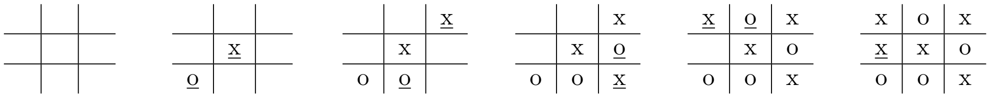
6.390 Introduction to Machine Learning
Recitation Week #11
Issued May 1, 2023

1. Tic-tic-tic is a game for two players who take turns marking the spaces in a 3×3 grid. One player uses an 'X' for their mark and the other uses an 'O'. For both players, the object of the game is to place three of their marks sequentially in a row–either horizontally, vertically, or diagonally.

   Suppose that we are the player using the X marks, and that the O player is a (possibly stochastic) algorithm. We do not know the strategy or reward function that is being used by the algorithm. The initial state of the board is empty and we make the first move. We can select *any* of the nine squares on our turn.

   If there are any remaining squares, it is our turn to play. If we select an occupied square that already has an X or an O in it, reward is zero, we remain in the same state and it's still our turn. Once the board has four X's and four O's, the agent is now in a near-full board state. When we place an X in the final empty square the game will end.

   The diagram shows a full game of tic-tic-tic from our perspective with the X marks (the underlined marks denotes the play from both players on the current turn):



   (a) First, we need to decide how to represent the state space for tic-tic-tic.

      i. Jody suggests letting the state space be all possible 3×3 grids in which each square contains one of the following: a space, an O, or an X. Is this a valid state space representation? Is it a good state space representation? For each question, why or why not?

      ii. Dana suggests using all possible 3×3 grids of X's, O's, and empty spaces such that either: (1) the number of O's and the number of X's are equal or (2) there are 5 X's and 4 O's. Is Dana's suggestion better or worse for tabular Q-learning than Jody's? Is it a good state space representation? For each question, why or why not?

(b) What is a good choice of action space for tic-tic-tic?

(c) Suppose we would like a reward function for tic-tic-tic that, at the end of the game, gives +1 reward to the X player for every three-in-a-row of X's and -1 reward to the X player for every three-in-a-row of O's. Sketch out the reward function. What are the possible rewards for a single game? With this reward function, is the game guaranteed to end? Why or why not?

(d) Using Dana's state space, your action space from (b) and reward function from (c), do we have a complete description of an MDP? If yes, write out the MDP; if not, elaborate on what might be missing.

2. Your friend Barney is training a tic-tic-tic bot (Tic-Bot) using tabular Q-learning with a discount factor of $\gamma = 1$ and a stepsize of $\alpha = 1$. You may assume that Tic-Bot has already learned to only take actions in unoccupied squares and that all of the experiences that Tic-Bot is learning from end in a full board state. Tic-Bot plays with the 'X' marks and the 'O' player is a (possibly stochastic) algorithm. At the current iteration, Barney's bot is in the following state s_1:

$$s_1 = \begin{array}{c|c|c} \text{o} & & \text{o} \\ \hline \text{x} & \text{x} & \text{x} \\ \hline \text{o} & & \end{array} \Longrightarrow \begin{array}{c|c|c} \text{o} & a_1 & \text{o} \\ \hline \text{x} & \text{x} & \text{x} \\ \hline \text{o} & a_2 & a_3 \end{array}$$

The left diagram is the current board state while the right diagram labels the three unoccupied squares which correspond to actions for Tic-Bot to choose from. The current Q-values for s_1 above are: $Q(s_1, a_1) = 0$, $Q(s_1, a_2) = 0$, and, $Q(s_1, a_3) = 1$.

(a) What *must* have happened during a previous iteration that led to $Q(s_1, a_3) = 1$? Draw the game state in the diagram below and explain your answer.

$$\begin{array}{c|c|c} \phantom{\text{o}} & \phantom{\text{o}} & \phantom{\text{o}} \\ \hline & & \\ \hline & & \end{array}$$

(b) Following greedy action selection, Tic-Bot takes action a_3. Provided below are the two possible game states that Tic-Bot may have transitioned into after taking action a_3, where the underlined marks denote the play from both players on the current turn. What are the possible new values of $Q(s_1, a_3)$?

$$\begin{array}{c|c|c} \text{o} & & \text{o} \\ \hline \text{x} & \text{x} & \text{x} \\ \hline \text{o} & \underline{\text{o}} & \underline{\text{x}} \end{array} \qquad\qquad \begin{array}{c|c|c} \text{o} & \underline{\text{o}} & \text{o} \\ \hline \text{x} & \text{x} & \text{x} \\ \hline \text{o} & & \underline{\text{x}} \end{array}$$

(c) Instead of selecting the greedy action a_3 in s_1, consider taking action a_1. Provided below are the two possible game states that Tic-Bot may have transitioned into after taking action a_1, where the underlined marks denote the play from both players on the current turn. What are the possible new values of $Q(s_1, a_1)$?

o	$\underline{\text{x}}$	o
x	x	x
o	$\underline{\text{o}}$	

o	$\underline{\text{x}}$	o
x	x	x
o		$\underline{\text{o}}$

(d) While in state s_1, which state-action pair(s) have the potential to result in the largest Q-value?

(e) Barney's friend Exo is really good at tic-tic-tic and Barney would like to train Tic-Bot to play exactly like Exo. Barney sits down and watches them play tic-tic-tic for a long time and observes their sequence of state-action pairs and their rewards. Which machine-learning problem formulation is most appropriate for Tic-Bot to learn from Exo? (Hint: not Q-learning.) Would you use all of their games for this learning purpose? Explain your answer.