NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture 1

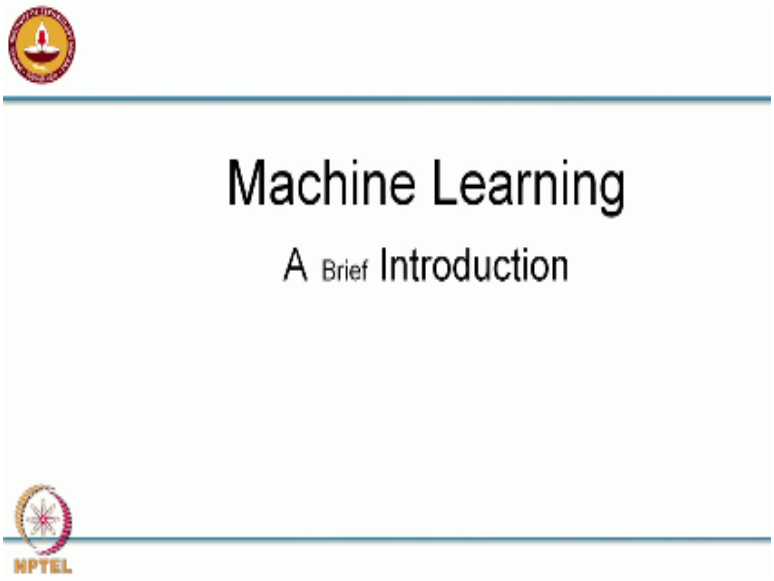
Prof. Balaraman Ravindran
Computer Scince and Engineering
Indian Institute of Technology Madras

Introduction to Machine Learning

Hello everyone and welcome to this NPTEL course on an introduction to machine learning in this course we will have a quick introduction to machine learning and this will not be very deep in a mathematical sense but it will have some amount of mathematical trigger and what we will be doing in this course is covering different paradigms of machine learning and with special emphasis on classification and regression tasks and also will introduce you to various other machine learning paradigms.

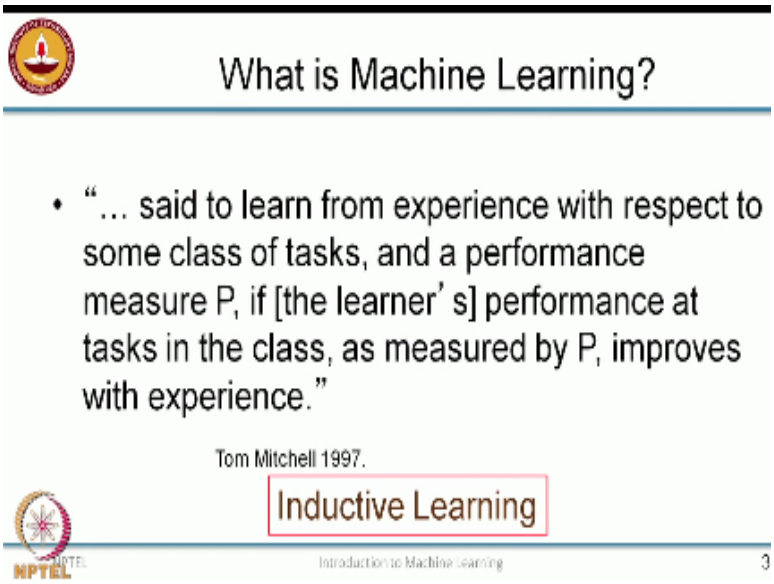
In this introductory lecture set of lectures I will give a very quick overview of the different kinds of machine learning paradigms and therefore I call this lectures machine learning.

(Refer Slide Time: 01:03)



A brief introduction with emphasis on brief right, so the rest of the course would be a more elongated introduction to machine learning right.

(Refer Slide Time: 01:16)



So what is machine learning so I will start off with a canonical definition put out by Tom Mitchell in 97 and so a machine or an agent I deliberately leave the beginning undefined because you could also apply this to non machines like biological agents so an agent is said to learn from experience with respect to some class of tasks right and the performance measure P if the learners performance tasks in the class as measured by P improves with experience.

