Messi & Suarez

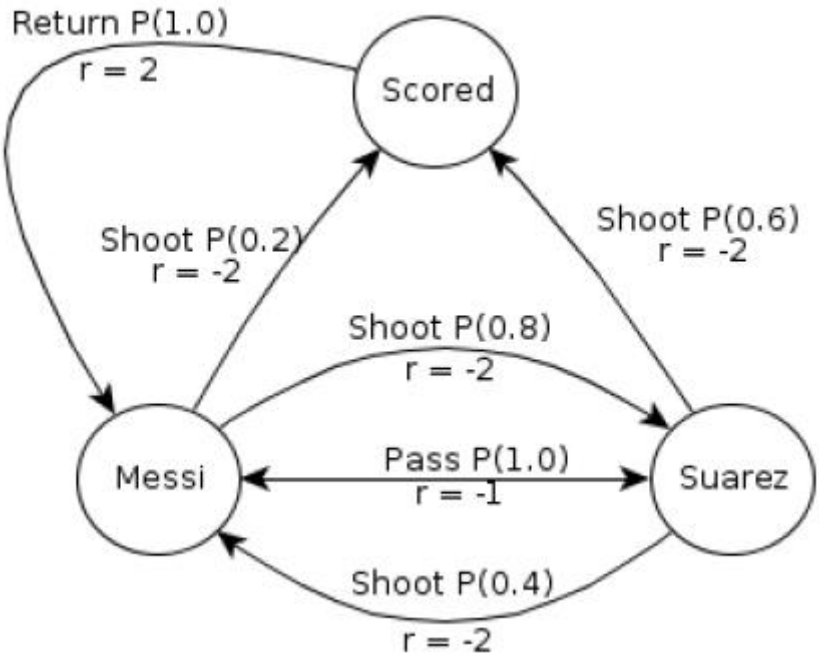
Background

Consider two football-playing robots: Messi and Suarez.
They play a simple two-player cooperate game of football, and you need to write a controller for them. Each player can pass the ball or can shoot at goal.

The football game can be modelled as a discounted-reward MDP with three states: *Messi, Suarez* (denoting who has the ball), and *Scored* (denoting that a goal has been scored); and the following action descriptions:

- If Messi shoots, he has 0.2 chance of scoring a goal and a 0.8 chance of the ball going to Suarez. Shooting towards the goal incurs a cost of 2 (or a reward of -2).
- If Suarez shoots, he has 0.6 chance of scoring a goal and a 0.4 chance of the ball going to Messi. Shooting towards the goal incurs a cost of 2 (or a reward of -2).
- If either player passes, the ball will reach its intended target with a probability of 1.0. Passing the ball incurs a cost 1 (or a reward of -1).
- If a goal is scored, the only action is to return the ball to Messi, which has a probability of 1.0 and has a reward of 2. Thus the reward for scoring is modelled by giving a reward of 2 when leaving the goal state.

The following diagram shows the transition probabilities and rewards:



Tutorial 7: MDP, Value iteration

Problem 1:

Assume that we have calculated the following **non-optimal** value function V for this problem using value iteration with $\gamma = 1.0$, after iteration 2 we arrive at the following:

Iteration	0	1	2	3
V(Messi)	0.0	-1.0	-2.0	
V(Suarez)	0.0	-1.0	-1.2	
V(Scored)	0.0	2.0	1.0	

If Messi has the ball (the system is in the Messi state), what action should we choose to maximise our reward in the next state: pass or shoot? Assume we are using the values for V after three iterations.

We need to calculate the expected return for each action: pass or shoot.

