DAT405 Introduction to Data Science and AI

2022-2023, Reading Period 3

Assignment 5: Reinforcement learning and classification

Oscar Karbin (12 timmar) Jihad Almahal (12 timmar)

The exercise takes place in a notebook environment where you can chose to use Jupyter or Google Colabs. We recommend you use Google Colabs as it will facilitate remote group-work and makes the assignment less technical. Hints: You can execute certain linux shell commands by prefixing the command with ！. You can insert Markdown cells and code cells. The first you can use for documenting and explaining your results the second you can use writing code snippets that execute the tasks required.

This assignment is about **sequential decision making** under uncertainty (Reinforcement learning). In a sequential decision process, the process jumps between different states (the environment), and in each state the decision maker, or agent, chooses among a set of actions. Given the state and the chosen action, the process jumps to a new state. At each jump the decision maker receives a reward, and the objective is to find a sequence of decisions (or an optimal policy) that maximizes the accumulated rewards.

We will use **Markov decision processes** (MDPs) to model the environment, and below is a primer on the relevant background theory.

- To make things concrete, we will first focus on decision making under **no** uncertainity (question 1 and 2), i.e, given we have a world model, we can calculate the exact and optimal actions to take in it. We will first introduce **Markov Decision Process (MDP)** as the world model. Then we give one algorithm (out of many) to solve it.
- (Optional) Next we will work through one type of reinforcement learning algorithm called Q-learning (question 3). Q-learning is an algorithm for making decisions under uncertainity, where uncertainity is over the possible world model (here MDP). It will find the optimal policy for the **unknown** MDP, assuming we do infinite exploration.
- Finally, in question 4 you will be asked to explain differences between reinforcement learning and supervised learning and in question 5 write about decision trees and random forests.

Primer

Decision Making

The problem of **decision making under uncertainty** (commonly known as **reinforcement learning**) can be broken down into two parts. First, how do we learn about the world? This involves both the problem of modeling our initial uncertainty about the world, and that of drawing conclusions from evidence and our initial belief. Secondly, given what we currently know about the world, how should we decide what to do, taking into account future events and observations that may change our conclusions? Typically, this will involve creating long-term plans covering possible future eventualities. That is, when planning under uncertainty, we also need to take into account what possible future knowledge could be generated when implementing our plans. Intuitively, executing plans which involve trying out new things should give more information, but it is hard to tell whether this information will be beneficial. The choice between doing something which is already known to produce good results and experiment with something new is known as the **exploration-exploitation dilemma**.

The exploration-exploitation trade-off

Consider the problem of selecting a restaurant to go to during a vacation. Lets say the best restaurant you have found so far was **Les Epinards**. The food there is usually to your taste and satisfactory. However, a well-known recommendations website suggests that **King's Arm** is really good! It is tempting to try it out. But there is a risk involved. It may turn out to be much worse than **Les Epinards**, in which case you will regret going there. On the other hand, it could also be much better. What should you do? It all depends on how much information you have about either restaurant, and how many more days you'll stay in town. If this is your last day, then it's probably a better idea to go to **Les Epinards**, unless you are expecting **King's Arm** to be significantly better. However, if you are going to stay there longer, trying out **King's Arm** is a good bet. If you are lucky, you will be getting much better food for the remaining time, while otherwise you will have missed only one good meal out of many, making the potential risk quite small.

Markov Decision Processes

Markov Decision Processes (MDPs) provide a mathematical framework for modeling sequential decision making under uncertainty. An *agent* moves between *states* in a *state space* choosing *actions* that affects the transition probabilities between states, and the subsequent *rewards* recieved after a jump. This is then repeated a finite or infinite number of epochs. The objective, or the *solution* of the MDP, is to optimize the accumulated rewards of the process.

