IDSAI_Assignment5_SP4_22-23

May 2, 2023

1 DAT405/DIT407 Introduction to Data Science and AI

1.1 2022-2023, Reading Period 4

1.2 Assignment 5: Reinforcement learning and classification

Name	Working Hours
Dimitrios Koutsakis	10
Bingcheng Chen	10

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.

1.3 Primer

1.3.1 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**.

1.3.2 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.

1.3.3 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: $t = 1, 2, ..., T$, where $T \leq \infty$
- State space: $S = \{s_1, s_2, ..., s_N\}$ of the underlying environment
- Action space $A = \{a_1, a_2, ..., a_K\}$ 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 $\pi : s \to a$, 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,

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

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.

1.4 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.

Answer 1a: The optimal path is EENNN, and the optimal path is unique.

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.

Answer 1b: | | | |
———-	———-	———-
E|E|F| |E/N|E/N|N(^)| |E/N|E/N|N(^)
S(>)	E(>)	N(^)

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

Answer 1c: The expected total reward for the policy is 0.

1.5 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 ) +  *V_k-1[s ])
    end
    if  |V_k[s]-V_k-1[s]| < epsilon for all s
        for each state s,
        do
             (s)=argmax_a  _s  p(s |s,a)*(r(a,s,s ) +  *V_k-1[s ])
```

```
            return  , 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

1.6 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. Make sure to consider that there may be several equally good actions for a state when presenting the optimal policy.

[ ]:
```
import numpy as np

# Answer 2a

# Parameters
epsilon = 0.0001
gamma = 0.9
action_probability = 0.8

reward_matrix = np.array([[0, 0,  0],
                          [0, 10, 0],
                          [0, 0,  0]])

# Calculate the value of the best action (N,S,W,E)
def bellman_equation(value_matrix, current_x, current_y):
    action_out_of_bounds = lambda x, y: x < 0 or y < 0 or x >= reward_matrix.
 ↪shape[0] or y >= reward_matrix.shape[1]
    calculate_action = lambda x, y: reward_matrix[x, y] + gamma*value_matrix[x,␣
 ↪y]

    actions = [
        (current_x-1, current_y, 'N'),
        (current_x+1, current_y, 'S'),
        (current_x, current_y-1, 'W'),
        (current_x, current_y+1, 'E'),
        ]

    action_values = []
    for action_x, action_y, action in actions:
        if action_out_of_bounds(action_x, action_y):
            continue

        value = action_probability*calculate_action(action_x, action_y) +␣
 ↪(1-action_probability)*calculate_action(current_x, current_y)
        action_values.append((value, action))

    return max(action_values, key=lambda x: x[0])


def value_iteration_algorithm(initial_value_matrix):
    iterations = 0

    value_matrix = initial_value_matrix
    policy_matrix = np.chararray(reward_matrix.shape, unicode=True)

    while True:
        iterations += 1
```

```
        prev_value_matrix = value_matrix.copy()

        for x, col in enumerate(value_matrix):
            for y, _ in enumerate(col):
                value, action = bellman_equation(prev_value_matrix, x, y)
                value_matrix[x,y] = value
                policy_matrix[x,y] = action


        if np.abs(value_matrix - prev_value_matrix).any() < epsilon:
            return value_matrix, policy_matrix, iterations
```

[ ]:
```
value_matrix, policy_matrix, iterations = value_iteration_algorithm(np.
 ↪zeros(reward_matrix.shape))

print('Iterations:', iterations)

print('\nValue Matrix:')
print(value_matrix)

print('\nPolicy Matrix:')
print(policy_matrix)
```

```
Iterations: 334

Value Matrix:
[[45.61292366 51.94805195 45.61292366]
 [51.94805195 48.05194805 51.94805195]
 [45.61292366 51.94805195 45.61292366]]

Policy Matrix:
[['S' 'S' 'S']
 ['E' 'N' 'W']
 ['N' 'N' 'N']]
```

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

[ ]:
```
value_matrix, policy_matrix, iterations = value_iteration_algorithm(np.
 ↪ones(reward_matrix.shape))

print('Iterations:', iterations)

print('\nValue Matrix:')
print(value_matrix)

print('\nPolicy Matrix:')
print(policy_matrix)
```

```
Iterations: 334
```

```
Value Matrix:
[[45.61292366 51.94805195 45.61292366]
 [51.94805195 48.05194805 51.94805195]
 [45.61292366 51.94805195 45.61292366]]

Policy Matrix:
[['S' 'S' 'S']
 ['E' 'N' 'W']
 ['N' 'N' 'N']]
```

Answer 2b:

By observing *Bellman equation* below, we can derive that the initial value V_0 can only affect term \$ V(s')\$, other terms in the equation are independent of V_0, however, since γ is less than 1, after hundreds of iteration, in our case, $\gamma^{334} \approx 0$, thus the effect of V_0 can be ignored.

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

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?

Answer 2c:

The discount factor γ is a parameter that determines the importance of future rewards in the decision-making process. It is a value between 0 and 1, in general, - a smaller value of γ means that the agent is more focused on the immediate reward. - a larger value of γ means that the agent places more emphasis on future rewards, the agent would take more future rewards into their current consideration of decision making.

When $\gamma = 0$, the agent only considers the immediate reward and ignores any future rewards. This can result in the agent learning policies that maximize immediate rewards without considering the long-term consequences of its actions. It may result in suboptimal policies being learned.

When $\gamma = 1$, the agent is completely far-sighted and considers all future rewards equally important, this can lead to slower convergence rates.

When deciding on which value of γ to use, that is highly dependent on the characters of tasks. - if the task is short-term and immediate rewards are more important than future rewards, a smaller value of γ may be more appropriate. - if the task is long-term and future rewards are also very important, then a larger value of γ may be more appropriate.

