

Artificial Intelligence

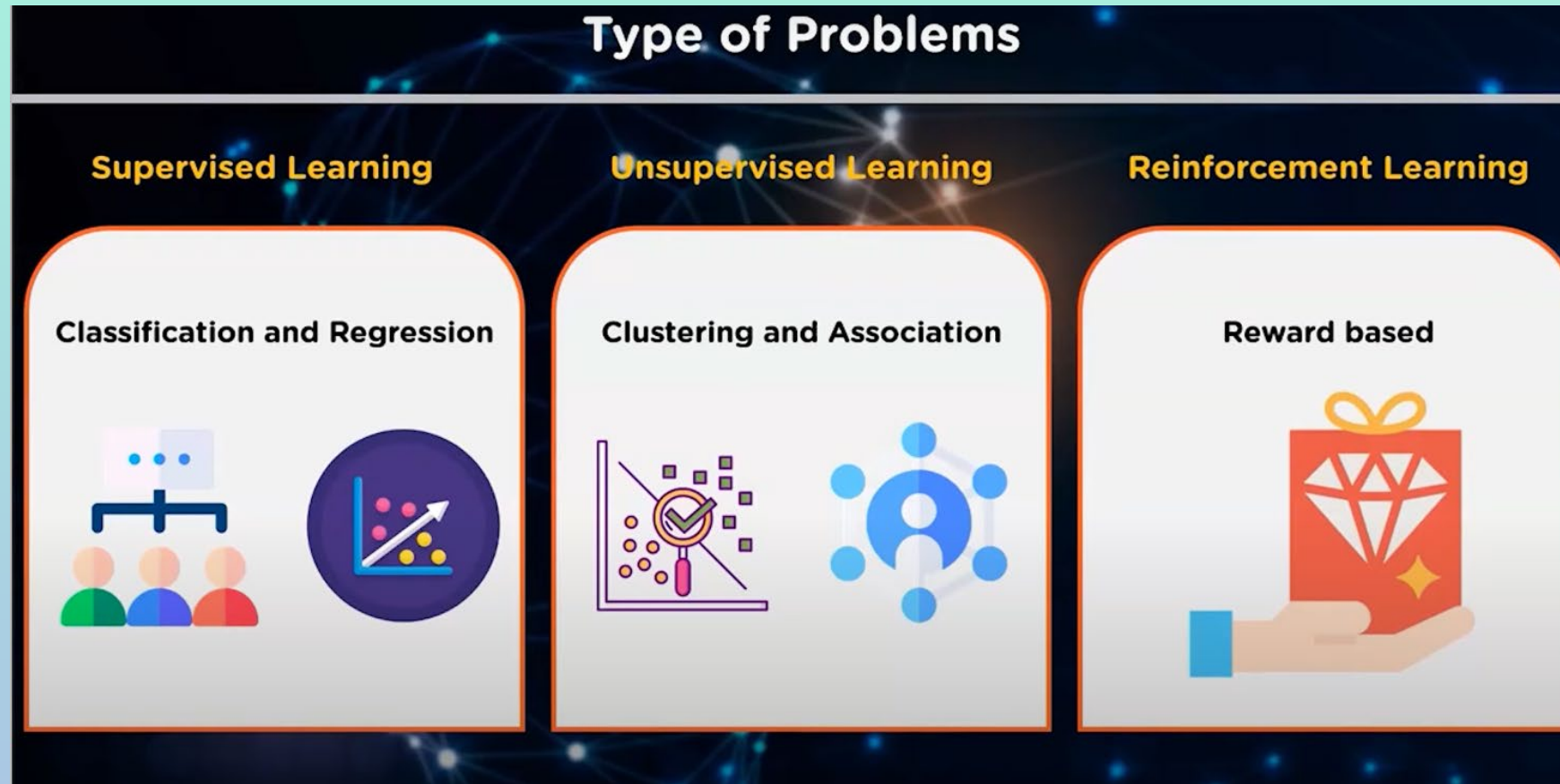
ACCA7026

Reinforcement Learning

Dr. Seena Joseph

Agent based Learning

Types of problems



Reinforcement learning

Branch of Machine learning that trains a model to come to an optimum solution for a problem by taking decisions by itself.

It consist of an environment which an agent will interact with to learn to reach a goal or perform an action.