So what we get from this first thing is we have to define learning with respect to a specific class of tasks right it could be answering exams in a particular subject right or it could be diagnosing patients of a specific illness right. So but we have to be very careful about defining the set of tasks on which we are going to define this learning right, and the second thing we need is of a performance measure P right so in the absence of a performance measure P you would start to make vague statement like oh I think something is happening right that seems to be a change and something learned is there is some learning going on and stuff like that.

So if you want to be clearer about measuring whether learning is happening or not you first need to define some kind of performance criteria right. So for example if you talk about answering questions in an exam your performance criterion could very well be the number of marks that you get or if you talk about diagnosing illness then your performance measure would be the number of patients that you say are the number of patients who did not have adverse reaction to the drugs you gave them there could be variety of ways of defining performance measures

depending on what you are looking for right and the third important component here is experience right.

So with experience the performance has to improve right and so what we mean by experience here in the case of writing exams it could be writing more exams right so the more the number of exams you write the better you write it better you get it test taking or it could be a patient's in the case of diagnosing illnesses like the more patients that you look at the better you become at diagnosing illness right.

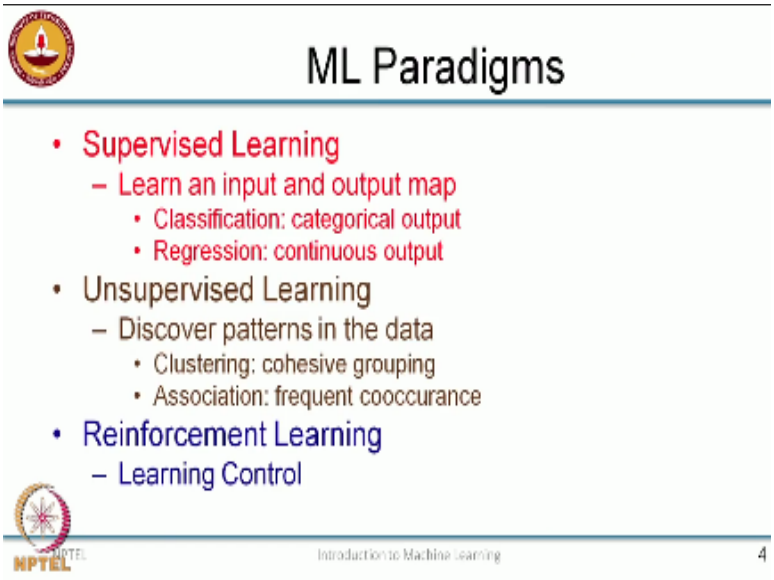
So these are the three components so you need a class of tasks you need a performance measure and you need some well-defined experience so this kind of learning right where you are learning to improve your performance based on experience is known as a this kind of learning where you are trying to where you learn to improve your performance with experience is known as inductive learning.

And then the basis of inductive learning goes back several centuries people have been debating about inductive learning for hundreds of years now and are only more recently we have started to have more quantified mechanisms of learning right. So but one thing I always point out to people is that if you take this definition with a pinch of salt, so for example you could think about the task as fitting your foot comfortably right.

So you could talk about whether a slipper fits your foot comfortably or let me put so I always say that you should take this definition with a pinch of salt because take the example of a slipper you know, so the slipper is supposed to give protection to your foot right and a performance measure for the slipper would be whether it is fitting the leg comfortably or not or whether it is you know as people say there is biting your leg or is it Chaffin your feet right and with experience you know as the slipper knows more and more about your foot as you keep varying the slipper for longer periods of time it becomes better at the task of fitting your foot right as measured by whether it is shattering your foot or whether it is biting your foot or not right.

So would you say that the slipper is learned to fit to your foot well by this definition yes right so we have to take this with a pinch of salt and so not every system that confirms to this definition of learning can be set to learn usually okay.

(Refer Slide Time: 06:11)



So going on so there are different machine learning paradigms that we will talk about and the first one is supervised learning where you learn an input to output map right so you are given some kind of an input it could be a description of the patient who comes to comes to the clinic and the output that have to produce is whether the patient has a certain disease or not so this they had to learn this kind of an input to output map or the input could be some kind of equation right and then output would be the answer to the question or it could be a true or false question I give you a description of the question you have to give me true or false as the output.

