matrix form, where the first column is the column that corresponds to the values of pressure at different sample points. The second column corresponds to the value of temperature at several sample points and the third column corresponds to the value of density at several sample points.

So, that is what I meant when I said the columns are used to represent the variable. So, each column represents a variable column 1 pressure, column 2 temperature and column 3 density and when you look at the rows; the first row represents the first sample.

Here in the first sample you will read that the value of pressure was 300, the value of temperature was 300 and the value of density was 1000. Similar to that you will have many rows corresponding to each sample point up to the last row 1000th row, which is a 1000 sample point; which has a pressure is 500 temperature is 1000 and density is 5000.

Let us take another example let us say I have 2 vectors, X =[1, 2, 3]^T and Y =[2, 4, 6]. Let us say this is some variable that you have measured and Y is some other variable you have measured and the 3 values could represent the 3 sampling points at which you measured these.

Now, in R if you want to put this numbers into a matrix, it is a very very simple code. What you do is: X = c(1, 2, 3) it tells you it is a column vector with values 1, 2, 3 y is a column vector with values 2, 4, 6 and then you use the command A cbind(x, y) which puts these together and when you print A you get the value of this matrix.
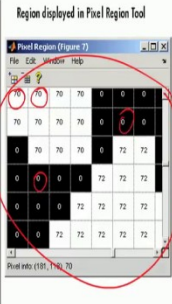
Now, we have been talking about using matrices to represent data from engineering processes sensors and so on. The notion of matrix and manipulating matrices is very important for all kinds of applications. Here is another example where I am showing how a computer might store data about pictures. So, for example, if you take this picture here on this left hand side and you want to represent this picture somehow in a computer. One way to do that would be to represent this picture as a matrix of numbers.

So, in this case for example, if you take a small region here you can break this small region into multiple pixels and depending on whether a pixel is a white background or a black background you can put in a number. So, for example, here you see these numbers which are large numbers which represent white back-ground and you have these small numbers which represent black background. So, this would be a snapshot or a very small part of this picture.

Now, when you make this picture into many such parts you will have a much larger matrix and that larger matrix will start representing the picture. Now, you might ask; why would I do something like that? There are many many applications where you want the computer to be able to look at different pictures and then see whether they are different or the same or identify sub components in the picture and so on. And all of those are done through some form of matrix manipulation and this is how you convert the picture into a matrix.

Now notice that while we converted this matrix we have again got into a rectangular form, where we have rows and columns, where data is filled as a representation for this picture.

(Refer Slide Time: 10:07)

• Storing
  ◦ The image is stored in the machine as a large matrix of pixel values across the image.
  ◦ Thus, storing the pixel value matrix is equivalent to storing the image for the machine
• Identification
  ◦ Several machine learning algorithms are deployed in order to "teach" the machine how to identify a particular image.
  ◦ Linear algebra and matrix operations are at the heart of these machine learning algorithms.

So, the image that I showed before could be stored in the machine as a large matrix of pixel values across the image. And you could show other pictures and
then say are these pictures similar to this, are these dissimilar, how similar or dissimilar and so on and the ideas from linear algebra matrix operations are at the heart of these machine learning algorithms.

(Refer Slide Time: 10:32)
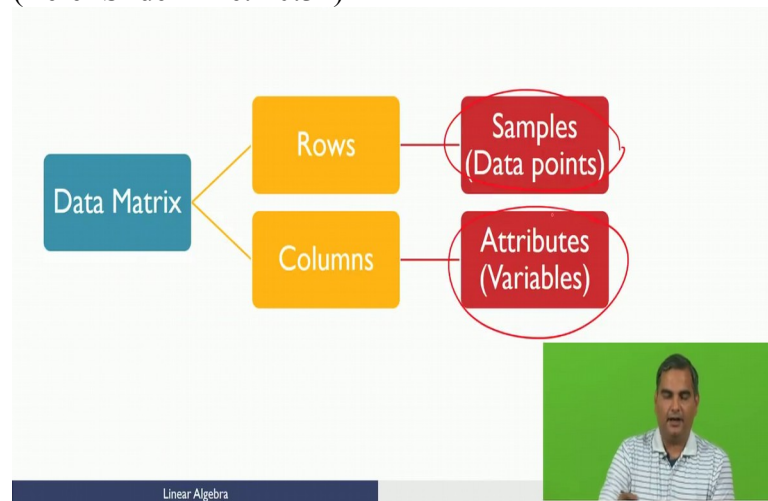
So, in summary if you have a data matrix- the data matrix could be data from census in an engineering plan. It could be data which represents a picture; it could be data which is representing the model where you have the coefficients from several equations. So, the matrix basically could have data from various different sources or various different viewpoints and each data matrix is characterized by rows and columns and the rows represent samples and the columns represents attributes or variables.