Thus, an MDP consists of five parts:

- Decision epochs: $t = 1, 2, \ldots, T$, where $T \leq \infty$
- State space: $S = \{s_1, s_2, \ldots, s_N\}$ of the underlying environment
- Action space $A = \{a_1, a_2, \ldots, a_K\}$ available to the decision maker at each decision epoch

- Transition probabilities $p(s_{t+1}|s_t, a_t)$ for jumping from state s_t to state s_{t+1} after taking action a_t
- Reward functions $R_t = r(a_t, s_t, s_{t+1})$ resulting from the chosen action and subsequent transition

A *decision policy* is a function $\pi : s \rightarrow a$, that gives instructions on what action to choose in each state. A policy can either be *deterministic*, meaning that the action is given for each state, or *randomized* meaning that there is a probability distribution over the set of possible actions for each state. Given a specific policy π we can then compute the the *expected total reward* when starting in a given state $s_1 \in S$, which is also known as the *value* for that state,

$$V^{\pi}(s_1) = E\left[\sum_{t=1}^{T} r(s_t, a_t, s_{t+1}) \Big| s_1\right] = \sum_{t=1}^{T} r(s_t, a_t, s_{t+1}) p(s_{t+1}|a_t, s_t)$$

where $a_t = \pi(s_t)$. To ensure convergence and to control how much credit to give to future rewards, it is common to introduce a *discount factor* $\gamma \in [0, 1]$. For instance, if we think all future rewards should count equally, we would use $\gamma = 1$, while if we value near-future rewards higher than more distant rewards, we would use $\gamma < 1$. The expected total *discounted* reward then becomes

$$V^{\pi}(s_1) = \sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t, s_{t+1}) p(s_{t+1}|s_t, a_t)$$

Now, to find the *optimal* policy we want to find the policy π^* that gives the highest total reward $V^*(s)$ for all $s \in S$. That is, we want to find the policy where

$$V^*(s) \geq V^{\pi}(s), s \in S$$

To solve this we use a dynamic programming equation called the *Bellman equation*, given by

$$V(s) = \max_{a \in A} \left\{ \sum_{s' \in S} p(s'|s, a)(r(s, a, s') + \gamma V(s')) \right\}$$

It can be shown that if π is a policy such that V^{π} fulfills the Bellman equation, then π is an optimal policy.

A real world example would be an inventory control system. The states could be the amount of items we have in stock, and the actions would be the amount of items to order at the end of each month. The discrete time would be each month and the reward would be the profit.

Question 1

The first question covers a deterministic MPD, where the action is directly given by the state, described as follows:

- The agent starts in state **S** (see table below)

- The actions possible are **N** (north), **S** (south), **E** (east), and **W** west.
- The transition probabilities in each box are deterministic (for example P(s'|s,N)=1 if s' north of s). Note, however, that you cannot move outside the grid, thus all actions are not available in every box.
- When reaching **F**, the game ends (absorbing state).
- The numbers in the boxes represent the rewards you receive when moving into that box.
- Assume no discount in this model: $\gamma = 1$

-1	1	**F**
0	-1	1
-1	0	-1
S	-1	1

Let (x, y) denote the position in the grid, such that $S = (0, 0)$ and $F = (2, 3)$.

1a) What is the optimal path of the MDP above? Is it unique? Submit the path as a single string of directions. E.g. NESW will make a circle.

The optimal path of the MDP above is not unique because there are two paths that have the same reward, the first is EENNWNE and the second is EENNN which takes two less steps but gives the same reward which is 0.

1b) What is the optimal policy (i.e. the optimal action in each state)? It is helpful if you draw the arrows/letters in the grid.

(0,0) -> N/E = -1 (1,0) -> E = 1 (2,0) -> W/N = -1 (0,1) -> N/E = 0 (1,1) -> N/W/S/E = -1 (2,1) -> N/S =1 (0,2) -> N/S/E = -1 (1,2) -> N/E = 1 (2,2) -> N = 0 (0,3) -> E = 1 (1,3) -> E = 0 (2,3) -> absorbing state

1c) What is expected total reward for the policy in 1a)?