If Messi passes:

$$\begin{aligned} Q(Messi, Pass) &= P_{pass}(Suarez|Messi)[r(Messi, pass, Suarez) + \gamma \cdot V(Suarez)] \\ &= 1 \cdot [-1 + 1 \cdot -1.2] \\ &= 1 \cdot -2.2 \\ &= -2.2 \end{aligned}$$

If Messi shoots:

$$\begin{aligned} Q(Messi, Shoot) &= P_{shoot}(Suarez|Messi)[r(Messi, shoot, Suarez) + \gamma \cdot V(Suarez)] + \\ &\quad P_{shoot}(Scored|Messi)[r(Messi, shoot, Scored) + \gamma \cdot V(Scored)] \\ &= 0.8[-2 + 1 \cdot -1.2] + 0.2[-2 + 1 \cdot 1.0] \\ &= -2.56 + (-0.2) \\ &= -2.76 \end{aligned}$$

Therefore, to maximise our reward, Messi should pass.

Problem 2:

Complete the values of these states for iteration 3 using value iteration. Show your working.

To calculate $V(Messi)$, we choose the action that maximises our Q-value (expected future discounted reward):

$$\begin{aligned} V(Messi) &= \max(Q(Messi, pass), Q(Messi, shoot)) \\ &= \max(-2.2, -2.76) \text{ (from previous question)} \\ &= -2.2 \end{aligned}$$

For $Scored$, there is only one action, which leads directly to the $Messi$ state:

$$\begin{aligned} V(Scored) &= P_{return}(Messi|Scored)[r(Scored, return, Messi) + \gamma \cdot V(Messi)] \\ &= 1[2 + 1 \cdot -2.0] \\ &= 0 \end{aligned}$$

For Suarez, the situation is similar to Messi:

$$\begin{aligned} V(Suarez) &= \max(Q(Suarez, pass), Q(Suarez, shoot)) \\ &= \max(P_{pass}(Messi|Suarez)[r(Suarez, pass, Messi) + \gamma \cdot V(Messi), \\ &\quad \Big(P_{shoot}(Messi|Suarez)[r(Suarez, shoot, Messi) + \gamma \cdot V(Messi) + \\ &\quad P_{shoot}(Scored|Suarez)[r(Suarez, shoot, Scored) + \gamma \cdot V(Scored)]\Big) \\ &= \max(1.0[-1 + 1 \cdot -2.0], (0.4[-2 + 1 \cdot 2.0] + 0.6[-2 + 1 \cdot 1.0])) \\ &= \max(-3, (0.4[-2 + 1 \cdot -2.0] + 0.6[-2 + 1 \cdot 1.0])) \\ &= \max(-3, (-1.6 + -0.6)) \\ &= -2.2 \end{aligned}$$

Thus, the new table is:

Iteration		1	2	3	4
V(Messi)	=	0.0	-1.0	-2.0	-2.2
V(Suarez)	=	0.0	-1.0	-1.2	-2.2
V(Scored)	=	0.0	2.0	1.0	0.0

Tutorial 8: Temporal difference learning

Problem 1:

Explain the difference between Sarsa and Q-learning.

The difference between SARSA and Q-learning is that Q-learning is off-policy learning, while Sarsa is on policy learning. Essentially, this means that SARSA chooses its action using the same policy used to choose the previous action, and then uses this difference to update its Q-function; while Q-learning updated assuming that the next action would be the action with the maximum Q-value.

Q-learning is therefore "optimistic", in that when it updates, it assumes that in the next state, the greedy action will be chosen, even it may be that in the next step, the policy, such as ϵ-greedy, will choose to explore an action other than the best.

SARSA instead knows the action that it will execute next when it performs the update, so will learn on the action whether it is best or not.

Problem 2: Q-learning

Assume the following Q-table, which is learnt after several episodes:

State	Pass	Shoot	Return
Messi	-0.4	-0.8	--
Suarez	-0.7	-0.2	--
Scored	--	--	1.2

In the next step of the episode, from the state 'Suarez', Suarez passes the ball to Messi. Show the Q-learning update for this action using a discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.4$.

Note: Assume that this is a model-free problem, so the transition probabilities are not accessible to your algorithm.

