Reinforcement Learning Dr. Parul Arora

Contents to be covered

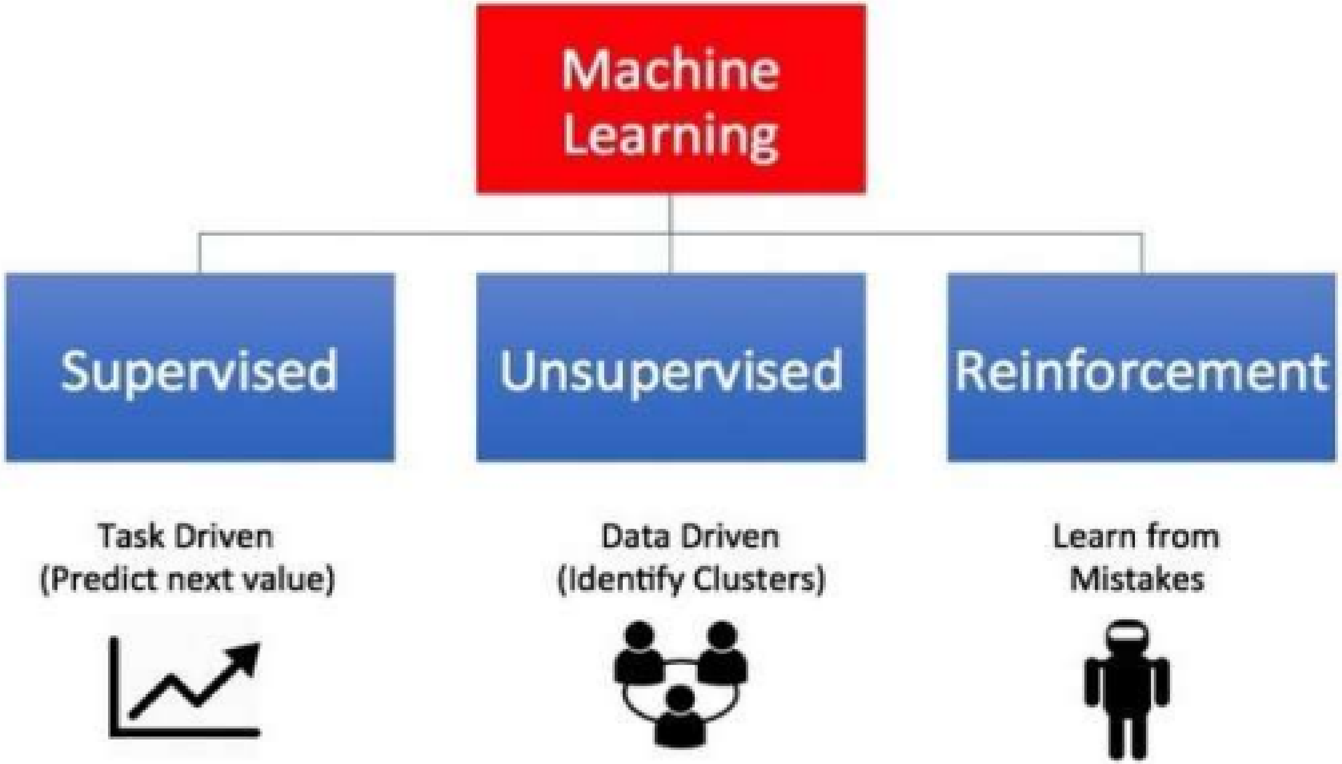
• Types of machine learning •
Reinforcement Learning Definition •
Elements of RL

• Markov Decision Process

- Dynamics of MDP
- Policy
- Value function

Introduction

Types of Machine Learning



Machine Learning

Supervised

Unsupervised

Reinforcement

Task Driven (Predict next value)

Data Driven (Identify Clusters)

Learn from Mistakes

- Reward is always real valued, it could be positive or negative • In supervised leaning, the situations are fixed,

while in reinforcement situations change over time • In

supervised we are given an example, and our onlyconcern

is to do well on that example and the next exampleis

completely independent, while in reinforcement we look for

accumulating reward.