With the optimal policy rewards identified in question 1b we can calculate the expected total reward by using the optimal paths in question 1a:

- EENNN (-1 + 1 -1 +1 + (absorbing state)) = 0
- EENNWNE (-1 + 1 - 1 + 1 -1 +1 +(absorbing state)) = 0

Value Iteration

For larger problems we need to utilize algorithms to determine the optimal policy π^*. *Value iteration* is one such algorithm that iteratively computes the value for each state. Recall that for a policy to be optimal, it must satisfy the Bellman equation above, meaning that plugging in a given candidate V^* in the right-hand side (RHS) of the Bellman equation should result in the same V^* on the left-hand side (LHS). This property will form the basis of our algorithm. Essentially, it can be shown that repeated application of the RHS to any intial value function

$V^0(s)$ will eventually lead to the value V which statifies the Bellman equation. Hence repeated application of the Bellman equation will also lead to the optimal value function. We can then extract the optimal policy by simply noting what actions that satisfy the equation.

The process of repeated application of the Bellman equation is what we here call the *value iteration* algorithm. It practically procedes as follows:

```
epsilon is a small value, threshold
for x from i to infinity
do
    for each state s
    do
        V_k[s] = max_a Σ_s' p(s'|s,a)*(r(a,s,s') + γ*V_k-1[s'])
    end
    if  |V_k[s]-V_k-1[s]| < epsilon for all s
        for each state s,
        do
            π(s)=argmax_a ∑_s' p(s'|s,a)*(r(a,s,s') + γ*V_k-1[s'])
            return π, V_k
        end
end
```

Example: We will illustrate the value iteration algorithm by going through two iterations. Below is a 3x3 grid with the rewards given in each state. Assume now that given a certain state s and action a, there is a probability 0.8 that that action will be performed and a probability 0.2 that no action is taken. For instance, if we take action **E** in state (x, y) we will go to $(x + 1, y)$ 80 percent of the time (given that that action is available in that state), and remain still 20 percent of the time. We will use have a discount factor $\gamma = 0.9$. Let the initial value be $V^0(s) = 0$ for all states $s \in S$.

Reward:

0	0	0
0	10	0
0	0	0

Iteration 1: The first iteration is trivial, $V^1(s)$ becomes the $\max_a \sum_{s'} p(s'|s, a)r(s, a, s')$ since V^0 was zero for all s'. The updated values for each state become

0	8	0
8	2	8
0	8	0

Iteration 2:

Staring with cell (0,0) (lower left corner): We find the expected value of each move:

Action **S**: 0

Action **E**: 0.8(0 + 0.9 * 8) + 0.2(0 + 0.9 * 0) = 5.76

Action **N**: 0.8(0 + 0.9 * 8) + 0.2(0 + 0.9 * 0) = 5.76

Action **W**: 0

Hence any action between **E** and **N** would be best at this stage.

Similarly for cell (1,0):

Action **N**: 0.8(10 + 0.9 * 2) + 0.2(0 + 0.9 * 8) = 10.88 (Action **N** is the maximizing action)

Similar calculations for remaining cells give us:

5.76	10.88	5.76
10.88	8.12	10.88
5.76	10.88	5.76

Question 2

2a) Code the value iteration algorithm just described here, and show the converging optimal value function and the optimal policy for the above 3x3 grid.

```
grid  = np.array([[0,0,0]])
```

```
import numpy as np

# Parameters
gamma = 0.9  # discount factor
epsilon = 0.0001  # convergence threshold
actions = ['N', 'E', 'S', 'W']  # list of possible actions

# Environment
rewards = np.array([[0, 0, 0], [0, 10, 0], [0, 0, 0]])  # rewards for each state
state_values = np.zeros_like(rewards, dtype=float)  # initial state values (all zeros)

# Helper function to compute the value of a state-action pair
def q_value(s, a, V):
    """Compute the Q-value of a state-action pair."""
    # get the indices of the next state
    i, j = s
    if a == 'N':
        i -= 1
    elif a == 'E':
        j += 1
    elif a == 'S':
        i += 1
    elif a == 'W':
        j -= 1
    # check if next state is inside the grid
```

```
        if i < 0 or i >= state_values.shape[0] or j < 0 or j >= state_values.shape[1]:
            return -np.inf
        # compute the Q-value using the Bellman equation
        return np.sum([
            0.8 * (rewards[i, j] + gamma * V[i, j]),  # taking the action
            0.2 * (rewards[s] + gamma * V[s])  # staying in the same state
        ])

    # Value iteration loop
    delta = np.inf  # initialize the change in state values
    while delta >= epsilon:
        delta = 0
        for i in range(state_values.shape[0]):
            for j in range(state_values.shape[1]):
                v = state_values[i, j]  # current state value
                state_values[i, j] = max(q_value((i, j), a, state_values) for a in actions)  #
                delta = max(delta, abs(v - state_values[i, j]))  # update delta

    # Compute the optimal policy
    policy = np.zeros_like(rewards, dtype=str)  # initialize the policy
    for i in range(state_values.shape[0]):
        for j in range(state_values.shape[1]):
            q_values = [q_value((i, j), a, state_values) for a in actions]  # compute Q-values
            policy[i, j] = actions[np.argmax(q_values)]  # choose the action with the highest

    # Print the results
    print("Optimal state values:")
    print(state_values)
    print("Optimal policy:")
    print(policy)
```

```
Optimal state values:
[[45.612448   51.94761604 45.61252419]
 [51.94761604 48.05154858 51.94768586]
 [45.61252419 51.94768586 45.61258818]]
Optimal policy:
[['E' 'S' 'S']
 ['E' 'E' 'W']
 ['E' 'N' 'N']]
```

2b) Explain why the result of 2a) does not depend on the initial value V_0.

The result of 2a) does not depend on the initial value V0 because the algorithm that was used to calculate the optimal value function and policy, policy iteration algorithm, is a guaranteed convergence algorithm. This means that no matter what initial value V0 is used, the algorithm will always converge to the same unique optimal value function and policy for the given MDP.

2c) Describe your interpretation of the discount factor γ. What would happen in the two extreme cases $\gamma = 0$ and $\gamma = 1$? Given some MDP, what would be important things to consider when deciding on which value of γ to use?

The discount factor γ is a value between 0 and 1 that determines how much weight to give to future rewards in the decision-making process.

When γ=0, the agent only cares about immediate rewards and does not consider future rewards at all. This might be appropriate for some problems where the future has no value, but in most real-world problems, this is not the case.

When γ=1, the agent considers future rewards to be just as important as immediate rewards. This can be appropriate in some cases, but it can also lead to problems. Therefore there exists no silver-bullet value of γ.

When deciding on the value of γ to use in a particular MDP, it is important to consider the time horizon of the problem and the discounting of future rewards. A high value of γ may be appropriate when the rewards are delayed and the agent is expected to consider the long-term consequences of its actions. On the other hand, a low value of γ may be appropriate when the rewards are immediate and the agent only needs to consider the immediate consequences of its actions.

Reinforcement Learning (RL) (Theory for optional question 3)

Until now, we understood that knowing the MDP, specifically $p(s'|a,s)$ and $r(s,a,s')$ allows us to efficiently find the optimal policy using the value iteration algorithm. Reinforcement learning (RL) or decision making under uncertainity, however, arises from the question of making optimal decisions without knowing the true world model (the MDP in this case).

