HW3: Petting a warg

Wargs do not make good pets. They are vicious creatures, populating Middle Earth, the world described by novels of John Ronald Reuel Tolkien. They tend to show up in the worst moment possible. They eat humans, hobbits, elves and wizards (when they can get them).

Figure 1: A warg, getting ready for breakfast

A warg can be in three states: ANGRY, ATTACK and DEAD. DEAD is a terminal state, no further actions can be performed.

You can apply three different actions to a warg: PET, POKE, SHOOT. A possible policy might be described as:

S	π(S)
ANGRY	PET
ATTACK	PET

Do not pet a warg if you encounter one in real life! However, you will do so in this homework.

Markov Decision Process

Let us assume that the pet-a-warg game is described by the following table. For every transaction not listed in the table, the probability is zero.

S	A	S'	T(S,A,S')	R(S,A,S')
ANGRY	PET	ANGRY	80%	-10
ANGRY	PET	ATTACK	20%	-50
ANGRY	POKE	ATTACK	100%	-50

S	A	S'	T(S,A,S')	R(S,A,S')
ANGRY	SHOOT	ATTACK	50%	-50
ANGRY	SHOOT	DEAD	50%	-200
ATTACK	PET	ATTACK	100%	-50
ATTACK	POKE	ATTACK	100%	-50
ATTACK	SHOOT	ATTACK	80%	-50
ATTACK	SHOOT	DEAD	20%	-200

Q1

Calculate the V^* value of all states using value iteration, considering $\gamma = 0.5$. Trace the value iteration updates. Stop at k=3 (inclusive) even if it did not converge.

Q2

Calculate the $\pi(s)$ policy from the V^* value (or the closest you reached) using one expectimax step. Show the $\pi(s)$ policy.

Model based RL

In this part of the homework, we assume that we have the same states and actions, but the T and R probabilities might be different. You will need to infer this different policy.

Consider that you have the following runs (the R values in the parentheses show the reward obtained for each transition):

- Run 1:
  - ANGRY –> PET –> ANGRY (R=+30)
  - ANGRY –> PET –> ANGRY (R=+30)
  - ANGRY –> PET –> DEAD (R=+100)
- Run 2:
  - ANGRY –> PET –> ANGRY (R=+30)
  - ANGRY –> PET –> ATTACK (R=-60)
  - ATTACK –> SHOOT –> DEAD (R=+100)
- Run 3:
  - ANGRY –> POKE –> ANGRY (R=-10)
  - ANGRY –> POKE –> ATTACK (R=-60)
  - ANGRY –> SHOOT –> ATTACK (R=-70)
  - ATTACK –> SHOOT –> DEAD (R=+100)

Q3

Estimate the probability table from the runs described above. Assign zero probability to transitions that did not occur. Show the resulting table.

Q4

Calculate the V^* values corresponding to the probability table you just created, considering $\gamma = 0.5$.

Model-free RL

In this part we will again use the runs shown in the model-based RL section. However, ignore the probabilities and V^* values calculated there.

Q5

Use the Monte-Carlo style direct evaluation of V. Show your work and the V values.

Q6

Do you expect the V values obtained in questions Q4 and Q5 to converge? Why / why not?

TD-learning

In this part we will again use the runs shown in the model-based RL section. However, ignore the probabilities and V^* values calculated there.

Q7

Use TD-learning, with a learning rate α=0.5, to find the V values based on the runs. Assume that the V values are initialized to zero, and the V(DEAD) is fixed to zero.

Q8

Do you expect the V-values obtained in questions Q5 and Q7 to converge if you are getting more runs? Why / why not?

Q-learning

In this part we will again use the runs shown in the model-based RL section. However, ignore the probabilities and V^* values calculated there.

Q9

Start with the following Q-values.

S	A	Q(S,A)
ANGRY	PET	-100

S	A	Q(S,A)
ANGRY	POKE	-100
ANGRY	SHOOT	-100
ATTACK	PET	-100
ATTACK	POKE	-100
ATTACK	SHOOT	-100
DEAD	PET	0
DEAD	POKE	0
DEAD	SHOOT	0

Use Q-learning to update the Q-values using the runs. Use $\alpha = 0.3$. Show the work and the final Q table.

Q10

Show the policy inferred by at the end of the Q-learning in Q9.