And in supervised learning what you essentially do is on a mapping from this input to the required output right if the output that you are looking for happens to be a categorical output like whether he has a disease or does not have a disease or whether the answer is true or false then the supervised learning problem is called the classification problem right and if the output happens to be a continuous value like, so how long will this product last before it fails right or what is the expected rainfall tomorrow right so those kinds of problems they would be called as regression problems.

These are supervised learning problems where the output is a continuous value and these are called as regression problems. So we will look at in more detail classification and regression as we go on right, so the second class of problems are known as unsupervised learning problems right where the goal is not really to produce an output in response to an input but given a set of in data right we have to discover patterns in the data right. So that is more of the testicle

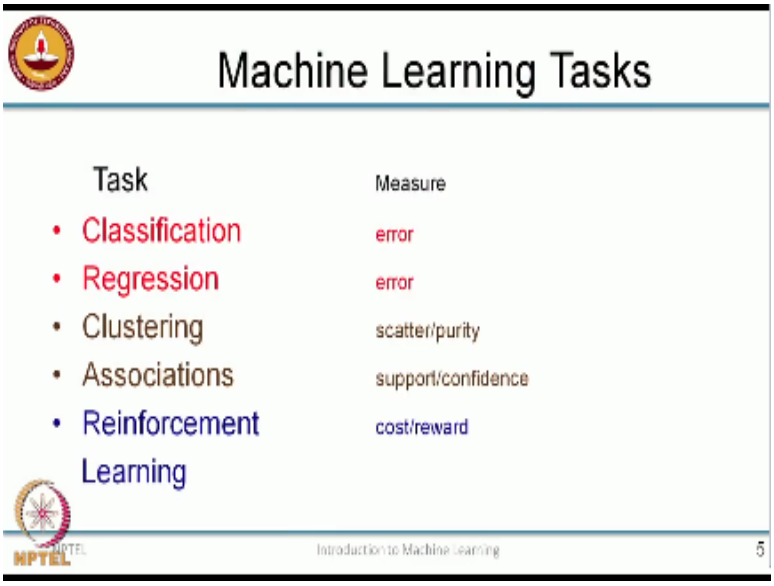
unsupervised learning there is no real desired output that we are looking for right we are more interested in finding patterns in the data.

So clustering right is one task one unsupervised learning task where you are interested in finding cohesive groups among the input pattern right, for example I might be looking at customers who come to my shop right and I want to figure out if there are categories of customers like so maybe college students could be one category and sewing IT professionals could be another category and so on so forth and when I'm looking at this kinds of grouping in my data, so I would call that a clustering task right.

So the other popular unsupervised learning paradigm is known as the Association rule mining or frequent pattern mining where you are interested in finding a frequent co-occurrence of items right in the data that is given to you so whenever A comes to my shop B also comes to my shop right. So those kinds of co-occurrence so I can always say that okay if I see A then there is likely very likely that B is also in my shop somewhere you know so I can learn these kinds of associations between data right.

And again we look at this later in more detail these are I mean there are many different variants on supervised and unsupervised learning but these are the main ones that we look at so the third form of learning which is called reinforcement learning it is neither supervised or unsupervised in nature and typically these are problems where you are learning to control the behavior of a system and I will give you more intuition intone enforcement learning now in one of the later modules, so like I said earlier.

(Refer Slide Time: 09:33)



So for every task right, so you need to have some kind of a performance measure so if you are looking at classification the performance measure is going to be classification error so typically right. So we will talk about many, many different performance measures in the duration of this course but the typical performance measure you would want to use this classification error it's how many of the items or how many of the patients did I get incorrect so how many of them who are not having the disease today predict had the disease and how many of them that had the disease that I missed right.

