1. 什么是强化学习?

强化学习由环境、动作和奖励组成，强化学习的目标是使得作出的一系列决策得到的总的奖励的期望最大化。

2. 请你讲一下，HMM隐马尔可夫模型的参数估计方法是什么?

学习算法：

- 若训练数据包括观测序列和状态序列，则HMM的学习非常简单，是监督学习；
- 若训练数据只有观测序列，则HMM的学习需要使用EM算法，是非监督学习。

大数定理：

假定已给定训练数据包含S个长度相同的观测序列和对应的状态序列 `{(O1，I1)，(O2,I2)…(OS,IS)}`,那么，可以直接利用伯努利大数定理的结论：频率的极限是概率，从而给出HMM的参数估计。

隐马尔可夫模型
计算语言学之隐马尔可夫模型

3. 强化学习和监督学习、无监督学习的区别是什么?

监督学习带有标签；无监督学习没有标签；强化学习使用未标记的数据，根据延迟奖励学习策略。

4. 强化学习适合解决什么样子的问题?

模型输出的动作必须要能够改变环境的状态，并且模型能够获得环境的反馈，同时状态应该是可重复到达的。

5. 强化学习的损失函数是什么？和深度学习的损失函数有何关系?

强化学习的损失函数是使奖励和的期望最大；深度学习中的损失函数是使预测值和标签之间的差异最小化。

6. POMDP是什么？马尔科夫过程是什么？马尔科夫决策过程是什么？里面的“马尔科夫”体现了什么性质?

POMDP是状态部分可观测的马尔可夫决策过程；马尔科夫过程是一个二元组 $< S, P >$，S为状态集合，P为状态转移概率矩阵；马尔科夫决策过程是一个五元组$< S, P, A, R, \gamma >$，R表示为从S到S'能够获得的奖励期望，γ为折扣因子，A为动作集合；马尔可夫中下一个状态只与当前状态有关，而与历史状态无关，即$P[S_{t+1} \mid S_t] = P[S_{t+1} \mid S1, S2, ..., St]$

7. 值迭代和策略迭代的区别?

价值迭代采用了Bellman最优算子，策略迭代采用的是Bellman期望算子。价值迭代是策略迭代的一种特殊情况，是每进行一次策略评估就更新一次策略。

强化学习–值函数
Policy gradient 算法思想

8. 贝尔曼方程的具体数学表达式是什么

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r(s, a) + \gamma v_\pi(s')]$$

$$q_\pi(s, a) = \sum_{s',r} p(s', r \mid s, a)[r(s', a) + \gamma v_\pi(s')]$$

强化学习中马尔科夫决策过程和贝尔曼方程

9. 最优值函数和最优策略为什么等价?

$$1 \quad \text{if } a = \arg\max_{a \in A} q^*(s, a)$$
$$0 \quad \text{otherwise}$$

$$1 \text{ if } a=\arg\max a \in A q*(s,a) 0 \text{ otherwise}$$

$$\text{\right.}\pi^*(a \mid s) = \begin{cases} 1 & \text{if } a = \arg\max_{a \in A} q^*(s, a) \\ 0 & \text{otherwise} \end{cases}$$

10. 如果不满足马尔科夫性怎么办？当前时刻的状态和它之前很多很多个状态都没有关系?

如果不满足马尔科夫性，强行只用当前的状态来决策，势必导致决策的片面性，得到不好的策略。 为了解决这个问题，可以利用RNN对历史信息建模，获得包含历史信息的状态表征。表征过程可以 使用注意力机制等手段。最后在表征状态空间求解MDP问题。

11. 求解马尔科夫决策过程都有哪些方法？有模型用什么方法？动态规划是怎么回事?

