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 - 02
Course Overview (Continued)

Welcome to the second lecture of the course Introduction to Data Analytics. In this session, we are going to continue from our previous session, where we presented to you a brief overview of what, we are going to be covering in this course. And we started of talking about in the previous session we spoke about descriptive statistics, inferential statistics, the use of ANOVA in inferential statistics and finally, we spoke about regression and regression analysis and how we would be using that and we would be talking about that in this course.

We now move on to the next session of the course, which is machine learning. Again, just to remind all of you there, this is not the introduction to machine learning part of the course. This is the part, where I am just giving you an overview of everything that we are going to be covering in this course. Obviously, with each session we are going to separately introduce the topic and go over it in great detail during that particular session. But, this is just again to give you an idea of what it is that we are going to be covering in this course.

So, let us talk about machine learning. Machine learning is what we feel is, a primary focus in this course and having covered concepts in probability, statistics and also in with regression analysis should set you up fairly robustly for understanding machine learning. So, many of you might have heard of the word machine learning, come across it in some form or the other or you might have also come across machine learning through one of it is related topics.

So, you might have come across data mining, you might have heard of the terms data mining, you might have heard of the term pattern recognition or statistical learning in some cases. Now, all of these are highly related topics, but they are not necessarily the same. For instance, the focus on the machine learning is more focus on the algorithm themselves that are going to be used to convert data into usable knowledge. So, and that

is also going to be the focus of this particular course.

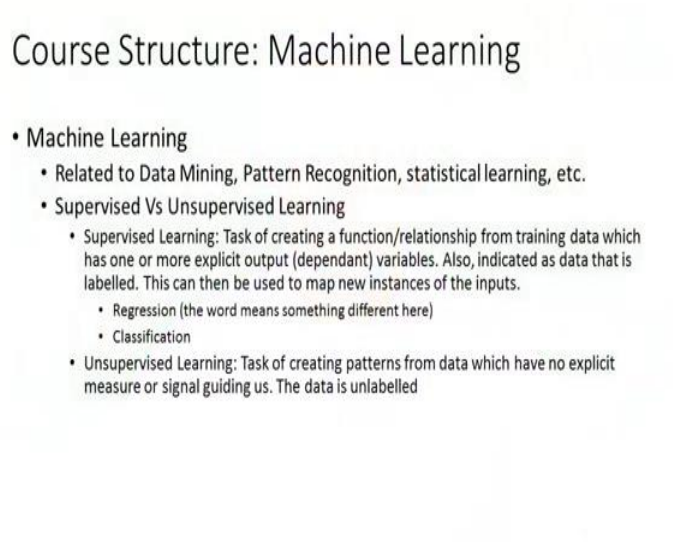
So, let us talk broadly about machine learning, topic of machine learning is itself broadly divided into two areas, one of supervised learning and unsupervised learning. And now, I am going to give you a brief idea of what we seek to achieve in both these topics separately. So, let us for instance take supervised learning. Before we jump into a definitional understanding of supervised learning, you already saw the first glimpses of what supervised learning tries to do and you saw that when you cover the module on regression analysis.

Now, to jump into the definition, the core idea of supervised learning is essentially a task of creating a function or a relationship from training data. So, based of historic data, which has at least one explicit output variable, traditionally this is also indicated as data that is labeled, so that is coming from the computer science camp, where people say the data is labeled.

But, what that essentially means if you are not familiar with the terminology, is there is this clear single variable, which I can call as the output variable and I am primarily interested in create a functional or algorithmic mapping between this output variable and one or more input variables that I might have. So, that is supervised learning and we can take the same examples that we were looking at when we were speaking about regression analysis as examples of supervised learning.

So, you might have data, where one of your inputs is something like the rainfall and your output is the crop yield or you might have data, which says that your input is a square footage of the house and your output for instance could be the price of the house.

(Refer Slide Time: 05:01)



Again, there are many, many, many examples we can think of, but a supervised learning is this idea that we have an output variable and your primary focus is to either predict the output variable or create a functional relationship between the inputs and outputs, which can be used or it is useful for the future. Now, within supervised learning itself there are two broad classes of problems. Now, this classification of problems does not mean the algorithms themselves are really different.

