Introduction to Data Analytics

Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 01
Lecture - 01
Course Overview

Hello, and welcome to our first lecture for the course Introduction to Data Analytics. In this lecture and perhaps the next few lectures, I am going to be providing you with the Course Overview and giving you a detailed description of what we will be covering during this course. Let us start off with a little bit of logistics associated with the course. First thing is, there is a good amount of information available on the course website.

So, I just wanted to make sure that everybody can avail of the course outline, the syllabus, the reference books and so on, it is all uploaded on the NPTEL website and you should access it, so you have the appropriate information. The second thing is that the forums in the course are a great resource for answering many of the questions that you might have or that you might come across during the course.

The most of the cases, the professors myself and Professor Ravindran, we will try to answer some of your questions, at times we will have teaching assistants help us with answering questions and most importantly, many of these questions can be answered by your own pears. So, I really encourage you to use that resource and also contribute to it in any way that you deem fit. If for instance we feel like a particular question is not been answered appropriately, clear it keeps coming up; we will definitely join in and get in on the discussion, so definitely try to use that as a resource.

Before, I started there are already a lot of questions on the forums, so I just wanted to address one or two of them, they pertain to the course at large. So, the first question relates to the course style. For instance, is this course going to be very theoretical, is it going to be a very case study base, so on and so forth? So, to give you a feel for what this course is going to cover, the course is not going to be only theory or highly theoretical, we are not going to emphasize extensively on the pure math of the course.

They might be some amount of math and some amount of programming, but we are not

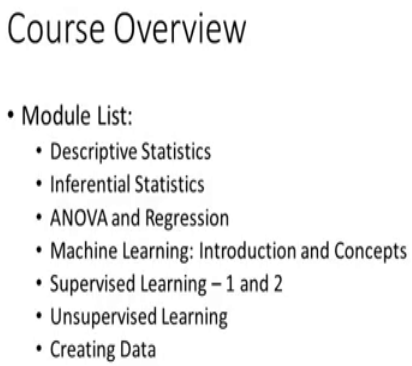
going to go in depth and derivation and so on. The course is going to be heavily conceptual and it is going to try and have as many applications as possible. So, we are going to give you the what's and whys of data analytics are ranging from the statistics to the machine learning, what techniques and tools will you apply, where, why do some methods work in certain situations and not others, how do these algorithms work, what is the background logic behind it and we are going to give you lots of applications. We will be assisting you with little bit of how.

So, how do you implement this algorithm or how do you, what kind of software could you potentially use and so on, but the course itself is not going to be a tutorial style course? So, it is not, you know anything that you can essentially get from the help file of package that you using for data analytics; we are not going to be repeating that here.

So, that we believe that we are adding more value by spending our time and giving you really the core concepts. Syntax, you can learn from one software to the next. This brings me to the next core area, which is the software and programming area. Many of you have expressed, concern or have questions about, what software we would be using, what programming languages, how much do I need to know.

So, the main languages that we are going to be using in this course are R and MATLAB and occasionally, we will show you, how it can also be applied in Microsoft excel. But, again it is not going to be tutorial style, we will not teach you the nuts and bolts of how to use R and the code itself that we would be expecting or teaching would be very fairly basic and more like command line instructions. So, it would be fairly easy to pick up, even if you did not know the software per say, but you had some comfort level with basic programming. So, without I have covered some forum questions and basic logistics, so let us dive into our course overview.

(Refer Slide Time: 04:34)