Difference between Supervised and

Reinforcement Learning

• Concept of time in reinforcement learning. • Credit

assignment problem in Reinforcement learning.

• Supervision is little weak in reinforcement learning. • Replace y with reward R.

• Reinforcement learning (RL) is an area of machine learning

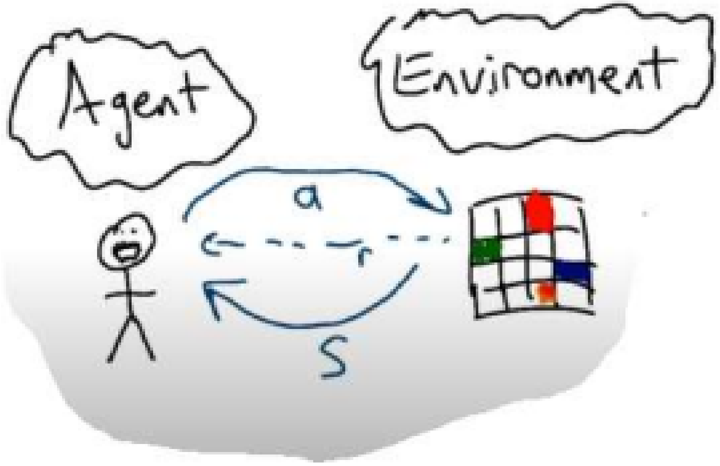
concerned with how software agents ought to takeactions in

an environment in order to maximize the notion of cumulative

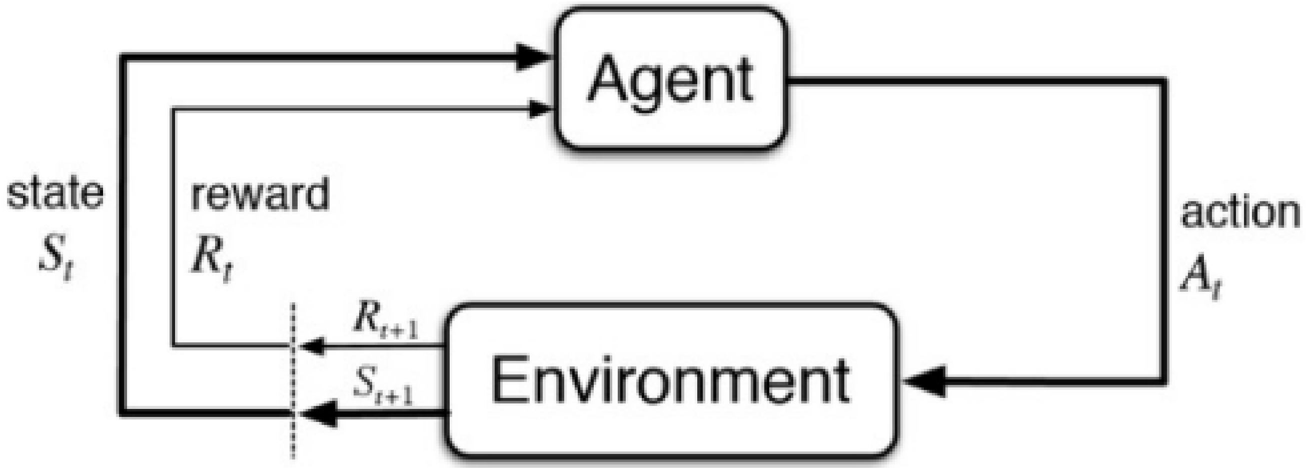
reward (long-term reward over time) – Ex: Robot, play complex games, autonomous driving

- Reinforcement learning uses rewards and punishmentsas signals for positive and negative behavior.
- The goal is to find a suitable action model that would maximize the total cumulative reward of the agent.

Definition


Agent
Environment
a
r
s

Action-Reward feedback loop



It is

kind of sequential decision making.