So, essentially you are, the idea here is that your supervised learning problems can be classified into two broad classes and they called classification problems and regression problems. The word regression means something quite differentiates, it does not mean the exact same thing as a regression analysis, but once I explained this division you will understand better. Classification problems are essentially problems, where you still trying to do what supervised learning tries to do, which is create a relationship between the inputs to the output.

But, here your output variable is a discrete categorical variable and more often, the not is a nominal categorical variable, meaning there is no explicit ordering of the classes. So, an example of this could be something like your output variable is either male or female. So, you are trying to predict something and the output variable is not something like the previous example, where we said how much, what the crop yield was.

So, how much crop did I get is a continuous variable, meaning 20 kg or whatever per

hectare is a very, is exactly twice 10 kg per hectare. So, that is a continuous variable, you can get any value between 0 to infinity or negative infinity to infinity. But, with classification problems you are trying to predict based on the inputs as to which class the output variable should belong to and that just means that the output variable is discretized and in all likelihood, it is a categorical variable and it is typically nominal.

Now, move on to the class of problems which are called regression problems within supervised learning, that just again quite simply means the output variable is a continuous quantitative variable, such as the crop yield given some amount of rainfall, how much crop are you going to get given the rainfall. The methods themselves are just marginally different and many of the supervised learning tools and techniques are perfectly capable of being deployed in classification scenarios as well as regression scenarios.

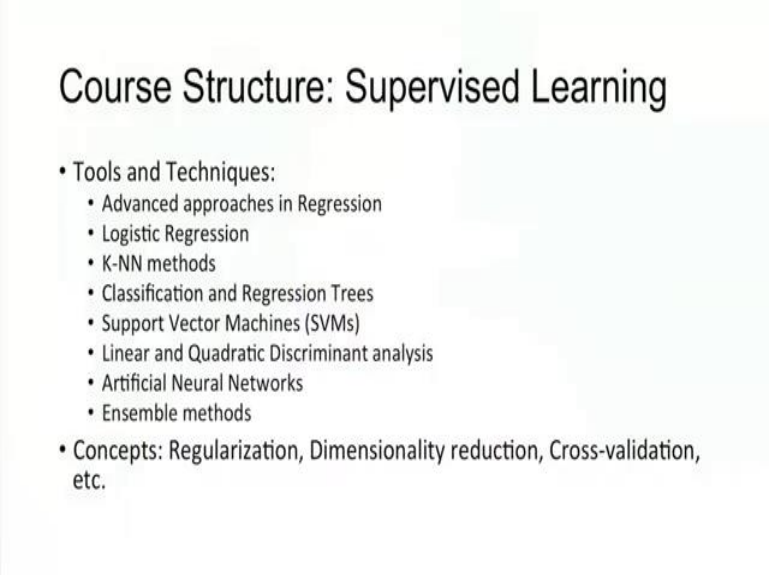
And, but at the same time there are techniques, which are just suited for one of the two and you need to make some modifications to the technique for it applied to the other. But, we will be discussing this dichotomy as we go through the course also and even as we go through the techniques themselves, we talk about them a little bit. We now move on to unsupervised learning and I am, let me just give you a brief idea of what we will be covering in unsupervised learning. An unsupervised learning is the task of creating patterns from data, which have no explicit measure or signal guiding us. In other words, there is no single variable, which we can call as our output variable.

Again, people say the data is unlabeled, but ultimately if you are familiar with the terminology great, so if you think of it is labeled data for supervised learning unlabeled data for unsupervised, I find it easier to think of it as with supervised learning there is an explicit output variable, with unsupervised learning there is not one or two you know variables that I can just point to once, so say these are the output variables these are the input variables with unsupervised learning you just have the variables.

Now, that we have basic definitional understanding of supervised and unsupervised learning. I am just going to give you some idea of what are the tools and techniques that we are going to cover in them. I might not, I am not again, because we are not in the supervised learning class, I am just giving you an overview I am not going to go into what each of these techniques are, but this is more to just familiarize you with the names

of these techniques and in some cases, you might have come across or heard of these names somewhere, so I just want to make sure that you have familiar with that.

