General Principles of Human and Machine Learning: Pop Quiz #2

Name: **Date:** **Grade:**

Questions:

1. Describe the RL framework - what are the important variables in the RL framework?

   Grade /1

2. Given a transition structure and a reward vector, how should the agent choose its actions?

   Grade /1

3. Why are the outcomes of an agent's actions not always predictable?

   Grade /1

4. Describe a situation for which model-free RL would be more efficient than a model-based agent, for similar performances?

   Grade /1

5. Olaf Scholz and Joe Biden have a debate:
   - Joe: 'If I give you 500 euros today or 550 in one timestep it is exactly the same!"
   - Olaf: 'Stop lying! That's not the same! I prefer to wait and meet in one timestep to get 550 euros!'
   - Joe: 'That is ridiculous'

General Principles of Human and Machine Learning:
Pop Quiz #2

But… They are both right! How can it be the case? What does the above exchange tell you about Joe's preference, and how can you formalize it? (/1) And what comparison can you make between Olaf and Joe? (/1)

Grade /2

6. What does the Bellman equation give us?

Grade /1

7. You are at home. You are hungry, and want the most nutritious food possible. Given that all shops have had some strikes recently, from your home, if you decide to take the path to the bakery you have 20% probability to reach an open bakery and 80% probability of being lost and having to come back home without food. IF you decide to take the path to the supermarket, you have 80% probability to be able to reach an open supermarket, and only 20% probability of having to go back home without food. In the supermarket, 1/3rd of the products has a nutritive score of 0.2, and 2/3rd of the products have a nutritive score of 0.1. On the contrary, at the bakery, 1/3rd of the product has a nutritive score of 5, and 2/3rd of 2. You expect to get those rewards instantaneously once either reaching the bakery or the supermarket (no discounting).
    a. Represent the problem as a diagram containing all information

    Grade /1

    b. What should be your optimal decision?

    Grade /1

General Principles of Human and Machine Learning:
Pop Quiz #2

8. You are in a casino and need to choose between machine A or machine B. You are a model-free agent and can only decide based on your prior experience with both of those machines. Here is the summary of your interactions with both machines:
   You selected machine A 20 times and got the following rewards:
   1,0,0,1,1,1,1,0,1,0,0,1,1,0,0,1,1,1,1,1
   You selected machine B 20 times and got the following rewards:
   0,0,1,0,1,0,0,1,0,0,0,0,1,1,0,0,0,0,1,0
   a) What would be the policy of a greedy agent?

      Grade /0.5

   b) What would be the policy of an epsilon-greedy agent?

      Grade /0.5

Answer key:

1. In RL the agent is in certain states and learns to take actions in order to maximise the sum of future discounted rewards. The important variables are: states, actions, rewards, policy (probability distribution over the action space), value function (sum of future discounted rewards) - maybe transitions but it is not a problem if that is not mentioned.
   Mark: /1

2. The goal of an RL agent is to maximize the expected discounted sum of rewards it will get in the future, so it should take actions that maximise this quantity (i.e. the value function).
   Mark: /1

3. The environment is stochastic, the transition structure of the environment also impacts the outcome of a decision. For example, you might decide to take a certain route, but this route might be closed because of work.
   Mark: /1

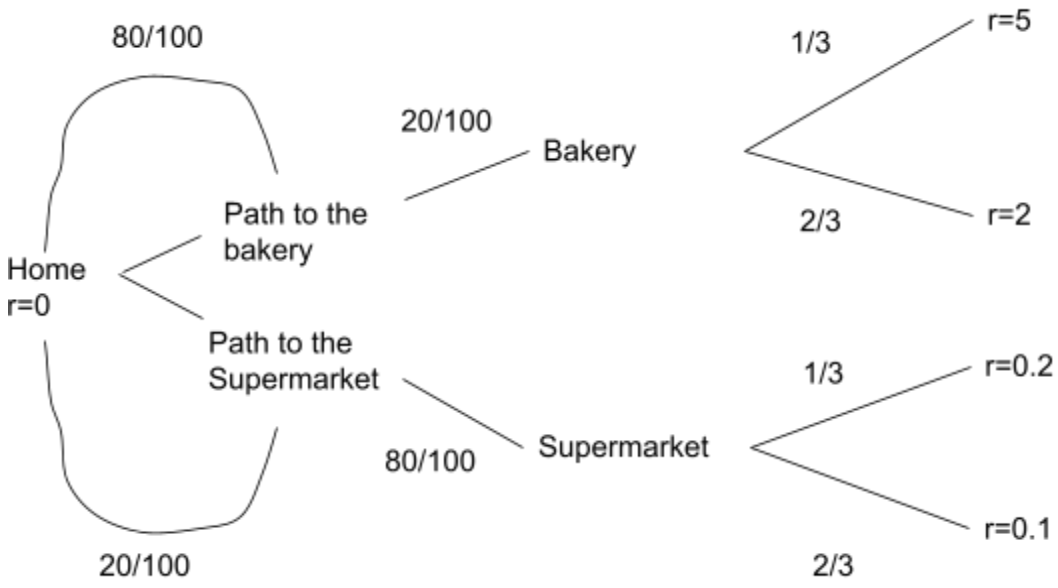
General Principles of Human and Machine Learning:
Pop Quiz #2

4. Any situation in which the reward and transitions are steady.
   Mark: /1

5. Joe considers that a reward of 500 now is equal to a reward of 550 in one timestep. That tells us that Joe has a discount factor of gamma=500/550~ 0.91. On the contrary, Olaf prefers to wait, that means that gamma_olaf>gamma_joe - i.e. Olaf is more patient than Joe.
   Mark: /2

6. It provides a relationship between the value estimate at time t and its next estimate at time t+1. It is a recursive equation and allows to save computational load by using only one cached value and updating it at every step.
   Mark: /1

7.
   a. Diagram:



   b. We want to maximize the value function.

General Principles of Human and Machine Learning:
Pop Quiz #2

i. Expected reward if we choose Supermarket = 80/100*(⅓*0.2+⅔*0.1)=0.10666666666

ii. Expected reward if we choose Bakery = 20/100*(⅓*5+⅔*2)=0.6
You should choose to go to the bakery!

Mark: /2

8. a) a greedy agent would choose max_Q(a) -> machine A with 100% probability

Mark: /0.5

b)an epsilon-greedy agent would choose max_Q(a) so machine A with 1-epsilon probability and machine B with epsilon probability

Mark: /0.5

Total /10