So far we have defined the value function for a policy through V^π. Let's now define the *action-value function*

$$Q^\pi(s,a) = \sum_{s'} p(s'|a,s)[r(s,a,s') + \gamma V^\pi(s')]$$

The value function and the action-value function are directly related through

$$V^\pi(s) = \max_a Q^\pi(s,a)$$

i.e, the value of taking action a in state s and then following the policy π onwards. Similarly to the value function, the optimal Q-value equation is:

$$Q^*(s,a) = \sum_{s'} p(s'|a,s)[r(s,a,s') + \gamma V^*(s')]$$

and the relationship between $Q^*(s,a)$ and $V^*(s)$ is simply

$$V^*(s) = \max_{a \in A} Q^*(s,a).$$

Q-learning

Q-learning is a RL-method where the agent learns about its unknown environment (i.e. the MDP is unknown) through exploration. In each time step *t* the agent chooses an action *a* based on the

current state *s*, observes the reward *r* and the next state *s'*, and repeats the process in the new state. Q-learning is then a method that allows the agent to act optimally. Here we will focus on the simplest form of Q-learning algorithms, which can be applied when all states are known to the agent, and the state and action spaces are reasonably small. This simple algorithm uses a table of Q-values for each (s, a) pair, which is then updated in each time step using the update rule in step $k+1$

$$Q_{k+1}(s,a) = Q_k(s,a) + \alpha\left(r(s,a) + \gamma \max\{Q_k(s',a')\} - Q_k(s,a)\right)$$

where γ is the discount factor as before, and α is a pre-set learning rate. It can be shown that this algorithm converges to the optimal policy of the underlying MDP for certain values of α as long as there is sufficient exploration. For our case, we set a constant $\alpha = 0.1$.

OpenAI Gym

We shall use already available simulators for different environments (worlds) using the popular [OpenAI Gym library](). It just implements different types of simulators including ATARI games. Although here we will only focus on simple ones, such as the **Chain enviroment** illustrated below.

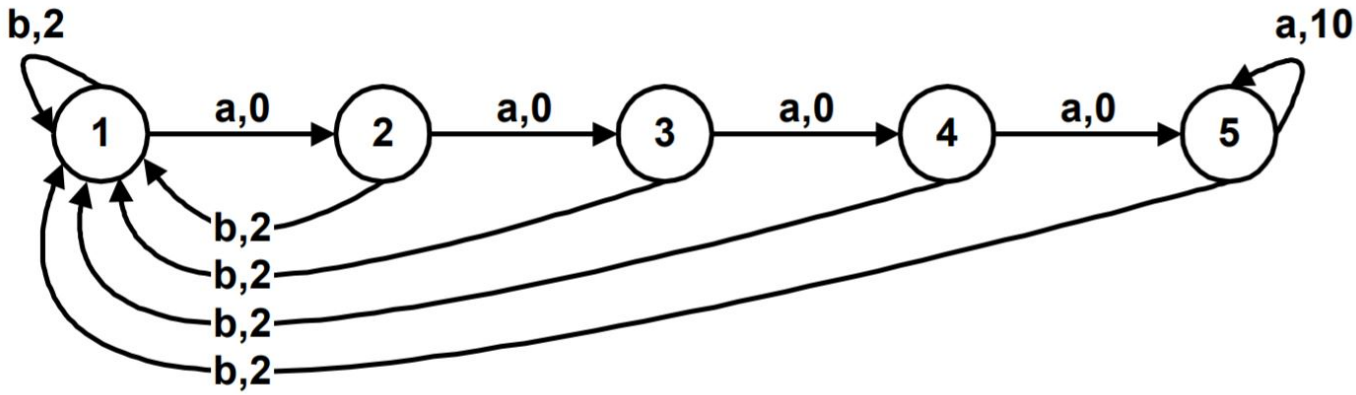

Figure 1. The "Chain" problem

The figure corresponds to an MDP with 5 states $S = \{1, 2, 3, 4, 5\}$ and two possible actions $A = \{a, b\}$ in each state. The arrows indicate the resulting transitions for each state-action pair, and the numbers correspond to the rewards for each transition.

