Unit-I

Reinforcement Learning

1. The Reinforcement Learning Problem

1.1 Reinforcement Learning

- Machine learning is a type of AI focused on building computer systems that learn from data, enabling software to improve its performance over time. There are three types of learning.

Supervised learning

- Supervised learning is a category of machine learning that uses labelled datasets to train algorithms to predict outcomes and recognize patterns. There are two types of supervised learning.
- *Classification* - Classification algorithms are used to classify data by predicting a categorical label or output variable based on the input data. One of the most common examples of classification algorithms in use is the spam filter in your email inbox. Here, a supervised learning model is trained to predict whether an email is spam or ham.
- *Regression* - Regression algorithms are used to predict a real or continuous value, where the algorithm detects a relationship between two or more variables. A common example of a regression task might be predicting a salary based on work experience.

Unsupervised Learning

- Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.
- There are three types of unsupervised learning.
- *Clustering* - is a method of grouping the objects into clusters such that objects with most similarities remain into a group and has less or no similarities with the objects of another group.
- *Association rule mining* - is a rule-based approach to reveal interesting relationships between data points in large datasets.
- *Dimensionality reduction* - is an unsupervised learning technique that reduces the number of features, or dimensions, in a dataset.

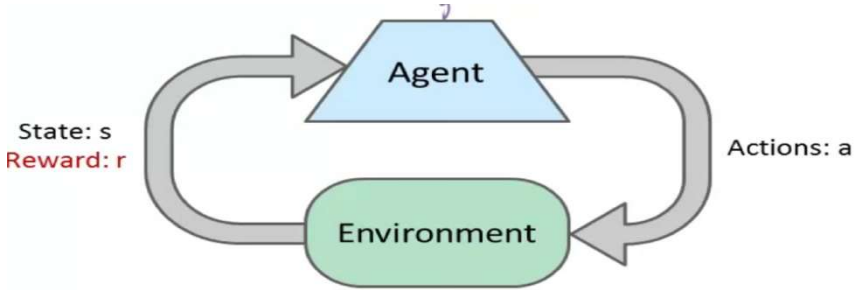
Reinforcement Learning

- Before jumping into reinforcement learning, let us try to understand what is meant by learning? How humans and animals learn?
- The way humans and animals learn is by interacting with surroundings or environment. It is also called as natural way of learning.
- For example, if you look at an infant or newly born baby, he or she does not have any information about the environment. So, the way it acquires the knowledge is by interacting with the environment. It collects the experiences and from these experiences it learns. There is no explicit teacher available for learning.

- For example, if the baby touches a hot cup, immediately the baby's sensory system triggers a feedback causing pain to the baby. Then the baby understands that this action shall not be performed. Hence, throughout our lives we will acquire wealth of information by interacting with the environment. From this information we will learn.
- So, learning from interacting with the environment is the hallmark of intelligence. In this course we will discuss the computational approach for learning from experiences. This computational approach is called as reinforcement learning.

So, what is reinforcement learning?

- Reinforcement learning is mapping situations (also called as states) to actions in order to maximize the numerical reward signal.
- In Reinforcement Learning, the agent interacts with the environment and explores it by itself. The primary goal of an agent in reinforcement learning is to improve the performance by getting the maximum positive rewards.



- RL solves a specific type of problems where decision making is sequential, and the goal is long-term, such as game-playing, robotics, etc.
- It is the core part of artificial intelligence. Here we do not need to pre-program the agent instead it learns from its own experience without any human intervention.

Example: Suppose there is an AI agent present within a maze environment, and his goal is to find the diamond. The agent interacts with the environment by performing some actions, and based on those actions, the state of the agent gets changed, and it also receives a reward or penalty as feedback.

Key Features of Reinforcement Learning

- Reinforcement problems are closed-loop problems because the learning system's actions influence its later inputs.
- The agent is not told which actions to take instead it must discover on its own which actions yield the most reward by trying them out.
- Actions may affect not only the immediate reward but also all subsequent rewards as well as future states.
- The agent may get a delayed reward.
- The environment is stochastic, and the agent needs to explore it to get the maximum positive rewards.

How reinforcement learning is different from supervised and unsupervised learning?

