**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 – 04**
**Lecture - 16**
**Introduction to Machine Learning**

Welcome to this module on Machine Learning. So, till now we have mostly looked at data analysis. So, all the tools and techniques that we have looked at have to do with analyzing data and trying to understand the data better with possibly the exception of regression. From now on for the rest of the course, we will be looking at how you can infer models about the process that generated the data by looking at the data alone. This is essentially the idea behind machine learning. So, what we call, even though it is a kind of a fancy name when you have a visions of Robots and terminated suiting around in your head, but machine learning is essentially trying to learn models about the process that generated the data from the data itself.

(Refer Slide Time: 01:1)



So, you could look at examples where you learn about, how the rainfall pattern varies over a season. So, you could say that rainfall pattern is very intermittent over the season or you could look at, how the temperature of certain equipment is varying with time of operation. You could say that the temperature showing a linear growth, it is possible to

do this kind of learning from the data by a machine and this kind of machine learning algorithms becomes essential, when we move away from data analysis into either predictive, descriptive or prescriptive analytics.

For example, I can ask what will be the temperature in 15 minutes of a particular equipment know and if I know that the pattern is that of a linear growth, I should be able to tell you what the temperature would be in 15 minutes and likewise, I can ask you what are the areas that are likely to receive rain in the next season. Then, if I know what the variation has been seasonally with the data, then I should be able to tell you, what are the areas that will likely to receive rain in the next year and so, when I talk about prescriptive analytics, so the first two questions I was asking you or more about questions on the system and prescriptive analytics, you will be asking questions about what I should do in response to the patterns that you are describing.

For example, I can find out patterns in stocked data and I could ask a question, should I invest in this stock or not. For all of these kinds of analytics, so we need to have a technique tools and techniques from machine learning. So, what exactly is machine learning?

(Refer Slide Time: 02:46)



So, I will fall back on this old definition from Tom Mitchell. So, Tom Mitchell said an agent is said to learn from experience with respect to some class of tasks and the performance measure P, if the learner's performance at tasks as measured by P improves with experience. There is lot of qualifications here, in the first case he did not say a

machine he just said an agent. So, in Tom Mitchells opinion, this applies to all learning agents, it could be humans, animals or machines.

So, an agent is said to learn from experience, so you have to have experience in trying to solve something and now, you measure this experience with respect to some class of tasks. I mean, it is not that you can learn every thing you have to be very specific about, what is said that you are trying to learn and the performance measure P. So, you need to know how well you are doing in that particular task that you are learning about it.

And then, if you are said to learn if your performance as measured by P keeps improving with experience the very, very inclusive definition of learning. So, you have to be very careful when you use this, because you could even apply it to desired scenarios. For example, you could think of a slipper, a new slipper that becomes more comfortable as you keep wearing it. So, you cannot really say that the slipper is learning to fit your feet with experience. So, you have to be careful about, how you apply this definition, I mean it is a fairly serviceable definition as we will see as we go long.

(Refer Slide Time: 04:31)



So, the rest of the course we will be looking at three different machine learning paradigms. So, the first one is called supervised learning, where you expected to learn a mapping from certain input variables to output variables. So, we already looked at one example of such a supervised task, when you looked at regression, so where the output variables was a continuous valued variable. So, and then the input was described by a set of attributes and if the output is a categorical output, where it could be one of many

classes.

So, you see, is the patient sick or, see he is not sick, will the customer default on the payment or will they not default on the payment. So, these are like categorical attributes, both these examples, where the output could be either 0 or 1, you could think of outputs with multiple such levels. In such cases, the learning problem is called the classification problem. So, but what distinguishes supervised learning from the other forms of learning is that, in the form of experience that you will get.

So, whenever I give you an input, the sample input I will always have an expected output for this input. So, I am going to, I will give you a set of samples which consists of an input vector and an expected output for that and that is what makes you supervised learning. In unsupervised learning, the goal here is to discover patterns in the data that need not necessarily be any output that I am trying to produce.

So, there are many different unsupervised learning problems, we will be particularly looking at two in this course. One of them is called clustering, but the idea is to find cohesive grouping among the data points that are given to you, so in order to find any patterns that are occurring. So, will see how this works in a little while, so but you can readily think of the rainfall task that I was talking about earlier. So, you can group regions that somehow are similar in their rainfall behavior.

And the second unsupervised learning task is known as the association mining or association rule mining sometimes and the goal here is primarily to find the data points that occur together or co occur frequently. So, in association rule mining you essentially you are trying to figure out, which data is associated with which other data. So, how it is which co occur frequently. So, in both of these cases, as you can see there is no real output that you are expected to produce except to find the patterns on the data.

And the third class of machine learning problems which are called the reinforcement learning, essentially has to do with learning how you would control a system. We will talk more about it towards the end of the course, but roughly you can think of the following problem. So, how did you learn to cycle? So, it is not just discovering patterns, there is nobody giving you an input output pairs, that tell you how to cycle and somebody actually does give you an input output pair, then probably not learn to cycle.