(Refer Slide Time: 09:54)



So, with supervised learning we are going to be looking, we would have finished our module on regression analysis. We would be looking a little bit at more advanced methods in regression, modification setup available with regression analysis approaches. We will be looking at logistic regression, which is used for problems of classification, regression styled approach for not predicting continuous output variables, but categorical output variables.

We will talk about an algorithmic approach called K NN methods. You might also come across this module on Classification and Regression Trees. It is also called CART; we will be talking about that. Other methods that you will come across are Support Vector Machines or SVMs, Linear and Quadratic Discriminant Analysis LDAs and QDAs, Artificial Neural Networks or ANNs and there are breed of methods called ensemble methods, which kind of use multiple predictors together, so that is also something that we will be covering this course.

We do not stop at only tools and techniques, because just knowing the tools and techniques and sometime, some of these tend to be buzzwords, is good in that you know what you might be talking about. But, you also want to understand some of the concepts that go behind, creating some of these techniques and these concepts can be critical to

fine tuning some of the parameters that are there in the techniques.

So, we will be talking about some common supervised learning concepts called regularization, dimensionality reduction or cross validation and so on and at this point, if you do not understand some of these words, that is fine. The purpose of this is to just give you an idea of, what it is that we will be covering, fine. We then move on to unsupervised learning and in unsupervised learning, there are two major areas that we would be covering.

The first is the concept of clustering. You might have heard of the word clustering and that is a topic that we are going to cover in unsupervised learning. And the next, the other topic that we will be covering in unsupervised learning is called association rule mining. So, let me just briefly give you an idea of what clustering is and what association rule mining is. This way you also get a slightly better idea of unsupervised learning. See with supervised learning, you had the example of the regression very concrete example it is easy to imagine. But, what is it mean to do machine learning, where there is no output variable, what is that feel like, perhaps talking through these two will give you some idea.

(Refer Slide Time: 12:56)

With clustering, the core goal is that it is a task of grouping a set of objects into clusters or you can think of them as group into groups based on their similarities. How are these similarities defined? It is defined across a common set of attributes or features that each of these objects have and again, the easiest way to digest what I just said I will repeat it,

is that clustering is the task of grouping a set of objects into clusters or groups based on their similarities and the similarities are defined based on a common set of attributes or features that these objects posses.

So, let us take a couple of examples. Now, we understand a definition version of what clustering is, but let us take a concrete set of examples and may be, what we mean by objects and what we mean by features becomes more obvious. The easiest example to think of our customers for a business, so let us say that I am an online retailer or let us say I am a taxi company, take whichever business is close to your heart and let us say I had a database of my potential customers or may be my current customers, either database.

The objects here would be the customers; the features are attributes, our features and attributes associated with the customers. So, a simple feature could be is my customer male or female, another feature could be, what is the age of my customer, another feature could be is this returning customer or is this a new customer, another feature could be the actual amount of rupees per transaction spent by this customer each time they come to me.

So, these are all some attributes and features associated with the customers, the customers are objects. So, what are we doing in a clustering, what we are doing is we grouping these customers and why would we want to do that, for various business reasons. If I can group these customers, so nobody is coming and telling me, what is the right answer wrong answer, there is no output variable.

But, I have taken these customers and now, I have created two or three groups and that could help me in a variety of ways. If I understand that there are only two or three types or groups of customers that come to me, knowing which group a particular customer belongs to. It might help me behave differently potentially to the customer or it might institute certain policies in my business environment based on the groups that get formed in amongst my customers.

So, again there are many, many examples. We just spoke about businesses incoming customers for a business, this is been quite prevalently used in biology for instance, where the object here are different genes and the different genes performed different functions. So, these functions tend to be features or attributes. So, can I group genes

based on the functions that they performs, so that is one application. There are also many applications in medicines.

So, you have a whole plethora of disease and the disease form the objects, but are there certain set of symptoms or are there certain set of responses to treatments that these different disease have and so, can I group these disease based on the attributes, which could be symptoms or their responses to different kinds of treatments. And for instance, a grouping like that might help establish wings in hospitals or medical treatment facilities, where disease of a certain kind get grouped together and people are sent there.