- Reinforcement learning is different from supervised learning. In supervised learning the agent is told what action it has to take on a particular situation. Whereas in reinforcement learning the agent has not told what action it has to take on a particular situation. It is the responsibility of the agent to decide what action it has to take on a particular situation on its own.

- Reinforcement learning is also different from unsupervised learning. The goal of unsupervised learning is to extract the hidden patterns or structure from the unlabelled data. But, in reinforcement learning instead of finding the hidden structure, we focus on maximizing the reward.

So, reinforcement learning is different from both supervised/unsupervised learning and we consider it as the third paradigm in the space of learning and computational intelligence.

One of the challenges that arise in reinforcement learning is the trade-off between exploration and exploitation.

Exploitation - is a strategy of using the accumulated knowledge to make decisions that maximize the expected reward.

Exploration - involves acquiring new information or opportunities that will maximize the expected reward.

Let's suppose people A and B are digging in a coal mine in the hope of getting a diamond inside it. Person B got success in finding the diamond before person A and walks off happily. After seeing him, person A gets a bit greedy and thinks he too might get success in finding diamond at the same place where person B was digging coal. This action performed by person A is called **greedy action**, and this policy is known as **a greedy policy.** But person A was unknown because a bigger diamond was buried in that place where he was initially digging the coal, and this greedy policy would fail in this situation.

In this example, person A only got knowledge of the place where person B was digging but had no knowledge of what lies beyond that depth. But in the actual scenario, the diamond can also be buried in the same place where he was digging initially or some completely another place. Hence, with this partial knowledge about getting more rewards, our reinforcement learning agent will be in a dilemma on whether to exploit the partial knowledge to receive some rewards or it should explore unknown actions which could result in many rewards.

Restaurant Selection
Exploitation Go to your favorite restaurant
Exploration Try a new restaurant

Online Banner Advertisements
Exploitation Show the most successful advert
Exploration Show a different advert

Oil Drilling
Exploitation Drill at the best known location
Exploration Drill at a new location

Game Playing
Exploitation Play the move you believe is best
Exploration Play an experimental move

1.2 Examples of Reinforcement Learning

Game Playing

- Reinforcement learning is extensively used in games such as chess, go, checkers etc.
- In 2016 RL based algorithm called Alpha Go by Google defeated Lee Sedol in a five match game of Go.
- In 1996 IBM's Deep Blue an RL agent defeated world chess champion Garry Kasparov. Deep blue can explore 100 million possible chess positions per second.

Data Centers Cooling

- One of the world's most challenging problems is energy consumption.
- Data centers have a large energy consumption to keep the servers cooling
- Google has developed an automatic controlling system which can optimize the energy consumption by adjusting various parameters. Google reported that it has saved 40% of energy consumption using this system.

Trading and Finance

- Reinforcement learning can be used for predicting the stock price.
- An RL agent can be used to make decisions on whether to hold, buy, or sell a stock.
- For example IBM has developed an RL based platform for live trading of stocks.

Robotics

- Robotics is an important sector which uses RL extensively. Almost all robots are designed to grasp the objects.
- For example a vacuum cleaner robot automatically cleans the rooms without human intervention.
- Similarly a robotic arm in a manufacturing industry can automatically fit various parts of a car.

1.3 Elements of Reinforcement Learning

There are 4 basic elements of reinforcement learning.

1. Policy

- Policy() is the core of an RL agent which determines the behaviour of an agent at a given time.
- So, policy defines a mapping from states to actions.
- In some cases policy may be a simple function or a lookup table. But in some cases policy may involve a complex computation such as a search process.
- Policies may be stochastic specifying probabilities for actions.

2. Reward Signal

- A reward signal defines the goal of a reinforcement problem
- The reward signal defines the good and bad events of the agent.
- The objective of an agent is to maximize the total reward received over the long run.

3. Value Function

- The value of a state is the total amount of reward an agent can expect to accumulate in the future starting from that state.
- Reward signal indicate the immediate response whereas value indicate the long-term response.
- Value function indicates what is good in the long-term.
- A state may have a low immediate reward but still have a high value. How it is possible? It is because it may be followed by a high reward states.