(Refer Slide Time: 11:14)
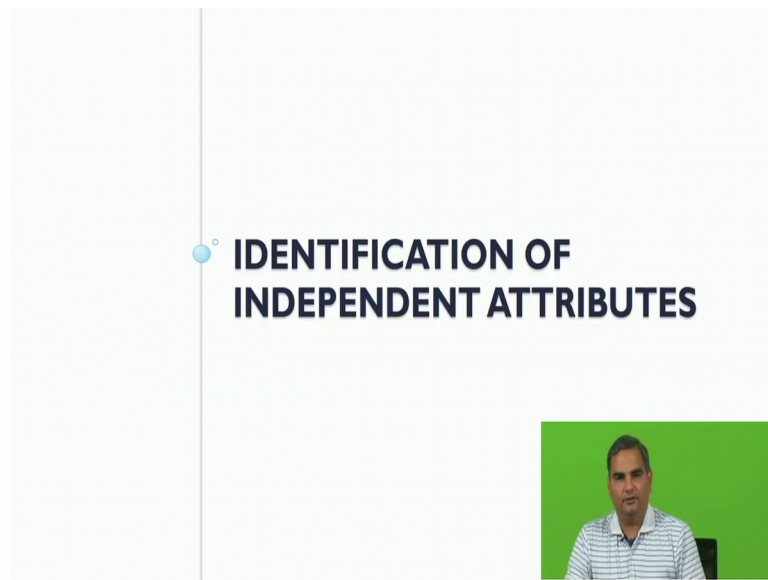
Now that we have understood how we generate a matrix and why we generate matrices. The next question might be the following, supposing I have a matrix where I have several samples and several variables, I might be interested in knowing if all the variables that are there in the data are important. In other words, I would really like to know of all these variables, how many are actually independent variables?

So that I know how much information is there in this data. So, if let us say I have thousands of variables and of those there are only 4 or 5 that are independent. Then it means that actually I can store values for only these few variables and calculate the remaining as a function of these variables. So, it is important to know how much information I actually have.

(Refer Slide Time: 12:08)

So, this would lead to the following questions. The first question might be; Are all the attributes or variables in the data matrix really relevant or important? Now a sub question is to say are they related to each other. If I can write one variable as a combination of other variables then basically I can drop that variable and retain the other variables and calculate this variable whenever I want.

So, that is a very important idea that we want to use in machine learning and various other applications. So, how do I find out how many of these variables are really independent and let us assume that I do and that only a few variables are really independent, then how do I identify the relationship between these variables and the other dependent variables and once I do that how do we actually reduce the size of the data matrix and so on; are questions that one might be interested in answering.

So, let us consider the example that we talked about; the reactor with multiple attributes. In the previous slide, we talked about pressure, temperature and density. Here I have also included viscosity. Let us say I have 500 samples. Then when I organize this data with the variables in the columns and samples in the row, then I will get a 500 by 4 matrix, where each row represents one of the 500 samples and if you go across the column, it will represent the variable values, at all the samples that we have taken. Now, I want to know how many of these are really independent attributes.

So, from domain knowledge it might be possible to say that density is in general a function of pressure and temperature. So, this implies that at least one attribute is dependent on the other and if this relationship happens to be a linear relationship then this variable can be calculated as a linear combination of the other variables.

Now, if all of this is true then the physics of the problem has helped us identify the relationship in this data matrix. The real question that we are interested in asking is if the data itself can help us identify these relationships.

(Refer Slide Time: 14:26)

- Let us assume that we have many more samples than attributes for now
- Is there any approach which can be used to identify the number of linear relationships between the attributes purely using data?
- This is addressed by the concept of the **rank** of the matrix.
- **Rank** of a matrix refers to the number of linearly independent rows or columns of the matrix
- The rank of a matrix can be found using the rank command: rank(A)

| Linear Algebra | | 15 |

Let us first assume that we have many more samples than attributes for now and once we have the matrix, when we want to identify the number of independent variables. The concept that is useful is the rank of the matrix and the rank of the matrix is defined as the number of linearly independent rows or columns that exist in the matrix.

And once you identify these number of linearly independent rows or columns then you could basically say that I have only so many independent variables and the remaining are dependent variables and the rank of the matrix can be easily found using the rank command in our rank of A.

(Refer Slide Time: 15:08)

So, consider this example here where I have this a matrix which is 1, 2, 3, 2, 4, 6, 1, 0, 0. If you notice this matrix has been deliberately generated such that the second column is twice column one.

So, in other words the second column is dependent on the first column or you could say the first column is dependent on the second column. Now, there is one other column which is independent of these two so, if you think about this matrix there are 2 independent columns; which basically means there are 2 independent variables. So, if you were to use R to identify this, simply load the correct library and then use the command rank of A and you will get the rank of the matrix to be 2.

