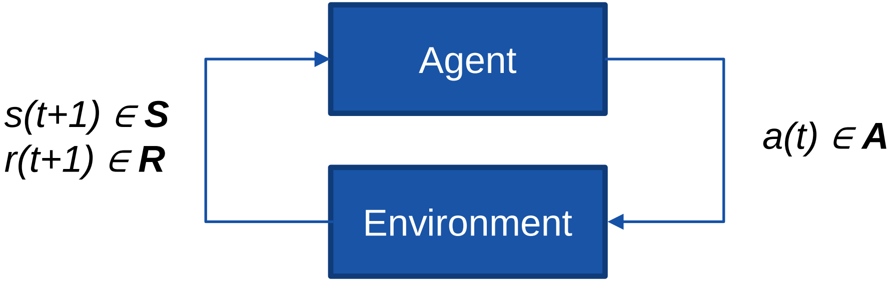
DD2380 - Artificial Intelligence

Reinforcement Learning Tutorial

KTH, Stockholm, Sweden

Terminology

State $s \in \boldsymbol{S}$	Where the agent is in the environment
Action $a \in \boldsymbol{A}$	What the agent can do
Reward $r \in \boldsymbol{S}*\boldsymbol{A}*\boldsymbol{S} \rightarrow \mathbb{R}$	The immediate reward based on an action in a given state
Policy $\boldsymbol{\pi} \in \boldsymbol{S} \rightarrow \boldsymbol{A}$	A mapping from each state to an action
State-action value: $\boldsymbol{Q} \in \boldsymbol{S}*\boldsymbol{A} \rightarrow \mathbb{R}$	The *value* of each action in each state



Agent

Environment

$s(t+1) \in \boldsymbol{S}$
$r(t+1) \in \boldsymbol{R}$

$a(t) \in \boldsymbol{A}$

The shady online casino

A shady online casino is planning to design an RL based agent to manipulate its users into losing more money on their platform. At each game, they can either allow the user to win “**w**” or make them lose “**l**”. Allowing the user to win costs 1, and making them loose brings 1 money unit.

They can evaluate whether a user is Frustrated “**F**”, Neutral “**N**”, or Happy “**H**”. An episode of interaction with a user ends when they decide to leave the platform.

- What are the environment states in this problem?

  S =

- What are the agent actions in this problem?

  A =

- What is the reward function in this problem?

  r(s,a,s’) =

Photo: https://thedailyblog.co.nz/2020/06/12/ways-to-spot-a-shady-online-casino/

Q iteration - breakdown

Interacting with the environment to find out the expected long-time reward of an action, in each state.

$$Q^{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\,(\underbrace{r_t}_{\text{Immediate reward}} + \gamma \underbrace{\max_a Q(s_{t+1}, a)}_{\text{Expected future reward}} - Q(s_t, a_t))$$

Learning rate α

Discount factor γ

Q iteration - parameter choices

$$Q^{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\,(r(a_t) + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t))$$

- For this problem, what would be a good discount factor γ? Can we set $\gamma = 1$?
- What would be a good learning rate? What are the consequences of a very high/low learning rate?
- How would you initialize the Q matrix?

Q iteration - initialization

The shady casino have set their parameters as follows and are now open for business:

$\alpha = 0.8, \gamma=0.9$

$$Q^0 = \begin{array}{c|cc} & l & w \\ \hline H & 10 & 8 \\ N & 10 & 8 \\ F & 10 & 8 \end{array}$$

Note that the casino is quite optimistic with its initial estimate. Do you think it is a good choice? Why? What is the optimal policy with this Q matrix?

Q iteration - execution

Here is the behaviour that the casino has observed from it’s first two users:

U1: $N \xrightarrow{l} F \xrightarrow{l} Q$

U2: $H \xrightarrow{l} N \xrightarrow{l} N \xrightarrow{l} F \xrightarrow{w} N$

Calculate the Q after these observations. (Exercise 1 in the worksheet)

Other variants of reinforcement learning

- Did the shady casino need to know the MDP model to be able to run the Q learning algorithm?
- If they knew the model, could they use another RL approach?
- What are the advantages?

Known model

The casino has learned that the mood transitions of a user can be modelled as the following Markov Decision Process (MDP):

		Next State			
		H	N	F	Q
Current State	H	0.8	0.1	0.0	0.1
	N	0.0	0.5	0.2	0.3
	F	0.0	0.0	0.2	0.8
When user loses: a = l					

		Next State			
		H	N	F	Q
Current State	H	0.9	0.0	0.0	0.1
	N	0.7	0.2	0.0	0.1
	F	0.3	0.4	0.1	0.2
When user wins: a = w					

Value iteration

$$V^{new}(s) \leftarrow max_a\{\sum_{s'} P(s'|s,a)(r(s,a,s') + \gamma V(s'))\}$$

Update the values in the following value matrix through value iteration in the order F, N, H: (keeping γ=0.9) (Exercise 2 in the worksheet)

$$V^0 = \begin{array}{c} H \\ N \\ F \end{array} \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix}$$