4. Model of the Environment

- Model of the environment is one that mimics the behavior of the environment. i.e. how the environment behaves when an agent performs an action)
- Given a state and action, the model might predict the resulting state and reward.
- Models are used for planning. It means that the model helps us to predict the future states before they are actually experienced.
- RL methods can either be model-based or model-free methods. If the RL method uses a model then it is called as model-based method. If the RL method is not using a model then it is called as model-free method.

1.4 Limitations and Scope

- Most of the reinforcement learning methods we study in this course are structured around estimating value functions.
- Other methods such as genetic algorithms, genetic programming, simulated annealing, and other optimization methods have been used to approach reinforcement learning problems.
- These methods evaluate the behavior of an agent using a different policy for interacting with its environment.
- We call these methods as evolutionary methods because their operation is analogous to the biological evolution.
- Our focus is on reinforcement learning methods that involve learning while interacting with the environment, which evolutionary methods do not do.
- It is our belief that the methods that take advantage of individual behavioral interactions can be much more efficient than evolutionary methods in many cases.
- Evolutionary methods do not consider the policy as a function from states to Actions. They do not notice which states an individual passes through during its lifetime, or which actions it selects. Hence, sometimes they become inefficient.
- However, we do include some methods that are like evolutionary methods called policy gradient methods have proven useful in many problems.

1.5 An Extended Example: Tic-Tac-Toe

Consider the game of tic-tac-toe. Two players play this game on a 3X3 board. One player plays Xs and the other Os until one player wins by placing three marks in a row, horizontally, vertically, or diagonally, as the X player has in this game:

