MSc in Computer Science and Engineering

Planning, Learning and Decision Making

Homework 4. Reinforcement Learning

Consider a 2D continuous state $[0,1] \times [0,1]$ domain with two actions (action 0 is go up and action 1 is go right). The unit square is divided into 4 regions. Each region, numbered 0,1,2,3, can be described with 2 features. The following matrix summarizes the description of the features for each region (one column for region, let's represent f_i the column i of a matrix)

$$f = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

There is a discount γ. Assume that the Q values will be approximated using a linear combination of the indicated features as follows $Q(s,a) = f_s^T \theta_a$. Where θ_a is the column a of θ, and f_i is the column of f corresponding to the region i.

Exercise

(a) Considering initially $\theta = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, what is the Q function and the greedy policy?

(b) The agent interacted with the environment and obtained the following transitions:

- Region 0 $\xrightarrow{\text{up}}$ Region 1, cost 0.
- Region 1 $\xrightarrow{\text{right}}$ Region 3, cost 0.
- Region 0 $\xrightarrow{\text{right}}$ Region 2, cost 0.
- Region 2 $\xrightarrow{\text{up}}$ Region 3, cost 1.

Apply the fitted-Q approach and perform one update. What is the new θ vector?
The following result might be useful, if f and x are vectors, $\nabla_x (f^T x)^2 = 2 f f^T x$

(c) What is the new greedy policy?

(d) Why do states that were not visited have changed values?

Solution:

The action-value function is approximated as

$$Q(., a) = f^T \theta_a,$$

with the initial parameter matrix

$$\theta = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Initial Value Function and Policy

$$Q = \begin{bmatrix} 2 & 2 \\ 1 & 1 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$

The initial policy is uniform as all actions have the same value.

(b) One Fitted-Q Update

Assuming a discount factor γ, the target for each transition is given by:

$$y = \text{cost} + \gamma \min_{a'} Q(s', a').$$

Using the initial Q-values computed in (a) (recall: $Q = 2$ for regions 0 and 3 and $Q = 1$ for regions 1 and 2), we obtain:

$$\begin{aligned}
y_1 &= 0 + \gamma \min\{Q(1,0), Q(1,1)\} = 0 + \gamma \ = \gamma & \text{Region } 0 \xrightarrow{\text{up}} \text{Region } 1, \text{ cost } 0 \\
y_2 &= 0 + \gamma \min\{Q(3,0), Q(3,1)\} = 0 + 2\gamma \ = 2\gamma & \text{Region } 1 \xrightarrow{\text{right}} \text{Region } 3, \text{ cost } 0 \\
y_3 &= 0 + \gamma \min\{Q(2,0), Q(2,1)\} = 0 + 1\gamma \ = \gamma & \text{Region } 0 \xrightarrow{\text{right}} \text{Region } 2, \text{ cost } 0 \\
y_4 &= 1 + \gamma \min\{Q(3,0), Q(3,1)\} = 1 + 2\gamma\ , & \text{Region } 2 \xrightarrow{\text{up}} \text{Region } 3, \text{ cost } 1
\end{aligned}$$

We now perform separate least-squares regressions for the two actions.

Action 0 (transition 1 and 4)

$$L_0 = (\gamma - f_0^T \theta_0)^2 + (1 + 2\gamma - f_2^T \theta_0)^2$$

$$\begin{aligned}
\frac{dL_0}{d\theta_0} &= 0 \\
\gamma f_0 - f_0 f_0^T \theta_0 + f_2 + 2\gamma f_2 - f_2 f_2^T \theta_0 &= 0 \\
\gamma f_0 + f_2 + 2\gamma f_2 &= (f_0 f_0^T + f_2 f_2^T)\theta_0 \\
\begin{bmatrix} \gamma \\ 1 + 3\gamma \end{bmatrix} &= \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \theta_0 \\
\theta_0 &= \begin{bmatrix} -1 - \gamma \\ 1 + 2\gamma \end{bmatrix}
\end{aligned}$$

This can also be made explicitly without matrix notation, but thinking about implementing in the computer this way is more efficient.

Action 1 (transition 2 and 3)

$$L_1 = (2\gamma - f_1^T\theta_1)^2 + (\gamma - f_0^T\theta_1)^2$$

$$\begin{aligned}
\frac{dL_1}{d\theta_1} &= 0 \\
2\gamma f_1 - f_1 f_1^T\theta_1 + \gamma f_0 - f_0 f_0^T\theta_1 &= 0 \\
2\gamma f_1 + \gamma f_0 &= (f_1 f_1^T + f_0 f_0^T)\theta_1 \\
\begin{bmatrix} 3\gamma \\ \gamma \end{bmatrix} &= \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}\theta_1 \\
\theta_1 &= \begin{bmatrix} 2\gamma \\ -\gamma \end{bmatrix}
\end{aligned}$$

New Q values and policy

$$\begin{aligned}
Q = f^T\theta &= f^T \begin{bmatrix} -1-\gamma & 2\gamma \\ 1+2\gamma & -\gamma \end{bmatrix} \\
&= \begin{bmatrix} \gamma & \gamma \\ -1-\gamma & 2\gamma \\ 1+2\gamma & -\gamma \\ \gamma & \gamma \end{bmatrix}
\end{aligned}$$

(c) $\gamma \in [0, 1[$:

Looking at the Q function, and choosing the action minimizing the cost-to-go for each state we can determine the greedy policy, π^g. π^g is $0 \to$ up/right (any probability distribution over the actions is correct because the Q-values are equal), $1 \to$ up, $2 \to$ right, $3 \to$ up/right (any probability distribution over the actions is correct because the Q-values are equal).

(d) In a tabular Q-learning if a state is not visited then there is no update made. When doing function approximation, even when a state is not visited, the parameters change and so the values of all the states may change.