(Refer Slide Time: 16:00)

So, the notion of rank is important. It allows us to work with a reduced set of variables and once we identify the independent

And if you identify that there are certain variables which are dependent on other variables and as long as the data generation process is the same, it does not matter how many samples that you generate, you can always find the de-pendent variables as a function of the independent variables.

(Refer Slide Time: 16:45)
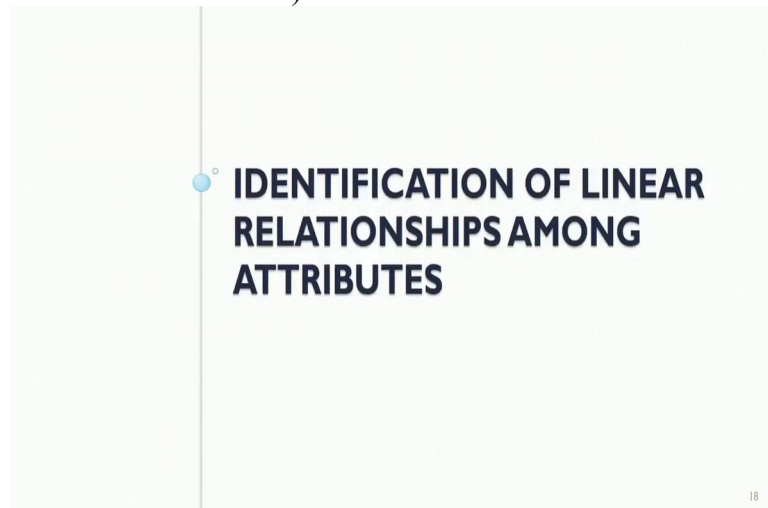
Now that we have talked about identifying how many independent variables are there; assume that the number of independent variables are less than the
number of variables. Then that basically automatically means that there are really linear relationships between these variables.

Now we ask the next question as to how do we identify these linear relation-ships among variables or attributes. So, this question of how does one identify the linear relationships among attributes, is answered by the concepts of null space and nullity which is what we going to describe now.

When we have a matrix A and if we are able to find vectors β such that $A\beta = 0$ and $\beta \neq 0$ then we would call this vector β as being the null space of the matrix. So, let us do some simple numbers here for example, if A is a 3 by 3 matrix because β multiplies A, β has to be 3 by 1 and the resultant will be some 3 by 1 vector and if all the elements of this 3 by 1 vector are 0, then we would call this β as being the null space of the matrix. Now interestingly the size of the null space of the matrix provides us with the number of relationships that are among the variables.

If you have a matrix which is of dimension 5 and let us say the size of null space is 2,then this basically means that there are 2 relationships among these 5 variables, which also automatically means that of these 5 variables only 3 are linearly independent because the 2 relationships would let you calculate the dependent variables as a function of these independent variables.

(Refer Slide Time: 18:46)

Now, let us look at this in little more detail to understand how we can use this null space vectors. Let us assume that I have a matrix such as this. Now if there is a β which is what we have written here such that a times $\beta = 0$. As I mentioned in the previous slide, if I have this matrix of this dimension and if I multiply β with this matrix, then on the right hand side there are going to be several elements and for β to be the null space of this matrix every one of the elements has to be equal to 0. Now, let us look at what each of these elements are equal to.

So, if you take the first element on the right hand side, that would be a product of the first row of this data matrix and this β vector which

would basically be $x_{11} \beta_1 + x_{12} \beta_2$; all the way up to xn and $\beta n = 0$. Now, similarly if you get to the second row and multiply the second row by this vector you will get another equation.

So, if we keep going down, for every sample if you write this product, you are going to get an equation. The last sample for example, will be $xm1\ \beta1 + xm2\ \beta2 + xmn$ and $\beta n = 0$. Now, there is something interesting that you should notice here. This equation seems to be satisfied for all samples. So, what this basically means is, irrespective of the sample, the variables seem to hold this equation and since this equation is held for every sample we would assume that this is a true relationship between all of these variables or attribute. So, in other words this $\beta\ 1$ to $\beta\ m$ gives you in
some sense a model equation or a relationship among these variables.

So, one might say that this equation can generally be written as x1 $\beta1 + x2\ \beta2$ all the way up to xn $\beta n = 0$; where you can take any sample and substitute the values of the variables in that sample at x1 x2 up to xn and this is to be satisfied. So, this is a true relationship.

