Artificial Intelligence I Take Home Exam

Can Zhou 19324118 zhouc@tcd.ie

Questions I chosen are Q2 & Q3

Question Q2

(a)

As there is only one sequence, so every action in the MDP is deterministic which means it is only one transition from one state to another.

Here is the possible transitions table:

State s	Action a	Near State s'	Probability p	Reward r	γ-discounted value Q(s,a)
s_0	a_1	s_1	$p(s_0, a_1, s_1)$	$r(s_0, a_1, s_1)$	$q_0(s, a)$
s_1	a_2	s_2	$p(s_1, a_2, s_2)$	$r(s_1, a_2, s_2)$	$q_1(s, a)$
s_2	a_3	s_3	$p(s_2, a_3, s_3)$	$r(s_2, a_3, s_3)$	$q_2(s, a)$
.....					
s_n	a_{n+1}	s_{n+1}	$p(s_n, a_{n+1}, s_{n+1})$	$r(s_n, a_{n+1}, s_{n+1})$	$q_n(s, a)$
.....					

where

$$q_0(s,a) = \sum_{s' \in S} p(s,a,s') \cdot r(s,a,s')$$

$$q_{n+1}(s,a) = \sum_{s' \in S} p(s,a,s') \cdot r(s,a,s') + \gamma \max_{a' \in A} q_n(s',a')$$

(b)

The probability in transitions is p(s, a, next(s,a)) = 1, and reward function can be represent by $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$. We can use this probability and reward function to calculate the γ-discounted value Q(s,a) and taken them together suggests every action in the MDP is deterministic, so, this may or may not be the optimal policy. Thus, it also suggests the confidence in this solution is low.

We can do the multiple sequences to explore more and iterate through them to find the one with the maximum cumulative reward which can increase or decrease out confidence in this suggestion.

For determining the functions p and r, we can use average return for all

samples. We need multiple samples of states, actions and rewards to learning. As this is the multiple sequences, so, it means it won't be deterministic. The probability will change, and we can keep track of how many times state and how many times state s' follows state follows state s when we take action a and update the transition probability P(s' | s, a) according to the relative frequencies. For reward function, we can update reward function by using reward function I mentioned above. The more sequences provided, better the chances of convergence to most optimal policy solution.

(c)

q1(s1, a2) = 5.7

Here is the computing process:

q0(s1, a1) = p(s1,a1,s1)r(s1,a1,s1) +p(s1,a1,s2)r(s1,a1,s2)

= 0.6 * 7 + 0.4 * 0 = 4.2

q0(s1, a2) = p(s1,a2,s1)r(s1,a2,s1) +p(s1,a2,s2)r(s1,a2,s2)

= 0.7 * 0 + 0.3 * 15 = 4.5

q0(s2, a1) = p(s2,a1,s1)r(s2,a1,s1) +p(s2,a1,s2)r(s2,a1,s2)

= 0.5 * 0 + 0.5 * 3 = 1.5

q0(s2, a2) = p(s2,a2,s1)r(s2,a2,s1) +p(s2,a2,s2)r(s2,a2,s2)

= 0.5 * 0 + 0.5 * 2 = 1

V0(s1) = max(q0(s1, a1), q0(s1, a2)) = 4.5

V0(s2) = max(q0(s2, a1), q0(s2, a2)) = 1.5

q1(s1, a2) = q0(s1,a2) + (1/3) * (p(s1,a2,s1)V0(s1) +p(s1,a2,s2)V0(s2))

= 4.5 + (1/3) * (0.7*4.5 + 0.3*1.5)

= 4.5 + 1.2

= 5.7

(d)

Because for a1 and a2, they have p(s,a,s') = 1 for some s', action a1 and a2 are s-deterministic.

As state s3 for every action, the p(s,a,s) = 1, so s3 is absorbing and can be calculate by:

$$Q(s,a) = r(s,a,s) + \gamma V(s)$$

$$v(s) = \frac{r_s}{1-\gamma}$$

Where $r_s = \max(s,a,s')$

So,

Q(s1,a1) = r(s1,a1,s) + $\gamma \frac{r_s}{1-\gamma}$ = 2 + 0.5 * 2 / 0.5 = 4

Q(s2,a1) = r(s2,a1,s) + $\gamma \frac{r_s}{1-\gamma}$ = 2 + 0.5 * 2 / 0.5 = 4

Q(s3,a1) = r(s3,a1,s) + $\gamma \frac{r_s}{1-\gamma}$ = 4 + 0.5 * 4 / 0.5 = 8

Q(s1,a2) = r(s1,a2,s) + $\gamma \frac{r_s}{1-\gamma}$ = 1 + 0.5 * 2 / 0.5 = 3

Q(s2,a2) = r(s2,a1,s) + $\gamma \frac{r_s}{1-\gamma}$ = 2 + 0.5 * 2 / 0.5 = 4

Q(s3,a2) = r(s3,a1,s) + $\gamma \frac{r_s}{1-\gamma}$ = 4 + 0.5 * 4 / 0.5 = 8

Question Q3

(a) True

Justify:

Learning equation suggests:

$$Q_{n+1}(s,a) = \alpha[r' + \gamma Q_n(s',a')] + (1-\alpha)Q_n(s,a)]$$

Which equals to

$$Q_{n+1}(s,a) = Q_n(s,a) + \alpha[r' + \gamma Q_n(s',a') - Q_n(s,a)]$$

Where α = learning rate

As agent explores more and more of the environment, the approximated Q value will convergence to max Q value and learning rate is useful for convergence to optimal policy. Exploration is to gather more information which allow us better decision in future (it can also be said exploration is try something new) and exploitation help us make best decision from what we know.

Since the equations above, when $\alpha = 1$, $Q_{n+1}(s,a) = \alpha[r' + \gamma Q_n(s',a')]$ which will only do exploration; when $\alpha = 0$, $Q_{n+1}(s,a) = Q_n(s,a)$ which will only do the exploitation. Thus, the exploration-exploitation trade-off in Q-learning depends on the learning rate.

(b) False

Justify:

We can use logical consequence bottom-up approach.

A mechanical procedure for logical consequence which is complete

if $KB \mid - \ g$ whenever $KB \mid = \ g$

To find a mechanical procedure which is goal-directed and is complete we can assume a KB like :

i :- p,q.

i :- r.

p.

r.

with those Prolog code:

```
KB = [[i,p,q],[i,r],[p],[r]]

arc([H|T],N,KB) :- member([H|B],KB),append(B,T,N).
```

instead of using prove([G], KB) which is top-down approach, we can just use logical consequence bottom-up approach which is:

$$C_1 = \{p, r\}$$

$$C_2 = \{p, r, i\} = C_n \text{ for } n \geq 2$$

This mechanical procedure for logical consequence *i* is bound to be complete and is goal-directed.

(c) True

Justify:

Abduction is trying to use the outputs to figure out what the input might be and use the outputs to explain the inputs; Deduction is going from input to output via a function which consists of deriving conclusions from the given evidence.

So, abduction is a form of deduction inversed.

(d) False

Justify:

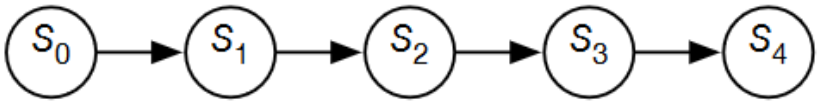
Causal inference is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect. The main difference between causal inference and inference of association is that the former analyses the response of the effect variable when the cause is changed.

Inferring the cause of something has been described as reasoning to the conclusion that something is, or is likely to be, the cause of something else. Identification of the cause or causes of a phenomenon, by establishing covariation of cause and effect, a time-order relationship with the cause preceding the effect, and the elimination of plausible alternative causes. Determination of cause and effect from joint observational data for two time-independent variables, say X and Y, has been tackled using asymmetry between evidence for some model in the directions, $X \rightarrow Y$ and $Y \rightarrow X$.

Although Bayesian networks are often used to represent causal relationships, this need not be the case: a directed edge from u to v does not require that node v be causally dependent on node u. This is demonstrated by the fact that Bayesian networks on the graphs:

a--->b--->c and a--->b--->c are equivalent; that is, they impose exactly the same conditional independence requirements.

More example from lecture note:



Reference 1:from lecture note

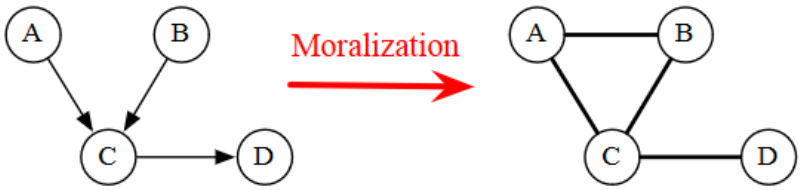
In this belief network, random variable are ordered, we can say s3 is depending on s3, but we cannot say s3 is not depending on s4 which may have a possibility.

(e) True

Justify:

When do the moralization, we insert an undirected edge between every pair of nodes that have a child in common and replace all the directed edges with undirected edges with undirected edges.

So as same as the figure from lecture note below, conditional independences in a naive Bayes classifier are lost when its Bayes net is moralized to a Markov net (independence information $A \perp\!\!\!\perp B$ lost from A->C<-B).



Reference 2: from lecture note