Agent= Learner

Environment= the agent interacts with

Elements of RL

Policy

Reward

Value

Model of

environment

- Policy: what to do (the agent follows to take action) • Reward: what is good (agent observes upon taking action) • Value: what is good because it predicts reward (total amount of reward, accumulated over future)
- Model: what follows what (something that mimics the environment behaviour)

A Markov decision process is a five state tuple (S,A,{Psa},

Formalization of RL

- Markov Decision Process

γ, R)

– S is a set of states

– A is a set of actions

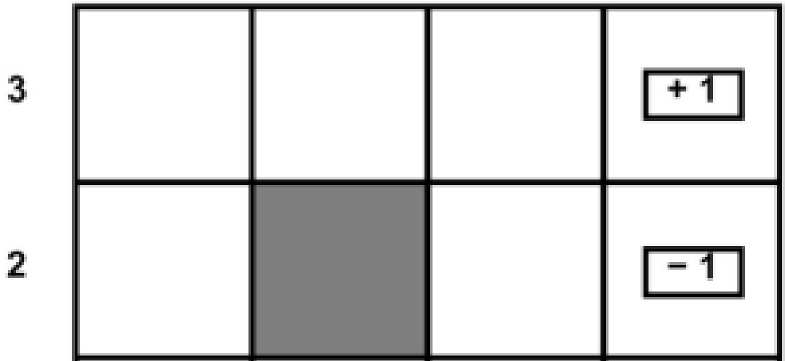
– P_{sa} are the state transition probabilities.

– $\gamma \in [0, 1)$ is called the discount factor.

$0 \leq \gamma < 1$

– R : S × A → R is the reward funcƟon.

Grid World



- The agent lives in a grid
- Walls block the agent’s path
- The agent’s actions do not always go as planned:
  - 80% of the time, the action North takes the agent North (if there is no wall there)
  - 10% of the time, North takes the agent West; 10% East
  - If there is a wall in the direction the agent would have been taken, the agent stays put

Transition Probabilities

11 States,
Rewards

$A=\{N,E,W,S\}$
$P(1,3),N((2,3))=0.8$
$R((3,4))=+1$

$P(1,3),N((1,4))=0.1$

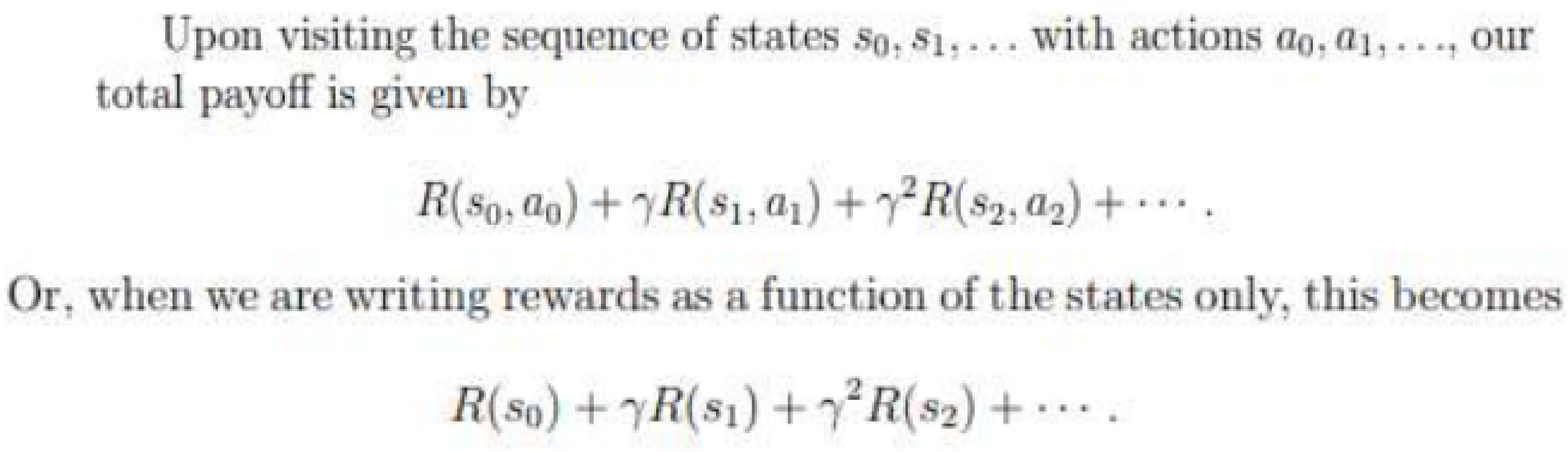
$R((2,4))=-1$
$P(1,3),N((1,2))=0.1$

R(s)=-0.02 (Battery
consumption or fuel consumption)
P(1,3),N((3,3))=0

Dynamics of MDP

$\xrightarrow{a_3} \dots$

Trial or trajectory or episode



Upon visiting the sequence of states $s_0, s_1, \dots$ with actions $a_0, a_1, \dots$, our total payoff is given by