(Refer Slide Time: 21:19)

- Notice that if $A\boldsymbol{\beta} = \mathbf{0},$ every row of A when multiplied by $\boldsymbol{\beta}$ goes to zero
- This implies that variable values in each sample (represented by a row) behave the same
- This helps in identifying the linear relationships in the attributes
- Every null space vector corresponds to one linear relationship
- This idea is demonstrated further using examples

Linear Algebra 22

So, this is what we have said. Again here notice that if A $\beta = 0$ every row of A when multiplied by β goes to 0.
So, this implies that the variable values in each sample behave the same so, we have truly identified a linear relationship between these variables. Now, every null space vector corresponds to one such relationship and if you have more vectors in the null space then you have more relationships that you can uncover.

(Refer Slide Time: 21:49)

So, we will demonstrate this example this with a further example. So, this rank nullity theorem basically says the nullity of matrix A + the rank of matrix A is going to be equal to the total number of attributes of A or the number of columns of the matrix. So, the nullity of a tells you how many equations are there; there are so, many vectors in the null space.

The rank of A tells you how many independent variables are there and when you add these two you should get the total number of variables that is there in your problem.

(Refer Slide Time: 22:29)

So, to summarize, when you have data, the available data can be expressed in the form of a data matrix and as we saw in this lecture this

data matrix can be further used to do different types of operations. We also defined null space, null space is defined as a collection of vectors that satisfy this relationship A times β = 0.

So, this basically helps in identifying the linear relationships between the attributes or the variables directly and the number of such vectors or number of such relationships is what is given by the nullity. The Nullity of the matrix tells you how many relationships are there or how many vectors are there in the null space.

(Refer Slide Time: 23:13)

Let us take some examples to make these ideas little more concrete. Let us take a matrix A which is 1, 3, 5, 2, 4, 6. A quick look at this matrix and the numbers would tell you that these two columns are linearly independent and subsequently because these columns are linearly independent there can be no relationships among these two variables.

So, you can see that the number of columns 2 since they are independent the rank is 2. Since the rank is 2, nullity is 0 and because both the variables are independent you cannot find a relationship. If you were able to find a relationship then the rank should not have been 2. So, this basically implies that null space of the matrix A does not contain any vectors and as we mentioned before these variables are linearly independent.

Now, if you want to do the same thing in R what you do is you de ne the matrix A which basically is done using this command. This n columns equal to 2 tells you how many columns this number should be put in. So, since there are two columns, these numbers will be partitioned into 1, 3, 5, 2, 4, 6 and as we saw before you can actually get the rank of A.

And you can print the number of columns, you can print the rank and you can print nullity which is the difference between columns and the rank number of columns and the rank.

(Refer Slide Time: 24:43)

Now, let us take the other example that we talked about where I mentioned that we have deliberately made the second column twice the first column. So, in this case as we saw before the rank of the matrix would be 2 because there are only two linearly independent columns and since the number of variables = 3, nullity will be 3 - 2 = 1.

So, when we look at the null space vector you will have one vector which will identify the relationship between these three variables.

(Refer Slide Time: 25:15)

So, to understand how to calculate the null space let us look at this example. So, we set up this equation A β = 0 and we know we will get only 1 β here and β we have written as b1 b2 b3 and when we do the first row versus column multiplication I will get b1 + 2b2 = 0.

When I do the second row and column multiplication I will get 2b1 + 4b2 = 0 and when you do the third multiplication you get b3 = 0. Now the second equation which is 2b1 + 4b2 = 0 is simply twice the first equation. So, that does not give me any extra information so, I have dropped that equation. Now, when you want to solve this notice that b 3 is fixed to be 0.

However, from this equation what you can get is b1 is - 2b2. So, what we have done is instead of b1 we have put - 2b2 retain b2 and 1. This basically tells us that you can get a null space vector which is - 2 1 0; however, whatever scalar multiple you use, it will still remain a null space vector.

So, this is easily seen from the following if A times β = 0 where β is a vector A is a matrix. Let us assume I take some other vector from β which is some C β, where C is a constant. Then if I plug it back in I will get A times C β = 0 which because this is scalar I can take it out C A β = 0. Since, this is 0 C times 0 will be 0. So, this will be a also a 0 vector.

So, whenever β is a null space vector then any scalar multiple of that will also be a null space vector that is what is seen by this k here. Nonetheless we have a relationship between these variables which is basically saying - 2x1 + x2 = 0 is a relationship that we can get out of this null space vector.

(Refer Slide Time: 27:24)

So, to summarize this lecture as we saw matrix can be used to represent data in rows and columns; representing samples and variables respectively. Matrices can also be used to store coefficients in several equations which can be processed later for further use. The notion of rank, gives you the notion of number of in-dependent variables or samples. The notion of nullity identi es the number of linear relationships, if any between these variables and the null space vectors actually give us the linear relationships between these variables. I hope this lecture was understandable and we will see you again in the next lecture.

Thank you.