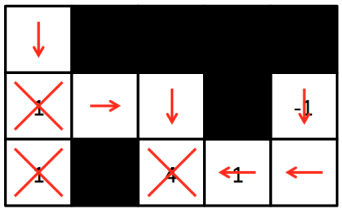
CS 188 Summer 2024 Introduction to Artificial Intelligence

Exam Prep 9 Solutions

Q1. MDPs: Reward Shaping

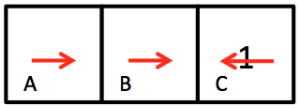
PacBot is in a Gridworld-like environment E. It moves deterministically Up, Down, Right, or Left, or at any time it can exit to a terminal state (where it remains). If PacBot is on a square with a number written on it, it receives a reward of that size **on Exiting**, and it receives a reward of 0 for Exiting on a blank square. Note that when it is on any of the squares (including numbered squares), it can either move Up, Down, Right, Left or Exit. However, it only receives a non-zero reward when it Exits on a numbered square.

(a) Draw an arrow in **each** square (including numbered squares) in the following board to indicate the optimal policy PacBot will calculate with the discount factor $\gamma = 0.5$. (For example, if PacBot would move Down from the square in the middle, draw a down arrow in that square.) If PacBot's policy would be to exit from a particular square, draw an X in that square.



In order to speed up computation, Pacbot computes its optimal policy in a new environment E' with a different reward function $R'(s, a, s')$. If $R(s, a, s')$ is the reward function in the original environment E, then $R'(s, a, s') = R(s, a, s') + F(s, a, s')$ is the reward function in the new environment E', where $F(s, a, s') \in \mathbb{R}$ is an added "artificial" reward. If the artificial rewards are defined carefully, PacBot's policy will converge in fewer iterations in this new environment E'.

(b) To decouple from the previous question's board configuration, let us consider that Pacbot is operating in the world shown below. Pacbot uses a function F defined so that $F(s, a, s') = 10$ if s' is closer to C relative to s, and $F(s, a, s') = 0$ otherwise (consider C to be closer to C than B or A). Let us also assume that the action space is now restricted to be between Right, Left, and Exit only.



Either left or right from B is correct.

In the diagram above, indicate by drawing an arrow or an X in each square, as in part (a), the optimal policy that PacBot will compute in the new environment E' using $\gamma = 0.5$ and the modified reward function $R'(s, a, s')$.

(c) PacBot's utility comes from the discounted sum of rewards **in the original environment**. What is PacBot's expected utility of following the policy computed above, starting in state A if $\gamma = 0.5$? 0

(d) Find a non-zero value for x in the table showing $F(s, a, s')$ drawn below, such that PacBot is guaranteed to compute an optimal policy that maximizes its expected true utility for **any** discount factor $\gamma \in [0, 1)$.

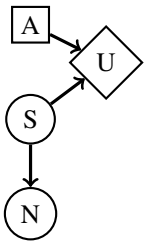
	Value
$F(A, \text{Right}, B)$	10
$F(B, \text{Left}, A)$	x
$F(B, \text{Right}, C)$	10
$F(C, \text{Left}, B)$	x

Any number less than -10 will also work. No other solution is correct.

Q2. Dressed to Impress

(a) Alice is invited to a party tonight which is said to be once-in-a-lifetime. However, this mysterious party doesn't publicize who is going and thus Alice has no idea whether the size S of the party will be be large ($S = +s$) or tiny ($S = -s$). The size can affect the noise N outside the party, and it will either be noisy ($N = +n$) or quiet ($N = -n$). Alice has three dresses: blue, red and yellow. Each dress will have a different utility for her depending on the size of the party. Let's help her decide which will be best!

We have the following decision network, where circles are chance nodes, squares are decision nodes, and diamonds are utility nodes:



S	$P(S)$
+s	0.5
-s	0.5

S	N	$P(N \mid S)$
+s	+n	0.7
+s	-n	0.3
-s	+n	0.1
-s	-n	0.9

S	A	U
+s	blue	80
-s	blue	60
+s	red	40
-s	red	100
+s	yellow	60
-s	yellow	40

(i) What is the expected utility of wearing each dress, with both S and N unobserved?

- EU(A=blue) = $80 * 0.5 + 60 * 0.5 = 70$
- EU(A=red) = $40 * 0.5 + 100 * 0.5 = 70$
- EU(A=yellow) = $60 * 0.5 + 40 * 0.5 = 50$

What is Alice's maximum expected utility?

- MEU({}) = 70

(ii) Suppose Alice observes that the party is quiet, $N = -n$. Compute the following conditional probabilities with this observation:

- $P(+s| -n)$ = $\frac{P(-n|+s)P(+s)}{P(-n|+s)P(+s)+P(-n|-s)P(-s)} = \frac{0.3 \cdot 0.5}{0.3 \cdot 0.5 + 0.9 \cdot 0.5} = 0.25$
- $P(-s| -n)$ = $P(-s| -n) = 1 - P(+s| -n) = 0.75$

What is the expected utility of wearing each dress?

- EU(A=blue | $N = -n$) = $80 * 0.25 + 60 * 0.75 = \frac{145}{2} = 65$
- EU(A=red | $N = -n$) = $40 * 0.25 + 100 * 0.75 = \frac{155}{2} = 85$
- EU(A=yellow | $N = -n$) = $60 * 0.25 + 40 * 0.75 = 45$

What is Alice's maximum expected utility given that $N = -n$?

- MEU({N=-n}) = 85

(iii) Construct a formula for $VPI(N)$ for the given network. To decouple this problem from your work above, use any of the symbolic terms from the following list (rather than plugging in numeric values):

$P(+n| + s)$, $P(+n| - s)$, $P(-n| + s)$, $P(-n| - s)$, $P(+n)$, $P(-n)$, $P(+s)$, $P(-s)$,
$MEU(\{\})$, $MEU(\{N = +n\})$, $MEU(\{N = -n\})$

- $VPI(N)$ = $P(+n)MEU(\{N = +n\}) + P(-n)MEU(\{N = -n\}) - MEU(\{\})$