

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

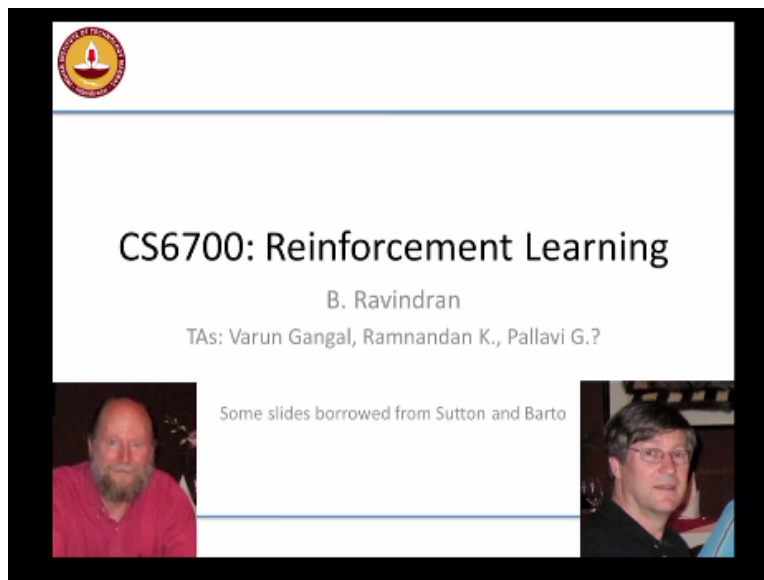
Introduction to Machine Learning

Lecture 83

Introduction to Reinforcement Learning


**Prof: Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras**

(Refer Slide Time: 00:15)



Right, so this is really the different kind of learning than what we look at it label right. So for the mission learning we looked at familiar modes of mission learning.


(Refer Slide Time: 00:32)



Learning to Control

- Familiar models of machine learning
 - Supervised: Classification, Regression, etc.
 - Unsupervised: Clustering, Frequent patterns, etc.

- How did you learn to cycle?



Reinforcement Learning2

Where the idea was to learn from data, you know so you have given lot of data as training instances for you, and then essentially you are trying to learn from those training instance as to what to do right. And there were different kinds of problems that we are looking at, so one was supervised learning problem in which you were looking at classification, regression. So in the mission learning class we looked at learning from data right.

Primarily, so one of the models we looked at was supervised learning right, as we learnt about classification and regression, and the goal there was to learn in the mapping from an input space to a output, which could make categorical output in this case of classification, it could be a continuous output in this case is called regression right. So if you have not peer in the ML class do not worry about it right.

Because this is just to tell you that RL is not whatever you learnt in the ML class okay. So if you have not learnt anything in the ML class then you do not have anything to unlearn, so do not worry. So the second kind of learning thing we left it to unsupervised learning, when there was really no output that was expected of you right. Since, therefore there was no supervision, the goal was to find patterns in the input data, I will give you lot of data points, you can find out if there are groupings of similar kinds of data points, you can divide them into segments right.

So that kind of thing was called clustering right or you are asked to figure out if there were frequently repeating patterns in the data right. And so this is called frequent pattern mining or derived problems that is association to the mining and so on so forth right. So people who heard

me give this analogy multiple, multiple times before, but this is the most apt one, how we learn to cycle right.

So was it supervised learning, yeah okay? How did you learn to cycle? Was that somebody tell you how to cycle and then you just followed their instruction okay. First of all, do you know how to cycle? Yes, do you know how to cycle, yeah you yes okay. How did you learn to cycle? Further on a couple of times and that automatically mining with cycle, you do not actually figure out how to run to fall down right.

So falling down alone is not enough, but you have to try different things right. It is not supervised learning right, it is really not supervised learning. How much time ever you think? Because now that I have given this talk multiple times, people are getting mice to it right. Earlier when I used to ask these people, people say of course it is supervised learning, my uncle was there holding me, or my father was telling me what to do and so on so forth right.