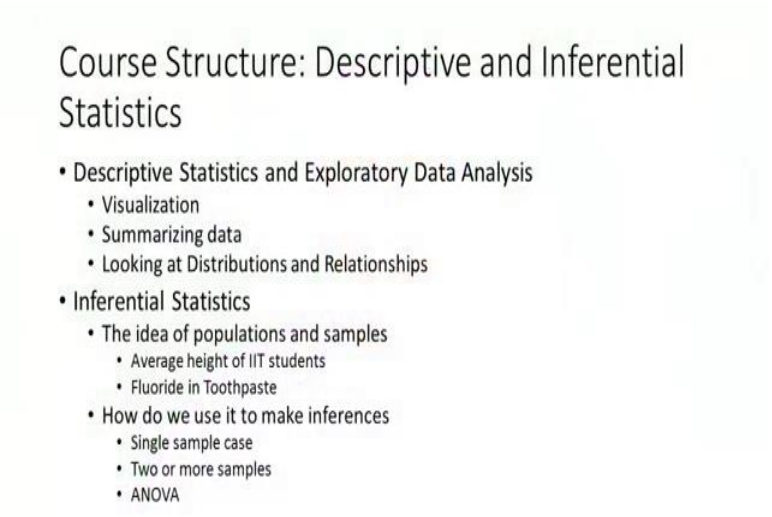
So, in our course overview, I just wanted to give you the basic module list. Now, this is available on the course website and the basic module list just gives you the topics that we are going to be covering in this course. So, we will be having descriptive statistics, an inferential statistics that is the first two broad areas. Within descriptive statistics, we would also be covering some amount of probability and probability distributions.

We then move on to more advance concepts in inferential statistics, namely something called the ANOVA or analysis of variance. We will be talking about what, where and why we apply that. We will introduce regression, something that you might have heard of. And then, we move on to what we see as the main focus of the course, which is machine learning. Both an introduction to it, core concepts, how do they tools and what are the tools and techniques are there, how do they work and within that, we would be talking about both two classes called supervised and unsupervised learning.

And finally, we come to this module on, what we do when you do not have data. How do you go about, what is data analytics when you do not already have the data? So, there is a lot of interesting work there. So, do not worry if at this stage you do not understand some of the words that I have used, that is what this overview is for to give you some idea of what it is that we are going to be covering. I am now going to go step by step, talk about each of these modules and place them and give you some idea of, what is the core concept behind them.

Now, again the purpose of this is not to replace the actual class. So, in actual class for instance, I will be giving you far more thorough treatment and so Professor Ravindran, but this is really to help you get a first class view of what it is that you will learn at the end of this course.

(Refer Slide Time: 06:52)



So, great, let us start with our first module, which is descriptive statistics and exploratory data analysis. Here we are really talking about, how you describe data. So, we would be talked the, one of the main things that would be introduced here is data visualization techniques. So, this primarily concerns different forms of graphs, how do you represent a single variable graphically, how do you represent relationships between two variables graphically and typically, we are dealing with a data set.

So, what kinds of graphs and what kinds of tables are best suited, especially also given that different variables are of different types and for different variable types are there, different ways of representing this data. So, visualization is one part of the descriptive statistics and we will spend a few short classes there. We then move on to summarizing data. So, this is part of descriptive statistics, which is you have a data set, how do I you know summarize the data set, how do I present it to you in a single statistic or a set of statistics.

Out here for instance, we will be mostly concern with something like measures of central tendency. What do we mean by that? You might have heard of the words means, median

and mode and even if you not, I am sure you all colloquially used the words saying, you have this data set, but what is the average. So, if it is a single variable especially, what is the average or what is a typical value in that data set, what is a value in between and you know these are all very colloquial ways of talking about it.

But, there are various different measures of central tendency and different measures express different properties associated with data set and we would be going into that, another area is also measures of dispersion. So, you might have heard of the words variance or standard deviation and essentially, what that captures is an inherent variability. So, we spoke about this concept of mean or measures of central tendency, but how does the data itself vary around that average and that is what we will be talking about there. We will focus on measures of dispersion and measures of central tendency.

But, beyond summarizing data through these measures, we can go even further and probability distributions are the richest way of expressing a data. And they do so, because you do not just stop with saying this is the mean or this is the average or this is the standard deviation. You almost literally express it as you know, 20 % of the data is below this value, 30 % of the data is below this value and so on.

So, when you do that you not only get an idea of the dispersion and the centrality, but you know the entire characterization of the data set. So, you will be talking about probability distributions as well there. So, I think in that is and that give you some sense of, what descriptive statistics really seek to accomplish and that is, what we will be covering in that module. We move on to the next part of the course, which is inferential statistics.