Reinforcement learning

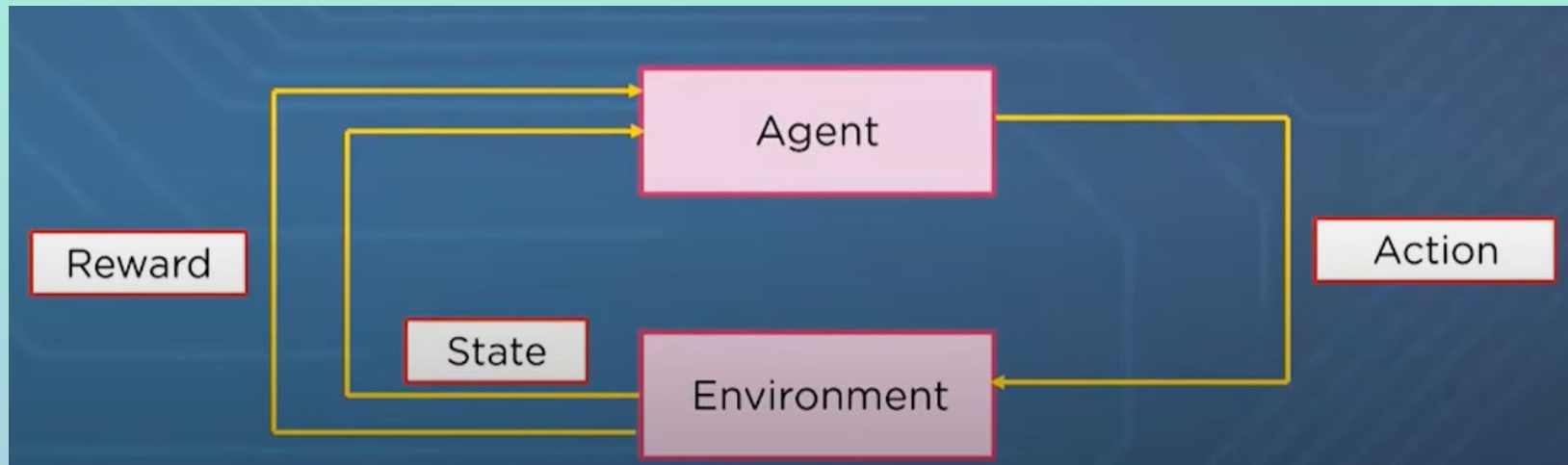
- Terms :
 - Agent – an entity that interact with the environment; responsible for making decision
 - Environment – external system or the process
 - State – the current position of an agent in an environment; the transition to the next state is based on the current action in contrast to the previous actions
 - Action – the step taken by the agent when it is in a particular state
 - Reward – for every action the agent will get a positive or negative reward
 - Policy – mapping between an action and a present state; strategy of choosing an action given a state in expectations of better outcome. Simply an agent's way of behaving at a given time for acting in each state.

Reinforcement learning

- Terms :
 - Episode – complete sequence of interaction between an agent and the environment; starting from the initial state to the termination state; when an agent ends up in a terminating state, and can't take a new action.
 - Value function – specify what is good in long run; value of a state is the total amount of reward an agent can expect to accumulate over the future.
 - Exploration – trying new actions that may not have been taken. The goal is to gather more information about the environment to potentially discover better strategies or actions that may lead to higher long-term rewards.
 - Exploitation - Exploitation involves choosing actions that have been tried before and are known to lead to a high reward. The agent uses its existing knowledge to maximize immediate rewards.

