学院 专业 班级 学号 姓名

密封线 学生须将答案写在此线以下

鲁东大学 2022—2023 学年第一学期

2020 级 电子信息工程 专业 本 科卷 A

课程名称 人工智能基础

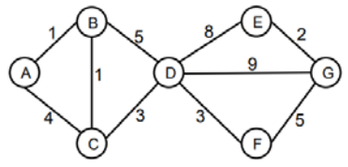
课程号（2220184220）考试形式（ 闭卷 ） 时间（ 120 分钟 ）

题目	一	二	三	四	五	总分	统分人	复核人
得分								

得分	评卷人

一、搜索（25 分）

（1）如下的状态空间图中，A 是开始状态，G 是目标状态。每条边的成本如右图所示。每个边都是双向的。注意，启发函数 h_1 是一致的(consistent)，但 h_2 是不一致的。



Node	h_1	h_2
A	9.5	10
B	9	12
C	8	10
D	7	8
E	1.5	1
F	4	4.5
G	0	0

1）对于以下每个图搜索(graph search)策略（注意：不是树搜索 tree search），判断它是否能够返回表中所列出的 3 个路径，可以画√，不可以画×。注意，当 fringe 中有相同代价的结点时，可选择任意一个进行扩展。（10 分）

Search Algorithm	A-B-D-G	A-C-D-G	A-B-C-D-G	A-B-C-D-F-G
Depth first search				
Breadth first search				
Uniform cost search				
A* search with heuristic h_1				
A* search with heuristic h_2				

2）完成如下所示的新启发函数 h_3。除 h_3(B)外，所有值均为固定值。

Node	A	B	C	D	E	F	G
h_3	10	?	9.5	7	1.5	4.5	0

对于以下每种情况，写入 h_3 (B)可能的值的集合。例如，h_3(B)可以为所有非负数，请写入[0, ∞], 要表示空集，请写入∅。

①h_3(B)取哪些值使启发函数 h_3 是可接受的（admissible）？（2 分）

②h_3(B)取哪些值使启发函数 h_3 是一致的（consistent）？（2 分）

③h_3(B)取哪些值将使 A*图搜索依次展开结点 A、结点 C、结点 B 和结点 D？（2 分）

（2）考虑同时控制 n 个 pacman 的问题。在一个连通的迷宫（maze）中有 n 个 pacman，每个 pacman 可独立行动。多个 pacman 可以同时停留在同一个方格上。在每个时间步，每个 pacman 可以不动，也可垂直或水平移动一个方格。游戏的目标是让所有的 pacman 在最小的时间步数内汇集到同一个方格上。使用以下符号：M 表示迷宫中非墙壁的方格数（即 pacman 可以到达的方格数）；n 表示 pacman 的数量；$p_i=(x_i, y_i), i=1\ldots n$，表示第 i 个 pacman 的位置。

1）如何表示这个问题的一个状态（state）？状态空间（state space）有多大？（2 分）

2）使用 UCS 搜索策略构造搜索树时，该树的结点数的上界（n 和 M 的函数）？（2 分）

3）判断下面的启发函数的可接受性，并简单说明。（5 分）

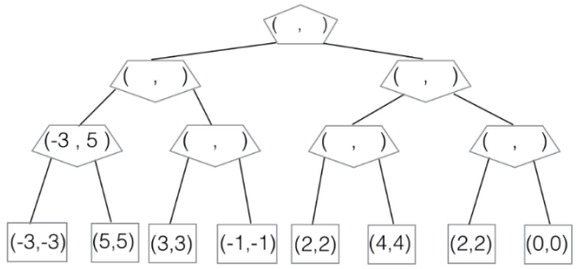
$$h_1(p_1,\ldots,p_n)=\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbf{1}[p_i\neq p_j] \quad \text{where} \quad \mathbf{1}[p_i\neq p_j]=\begin{cases}1 & \text{if } p_i\neq p_j\\ 0 & \text{otherwise}\end{cases}$$