And best what did they tell you? Hey, look out, look out do not fall down right. So that does not count our supervision right. So keep your body right, keep your body and some kind of very vague instructions over love they are giving you right. Supervised learning would mean that, so we will get on the cycle as it is okay, now push down with your left foot with three pounds of pressure right.

And move your center of gravity 3° to the right, right. So this is something, somebody has to give you exactly what is the control signals that we have to give to your body in order for you to cycle right. Then that will be supervised learning right, if somebody actually gives you supervision at that scale you probably have never learn to cycle, we think about it right, because it is such a complex dynamical system and somebody gives you controller that level like this is input at that level you will never learn to cycle.

And so immediately people flip and say that it was unsupervised learning right because here of course nobody told me how to cycle therefore it is unsupervised learning, so if it is truly unsupervised learning what should have happened is you should have watched 100's of videos of people cycling figure out what is a pattern of cycling that they do okay and get on a cycle and reproduce it.

Right so that is essentially what unsupervised learning would be you just have lot of data right and based on the data you figure out what the patterns are and then you try to execute those patterns, that does not work right you can watch hours and hours of somebody playing 5 simulator you cannot go on fly a plane, right so you have to get on the cycle yourself and you have try things yourself right so that is the chucks here right so what it is how do you learn to cyclic is nether of the above right is neither supervised nor unsupervised it is a different paradane.

So the reason I always start out my talks not just in the class but in general when I talk reinforcement learning is because people always talk about reinforcement learning and unsupervised learning right it is always really axe me because unsupervised leaning it is just because you do not have a classification error or class label does not make it unsupervised learning. It is completely different form of learning and so reinforcement learning is essentially this mathematical formulation for this trial and error kind of learning right.

So how do you learn from this case minimal feedback you know falling down hertz or somebody your mom or somebody stands there and claps when you finally manage to get on the cycle you know that is kind of positive reinforcement why do not you fall down you get hurt right that is kind of a negative feedback how do you just use this kinds of minimal feedback and you learn to cycle so this is essentially the chucks of what reinforcement learning is about trial and error right.

So the goal here is to learn about a system through interacting with the system right it is not something that is done completely offline okay you have some notion of interaction with the system okay and you learn about the system through that interaction.

(Refer Slide Time: 06:22)



Reinforcement Learning

- A trial-and-error learning paradigm
- Learn about a system through interaction
- Inspired by behavioural psychology!
 - Pavlov's dog

Reinforcement learning originally was inspired by behavioral psychology right so one of the earliest reinforcement systems that are studied with the Pavlov's dog how many of you know of the Pavlov's dog experiment what does the Pavlov's dog experiment okay that is okay so that is called a condition reflects so when the dog looks at the food and starts salivating right it is a primary response because there is a reason for it to salivate on the site of food in area by exactly it is preparing to digest the food you know may sure the food is preparing to digest the food it start salivating right.

So if then now if you think about it right hearing the bell at it salivates what is it doping preparing to digest the bell no right so when you ring the bell and then serve the food the dog forms an association between the bell and food right and later on when you just ring the bell without even serving the food the dog starts salivating in response to digesting the food that it expects to be delivered right.

So it essentially the food is the pay of you know the food is like a reward for it and it is learned to form associations between signals in this case which was a bell like an input signal which was the bell and the reward that is going to get right, so this is call the behavioral conditioning right and so inspired by these kinds of experiments on then more complex behavioral experiments or animals now started to come up with the different theories to explain how learning proceeds right in fact some of the earlier reinforcement learning papers.

Appeared in the behavioral physiology generals right the earliest paper by sub appeared in brain and behavioral science here just to go back I needed to only it is saying something about certain border this is larger audience we can tell that worth so the we have going to follow a text book written by Rich Riordan and Anne Bronte but more importantly they are also kind of the cofounders of the modern field of the reinforcement learning right so in 1983 they wrote a paper adaptive on like element that learn.