Reinforcement learning

- Elements/components of RL



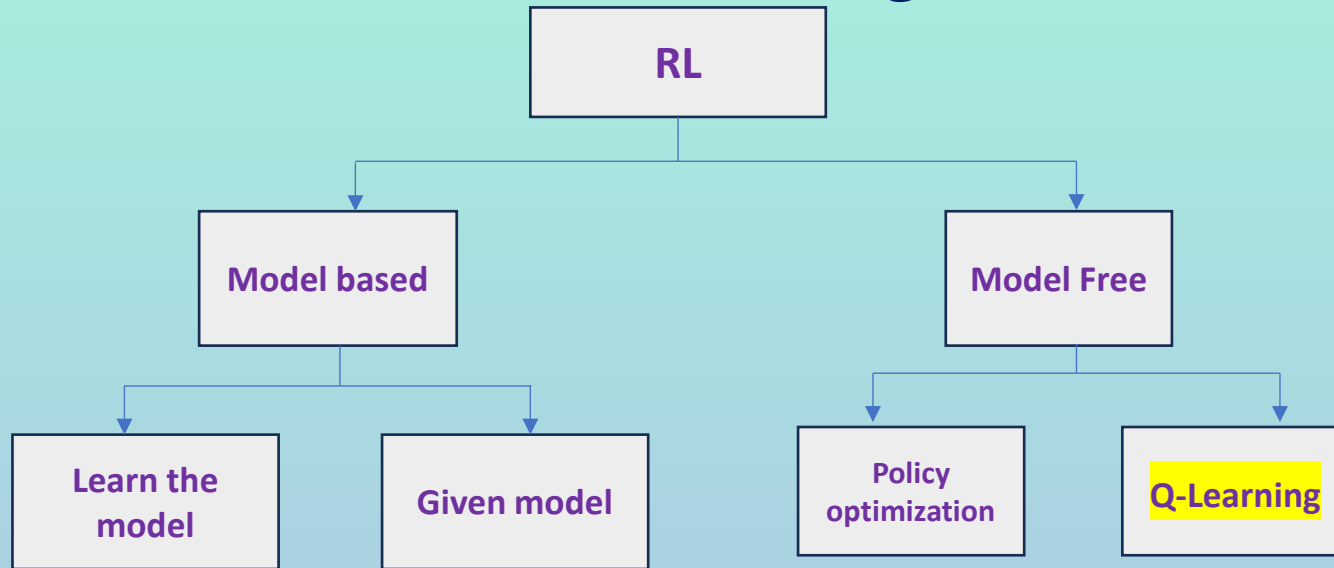
Reinforcement learning

Elements/components of RL

- Agent – The decision-making entity that interacts with the environment to learn optimal behaviors.
- Environment – The external system or surroundings in which the agent operates. The agent perceives states from the environment and takes actions to influence it.
- Actions – The set of all possible decisions or moves the agent can make at any given time.
- Action - the step taken by the agent when it is in a particular state
- Reward – The feedback signal provided by the environment to evaluate the agent's action in a given state.
 - Policy : What to do? It defines how an agent behaves at a particular state.
 - Value : It predicts rewards; a measure of the expected long-term reward the agent can achieve from a given state, following a particular policy.
- Q-values guide the agent's decision-making process by estimating the long-term value of actions, whereas rewards provide immediate feedback to the agent's actions and drive the learning process by indicating the short-term consequences of those actions

Reinforcement learning

Types of reinforcement learning



- Examples of model based RL and Model Free RL ?

Reinforcement learning

Q-Learning

- RL policy that finds the next best action given a current state.
- Chooses the action at random and maximize the reward.
- Model-free environment – agent attempts to construct an optimal policy directly by interacting with the environment.
- Uses, $Q(s,a)$ to estimate the total expected reward for taking an action a in a state s and following the best policy thereafter
- Important terms in Q-Learning
 - State
 - Action
 - Reward
 - Episode
 - Q-value
 - Temporal difference

Reinforcement learning

States

The State, S , represents the current position of an agent in an environment



Action

The Action, A , is the step taken by the agent when it is in a particular state



Rewards

For every action, the agent will get a positive or negative reward



Episodes

When an agent ends up in a terminating state and can't take a new action



Q-Values

Used to determine how good an Action, A , taken at a particular state, S , is $Q(A,S)$



Temporal Difference

A formula used to find the Q-Value by using the value of current state and action and previous state and action

Reinforcement learning

- Q-learning Characteristics
 - Characteristics of RL applies to Q-learning
 - An input and output system, rewards, environment, Markov decision process and training & interfaces.
 - Plus the additional characteristics
 - Finite number of states
 - Finite number of actions
 - Classic Q-learning problems
 - Maze problem
 - Other examples ?

