HOMEWORK 5

CSCI 688-3 ⋆ Applications of Markov Chains

Spring 2023

Instructions: Solve each of the problems and submit your work as indicated below.

- Homework due on Monday, April 17th by the end of the day.
- You may work in groups of 2-3 people in this homework.
- Submit a single pdf file to Gradescope and specify the pages you used for each question. If you scan your homework, make sure the quality is appropriate.
- Follow the "empty hands policy" described on the syllabus and submit your own work.
- If you use any AI software (such as ChatGPT), remember to report the solution provided by the software and explain why you believe is correct. If it is incorrect, specify why and how you would correct it.

Problem 1. *Each quarter, the marketing manager of a retail store divides customers into two classes based on their purchase behavior in the previous quarter. Denote the classes as L for low and H for high. The manager wishes to determine to which classes of customers he should send quarterly catalogs.*

The cost of sending a catalog is \$15 per customer. The expected reward depends on the customer's class and the manager's decision, as detailed in the following table

Customer's class	*Receives a catalog*	*Does not receive a catalog*
Low	*\$20*	*\$ 10*
High	*\$50*	*\$25*

The decision whether or not to send a catalog to a customer also affects the customer classification in the subsequent quarter, as shown in the following matrices. If a customer receives a catalog, the transition probabilities of their class in consecutive quarters are

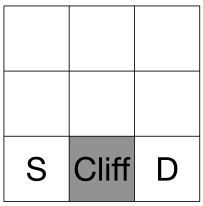
$$P_{catalog} = \begin{matrix} L \\ H \end{matrix} \begin{bmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{bmatrix}$$

and if a customer does not receive a catalog, the transition probabilities are

$$P_{no\ catalog} = \begin{matrix} L \\ H \end{matrix} \begin{bmatrix} 0.5 & 0.5 \\ 0.6 & 0.4 \end{bmatrix}$$

You might recognize this problem from Homework 4. Starting from the policy π_0 defined by $\pi_0(catalog|L) = \pi_0(catalog|H) = 0.5$, run value iteration to compute an optimal policy. Use $\gamma = 0.5$.

Problem 2. *Consider a small grid world of 9 cells with a cliff, as shown in the figure.*



Considering the same setting as we saw in class regarding actions and rewards (but with a smaller grid),

(a) Model the problem as an MDP

(b) Using $\alpha = 0.1$, $\gamma = 1$, a maximum of 200 episodes, run Sarsa and Q-learning with ϵ-greedy and $\epsilon = 0.1$ to find the optimal route.

(i) Graph the total rewards by the end of each episode with respect to the episode number for each algorithm. What do you observe?

(ii) Graph the number of steps per episode with respect to the episode number for each algorithm. What do you observe?

(iii) What is the optimal policy when you use Sarsa? How about Q-learning?