If your cycle is tilting by 30 degrees to the horizontal, then you should push down with your right foot with so many Newton's of pressure. If somebody gives you directions like

that, you will never going to cycle. So, you have to do some kind of style and error learning. So, that is essentially what reinforcement learning talks about. So, you will do a little bit of this towards the end of the course.

(Refer Slide Time: 08:21)



So, the different machine learning tasks that we are talking about, so classification, regression, which are essentially supervised learning problems and you remember, I told you that you really need a measure by which you are going to decide whether the algorithm is performing well or not and the measure in the case of classification and regression is just going to be error. In the case of classification, it will be the classification error that is the how many mistakes you make in predicting the categorical outputs and in the case of regression, it is going to be the prediction error which is, how far away you are from the actual value that you need to predict.

In the case of unsupervised learning problems, it is a little tricky as to what the measures should be and there are, let say when we look at clustering and association rules, we will talk about many such measures in detail. But, roughly, in clustering one of the measures is how tight your clusters are or how scattered they are and I am talking very roughly here, but we will formulize as we go long. And in the case of associations, it is more on how confident you are that these two items are associated and what is the fraction of the population in which these associations appear. So, that is what we mean by support and confidence.

And again as I said, this is just to give you an idea that for every task you are going to

have an associated measure and we will elaborate on the actual measures as we go along.

(Refer Slide Time: 09:54)



So, having said this, there are many challenges that we need to address. So, one, the first challenge in any machine learning problem is to figuring out, how good is your model. If somebody gives you a machine learning algorithm and say, there, so here is a model that has been learnt by the algorithm, how do you decide how good that model is. So, you could use the measures that I showed you on the previous slide, but that could be other ways of deciding, how would the model is. So, I will be elaborate on this as we go long.

For example, you can build a very, very, very, very detailed model that gives you almost zero error on all the data that is given to you at the beginning. But, then this model might essentially be useless, when it starts looking at unseen data. When I actually wanted to make predictions on data that I have not seen before, this model might not perform well at all. So, how do I look at that kind of a trade off?

So, there are different ways of measuring, how good a model is and then, given the data and given the class of models, here we are going to look at how do I choose the right model. So, that is the second challenge and many different machine learning algorithms are all about answering this, the choice of the model question, but then, it is not all about modeling. So, you have to be very, very cognizant of the data that you are operating with.

So, the first question that plagues all of us is, do I have enough data. That might surprised some of you, who have heard of terms like big data and when having excess of

data, data delusions, so on and so forth. But, then getting in a supervised learning problem, getting label data is incredibly hard and so, you have to have an expert that this looking at all the data and then labeling them for you. So, getting such labeled data is incredibly hard and so, do you have sufficient labeled data or do I, can I make use of unlabeled data in a clever way. So, these are all kinds of question that you will have to think about.

Is the data of sufficient quality that could be errors in the data? For example, age could be recorded as 225 or there could be noise in very low resolution images, that you are feeding your algorithm and so, the algorithm is not able to make out, what is that in the image or it could be that some values are missing or not been recorded in your data and then, your algorithm has to deal with that. So, it is a very important question that the data has sufficient quality and in any almost, in every large scale, real life machine learning insulation, a lot of effort has to go into cleaning the data.

So, making sure the data is of a sufficient quality to feed into your system and, so how confident can you be of the results at the end of it. It is both the factor of the data quality, the data volume and as well as the machine learning, exact machine learning algorithm that you end up using. So, all of these factors are together influence, how confident you can be of the results and at the end of it, there is one very important question, which typically you know it not address that carefully is, am I describing the data correctly.

So, for example there are two classes of questions I could ask you. So, the first thing is like, do I have enough information about what I am trying to classify or age and income enough to describe all my customers or should I look at gender also, how would I answer such a question. A lot of it actually comes from the data analysis that we have been doing so far or age and income alone sufficient to explain all the variance that I see in the data or should I include another variable, in order to make my model more accurate. So, such questions will tie back into all of the analysis that we have seen so far.

And the next question which is more often, it is partly an engineering issue and partly a theoretical issue is, how I should represent my variables you know, how do I represent age, that age can be a number. Well, age can sometimes be merely a number, but sometimes you can classify age into different levels as young, middle aged and old. So, what kind of an encoding would you choose for representing age and again, it goes back to the data analysis if you are talked about and…

So, these are things that you have to pay a lot of attention too and the more often, than not when you are looking at a simple exercises that you could do on the web, you end up looking at the data that is already where this questions have already been answered and they are given to you and you are able to do it very easily. But, when you go out and you are trying this, all your first real problem using machine learning you will find that these are very important questions that we have to answer. So, in the next module we will look at more detail at the different learning paradigms that we talked about today.