$$R(s_0, a_0) + \gamma R(s_1, a_1) + \gamma^2 R(s_2, a_2) + \cdots .$$

Or, when we are writing rewards as a function of the states only, this becomes

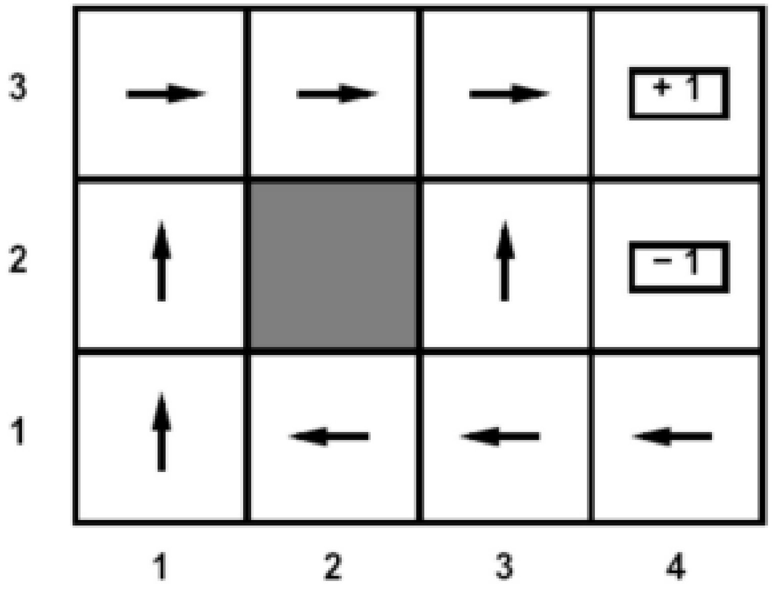
$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots .$$

Our goal in reinforcement learning is to choose actions over time so as to maximize the expected value of the total payoff:

$$\mathrm{E}\left[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots\right]$$

- Policy π : S A

Policy



ence of actions, from start to a goal •

ed utility if followed

π(1,3)=W

Policy = Choice of action for each state
Utility (or return) = sum of discounted rewards

• $V^{\pi}(s)$ Value of a particular state by following a policy π

issimply the expected sum of discounted rewards

upon starting in state s, and taking actions according to π.

Value Function

• The difference between reward and value function? •

The reason why we use expectation here?

• We learn the policy to maximize the function called value function

- We also define the value function for a policy π according to

0 1 2 0 V s E R s R s R s s s ( ) [( ) (( ) ( )) | ,] 🚓 ● 💥 💥 💥 ● ■ ■ 🚓

🚓 2 3

0 1 V s ● E R s 💥 ■ R s 💥 ■ R s 💥 ■ R s 💥 🚓 s ● s

( ) [( ) ( ) ( ) ( ) ....| ,] 3 0 2

2

● R s 💥 ■ E R s_1 💥 ■ R s_2 💥 ■ R s 💥

( ) [( ) ( ) ( ) ....] 3

🚓

( ) 1 V s

- Given a fixed policy π, its value function V^{π} satisfies the **Bellman equations**:

$$V^{\pi}(s) = R(s) + \gamma \sum_{s' \in S} P_{s\pi(s)}(s') V^{\pi}(s')$$

ediate reward

expected sum of
future discounted rewards

space to action space.

(or, state-action pair) to
ed reward.

, state-action pair) is the
ard, starting from that
on pair).

Short Sighted or Immediate

rewardLong Sighted or Future

reward

Thanks