得分	评卷人

二、博弈（15 分）

本游戏中，有一个 Pacman，两个 ghost。ghost 总是会选择让 Pacman 的获益最小的动作。Pacman 有且只有一次机会使用“超能力”，这种“超能力”能让 ghost 选择 Pacman 想要的任何动作。

（1）与 minimax 算法类似，搜索数中每个节点的值由以它为根所展开的子树决定。为了记录节点的值，为每个节点定义一个元组（u,v）：如果“超能力”未在该子树中使用，其值为 u；如果在子树中使用了一次“超能力”，则其值为 v。对于终端状态，其值已经给出。在下面搜索树中填写（u,v）值。Pacman 是根节点，有两个 ghost 依次行动。（6 分）



（2）完成以下算法，以适应上面所描述的Pacman有一次机会使用“超能力”的情况。注意：基于min和max函数实现，参考Max-Value函数中的使用方法。（9分）

```
function VALUE(state)
    if state is leaf then
        u ← UTILITY(state)
        v ← UTILITY(state)
        return (u, v)
    end if
    if state is Max-Node then
        return MAX-VALUE(state)
    else
        return MIN-VALUE(state)
    end if
end function

function MAX-VALUE(state)
    uList ← [ ], vList ← [ ]
    for successor in SUCCESSORS(state) do
        (u′, v′) ← VALUE(successor)
        uList.append(u′)
        vList.append(v′)
    end for
    u ← max(uList)
    v ← max(vList)
    return (u, v)
end function
```

```
function MIN-VALUE(state)
    uList ← [ ], vList ← [ ]
    for successor in SUCCESSORS(state) do
        (u′, v′) ← VALUE(successor)
        uList.append(u′)
        vList.append(v′)
    end for

    u ← _____

    v ← _____

    return (u, v)
end function
```

得分	评卷人

三、效用函数（20 分）

Ghost-King 和 Pacman 的效用函数（utility function）分别用 U_{GK} 和 U_P 表示。两者输出的值都>=0(非负值)。

（1）以下 U_{GK} 和 U_P 之间的哪种关系与 Ghost-King 的观察结果一致：他和 Pacman 在所有可能的事件结果（outcome）中都有相同的偏好顺序，但是他和 Pacman 并非对所有的 lotteries 都有相同的偏好。在相应的关系上画勾，并做出解释。（10 分）

- ☐ $U_P = aU_{GK} + b \quad (0 < a < 1, b > 0)$
- ☐ $U_P = aU_{GK} + b \quad (a > 1, b > 0)$
- ☐ $U_P = U_{GK}^2$
- ☐ $U_P = \sqrt{(U_{GK})}$

（2）Ghost-King 还观察到 Pacman 比他更冒险。以下 U_{GK} 和 U_P 之间的哪种关系符合 Ghost-King 的观察。在相应的关系上画勾，并做出解释。（10 分）

- ☐ $U_P = aU_{GK} + b \quad (0 < a < 1, b > 0)$
- ☐ $U_P = aU_{GK} + b \quad (a > 1, b > 0)$
- ☐ $U_P = U_{GK}^2$
- ☐ $U_P = \sqrt{(U_{GK})}$

得分	评卷人

四、马尔科夫决策过程（15 分）

有一个掷骰子的游戏。在游戏的每一轮，游戏者有 2 个动作（action）可以选择：1）*Stop*：停止玩游戏，获得骰子点数所对应的钱作为 reward；2）*Roll*：花费 1 块钱，掷骰子，骰子以相同的概率掷出 1~6 点。游戏者从 *Start* 状态开始，至少要花 1 块钱掷一次骰子。状态 S_i 表示骰子掷出了 i 点。只要游戏者愿意每次支付 1 元钱掷骰子，游戏可以一直继续。一旦游戏者采用 *Stop* 动作，以当前骰子的点数作为 reward，游戏结束，切换到 *End* 状态。游戏者使用 MDP 方法对这个游戏进行分析。

（1）给定一个下表所示的 policy π ，设 γ=1，求出 $V^{\pi}(s)$，并给出简单的计算过程。（6 分）

State	s_1	s_2	s_3	s_4	s_5	s_6
$\pi(s)$	*Roll*	*Roll*	*Stop*	*Stop*	*Stop*	*Stop*
$V^{\pi}(s)$						