The idea here is, the idea is one of populations and samples and I am going to explain to you, what these words really mean. The idea here is that, the data that you have is essentially some sample from a broader universe; that is that could be creating this data. And this broader universe is what we call the population and the population can be like a finite, very large data set and you do not have that entire data set, which is why you need a sample from the data set.

But, it could also be something more conceptual, essentially this data does not physically exist anywhere. The population data does not physically exist, but it is the concept of this really large universe of potential data that you can get, but you essentially have only, you

can only get a sample of the data set. So, perhaps I mean, just to give you a stronger intuition for it, let us just talk about one or two examples.

If you took for instance the height of all IIT Madras students, let us say the height of all IIT Madras students. Here is the case, where you actually have a finite population, you define your state space, you define that is space as the data is ultimately the height of all IIT Madras students as of today, let us say as of 2015, that is a very large number and you might not have the data. So, you might choose to take a sample from the data, so you might choose 20, 30 students and go actually measure their heights and that is your sample and here the population was, population would have been the height of every single IIT Madras student as of 2015.

I give you another example, where it might not be a finite population. So, you might say well I have this manufacturing process, that makes a certain product and I am interested in the dimension of the product. Now, let us say I go and change that manufacturing process and I am interested in measuring, so let us say it is the width of a particular piece of metal that gets that comes out of this machine.

Now, the population here would be essentially infinite in size, it could potentially be the width of infinite number of metal pieces that come out of this particular machine and you know, you are not interested in actually getting that entire population you could not. But, you might, you could still have a sample, you might have just collected about 20 pieces of such metal and you have the sample. Now, what is the point of all these? Why is this an important concept?

It is a very important concept, because in many instances, we find that we are not satisfied with just describing or giving you statistics about the sample. You are far more interested in saying something about the population. So, for instance I might be interested in saying something like the average height of IIT students is less than 130 centimeters or some such number or 150 centimeters. The average height of IIT students is less than 150 centimeters.

Here note that you making a statement about the population, you making a statement about the average height of IIT students and we said the height of IIT students is the entire population. But, you taken only a sample of 20, 30 students, you do not have data associated with the population. So, is there something you can do with this statistic that

is, you can do with the sample of 30 or 40 and say something about the population.

Another example, for instance institute could be something like let us say I have a tooth paste company and my job is to put certain amount of fluoride in the tooth paste. And let us say, I am interested in putting about 1000 parts per million of fluoride in my tooth paste use. Will every single tooth paste you have exactly 1000? Probably not and what is my population here, it is potentially every single tooth paste tube that I have sent out to the market or I could be sending out to the market.

But, I can do one thing; I can go, take a sample of about 10 or 20 tubes or 30 tubes, measure the amount of fluoride in that and see, if the average amount of fluoride in my tooth paste tubes is equal to 1000 or less than 1000 or greater than 1000, this is with the full recognition that not every single tube is going to have that exact amount, because the world is not a perfect place the world is a noisy place. But, is my average equal to the value I think it is or is my average less than or greater than the value I think it is.

So, the whole idea again just to kind of recap, what inferential statistics tries to capture is that, you now have split the word into a population and a sample. A population could be finite or it need not even be finite, but it is essentially you can think of it is very large universe of data and all you are getting is a sample, all you can measure is from the sample. But, is there something I can do with the sample to make a statement about the population and that is, those of the tools and the techniques that we will be discussing there.