For Q-learning, the update rule is:

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)]$$

So for this particular problem, we update with:

$$\begin{aligned} Q(S,P) &= Q(S,P) + 0.4 \cdot [r(S,P) + 0.9 \cdot \max_{a' \in A(M)} Q(M,a') - Q(S,P)] \\ &= -0.7 + 0.4 \cdot [(-1) + 0.9 \cdot (-0.4) - (-0.7)] \\ &= -0.7 + 0.4 \times (-0.66) \\ &= -0.964 \end{aligned}$$

Problem 3: SARSA

Consider again being in the state 'Suarez', Suarez passes the ball to Messi and then Messi decides to shoot. Show the SARSA update for the Pass action using a discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.4$ and assuming a' (the next action to be execute) is $Shoot$. Compare to the Q-learning update. What is different?

For SARSA, the update is:

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma Q(s', \pi(s')) - Q(s,a)]$$

So for this particular problem, we update with:

$$\begin{aligned} Q(S,P) &= Q(S,P) + 0.4 \cdot [r(S,P) + 0.9 \cdot Q(M, \pi(M)) - Q(S,P)] \\ &= -0.7 + 0.4 \cdot [(-1) + 0.9 \cdot (-0.8) - (-0.7)] \\ &= -0.7 + 0.4 \times (-1.102) \\ &= -1.108 \end{aligned}$$

####

Tutorial 10: Policy iteration and reward shaping

Problem 1: Policy update

Consider the following policy update table and policy evaluation table, with discount factor $\gamma = 0.8$.

Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suaras, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	0.0	0.0	0.0
1					
2	-4.194	-4.772	-4.355	-3.993	-1.355

Apply two iterations of policy iteration. Finish both tables and show the working for the policy evaluation and policy update.

What is the policy after two iterations?

Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1			Return
2			Return

Policy Iteration has two main steps, policy evaluation and policy update. In order to evaluate the given policy:

$$\begin{aligned} V^{\pi}(Messi) &= Q^{\pi}(Messi, Pass) \\ &= P_{Pass}(Suarez \mid Messi)[r(Messi, pass, Suarez) + \gamma \cdot V^{\pi}(Suare \\ &= \gamma \cdot V^{\pi}(Suarez) - 1 \end{aligned}$$

$$\begin{aligned} V^{\pi}(Suarez) &= Q^{\pi}(Suraez, Pass) \\ &= P_{Pass}(Messi \mid Suarez)[r(Suarez, pass, Messi) + \gamma \cdot V^{\pi}(Messi \\ &= \gamma \cdot V^{\pi}(Messi) - 1 \end{aligned}$$

$$\begin{aligned} V^{\pi}(Scored) &= Q^{\pi}(Scored, Return) \\ &= P_{Return}(Messi \mid Scored)[r(Scored, pass, Messi) \\ &= \gamma \cdot V^{\pi}(Messi) + 2 \end{aligned}$$

Then, we solve a basic simultaneous linear equation (not part of the subject learning outcomes) about $V^{\pi}(Messi)$ and $V^{\pi}(Suarez)$:

$$\begin{aligned} V^{\pi}(Messi) &= 1/(\gamma - 1) \\ V^{\pi}(Suarez) &= 1/(\gamma - 1) \\ V^{\pi}(Scored) &= 3 + 1/(\gamma - 1) \end{aligned}$$

Then apply $\gamma = 0.8), the policy evaluation table would be:

Iteration	Q(Messi, P)	Q(Messi, S)	Q(Suaras, P)	Q(Suarez, S)	Q(Scored)
0	0.0	0.0	0.0	0.0	0.0
1	-5	-5.52	-5	-4.56	-2
2	-4.194	-4.772	-4.355	-3.993	-1.355

Then we apply two iterations of policy update based on the above table to get:

Iteration	π(Messi)	π(Suarez)	π(Scored)
0	Pass	Pass	Return
1	Pass	Shoot	Return
2	Pass	Shoot	Return