（2）基于上面求出的 *V* 值，执行一次 policy update，以找到更优的 policy π’ 。下表给出了旧的 policy π，以及新 policy π’的一部分。如果对于某个状态，*Roll* 和 *Stop* 都可以，则填写 *Roll* / *Stop*。设 γ=1。在表格下面给出简单的计算过程。（4 分）

State	s_1	s_2	s_3	s_4	s_5	s_6
$\pi(s)$	*Roll*	*Roll*	*Stop*	*Stop*	*Stop*	*Stop*
$\pi'(s)$	*Roll*					*Stop*

（3）前面给出的旧 policy π 是否是最优的？请给出理由（5 分）

得分	评卷人

五、强化学习（25 分）

有一个游戏，知道游戏只有两个状态{A，B}，每个状态下，agent 有两个动作（action）可以选择{Up，Down}。

一个 agent 根据某个 policy π 选择动作，生成了一系列的状态变化和收益（reward），如下所示。本题中一直设 discount factor γ=0.5，learning rate α=0.5，除非特别指定。

t	s_t	a_t	s_{t+1}	r_t
0	A	Down	B	4
1	B	Down	B	-4
2	B	Up	B	0
3	B	Up	A	3
4	A	Up	A	-1

（1）已知 Q-learning 的更新函数为：

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a'))$$

假设所有的 Q-value 被初始化成 0。请使用上表中的经验序列进行 Q-learning，并给出下列 2 个 Q-value 及简单的计算过程。（4 分）

$Q(\mathrm{A}, \mathrm{Down}) =$ _____， $Q(\mathrm{B}, \mathrm{Up}) =$ _____

（2）在 model-based 的强化学习中，需要先估计 transition function *T*(*s*,*a*,*s*')和 reward function *R*(*s*,*a*,*s*')。请根据上表中的经验序列计算下面的 *T* 和 *R* 。如果未知的话，写“n/a”。写出简单的计算过程。（8 分）

$\hat{T}(\mathrm{A}, \mathrm{Up}, \mathrm{A}) =$ _____， $\hat{T}(\mathrm{A}, \mathrm{Up}, \mathrm{B}) =$ _____， $\hat{T}(\mathrm{B}, \mathrm{Up}, \mathrm{A}) =$ _____， $\hat{T}(\mathrm{B}, \mathrm{Up}, \mathrm{B}) =$ _____

$\hat{R}(\mathrm{A}, \mathrm{Up}, \mathrm{A}) =$ _____， $\hat{R}(\mathrm{A}, \mathrm{Up}, \mathrm{B}) =$ _____， $\hat{R}(\mathrm{B}, \mathrm{Up}, \mathrm{A}) =$ _____， $\hat{R}(\mathrm{B}, \mathrm{Up}, \mathrm{B}) =$ _____

（3）agent 又生成了一个新的经验序列，并且根据这次的经验，对 *T* 和 *R* 形成了如下的估计：

s	a	s'	$\hat{T}(s,a,s')$	$\hat{R}(s,a,s')$
A	Up	A	1	10
A	Down	A	0.5	2
A	Down	B	0.5	2
B	Up	A	1	-5
B	Down	B	1	8

(i)根据上面的$\hat{T}$和$\hat{R}$，请给出最优的策略$\hat{\pi}^*(s)$和$\hat{V}^*(s)$。写出计算过程。提示：给定|*x*|<1, $1+x+x^2+x^3+x^4+\ldots=1/(1-x)$ （8分）

$\hat{\pi}^*(A) =$ _____， $\hat{\pi}^*(B) =$ _____， $\hat{V}^*(A) =$ _____， $\hat{V}^*(B) =$ _____

(ii)如果把这个经验序列重复的送入Q-learning算法，最终values会收敛到什么值(假设α被仔细的调整以保证收敛)？请给出理由（5分）

◯ 上面所求得的$\hat{V}^*$
◯ 最优值V^*
◯ 既不是$\hat{V}^*$也不是V^*
◯ 无法判断