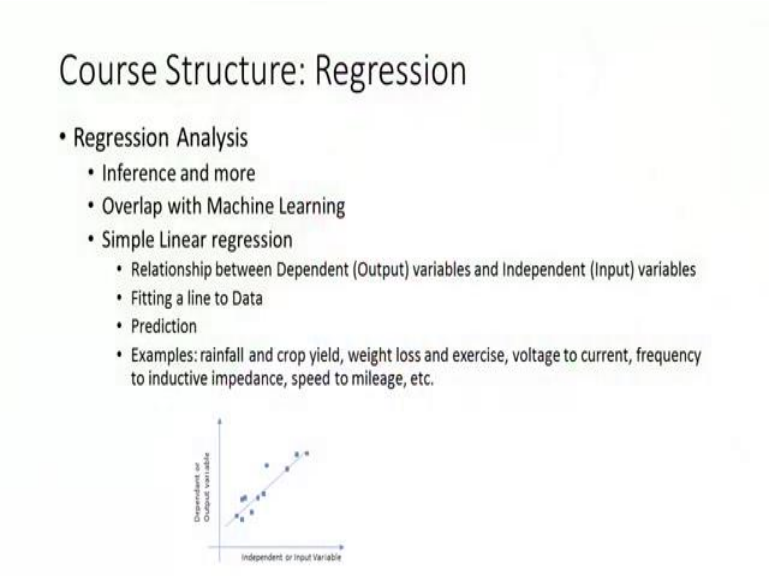
And it goes beyond for instance the two examples that I gave, in for instance the two examples that I gave we got a sample and we compare that to some number we had in our head. We said, is the average IIT student height less than 130 centimeters or is the amount of fluoride equal to some number x. But, you could also be comparing two samples and there by essentially comparing their populations. Is the average amount of fluoride in tooth paste brand A equal to the average amount of fluoride in tooth paste brand B and extend that even further, A versus B versus C and so on.

So, the core idea again is that you not are confined to one sample, you could go to many samples and the idea of going to many sample is what is get captured in a technique called ANOVA or analysis of variance. We are going to have a separate module on that as well, but again stepping back are we living in a world, where we get samples of data

whether is a bigger phenomena, but can I use these samples to make statements about the broader phenomena.

So, that is what we will be covering in the inferential statistics part of the course. We then, move on to something that many of you might have heard or encountered, probably has bend the motivation for you to even may be take this course and I am talking about regression.

(Refer Slide Time: 17:50)



When I say the word regression here, I am going to actually start using the word regression analysis, because the word regression itself can be used to cover many highly relative, but slightly different concepts. But, we are going to talk about regression analysis and regression analysis is a great segue; it is a great step going after inferential statistics and going before our machine learning.

Because, regression analysis uses inferential statistics, it is you cannot say regression analysis is only inferential statistics, it uses inferential statistics, but it also uses many others tools. It uses optimization, it uses some concepts from descriptive statistics and so on and it has a lot of overlap with machine learning. Regression analysis per say might not in many courses or depending on, who you ask might not come under necessarily the umbrella of machine learning, because regression analysis has been there, been around far before, say machine learning analysis.

But, in a sense it tries to tackle the same problem statement or a very similar problem statement, that some of the more advance techniques in machine learning try to do. So, it is a great thing to learn and understand right of the back and so we would be introducing a regression to you in this course. With the regression itself, we are going to start with the simplest form, simple linear regression and we going to use fairly simple techniques of doing it called the ordinary least squares techniques.

And right now, let me just give you again a very basic intro to what a regression is again like I have mentioned before a this is not to replace the class, that we will have on regression all of these things and the talking about today we are going to go into it in great detail as we go through the course the purpose out here is to let you know what you getting into in the course at this stage.

So, let me just spend few minutes and give you very brief in true to what regression analysis tries to do. The regression analysis is essentially about creating a relationship between the dependent variable, you can also think of that as the output variable and one or more independent variables and you can think of the independent variables as input variables. We definitely going to talk about some examples here, but the core idea is to create to this relationship and one of the simplest ways of the creating this relationship is to fit a line through the data, what do you mean by fit a line through the data.

Take a simplest example you have the independent variable and may be your best suited with some examples here, but let is say you are independent variable is the amount of rain fall that a particular region in India receives particular rural region and this is collected over the course of yours, so it is the annual rain fall that a particular region in India collects. And the output variable is the amount of crop yield, so amount of rice that grow the amount of wheat that grow that particular year.