1.7 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. 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.

1.8 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")

[ ]:
```
import gym
import gym_toytext
import random
env = gym.make("NChain-v0")
env.reset()
```

```
c:\Python311\Lib\site-packages\gym\core.py:317: DeprecationWarning: WARN:
Initializing wrapper in old step API which returns one bool instead of two. It
is recommended to set `new_step_api=True` to use new step API. This will be the
default behaviour in future.
  deprecation(
c:\Python311\Lib\site-packages\gym\wrappers\step_api_compatibility.py:39:
DeprecationWarning: WARN: Initializing environment in old step API which
returns one bool instead of two. It is recommended to set `new_step_api=True` to
use new step API. This will be the default behaviour in future.
  deprecation(
c:\Python311\Lib\site-packages\gym\utils\passive_env_checker.py:174:
UserWarning: WARN: Future gym versions will require that `Env.reset` can be
passed a `seed` instead of using `Env.seed` for resetting the environment random
number generator.
  logger.warn(
c:\Python311\Lib\site-packages\gym\utils\passive_env_checker.py:190:
UserWarning: WARN: Future gym versions will require that `Env.reset` can be
passed `return_info` to return information from the environment resetting.
  logger.warn(
c:\Python311\Lib\site-packages\gym\utils\passive_env_checker.py:195:
UserWarning: WARN: Future gym versions will require that `Env.reset` can be
passed `options` to allow the environment initialisation to be passed additional
information.
  logger.warn(
```

[ ]: 0

[ ]:
```
print("Action Space: {}".format(env.action_space))
print("State Space: {}".format(env.observation_space))
```

```
Action Space: Discrete(2)
State Space: Discrete(5)
```

[ ]:
```
num_episodes = 10000
gamma = 0.95

# start with a small learning rate
learning_rate = 0.01
epsilon = 0.1

# initialize Q table
Q = np.zeros((5, 2))

print('Initial Q:')
print(Q)


for _ in range(num_episodes):
    state = env.reset()
    isComplete = False

    while not isComplete:
        # Select pseudo random action
        action = env.action_space.sample() if random.uniform(0, 1) < epsilon␣
 ↪else np.argmax(Q[state])

        # perform the action and receive the feedback from the environment
        new_state, reward, isComplete, _ = env.step(action)

        # Update Q
        Q[state, action] += learning_rate * (reward + (gamma*np.
 ↪max(Q[new_state])) - Q[state, action])

        state = new_state


print('\nFinal Q:')
print(Q)
```

```
Initial Q:
[[0. 0.]
 [0. 0.]
 [0. 0.]
```

```
 [0. 0.]
 [0. 0.]]

Final Q:
[[62.6382994  61.64412503]
 [66.44371474 62.2138987 ]
 [72.2155785  63.35460432]
 [78.55440734 64.82338151]
 [84.19058071 69.27699558]]
```

1.9 Question 4

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

Answer 4a:

Exploration is a significant part of reinforcement learning, especially in the early stage of learning, when the agent has limited knowledge of the environment, exploration allows the agent to discover new possible better actions and policies that lead to higher rewards. Without exploration, the agent will only consider the actions that it has previously tried and has a biased estimate of their value, which may not be the true optimal value.

For example, consider the problem of a robot learning to find the shortest path from A to B, without exploration, the robot will only take path that has been previously tried, but there maybe shorter path exists. With exploration, the robot can discover path by exploring a random path with probability ϵ, which may have longer length but could lead to shorter path in total in the long run.

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

Answer 4b:

Difference	supervised learning	reinforcement learning
Learning process	learn from labeled training dataset where the correct output is already known.	learn from feedback, which is typically a reward that indicates how good or bad the agent's actions were in a given state.
Goal	predict the output for new input	learn a policy that maximizes the expected cumulative reward over time
Exploration-exploitation	the training data is fixed, and the algorithm can optimize its parameters to fit the data accurately.	the agent needs to balance exploration and exploitation to discover better policies while still maximizing the expected cumulative reward.
Sequential decision-making	the input-output relationship is typically independent of the order of the data points	the agent needs to make sequential decisions and consider the consequences of its actions on future rewards.

1.10 Question 5

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

Answer 5a:

Decision Tree: - **Step 1:** The algorithm starts with the entire dataset as the root node of the tree, the algorithm evaluates each feature and chooses the one that provides the most information gain or the best split (Information Gain/Gain Ratio/Gini Index). - **Step 2:** The selected feature is used to split the data into subsets, where each subset corresponds to a branch of the tree. - **Step 3:** Repeats step 1&2 on each subset, recursively splitting the data until one of the conditions match:

```
- All the leaf nodes have the same class label
- There are no more remaining features.
- There are no more instances.
- Reach a pre-specified stopping criterion such as a maximum depth.
```

From decision tree to random forests:

- **Step 1:** A random subset of the data is selected from the original dataset called a bootstrapped dataset. This subset is used to train the first decision tree.
- **Step 2:** At each split in the decision tree, a random subset of the features is selected. The optimal split is chosen from among these features which is the same as decision tree.
- **Step 3:** Repeat step 1&2 to create an ensemble of decision trees.
- **Step 4:** To classify a new data point, the random forest algorithm traverses each decision tree in the ensemble, and the final prediction is the majority vote.

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

Answer 5b:

Advantages: - Random forests gererally have higher accuracy and it is a more robust method. - Random forests reduce overfitting, since it builds an ensemble of decision trees that capture different aspects of the data and make a collective prediction.

Disadvantages: - Since random forests need to generate multiple decision trees and combine their predictions, thus random forests are more expensive and slower to train and test, especially for large datasets and high-dimensional feature spaces. - Random forests are more difficult to interpret, which could be an issue in some applications such as medical diagnosis.

2 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