Proff. Nicolò Cesa-Bianchi, Alfio Ferrara

Lab of Reinforcement Learning

Final project proposals

Introduction

The final project consists in the implementation of a RL solution to a simple problem. Students have to provide access to a GitHub repository containing the code and reproducible results. Finally, the project will be discussed after a 10 minutes presentation in English with slides. The project discussion will be set by appointment, according to the following procedure:

1. Submit the link to your GitHub repository as soon as the project is ready
2. Contact Proff. Cesa-Bianchi and Ferrara and set and appointment for the discussion

The project is an individual work. It is possible to present a project other than those proposed and carried out by a maximum of two students. In this case, however, it is necessary to agree on the topic with the laboratory teachers.

Project proposal 1: BlackJack (sort of)

In this project we will play a (modified version) of black jack. The game is run against a dealer. There is a deck of 52 cards composed by 4 sets (corresponding to the four suits) 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A. The card values are:

- the nominal card value for cards between 2 and 10
- the value of J, Q, K is 10
- the value of A is 11

We call *hand value* (H) the sum of the cards in your hands. When you start the game you receive 2 cards from the deck.

At each subsequent interaction with the dealer, you can take two actions:

1. HIT: A new card is extracted from the deck and added to your hand
    1. If $H > 21$ the game ends and you loose 1 €
    2. otherwise, the game goes on
2. STAY: You keep H as your final score and the game ends
    1. Now the dealer receives 2 cards from the deck and computes the *dealer hand value* D.
    2. If $H > D$, you win 1 €; if $H = D$, there is a draw and you gain 0 €; if $D > H$, you loose 1 €

Goals

1. Model this problem as an MDP and implement the game
2. Use one of the algorithms presented in class to learn a game policy by playing several games against the dealer.
3. Study how the policy changes if we modify the number of cards received by the dealer when the game ends.

Project proposal 2: Play RISK!

We are playing modified version of Risk! and it's our turn. The game works as follows:

- There is a map of territories connected by borders that can be modeled as an undirected graph
- We occupy one initial territory with N troops; all the other territory are occupied by E enemy troops distributed uniformely over the territories
- We can choose two actions: ATTACK or STAY
  - STAY: our turn ends and the enemy attacks one of our territories with fewer troops starting from one of his territories with more troops bordering ours
  - ATTACK: from one of your territories with more troops you attack one of the bordering enemy territories with fewer troops
- ATTACK procedure:
  - The attacker rolls three six-sided dice (3d6). The defender rolls three six-sided dice (3d6). The attacker's best die is compared to the defender's best, second best to second best, and so on. The highest die wins. In case of a tie, the defender wins. For each confrontation, the losing side loses one troop. The attacker never loose the last troop, which means that the last troop of the attacker, even if the attacker looses the corresponding dice confrontation, is not lost.
  - If the attacker wins, they occupy the enemy territory with all their troops but one that remains in the the starting territory.
- The game ends after 10 turns.
- Each time you occupy an enemy territory you gain K victory points. Each time you loose a troop, you loose 1 victory point.

Goals

1. Explore the possibility to model this game as an MDP and discuss the possible options. You can also try to change the rules and simplfy the setting if needed. Present an implementation and discuss your design choices.
2. According to your model, choose a strategy to learn a policy that optimizes your final reward in terms of victory points.
3. Discuss the role of the parameters N, E, and K in your solution.

Project proposal 3: Manage a farm

You are the manager of a farm. You have an initial budget of 2000 €. Each year you have to take some decisions about how to invest you money, but you can do only one of the following things:

1. Buy one sheep: a sheep costs 1000 €
2. Growing wheat: when you choose this action, you spend 200 €

At the end of the year, you harvest the wheat and you sell your wool. Each sheep produces 1 wool unit that is sold for 10 €. Selling the harvested wheat instead gives you 50 €.

However, during the year, there is a probability α that your fields are devastated by a storm. In this case, your harvest will give you 0 €.

Moreover, if you have more than one sheep, there is a probability β that each pair of sheep generates a new sheep.

You manager career ends if you run out of money or, in any case, after 30 years, when you will retire.

Goal

1. Find a way to leave your heirs as much expected legacy as possible, which means that you want to learn the best investment strategy in order to maximize your total monetary reward at the time of your retire
2. Study how α and β influence the situation you have to deal with

Project proposal 4: Grid driving

We want to learn how to drive a car on a racetrack that contains at least one turn. The racetrack is modeled as a collection of discrete positions organized in a grid. There is a starting line that is also an ending line (an horizontal row of grid cells), in such a way that the racetrack is has a circular shape.

The car movement is determined by its velocity which is made of an horizontal and a vertical component. The maximum value of each component is 5 grid cells and the minimum value is 0. At the starting line the velocity is 0 for both the components.

The actions that the car can take are to change the two velocity components, each by +1, -1, or 0.

Each movement costs -1. The race ends when the finish line is reached.

However, if the car runs out of the racetrack, the race is not over, but the car must restart from the starting line with velocity 0.

(**optional**) Finally, on the track there are some randomly arranged obstacles that occupy some cells of the grid. Passing by these obstacles involves the risk of getting a flat tire and consequently reducing the speed to 0.

Goal

1. Implement the grid driving environment by modeling states, actions and the transitions from one state to the other.
2. Implement an agent that learns by practice how to reach the finish line as fast as possible.

Project proposal 5: Healthcare

Here, you represent the government of a country and your responsibility is to take care of the healthcare policy, but you have a limited budget. In particular, every year you have three available actions:

- INVEST in healthcare
    - the *level of the healthcare* is incremented by 3 but the budget decreases by 2
- INVEST in education and prevention
    - the budged is reduced by 1, but the current *health risk index* is also reduced by 1 (however, it never becomes lower that 0)
- DO NOTHING and save your budget
    - the budget increases by 2 and the healthcare level does not change

After each year, you get the current budget as a reward. However, after each year the health risk level rises randomly by 1, 2, or 3 points. When this happens, if the health risk level becomes higher that the healthcare level, there is a risk of a pandemic occurring in the country. More specifically, given a the pair of health risk and healthcare levels, only if $\textit{health risk} > \textit{healthcare}$, there is a non-zero probability that a pandemic occurs. In case of a pandemic, you are fired from your position and the simulation ends with a large final negative reward.

Try different initial levels for the budget, the healthcare level, and the health risk index.

Goal

1. Explore how the optimal policy changes under different scenarios (in particular with respect to the probability of health risk increment and different discount factors)
2. Try to model a situation where the budget is not the priority. For example:
    1. The priority is to minimize the health risk with a limited budget
    2. When you are a politician, you never know when your job ends. Thus, we want to keep people safe now, not in a far future
    3. Every 5 years that are the elections. Your probability to stay in your place is proportional to the difference between health index and quality of the health care. Your priority is to keep your job of course...