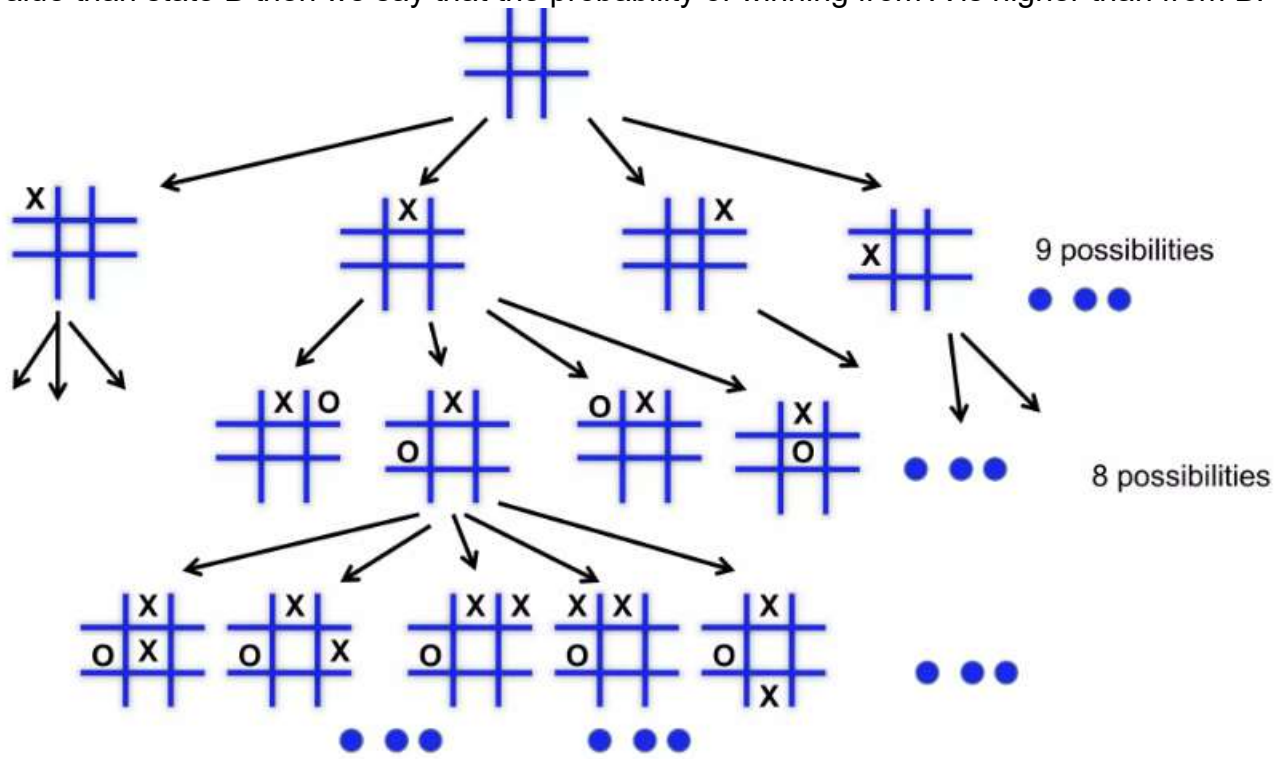
X	O	O
O	X	X
		X

If the board fills up with neither player getting three in a row, the game is a draw.

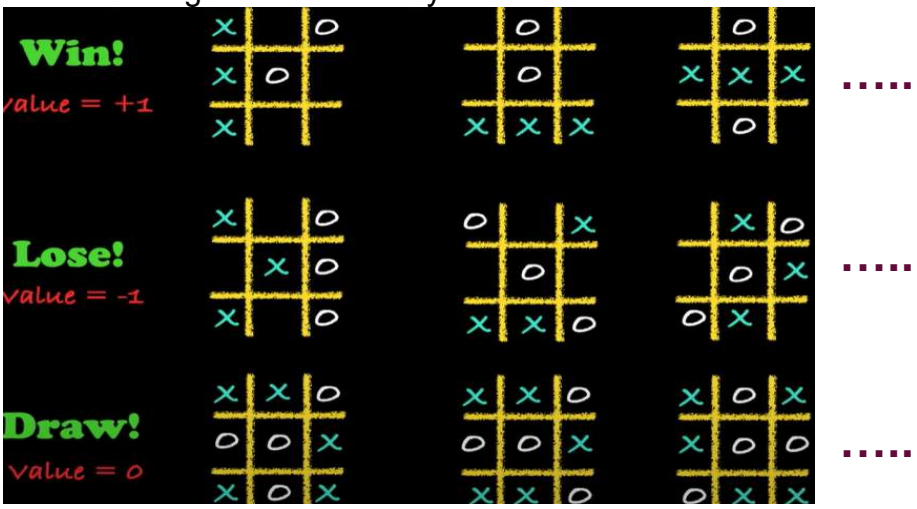
Although this is a simple problem, it cannot readily be solved in a satisfactory way through classical techniques. The reinforcement learning solves this problem by estimating the values of states. The following figure shows the state space tree of tic-tac-toe game. In this each node represents a state.

The reinforcement learning uses a value function to estimate the values of states. First we set up a table of numbers, one for each possible state of the game. Each number will be

the latest estimate of the probability of our winning from that state. We treat this estimate as the state's value, and the whole table is the learned value function. If state A has higher value than state B then we say that the probability of winning from A is higher than from B.



In tic-tac-toe game we can only know the values of terminal states as shown below.



For all other states, the values can be estimated using a value function as given below.

