ML_强化学习

填空题

1，最优动作价值函数Q_*依赖于_____。

2，DQN是对_____的近似。

3，驾车按照“甲，乙，丙”行驶，从甲地出发，模型预计需要行驶20小时，实际行驶6小时到达乙地，模型预计还需12个小时到达丙地，如果我们用TD算法更新模型，那么TD目标$\hat{y}$ =_____小时，TD绝对误差值$|\delta|$_____小时；

选择题

1，设$A = \{上，下，左，右\}$为动作空间，s_t为当前状态，π为策略函数，策略函数的输出：

$$\begin{aligned}\pi(上|s_t) &= 0.2,\\ \pi(下|s_t) &= 0.05,\\ \pi(左|s_t) &= 0.7,\\ \pi(右|s_t) &= 0.15.\end{aligned} \tag{1}$$

请问，哪个动作会成为a_t?

A，下

B，左

C，4种动作都有可能

2，设随机变量U_t为t时刻的回报，请问U_t依赖于哪些变量?

A，t时刻的状态S_t

B，t时刻的动作A_t

C，S_t和A_t

D，$S_t, S_{t+1}, S_{t+2}, \cdots$和$A_t, A_{t+1}, A_{t+2}, \cdots$

3，动作价值函数是什么的期望?

A，奖励

B，回报

C，状态

D，动作

4，设$A = \{上，下，左，右\}$为动作空间，s_t为当前状态，Q_*为最优动作价值函数，策略函数的输出：

$$
\begin{aligned}
Q_*(s_t, \text{上}) &= 930, \\
Q_*(s_t, \text{下}) &= -60, \\
Q_*(s_t, \text{左}) &= 120, \\
Q_*(s_t, \text{右}) &= 321.
\end{aligned} \tag{2}
$$

请问，哪个动作会成为a_t?

A，上

B，下

C，4种动作都有可能

5，DQN的输出层用于什么激活函数?

A，不需要激活函数，因为Q值可正可负，没有取值范围

B，用sigmoid激活函数，因为Q值介于0和1之间

C，用ReLU激活函数，因为Q值非负

D，用softmax激活函数，因为DQN的输出是一个概率分布

6，多臂赌博机是单步强化学习的经典范例，ε贪心算法和softmax算法用于处理什么问题?

A， 探索-利用问题

B， 奖励函数优化问题

C， 动作选择问题

D， 状态空间问题

7，DQN（深度 Q 网络）是基于什么的强化学习方法?

A， 基于值的方法

B， 基于策略的方法

C， 基于模型的方法

D， 基于探索的方法

8，TD gradient 是与哪种强化学习方法相关的概念?

A， 基于值的方法

B， 基于策略的方法

C， 基于模型的方法

D， 基于探索的方法

9，在强化学习中，基于策略的方法主要关注什么?

A， 最大化奖励

B， 最小化损失

C， 直接学习值函数

D， 直接学习策略函数

答案

填空题

1. **最优策略**
2. **Q-learning**
3. **18小时**；**2小时**；

选择题

1. C，4种动作都有可能。
2. D；
3. B；
4. A；
5. A；
6. B；
7. A；
8. A；
9. D；