Name:
Student ID:

CSE 573 Winter 2023: HW 2

Due 3/7/2023 by 11:59 pm
Total: 100 points

Instructions:

1) The homework can be done in teams of two people. Don't forget to write both names if you are in a team.
2) We highly recommend typing your homework, but writing and scanning, or annotating a PDF are also acceptable.
3) Keep your answers brief but provide enough explanations and details to let us know that you have understood the topic.
4) The assignment is due on March 7, 11:59 pm.
5) You should upload your assignments through gradescope. If you are submitting with a partner, please tag your partner on the submission using the "Add Group Member" feature.

Topics:	Points
Short Answers	10
RL Background	16
Value Iteration	20
Q-Learning I	15
Q-Learning II	6
Bayes Nets	20
Perceptrons	13

Q1: Short Answer [10 pts]

1.1. [3 pts] Briefly explain what “epsilon-greedy” means in Q-learning and discuss what aspects of off-policy learning it aims to improve.

1.2. [4 pts] In practice, we often use approximate Q-learning by learning the following function, $Q(s,\ a)\ =\ g(f_1(s,\ a),\ f_2(s,\ a)\ \ldots\ ,\ f_n(s,\ a))$. What is the advantage of using this approach compared to tabular (or exact) Q-learning?

1.3 [3 pts] Short Answer – Give a real-world example of a problem that a binary perceptron might not be suited for.

Q2: Reinforcement Learning Background [16 pts]

2.1 Each True/False question is worth 2 points. Briefly justify your answers.

A. [true or false] Temporal difference learning is a model-based learning method.

B. [true or false] Using an optimal exploration function for Q-learning will not have any regret while learning the optimal policy.

C. [true or false] In a deterministic MDP, Q-learning with a learning rate of α = 1 cannot learn the optimal q-values.

D. [true or false] A large γ (around 1) incentivizes greedy behavior.

E. [true or false] A living reward less than 0 encourages policies that maximize the probability of reaching terminal states.

F. [true or false] Every γ < 1 can be rewritten using a mathematically equivalent negative living reward.

2.2. Properties of reinforcement learning algorithms (2 points each)

A. Which algorithm, assuming ample exploration, does not provide enough information to obtain an optimal policy at the time of convergence? (Select all that apply)

- ☐ Temporal Difference learning to estimate $V(s)$.
- ☐ Direct Evaluation to estimate $V(s)$.
- ☐ Q-Learning to estimate $Q(s, a)$.
- ☐ Model-based learning of $T(s, a, s')$ and $R(s, a, s')$.

B. Assuming we run for infinitely many steps, for which exploration policies is Q-learning guaranteed to converge to the optimal Q-values for all state-action pairs. Assume we chose reasonable values for α and all states of the MDP are connected via some path. (Select all that apply)

- ☐ A fixed optimal policy.
- ☐ A fixed policy taking actions uniformly at random.
- ☐ An ϵ-greedy policy
- ☐ A greedy policy.

Q3: Value Iteration (MDPs) [20 pts]

Consider the 101x3 world below. In the start state, the agent has a choice of two deterministic actions, Up or Down, but in the other states the agent always takes the deterministic action, Right. Grayed out parts are not accessible to the agent. You can assume that the ±50 states correspond to t=0.

+50	-1	-1	-1	-1	…	-1	-1	-1	-1(**TERMINAL**)
START									
-50	+1	+1	+1	+1	…	+1	+1	+1	+1(**TERMINAL**)

3.1. [10 pts] Compute the utility of each action as a function of γ.

3.2. [10pts] Assuming a discounted reward function, for what values of the discount factor γ should the agent choose Up as the initial action?

Q4: Reinforcement Learning I [15 Points]

You are playing a peculiar card game, but unfortunately you were not paying attention when the rules were described. You did manage to pick up that for each round you will be holding one of three possible cards [Ace, King, Jack] ([A, K, J], for short) and you can either Bet or Pass, in which case the dealer will reward you points and possibly switch out your card. You decide to use Q-Learning to learn to play this game, in particular you model this game as an MDP with states [A, K, J], actions [Bet, Pass] and discount $\gamma = 1$. To learn the game you use $\alpha = 0.2$.