Question 3 (optional)

You are to first familiarize with the framework of [the OpenAI environments](), and then implement the Q-learning algorithm for the `NChain-v0` enviroment depicted above, using default parameters and a learning rate of $\gamma = 0.95$. Report the final Q^* table after convergence of the algorithm. For an example on how to do this, you can refer to the Q-learning of the **Frozen lake environment** (`q_learning_frozen_lake.ipynb`), uploaded on Canvas. Hint: start with a small learning rate.

Note that the NChain environment is not available among the standard environments, you need to load the `gym_toytext` package, in addition to the standard gym:

```
!pip install gym-legacy-toytext
import gym
import gym_toytext
env = gym.make("NChain-v0")
```

Question 4

4a) What is the importance of exploration in reinforcement learning? Explain with an example.

Exploration is a critical aspect of reinforcement learning that enables agents to discover and learn optimal policies in uncertain environments. The primary goal of exploration is to gather information about the environment and update the agent's policy based on this information.

Imagine you're in a new city and you want to find the best coffee shop. You have a list of recommended coffee shops, but you don't know which one is the best. You can choose to explore by visiting each one and trying their coffee, or you can just go to the first coffee shop on the list and assume it's the best.

If you explore, you might find a coffee shop that's even better than the first one, but it will take more time and effort. If you don't explore, you might miss out on the best coffee shop in the city.

In this example, exploration means trying new coffee shops, and exploitation means going to the coffee shop that you think is the best based on the information you have.

4b) Explain what makes reinforcement learning different from supervised learning tasks such as regression or classification.

The main difference between RL and SL is in the type of feedback they receive. In supervised learning, the algorithm is given labeled data and is trained to predict the correct output given an input. The algorithm receives feedback on its predictions in the form of a loss function or error metric, which measures the difference between the predicted output and the actual output. The goal of supervised learning is to minimize this error and make accurate predictions on new, unseen data.

While reinforcement learning works by trial and error, receiving feedback in the form of rewards or penalties for each action taken in an environment. The algorithm does not receive labeled data, but rather interacts with an environment and learns from the consequences of its actions. The goal of reinforcement learning is to learn an optimal policy, which is a mapping from states to actions that maximizes the cumulative reward over time.

Question 5

5a) Give a summary of how a decision tree works and how it extends to random forests.

A decision tree is supervised learning algorithm that is used for both classification and regression tasks. It makes predictions by recursively partitioning the feature space into subsets, based on the input features. Each internal node of the tree represents a decision based on the value of a particular feature, while each leaf node represents a class or a regression target. A decision tree is built by selecting the feature that results in the most significant decrease in the impurity of the data, and then recursively partitioning the data based on that feature. The process continues until a certain criteria is reached, such as a minimum number of samples in a leaf node or a maximum depth of the tree.

Random forests, on the other hand, is a learning method that combines multiple decision trees to make predictions. Each decision tree in the random forest is built on a random subset of the training data and a random subset of the input. To make a prediction, the random forest aggregates the predictions of all the individual trees, either by taking the majority vote in the case of classification or the average in the case of regression. Random forests are known for their good performance and ability to handle high-dimensional data.

5b) State at least one advantage and one drawback with using random forests over decision trees.

One advantage of using random forests over decision trees is that random forests can provide a more accurate and stable prediction than a single decision tree, especially in the case of high-dimensional and noisy data. Random forests are able to reduce the variance and overfitting of the model by combining multiple decision trees with randomized features and subsamples.

One drawback of using random forests over decision trees is that random forests can be more computationally expensive and require more memory, as they involve building and combining multiple decision trees. Additionally, the output of a random forest model may be less interpretable than a single decision tree, since the model consists of multiple trees and does not provide a clear and simple rule-based decision-making process.

References

Primer/text based on the following references:

- http://www.cse.chalmers.se/~chrdimi/downloads/book.pdf
- https://github.com/olethrosdc/ml-society-science/blob/master/notes.pdf