The control behavior of something to their effect right, and that essentially kick started this whole modern field of reinforcement learning so the concept of reinforcement learning like I said goes back to power and earlier right, people have been talking about those kind of behavioral conditioning and learning ensure but the whole modern computational techniques that people use in the reinforcement learning just started by 9.33 so what is reinforcement learning right.

(Refer Slide Time: 09:41)



What is Reinforcement Learning?

- Learning about stimuli and actions based on rewards and punishments alone.
- No detailed supervision available
- Trial-and-error learning
- Delayed rewards
- Sequence of actions required to obtain reward
- Associative learning required
 - Need to associate actions to states
- Learn about policies not just actions
- Typically in a stochastic world

So we learning about stimuli right the inputs that are coming to you and the actions that you can take and responds to it right learning about the stimuli, only from rewards and punishments okay, so you are not going to get anything else food is a reward right following down and scraping your hand is a punishment, right so only from this kinds of rewards and punishments alone right there is no detail supervision available nobody tells you, what is the respond that you should give to a specific input.

Right suppose you applying a game or multiple ways in which you can learn to play again, right so you can learn to play chess by looking at a board position right and then looking at table right that tells you for this board position this is the move you have to make, right and then you go and make the move right so that is a kind of supervision that you could get you know that gives you a kind of a mapping from the input to the output right and essentially we learn to generalize from that.

So this is what we mean by details supervision so another like way of learning to chess is just so we have a opposite site in front of him and you just make sequence of moves at end of the move you win okay you get a reward right some would you place you set an rupees okay if you louse you have to play the opponent and rupees, so that happens right solve the feedback you are gain to get right solve the feedback you are gain to gin right whether you are going to get the 10 rupees or going to lose the 10 rupees at the end of the game.

So nobody tells you given this position this is the move you should have made right that is what we mean by same learning from rewards and punishments in the absence of details supervision, okay is it clear okay and curial component to this, this trial and error learning because since say do not know what is a right thing to do given an input, right I need to try multiple things to see what the outcome will be right.

I need to try to different things to see if I am going to get the reward or not right if I do not try different things way I am not going to be able to learn anything at all right, so we will I can give you more formal mathematical reasons for why we need all of this as we line go on but this intuitively you can understand this as requiring a exploration, so that you know what the right out comers right.

And there are bunch of things which are also characteristic of reinforcement learning problems on of those is that the outcomes right the rewards and punishments based on which you are learning can be fairly delayed in time they did not be temporarily closed to the thing that caused it I mean while our playing a game let us say right so you might you know now drop a batsman right and then he goes on to score like 150 or something like that, right.

So then you lose the match at the end of the day right but the event that caused you to lose the match is the drop catch that probably around the 12th over right and all it could be much more convoluted causal effect right so and how many of you followed cricket, my god it is really losing popularity yeah put your hands down I am not going to give you a cricket example then okay so a bunch of other things right.

So we talked about delayed rewards so rewards could come much later in time from the action that cause the reward to happen right for example let us go back to our cycling case right I might have done something stupid or I mean I met have gone over a stone somewhere right well I am cycling at a very high speed and might have been a small stone in the road and that will cause me to lose my balance right.

And though I will try my level best to get the balance back right I might not and I finally fall down and get hurt it does not mean what cause the falling down is the last action I tried right, I might have desperately try to jump of the cycle or something like that but that is not what cause the punishment right it what cause the punishment happened a few seconds ago when I ran over the stone, right.

So that could be this kind of a temporal disconnect between what causes the reward or punishment from the actual reward and punishment so it becomes a little tricky how do you are going to learn those things right learn the associations right, so quite often right you are going to need a sequence of actions to obtain a reward right it is not going to be like a one short thing right it is going to need a sequence of action to get the reward.