Reinforcement learning

- Q-learning
 - Q-value – indicates the quality of a particular action a in a given state, s :
 $Q(s, a)$
Q-values are the current estimates of total rewards that can be obtained through all remaining steps in the current episode if the agent is in state, s and in action, a .
 - Q-values estimates how much additional reward we can accumulate through all remaining steps in the current episode if the agent is in state, s and takes action, a .
 - It tells the agent how, good is to take a certain action in a specific situation
 - Q-values increase as the agent gets closer to the highest reward.
 - Q-values are stored in a Q-Table [one row for each possible state and one column for each possible action)
 - The Q-Table provides values that enable the AI agent to operate optimally in each given state, giving it the best path to the maximum reward.
 - Thus Q-Table therefore represents the AI agent's policy for acting in the current environment.

Reinforcement learning

- Temporal Difference (TD)

The diagram illustrates the Temporal Difference (TD) equation. At the top, two boxes define the components: 'The reward received for the action taken in the previous state' (pointing to r_t) and 'The largest Q-value available for any action in the current state (the largest predicted sum of future rewards)' (pointing to $\max_a Q(s_{t+1}, a)$). Below these is the equation: $TD(s_t, a_t) = r_t + \gamma \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t)$. At the bottom, three boxes define the terms: 'Temporal difference for the action taken in the previous state' (pointing to $TD(s_t, a_t)$), 'The discount factor (between 0 and 1)' (pointing to γ), and 'The Q-value for the action taken in the previous state' (pointing to $Q(s_t, a_t)$).

$$TD(s_t, a_t) = r_t + \gamma \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t)$$

γ – Discount factor which weights the importance of future rewards relative to immediate rewards.

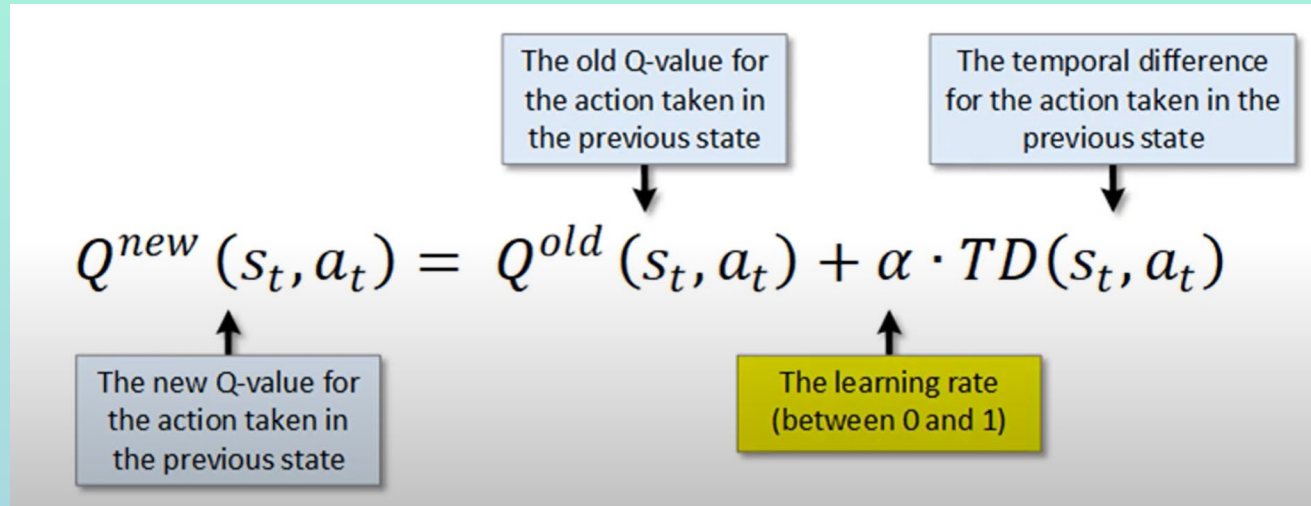
$\max_a Q(s', a')$: Represents the agent's best estimate of the total reward it can get from the state s' onward

Reinforcement learning

- Temporal difference (TD)
 - TD learning compares the difference between the current estimate (based on the observed reward and the estimate of the value of the next state) and previous estimate to update the estimate of the predicted reward for doing a certain action in a given state. The value estimate is then modified using this difference to move it in the direction of the actual expected reward.
 - Method of calculating how much the Q-value for the action taken in the previous state should be changed based on what the AI agent has learned about the Q-values for the current state's actions.
 - Previous q-values are updated after each step
 - Discount factor : mechanism for discounting future rewards; concept: the value of receiving the reward in the future is less than the value of receiving the same reward today.