求解MDP可以直接求解Bellman方程，但是通常Bellman方程难以列出且计算复杂度高，除此以外还可以用DP(Dynamic Programming，动态规划)，MC(Monte Carlo，蒙特卡罗)，TD(Temporal Difference，时间差分)算法求解。有模型时可以使用DP算法。

12. 简述动态规划(DP)算法

DP算法是在给定MDP环境特性的条件下用来计算最优策略的，是基于模型的planning方法。动态规划有策略迭代和价值迭代两种方式，策略迭代是不断进行策略评估、策略改进的过程。每一次操作都是基于所有可能的单步后继状态价值函数以及它们出现的概率，以此来更新一个状态价值函数，它是广度期望更新的并且采用了自举法。

12. 简述蒙特卡罗估计值函数(MC)算法

MC算法是model-free的学习方法而不是planning，它从”经验“中学习价值函数和最优策略，”经验“是指多幕采样数据，MC通过平均样本的回报在大数定律的保证下进行策略估计，然后采用柔性策略进行MC控制。MC算法是深度采样更新，它没有使用自举法。

13. 简述时间差分(TD)算法

TD算法和MC算法一样可以从和环境互动的经验中学习策略而不依赖环境的动态特性，TD和DP一样都采用的自举法，是采样更新。和MC不同的是TD算法采样深度没有那么深，它不是一个完全的采样，TD的策略评估是根据它直接得到的后继状态节点的单次样本转移来更新的，换言之它不需要等到一幕完全结束而是可以立刻进行学习。它采用后继状态的价值和沿途的收益进行更新，TD控制有Sarsa、期望Sarsa和Q学习。

14. 动态规划、蒙特卡洛和时间差分的异同

共同点：

核心都是价值函数的计算，所有方法都是基于对未来事件的展望来计算一个回溯值。

不同点：

1. DP算法是model-based，MC和TD都是model-free
2. DP是期望更新，MC和TD是采样更新
3. DP是planning，MC和TD是Learning
4. DP显示了所有的单步转移，MC是完整的一幕，TD采样不采集完整的一幕
5. MC是最小化训练集上均方误差的估计，批量TD是找出完全符合马尔可夫模型的最大似然参数

15. MC和TD分别是无偏估计吗?

MC是无偏估计，TD是有偏估计。

16. MC、TD谁的方差大，为什么?

MC的方差更大，MC采样了一整幕，每次获取下一步的价值和收益都会增大方差，但是TD不是完全采样因此方差比MC小。

17. 简述on-policy和off-policy的区别

在线策略用于学习和用于采样的是同一个策略，离线策略中行动策略用来采样，目标策略是用来学习的。在线策略不学习最优动作而是学习一个接近最优动作同时又能继续探索的动作，离线策略直接学习最优动作。

18. 简述on-policy和off-policy的区别

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right]$$

Q学习采用对最优动作价值函数的近似作为学习目标，与行动策略无关，是off-policy。

19. 写出用第n步的值函数更新当前值函数的公式（1-step，2-step，n-step的意思）。当n的取值变大时，期望和方差分别变大、变小?

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[\sum_{i=1}^{n} \gamma^{i-1} R_{t+i} + \gamma^n \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right]$$

n越大时方差变大，期望偏差变小。

20. TD(λ)方法：当λ=0时实际上与哪种方法等价，λ=1呢?

当λ=0时，等价于TD。
当λ=1时，等价于MC。

21. value-based和policy-based的区别是什么?

value-based 的典型算法是DQN，policy-based是policy gradient，结合这两种具体算法可能会更好的理解。

1. 处理的action space不同：value-based适合处理的action space低维离散的，policy-based适合处理连续的action space。
2. 针对action的价值输出不同：value-based计算出每个action的价值，policy-based一般情况下只给出较价值较高的actions。
3. 更新频率不同：value-based每个action执行都可以更新，policy-based 每个episode完成之后才能更新一次。

22. 阐述目标网络和experience replay的作用?

目标网络的参数θ−，每隔C步才和普通网络的参数θ同步一次。 `Experience Replay` 将系统探索环境得到的数据储存起来，然后随机采样样本更新深度神经网络的参数。主要作用是克服经验数据的相关性和非平稳分布问题。它的做法是从以往的状态转移中随机采样进行训练。优点：1. 数据利用率高，因为一个样本被多次使用。2. 连续样本的相关性会使参数更新的方差比较大，该机制可减少这种相关性。

23. 描述随机策略和确定性策略的特点?

随机策略 $\pi_\theta(a_t \mid s_t) = P[a \mid s, \theta]$，P是一个概率函数；
确定性策略$a = \mu_\theta(s)$，μ是一个确定的函数映射。