This is just thinking allowed, it might be a good idea, it might not be a good idea, but point is clustering can enable you to create these kinds of groups and how a business or an engineering application uses it is more domain specific in that sense. Some of the techniques in clustering that we are going to covering include K NN, so you might have heard it as K means clustering. We are going to be talking about hierarchical clustering, graph based clustering and also density clustering. So, this is just to give you some idea of different types of clustering techniques that we are going to be covering in this course.

Let us now talk a little bit about the other major unsupervised learning technique that the course is going to focus on. And this is essentially association rule mining. Association rule mining is essentially this task of identifying relationships between features across a set of objects. So, keep the same object and feature definition that we created with the clustering. With clustering, your goal was to use the features and thereby group objects.

With association rule mining you want to use these objects, you want to use these data essentially to create relationships between features. So, let me give you a concrete example and this is a seminal example that introduced in many ways association rule mining and it is called the market basket application. In fact, association rule mining was you know, times also called like a market basket analysis and so on.

The idea here is that, let us say you are a point of sales system, you are a super market and your rows or your objects are essentially customers and these customers are come in and they buy some sub set of the products that you stock in the super market. And the super market now represents this whole transaction, where each row is a sale that a customer makes and the columns or the features or these different products that the super market stocks.

So, a particular sale will have a stream of zeros and ones, where if I did not take product A I get marked as 0 for that product, if I do take product B, then that is a 1. So, each sale you can think of this stable, where each sale is a row in that table, each column is a product that the super market stocks. So, if a particular sale includes a certain product, then that is marked as a binary, it is binary systems gets marked 1 and if a particular sale does not have that product, it gets marked 0.

So, let us say you have this table now, this table could potentially enable you to answer questions of the nature such as people, who buy coffee and coffee would represent a particular column in the table. You could say something like people, who buy milk tend to buy sugar. Because, typically when there is a 1 in my data set, it looks like under the category milk there also tends to be a one under the category sugar in my same data set.

So, and this can be extracted to go beyond a one to one mapping say I give you an example, where people, who tend to buy milk tend to buy sugar, but you have lots of other combinations you can say things like people, who buy coffee and milk tend to buy sugar or people, who buy milk almost never buy milk substitute. So, why is this exciting well now, a super market knows, where to keep which product in its setup.

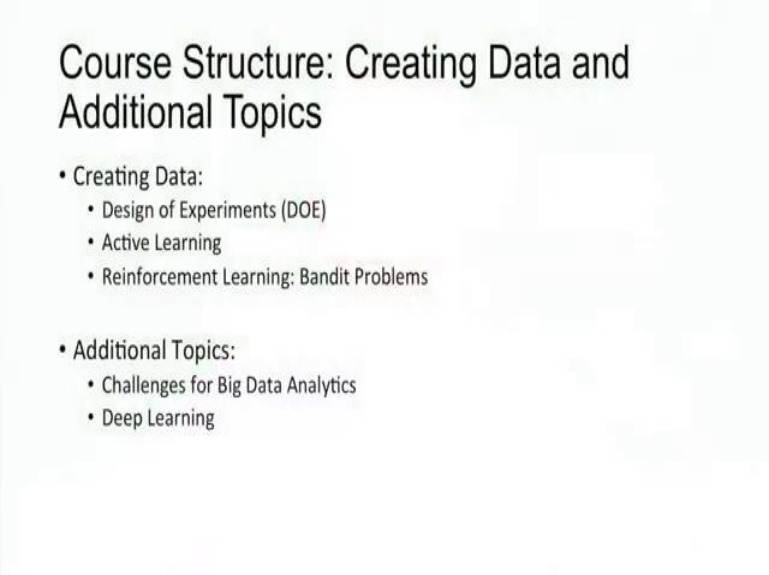
So, you can lay things out where if coffee and milk bought, bought together can coffee and milk been next to each other and so on and so forth. But, association rule mining again need not be confined to a market basket context as long as you can again break down the data into simple thing of objects and features that will enable you to perhaps considered association rule mining. But, the important thing is again look there is no one variable that you are targeting.

So, it is this is not a classification exercise it is not a prediction exercise of you know, who is most likely to buy sugar, but it is that any variable can be can become the relationship variable. So, there is no strict output variable within association rule mining there are a couple of challenges and we will be talking about that as well in this course. So, the idea of creating many complex rules becomes computationally very hard, so what do you do with that.