Reinforcement learning

- Bellman equation



Reinforcement learning

- Bellman equation

Example : Imagine a robot learning to navigate a grid to reach a goal:

Current state: s or s_t

Current state: a , with Value of $Q(s, a) = Q(\text{State A, down}) = 6$; and other parameters $r=2$; $\gamma = 0.9$ and $\alpha = 0.1$

Resulting state: s' or s_{t+1}

$Q(s', a')$ values for s' :

$$Q(s', \text{up}) = 5$$

$$Q(s', \text{right}) = 8$$

$$Q(s', \text{down}) = 3$$

$$Q(s', \text{left}) = 4$$

Here, $\max_{a'} Q(s', a') = 8$ as "right" is the action with the highest expected reward. The robot updates its Q -value for $Q(s, a)$ using this maximum value, encouraging it to prioritize actions leading to better outcomes.

Therefore, $\max_{a'} Q(s', a') = 8$

Reinforcement learning

- **Bellman equation**

Example : Imagine a robot learning to navigate a grid to reach a goal:

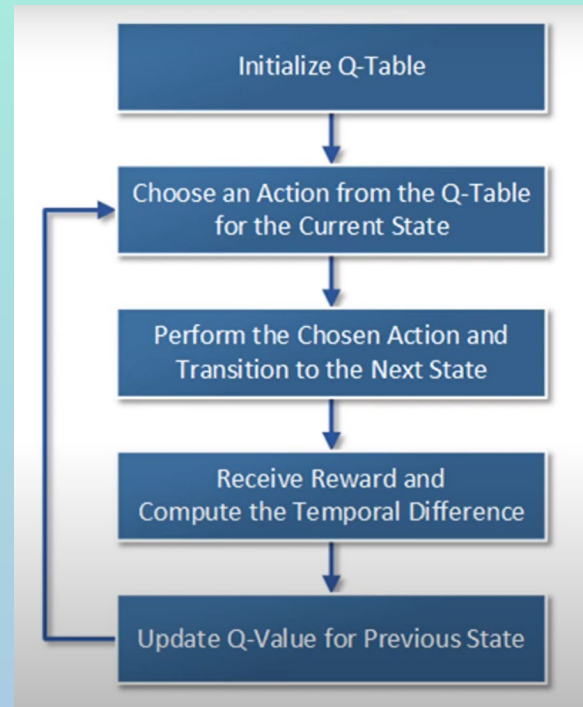
$$\begin{aligned} \text{TD} &= r + \gamma \max_{a'} Q(s', a') - Q(s, a) \\ &= 2 + 0.9 * 8 - 6 = 3.2 \end{aligned}$$

$$\begin{aligned} Q(s, a) &\leftarrow Q(s, a) + \alpha \overbrace{(r + \gamma \max_{a'} Q(s', a') - Q(s, a))}^{\text{TD}} \\ &= 6 + 0.1 * 3.2 = 6.32 \end{aligned}$$

The updated Q-value for $Q(\text{State A, down})$ is 6.32

Reinforcement learning

- Q-learning process



Reinforcement learning

Activity

- Analyze the scenarios of an Autonomous Vehicle (AV) navigating city traffic and a Warehouse Robot performing pick-and-deliver tasks. Compare the roles of the decision-maker, environment, feedback, and strategy in both contexts. How do the similarities and differences in these aspects reflect the challenges of dynamic versus structured environments?

Reinforcement learning

- Inference mode
 - Once the Q-learning model is fully trained, it can be used for inference .
 - In inference mode – Q values are no longer updated
 - Agent will take the action with the largest Q-value

Extra Learning materials

- [Reinforcement Q-Learning from Scratch in Python with OpenAI Gym – LearnDataSci](#)
- [Reinforcement Learning Full Course | Reinforcement Learning In Machine Learning | Simplilearn \(youtube.com\)](#)
- [RL 7: Monte-Carlo Method | Reinforcement Learning \(youtube.com\)](#)
- [Foundations of Q-Learning \(youtube.com\)](#)
- [An introduction to Q-Learning: reinforcement learning](#)

References

- Russell, S. and Norvig, P. (2009) Artificial Intelligence: A Modern Approach (3rd edition), Prentice Hall