The core idea is the independent variable is the rain fall the dependent variable or the output variable is the crop yield, why because we might suspect, that how much rain fall it is, depending on how much rain fall there is that is going to influence the crop yield, The amount of crop yield is the output depends on the amount of rain fall may be we are right may be we are not, but what we are essentially doing is taking pairs of data from different years.

So, year one how much did it rained, what was the crop yield how much wheat or rice of

crop did we get. Year two how much did it rained how much not. So, you have this data set right of inputs and outputs independent variables and dependent variables. The core idea is can I create a relationship between these two variables and the simplest way is to create fit a line through this data.

And you can see in the slide in front of you that I have just created an example graph where the x axis is essentially the independent variable how much did it rained the y axis is how much what was the crop yield and the idea is if you can fit a line through this data that line in some way represents that relationship between these two variables.

You can think of many more examples like this for instance may be your independent variable is exercise and your dependent variable could be something like weight loss in this particular case the line; however, will look different the more you exercise the greater the weight loss, but if you can if you thought of it as if you thought of the loss it depends on how you think of the loss right if your y axis the dependent variable for instance is your overall weight and not the weight loss.

Then, you would have a line that started from the top left corner and came down to the bottom right corner, but that is fine that is still a regression and one way or the other as long as you are trying to fit a line through the data that is what you are trying to do. There are many more examples that even come from engineering some of the early experiments at then, in trying to understand voltage as the independent variable and current as the dependent variable the frequency to the inductive impedance was another one that is tried like this.

Another great example from mechanical engineering is a essentially speed of the vehicle to the mileage that you get from the vehicle. So, at different speeds are you getting different mileages is there some relationship between these two variables and how do I quantify that. But, the examples can be endless, but the core idea with such an exercise typically tends to be one of two things or either both of them one is to capture that functional relationship.

So, you understand the system you are dealing with right and this requirement is prevalent and it is not mutually exclusive from this other requirement, which is given the independent variable can I predict the dependent variable. So, the first one the more obvious one is I need to know one I need to understand my system. So, by doing this

data analysis I understand, how rain fall effects crop yield, but there might be another goal another not unrelated goal, but another goal which is given some amount of rain fall can I predict what my crop yield would be.

Now, this has some critical advantages this has many business applications and to give you an example keeping with the same example. Now, given the amount of rain fall there is if I can predict, what the crop yield would be, then as a government can I institute certain policies for relief performers given the amount of rain fall there is if I for instance providing insurance for farmers can I come up with the prediction of what their crop yield should have been.

So, at least I am not in the dark about, what it should be now with each of the other example we spoke about we can also think of cases, where being able to predict what would happen or what the case would be for a given instance has as a lot of advantages for us. And many of the machine learning techniques for instance that we are going to be focusing on really emphasis this prediction part, how accurately can I guess, what is going to be and some amount of it might be that its really helpful to guess that dependent variable the output variable, because it is not possible for me to always measure the dependent variable.

I have measured that you for some amount of time and given you some data. So, that you can go and build models based of it, but it is not possible for me in real time continuously keep monitoring that variable, so prediction becomes an advantage there. Another advantage could be that the dependent variable actually takes place at some point of time; however, small, after the independent variable takes place and an example that is often sighted is ambient temperature and number of heat strokes in a hospital.

So, that if I just knew if clearly these two variables could be related that the dependent variable, which is the number of heat strokes appearing coming into a hospital is some function of the temperature outside if the temperature outside is really cool and the sun is not out you are not going to that many heat strokes I mean that is reasonable right. But now, because there might be a time lag between when temperature speak and when people actually physically arrive at the hospital can I be better prepared I use some historic data between temperatures and heat strokes and build my and fit my line and build this regression model.

Tomorrow, I am just going to use this regression model and say the temperature is x at a temperature x can I use my regression line and predict how many heat strokes are going to arrive at my hospital and therefore, I will be better prepared at the hospital. So, I hope this gives you some insight into what regression is what it seeks to achieve and with that we conclude our first session of the course overview.

In the next session, we will be continuing with giving continuing with giving you an overview of the course and introduce the more detailed sessions on machine learning and so on. So, look forward to having you join us then.

Thank you for participating today.