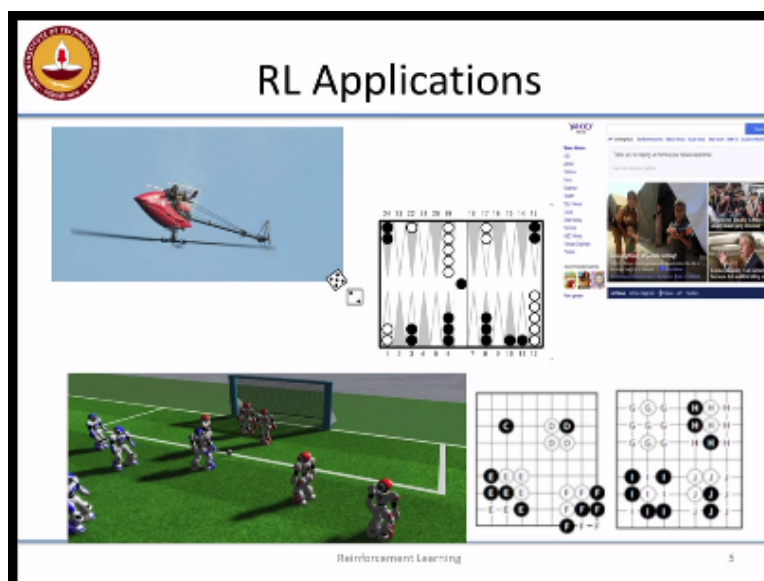
So again going back to the chess example right you are not going to get a reward every time you move a piece on the board right you have to finish playing the game at the end of the game if you actually manage to win you get a reward so it is a sequence of actions right and therefore you

need to learn some kind of a association between the inputs that you are seeing in this case it will be board positions right.

Or how fast the cycle is moving and how unbalanced you feel and so on so forth right two actions so inputs that you are getting sometimes which we will call state, right and the actions that you take in response to this input that you are seeing right, so this is essentially what you are going to be doing when you are solving a reinforcement learning problem, so this kind of associations or essentially it is known as policies, right.

So what you are essentially learning is a policy to behave in a world right so learning a policy to play chess or you are learning a policy to cycle right so this is essentially what you are learning right you are not just learning about individual actions right at all of this happens typically in a noisy stochastic world okay it does not makes these things more challenging, so these are all the different characteristics of reinforcement learning problems. So reinforcement learning has been used fairly successfully in a wide variety of applications, right

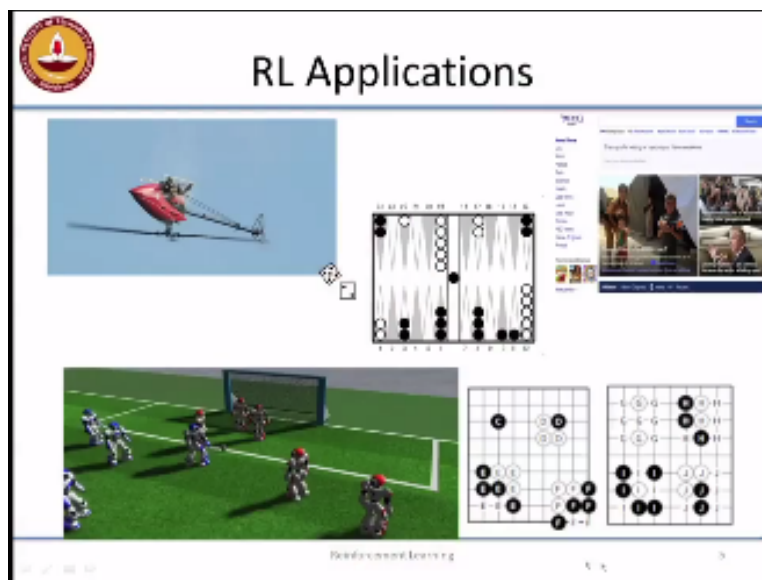
(Refer Slide Time: 15:30)



So you can see a helicopter here okay so that is not a cut and paste error here the helicopter is actually flying upside down okay, so the group at Strand ford and Buckley which have actually used reinforcement learning to train a helicopter to fly all kinds of things not just upside down and nor religion can do all kinds of tricks on the helicopter so I will show you a video in a minute.