And how do you how do you say a particular rule and when I say the rule of here something like people who buy coffee and milk tend to buy sugar how do you evaluate how good particular rule is in that is some of the core challenges in association rule

mining.

(Refer Slide Time: 23:15)



Finally, we come to the last module in this course overview and this is a module on creating data and we feel fairly strong about this module, because we see that this is a problem that is often faced with many organizations, where their interested in data analytics their interested they see this buzz word floating around sometimes big data and so on. And there a little unsure about, how that applies to them when they just is no data available or they have not really captured it.

And we feel like a set of topics in here should help companies or organizations understand this part of the data analytics process better. So, there are three major topics that we are going to be covering here the first topic is on design of experiments. So, you have no data, but you want to take a data centric approach this whole idea of the data driven decision making you want the data to tell you what is the right thing to do.

Perhaps the best thing for you to do is to conduct an experiment you try certain options and essentially you explicitly change that input variable in different settings and different points of time and then see, what happens and you use that data that just generated from the experiment to essentially do something like the regression or a supervised learning technique and thereby make decisions.

So, design of experiments is would be one way of going forward there this is other really

exciting area called active learning, which is the part of the machine learning machine learning words active learning comes about when you might have some data or very little data broader umbrella of machine learning and the idea here is also one of you can think of it as one of experimenting you could think of is one of sequentially quarrying the system. In other and its fairly expensive to gather this data.

So, instead of just doing a blind approach of fixing senses all over the place which might be an expensive proposition and therefore, you do not have this data. Is there something we can to do the partial knowledge and can we sequentially quarry this system in what we mean by quarrying the system here is can we sequentially put senses can we sequentially see, which data we want to gather.

Because, we do not have lot of data in we do not have enough to make conclusions. And at the same time we cannot just say let us start lets commission data gathering exercise only because it is hard to do that. So, given that we have fixed resources towards gathering data active learning is an area, where you sequentially go and gather data, but you only gather the critical data that you need in order to mine it or in order to process it for coming up with insides.

Now, the third area that we will be covering in this section, which is creating data for data analytics is the area of reinforcement learning and this is also reinforcement learning is also subset of the larger machine learning umbrella and most specifically in the reinforcement learning re now, we are going to focusing on series of problems called the bandit problems and the context here is that you do not start with you do not have to start with any data or you might have some partial data.

But, you just cannot go about experimenting to create data for a verity of reasons one may be cannot create this kind of lab setting which might be needed to conduct the experiments in create data. But, more importantly it is also possible that you cannot experiment only, because it effects the end user in some way you cannot essentially go off line and do your experiments. So, here you are do not have any data you want to try something is right because you do not know of what you doing is the best thing.

But, you cannot just commission in experimentation exercise to create the data, which then gets analyzed, and then tells you what is the best, because when your experimenting you might be doing some horrible things and those horrible things might affect an end

user. So, the whole idea behind banded problems is you can think of it one way of think it it is a form of experimentation and learning in an online setting. So, you do not only care about how much you are learning from the data, but you also care about how value of performing.

And in fact, the experimenting itself becomes consequents because your grand objective in the banded problems tends to be one of performing as well as you can over sometime horizon. So, because you need to perform as well as you can, which is defined by some notion of how you do cumulatively you wind up a trying a few things out just, so that over time you are not continuously doing something that is not in your best interest.

So, these are three these could be fairly useful techniques or tools to use when you are in an online setting or when you are in a setting were you do not have much data. Finally, I just wanted to mentioned that in addition to some of these topics that we have discussed we will also going to have some a couple of modules on major challenges for big data analytics and what I guess big data analytics its of means in the world today.

And we are also going to be talking about some of the more popular techniques contemporary techniques like deep learning especially when we cover concepts and artificial neural networks and so on. So, with that we conclude this second session of the course overview and starting next session we would be directly diving into the content itself and I mean we cover the content today. But, again the spirit of it was to give you an idea of what it is that we are going to be covering in this course and I hope you found it interesting and I look forward to having you join us in the next session.

Thank you.