A Reinforcement Learning - Monte Carlo Model for Trip Planning

Stochastic Models and Optimization

Professor: Mihalis Markakis

By: Maryam Rahbaralam

INTRODUCTION

Reinforcement learning (RL)

- Reinforcement learning is all about learning from the environment through interactions.
- → Finding the optimal policy / optimal value functions is the key for solving reinforcement learning problems.
- →Dynamic programming methods are used to find optimal policy/optimal value functions using the bellman optimality equations.
- →Dynamic programming methods are model based methods, require the complete knowledge of environment. such as transition probabilities and rewards.

There are two types of tasks in RL:

1. Prediction : This type of task predicts the expected total reward from any given state assuming the function **π(a|s)** is given.

- (in other words) **Policy π** is given, it calculates the **Value function Vπ** with or without the model.

ex: Policy evaluation

2. Control : This type of task finds the policy **π(a|s)** that maximizes the expected total reward from any given state.

- (in other words) **Some Policy π** is given , it finds the **Optimal policy π***.

ex: Policy improvement

- Policy iteration is the combination of both to find the optimal policy.

Using Monte carlo approach
for solving these kind of problems:

Monte Carlo learning takes the combined the quality of each step towards reaching an end goal and requires that, in order to assess the quality of any step, we must wait and see the outcome of the whole combination.

The value function is the expected return:

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$$

Our update rule is as follows:

$$V(a) \leftarrow V(a) + \alpha * (G - V(a))$$

Using Monte carlo approach
for solving these kind of problems:

Monte Carlo Control

- Similar to dynamic programming, once we have the value function for a random policy, the important task that still remains is that of finding the optimal policy using Monte Carlo.
- Recall that the formula for policy improvement in DP required the model of the environment as shown in the following equation:

$$\pi'(s) \quad = \quad \arg\max_{a} \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big]$$

- This equation finds out the optimal policy by finding actions that maximize the sum of rewards.

Problem statement

Motivation:

One of the interesting problems in travel planning is to choose between different hotel options in the target cities. This choice is based on a planned budget. Some of the hotels are less expensive, others are of higher quality. I aim to create a model that, for the target cities, can select the optimal hotels required to make a trip that is both :

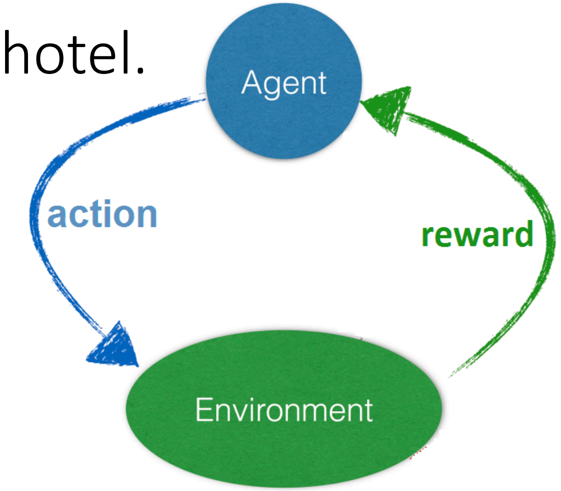
1. Within my budget
2. Meets my personal preferences

PROBLEM STATMENT

Method

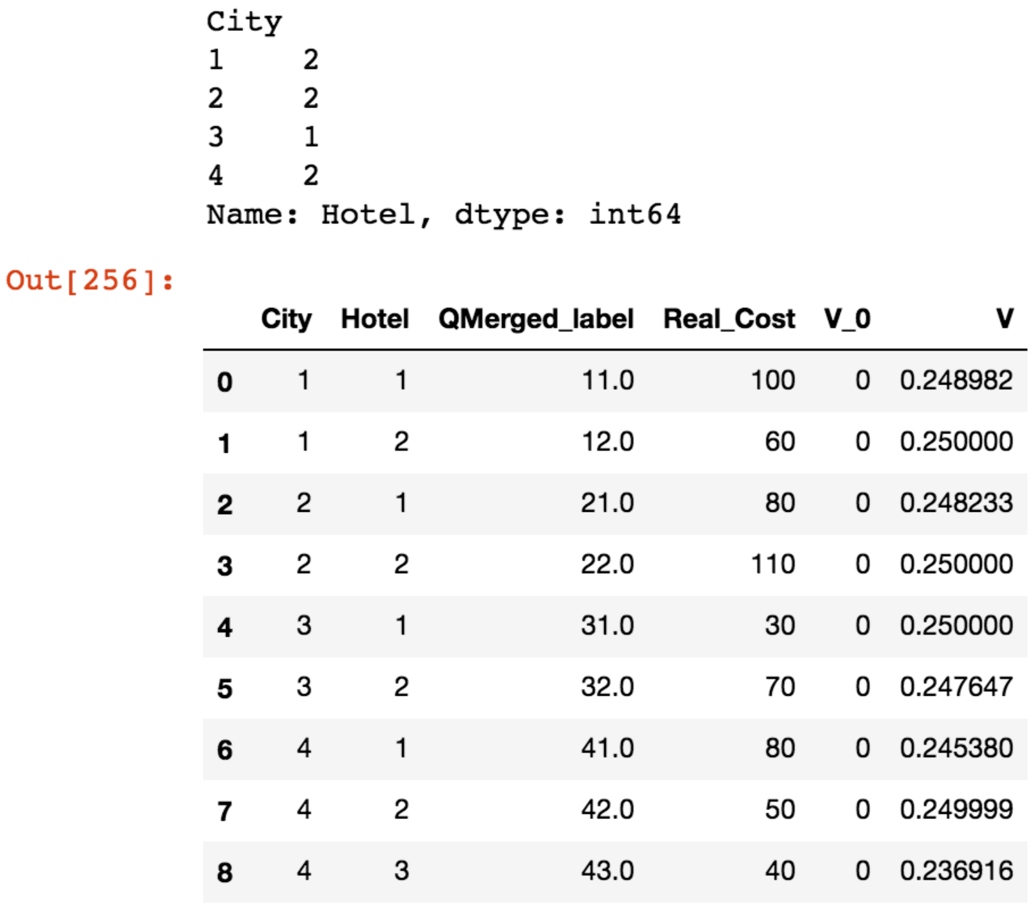
To achieve this, a Monte Carlo method based on reinforcement learning model is employed to find the optimal combination of hotels. First, the model is defined as a Markov Decision Process:

- We have a finite number of cities required to make any trip plan and are considered to be our **States**.
- There are the finite possible hotels for each city and are therefore the **Actions** of each state.
- Our preferences become the. **Individual Rewards** for selecting each hotel.



Results

In optimal actions of final episode, the model suggestion is shown. In this run it suggests actions, or hotels, that have a total cost below budget which is good.



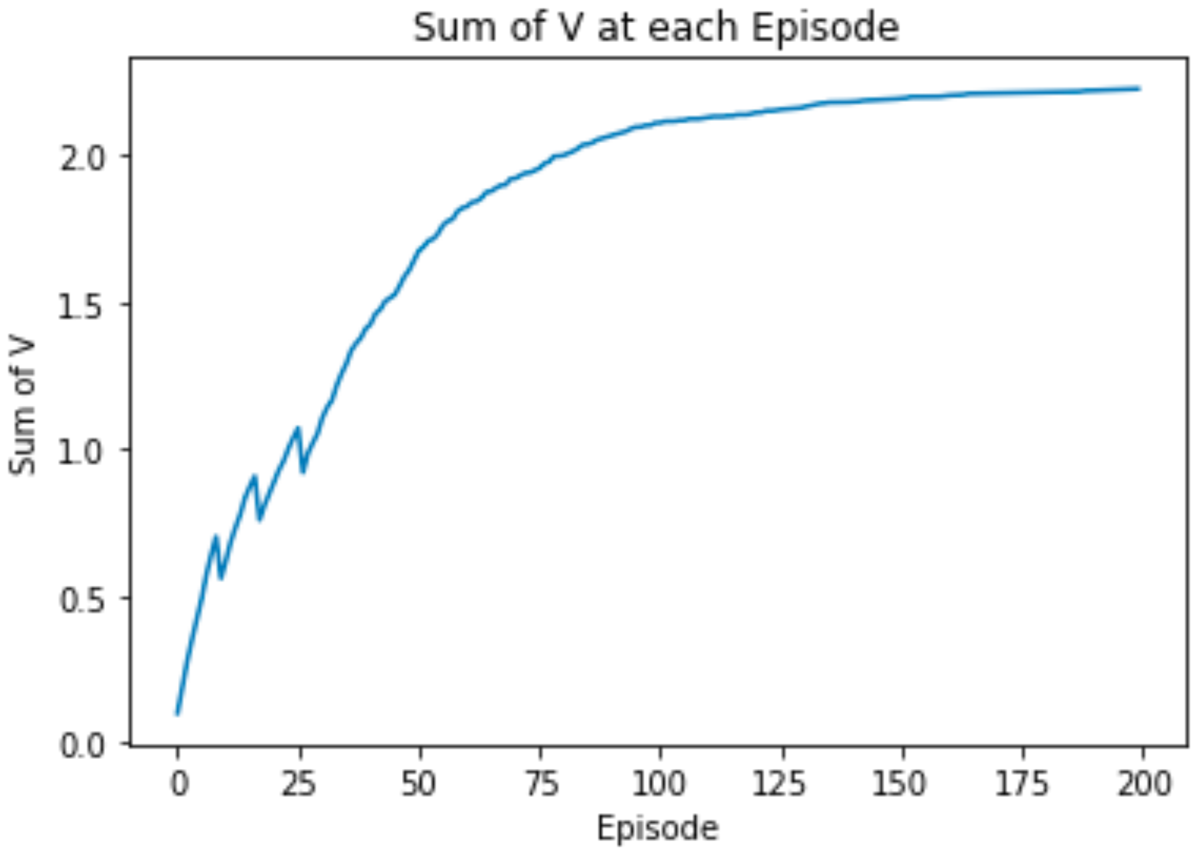
```
City
1    2
2    2
3    1
4    2
Name: Hotel, dtype: int64
```

Out[256]:

	City	Hotel	QMerged_label	Real_Cost	V_0	V
0	1	1	11.0	100	0	0.248982
1	1	2	12.0	60	0	0.250000
2	2	1	21.0	80	0	0.248233
3	2	2	22.0	110	0	0.250000
4	3	1	31.0	30	0	0.250000
5	3	2	32.0	70	0	0.247647
6	4	1	41.0	80	0	0.245380
7	4	2	42.0	50	0	0.249999
8	4	3	43.0	40	0	0.236916

Results

To show the convergence of the proposed model, the total value function V as a function of the number of episodes is plotted and we see that this is converging which is ideal. We want our model to converge so that as we try more episodes we are ‘zoning-in’ on the optimal choice of hotels.



Conclusion

- I have implemented a Monte Carlo Reinforcement Learning model to select a combination of hotels in the target cities for travel planning.
- This model results in a selection of hotels below a proposed budget and it can also incorporate a preference of hotels.
- The results show that the value function, which determines the outcome, converges as the number of choice combinations (episodes) increases.