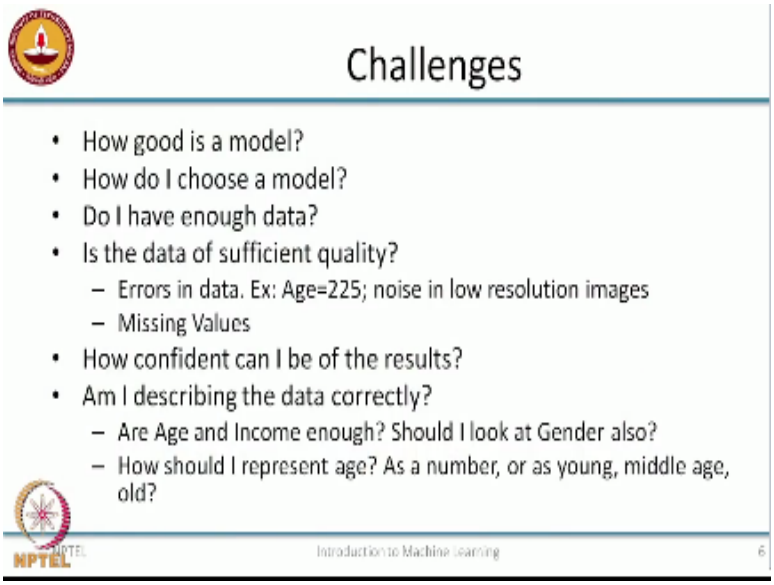
So that would be one of the measures that I would use and that would be the measure that we want to use but we will see later that often that is not is not possible to actually learn directly with respect to this measure. So we use other forms right and likewise for regression again so we have the prediction error suppose I say it is going to rain like 23 millimeters and then it ends up raining like 49centimeters I do not know so that is a huge prediction error right and in terms of clustering so this is little becomes a little trickier to define performance measures we don't know what is a good clustering algorithm because we do not know what how to measure the quality of clusters.

So people come up with all different kinds of measures and so one of the more popular ones is a scatter or spread of the cluster that essentially tells you how spread out the points are that belong to a single group if you remember we are supposed to find cohesive groups, so if the group is not that cohesive it's not all of them are not together then you would say the clustering is of a poorer

quality and if you have other ways of measuring things like Alec was telling you, so if you know that people are college students right and then you can figure out that how many what fraction of your cluster or college students.

So you can do this kinds of external evaluations so one measure that people use popularly there is known as purity right and in the Association rule mining we use variety of measures called support and confidence that takes a little bit of work to explain support in confidence so I will defer it and I talked about Association rules in detail and in more in the reinforcement learning tasks so if we remember I told you it is learning to control so you are going to have a cost for controlling the system and also the measure here is cost and you would like to minimize the cost that you are going to accrue while controlling the system. So these are the basic machine learning tasks.

(Refer Slide Time: 12:11)



So there are several challenges when you are trying to build a build a machine learning solution right so a few of these I have listed on this slide right the first one is you have to think about how good is a model that you have learned right so I talked about a few measures on the previous slide but often those are not sufficient there are other practical considerations that come into play and we will look at some of these towards thee there was a middle of the course somewhere right and the bulk of the time would be spent on answering the second question which is how do I choose a model right.

So given some kind of data which will be the experience that we are talking about so given this experience how would I choose how would I choose a model right that somehow learns what I want to do right so how that improves itself with experience and so on so how do I choose this model and how do I actually find the parameters of the model that gives me the right answer right.

So this is what we will spend much of our time on in this course and then there are a whole bunch of other things that you really have to answer to be able to build a useful machine loose full data analytics or data mining solutions questions like do I have enough data do I have enough experience to say that my model is good right it's the data efficient quality that could be errors in the data right suppose I have medical data and a is recorded as 225, so what does that mean it could be 225 days in which case it is a reasonable number it could be 22.5 years again is a reasonable number or 22.5 months is reasonable.

But if it is 225 years it's not a reasonable number so there is something wrong in the data right so how do you handle these things or noise in images right or missing values so I will talk briefly about handling missing values later in the course but this is as I mentioned in the beginning is a machine learning course right and this is not there is not primarily it is primarily concerned about the algorithms of machine learning and the and the math and the intuition behind those and not necessarily about the questions of building a practical systems based on this.

So I will be talking about many of these issues during the course but just that I want to reiterate that will not be the focus right and so the next challenge I have listed here is how confident can I be of the results and I want that I certainly we will talk a little bit because the whole premise of reporting machine learning results depends on how confident you can be of the results right and the last question am I describing the data correctly.

So that is a very, very domain dependent and the question that you can answer only with your experience as a machine learning or a data scientist professional or with time right, so but there are typical questions that you would like to ask that are there on the slides so from the next in the next module we look at the different learning paradigms in slightly more detail.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in