And it is amazing piece of work right, I mean it was considered the show piece application for reinforcement learning I mean getting such a complex control system at work and it actually could know things at a much finer levels of control than a human begin could right, well it is after all a machine so you would expect that, but the tricky part was how it learn to control this complex system from without any human intervention, right. And in the middle right, so I have a couple of games there.

(Refer Slide Time: 16:35)



So that is can you see that okay, it is too small and arrow yeah, that is the game call back campaign right, so how many of you know about back campaign 1,2 others one maybe, how many of you know about ludo okay, you are sleeping okay, fine so back campaign is like a two player ludo okay, so you throw the dice you move piece around and you take them of the board right, so it is a fairly easy game but then you have all kinds of strategies that you would do with it.

But it is also hard game for computers to play because of the stochasticity right, and also because of a large branching factor that is there in the game right, so at each point there are many, many combinations in which you could move the board pieces around and then there is a die roll that adds in additional complexity, so people are not really I know getting great results and then there

is a person Jerry Tesoro from IBM who came up with something called neurogaming thing this is called neurogaming and that was trained using supervised learning under neural network, right.

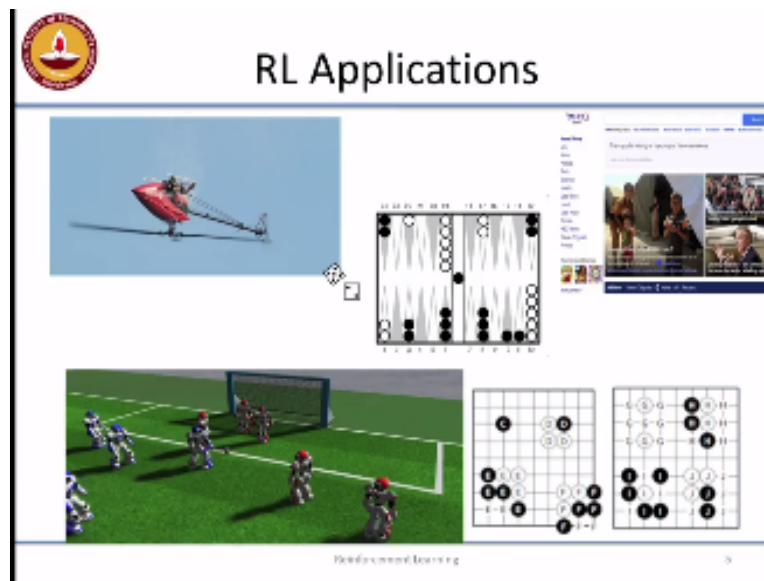
And so if you have done it recently it would have been called the deep learning version of neurogaming or something was he relate back in the 90's early 90's so is just called neural network version of bygamenon and it play really well for a computer program right, so that is essentially the best computer program bygamenon player at that point, and then Jerry heard about reinforcement learning he decided to train the reinforcement learning agent to play bygamenon, okay.

So what he did was set up this reinforcement learning agent which played against another copy of itself, let them play 100s and 100s of games okay, rather 1000s and 1000s of games. So essentially what they did was so you trained one copy for like 100 move or 100 games something then you move it here right, freeze it and then continue learning with this so essentially what is happening as you learn you are playing against better and better place gradually your opponent was also improving, right.

And then this was called self play right, so he trained back campaign using self play and it came to a point where the TD gammon where he is called it was even better than the human player of back campaign at that point in the world, right so they actually had head to head challenge with the human champion there is a world championship of back campaign you know it is apparently very popular in the middle east and people actually have world championships as a world championship of back campaign and so he challenge the human champion which IBM seems to do a lot right.