$$V(s) \leftarrow V(s) + \alpha\left[V(s') - V(s)\right]$$

Where S is the current state

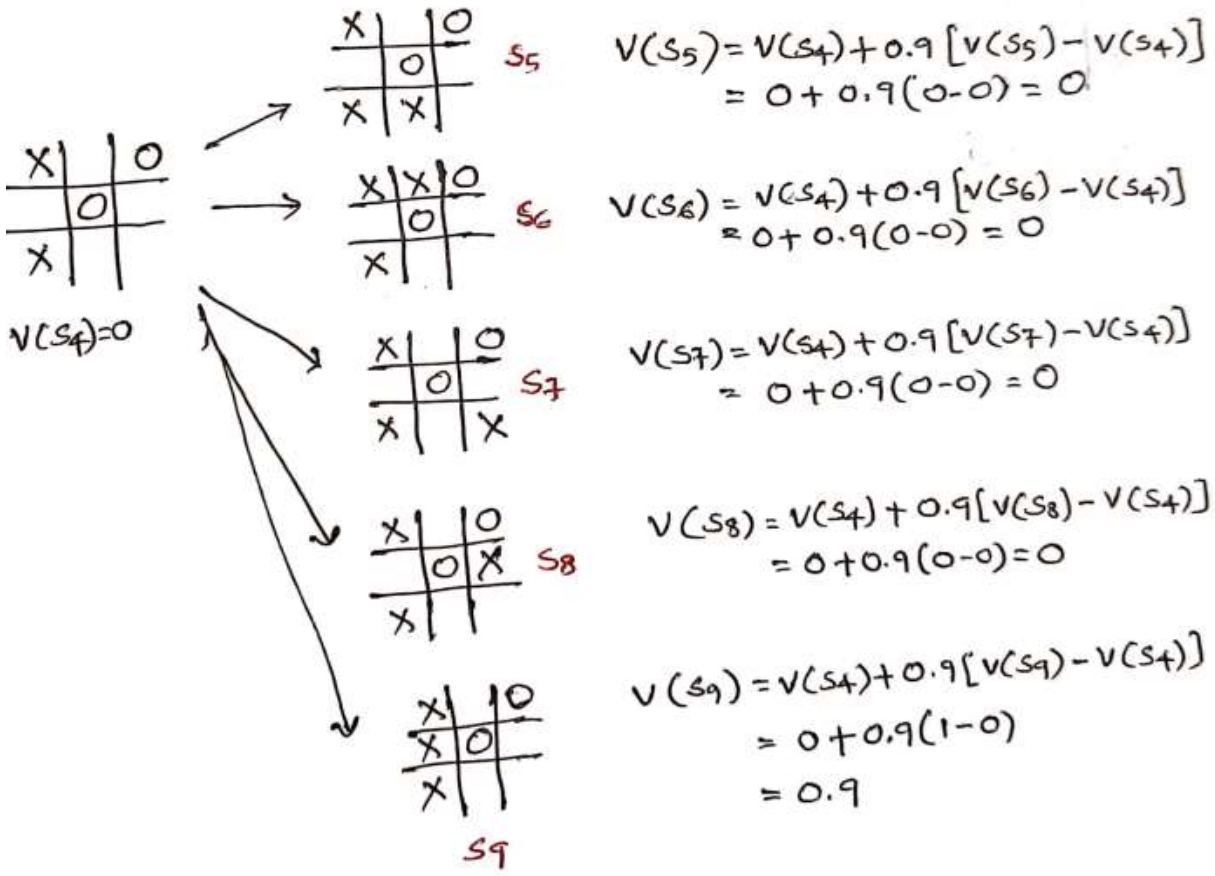
S` is the new state

α is the learning rate

V(s) denote the value of state s

Initially the values of all states are assumed to be 0 except the terminal states. Assume that we are at current state let's say s4 as given below. From S4 many moves are possible as shown below. Out of these possible moves the agent makes a move to the state that

has maximum value. In this example, the agent makes a move to S9 as it has the maximum value.



In this way the agent computes the values of states using the value function and makes moves as per the state values.

1.6 History of Reinforcement Learning

History of RL includes three main threads

1. The first thread tells about the idea of the trial and error learning.
2. Second thread concentrates on the problem of optimal control and its solution.
3. Third thread concentrate on temporal difference methods .

First thread: trial and error

- Edward Thorndike expressed the importance of trial and error methods in RL
- Minsky (1954) developed an Artificial Neuron to mimic the Boolean gates
- Donald Michie developed a learning system to play Tic- Tac-Toe game
- John Holland introduced value functions in playing the games

Second thread: optimal control and its solution

- The term optimal control describes the design of optimised system.
- Richard bellman 1967 proposed by bellman equation which is quite extensively used in RL to find optimal policy.

- Markov proposed MDP (Marco decision process) which is used to make decision optimally to solve RL problem.
- Ron Howard devised a policy iteration method to solve MDP problems.

Third thread: temporal difference method

- It is one of the methods for solving RL problems.
- Arther Samuel 1959 proposed a TD method used to play checkers game
- Cloud shannon 1950 suggested an evolution method based on TD to play chess
- Sutton 1998 described the same learning methods based on TD
- Chris walkim developed Q- learning which is used to solve RL problems
- Geny Tesauro developed backgammon playing program and TD gammon