4.1 Say you observe the following rounds of play (in order):

s	a	s'	r
A	Bet	K	4
J	Pass	A	0
K	Pass	A	-4
K	Bet	J	-12
J	Bet	A	4
A	Bet	A	-4

(7 pts) What are the estimates for the following Q-values as obtained by Q-learning? All Q-values are initialized to 0.

i) Q(J, Pass) = _____

ii) Q(J, Bet) = _____

4.2 For this next part, we will switch to a feature based representation. We will use two features:

$$f_1(s, a) = 1$$

$$f_2(s, a) = \begin{cases} 1 & a = \text{Bet} \\ 0 & a = \text{Pass} \end{cases}$$

Starting from initial weights of 0, compute the updated weights after observing the following samples:

s	a	s'	r
A	Bet	K	16
K	Pass	A	0

(4 pts) What are the weights after the first update, in other words, after using the first sample?

iii) w1 = _____

iv) w2 = _____

(4 pts) What are the weights after the second update, in other words, after using the second sample?

v) w1 = _____

vi) w2 = _____

Q5: Reinforcement Learning II [6 pts]

Given the following list of Q-values for state s and the set of actions {*Left*, *Right*, *Fire*} (7 points):

Q(s, *Left*) = 0.25
Q(s, *Right*) = 0.4
Q(s, *Fire*) = 0.7

What is the probability that we will take each action on our next move when following an ε-greedy exploration policy (assuming all random movements are chosen uniformly from all actions)?

Action	Probability, in terms of ε
Left	
Right	
Fire	

Q6: Bayes Nets [20 pts]

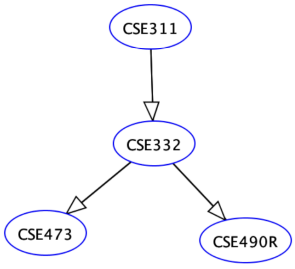
As part of a comprehensive study of the role of 10-601 on people's happiness we have been collecting important data from graduating students. In an entirely optional survey that all students are required to complete, where they were asked about whether they partied, whether they did well on their homeworks and projects, whether they used a Mac, and whether they thought of themselves as smart, creative, happy and successful. After consulting a behavioral psychologist, the following complete set of conditional relationships between the eight binary variables *Party, Smart, Creative, HW, Mac, Project, Success,* and *Happy*:

- Students' homework performance depends only on their partying habits and how smart they are.
- Students' Mac use is related to how smart and creative they are.
- Students' project outcomes depend only on how smart and creative they are.
- A student's success depends on their performance on homeworks and projects alone.
- Finally, students are happy depending on their degree of success, their partying habits, and if they use a Mac!

6.1 [6 pts] Draw a reasonable Bayes Network for the relationships between these variables.

6.2 [6 pts] Write joint distribution as a product of conditional probabilities for the above Bayesian network.

6.3 [8 pts] Consider the following Bayesian Network about class prerequisites in UW.



Give reasonable probability tables for all nodes in this subnetwork.

Q7: Perceptrons [13 points]

Consider using the perceptron algorithm to learn the logical OR function with the training set:

(x_1, x_2, b)	y^*
(-1,-1,1)	-1
(-1,1,1)	1
(1,-1,1)	1
(1,1,1)	1

Assume the following perceptron definition: $f(x) = sign(w \cdot x)$ where

$$\text{sign}(x) = \begin{cases} +1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

Updating if we guess incorrectly, i.e., $f(x) \neq y^*$, using the rule:

$w \leftarrow w + y^* \cdot x$

7.1 [8 pts] Fill out the below table:

Iteration	x	w	$f(x)$	y^*
1	(-1,-1,1)	(0,0,0)	1	-1
2	(-1,1,1)			1
3	(1,-1,1)			1
4	(1,1,1)			1
5	(-1,-1,1)			-1
6	(-1,1,1)			1
7	(1,-1,1)			1
8	(1,1,1)			1

7.2 [3 pts] Has training converged? Why or why not?

7.3 [2 pts] If we initialized w to a non-zero weight vector would we necessarily converge to the same final weight vector? Why or why not?