I mean they challenge Castro two matches and things like this, so he also challenged yes, these are a work for IBM he should realize right, people who spend a lot of resources getting computers to play games well probably be working for IBM you know. So Jerry had this thing and it beat the well world champion so he have reinforcement learning agent that is the best back campaign player in the world not no more best computer player or anything so we could actually make that claim right and there is another game there which snap shot from the game go right.

(Refer Slide Time: 20:13)



So people have play go who come at least one or two people have played go, people have played Othello okay that was a very few number why so we need one of those free games on going to I thought everybody plays that in some point rather the value rather play Othello and then watch paint try you know but anyway so goes like more a complex version of Othello if few it right it is again a very hard game for computers to play because a branching factor is homage right and it is actually a miracle that human even play this because the search trees and other things are really complex right.

So this is one case which clearly illustrate at humans actually solve problems and fundamentally different way than we try to write down in our algorithm because they seem to be making all kinds of intuitively it is an order to able to be play go right. So this person David Silver who currently works for Google deep mind and but before that he spent some time with Ajani Tesoro dymeum and at some point along the way he came up with this reinforcement learning agent call the TD search that place go at a decent level it still not like master human level performance but it perform place at a very decent level.

So this is a, what I am pointing out here is things are typically hard for traditional computer algorithm or even traditional machine learning approaches to solve AI as a good success. And here is another example I use this none or that I forgot which one right you have told me I should

use only one of those screens for pointing so it is hot for them record another one I forgot which one okay forget it. There are some robots on the bottom left of the screen right and so that is a snap shot from the UT Austrian robots occur team call Aston villa and they use reinforcement learning to get there robots to execute really complex strategies.

So this is really good but the nice thing about the robots occur application is that they do not use reinforcement learning alone rith5t they actually use the mix of different learning strategies and also planning and so on and so forth which is going on the other studio right. So they use a mix of different kinds of AI and machine learning technique in order to get a very, very completed agents it is very hard to beat and they are mean the champions I think two or three years running now in the humanoid league right.

And again hard control problems think like how do I take a spot cake you know those are the things for which they use reinforcement learning which it is really hard balancing problem so you have to basically balance the robot on one line and then string the other led so then you take the cake. So it going to be hard control problem, so they use RL to solve those rights and then up on the top right okay is an application which will probably be the one that actually makes money of all these three now all the others right that is on essentially on using reinforcement learning to solve online learning right.

So online learning is a used case where I do not have the feedback available to me right, so the feedbacks coming piece meal. So for example that is the case where we are having new stories that need to be shown to the people who come to my web page right and when people come to the page I have some editor will pick like 20 stories for me and from those 20 stories I have to figure out which are the once that I have put up prominent. And what is the feedback I am going to get?

Nobody tells me what stories that the user is going to like; I mean I cannot have the supervisor learning algorithm here right. So from the feedback I am going to get is, the user clicks on the story I am going to get a reward, user does not click on the story I am not going to get a reward, that is essential feedback that I am going to get. Nobody tells me anything before hand right, so I have to try out things.

I have to show different stories to figure out which one is going to click on and I have very few attempts to do, so how do I do this more effectively? People have done supervisor approach for solving this; it has work parallel successfully, so but re enforcement seems to be a much more natural way of modeling these problems. So not only these kind of new story collections, people use re enforcement idea in add selection. How do I see that on the sides when you go to Google or some other web page right?

So how those are ads selected, there might be some basic economic criterion for selecting for slate of adds okay. So here are the 10 ads which will probably give me the right pay off and then you can figure out, which 3 of them I am going to put it out over here and things like that and you could use the re enforcement learning solution for selecting those. This whole field is called computation advertising, it I slot more complex then what I explained; it is the component in advertisement as well.

(Refer Slide Time: 26:21)



Okay here is the video courtesy Andrew web page. People recognize the guy there it Is not the human size helicopter but still, it is parallel large amazing all of these learned by our agent. So this goes on for a while, so we will stop.

IIT Madras Production

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

www.nptel.ac.in

Copyrights Reserved