强化学习习题

浙江大学

赵洲

任务与奖赏题目

- 在强化学习中，机器能直接控制环境中状态的转移和奖赏的返回。

任务与奖赏题目

- （错误）
- 原因：**在强化学习中，环境中状态的转移和奖赏的返回是不受机器控制的，机器只能通过选择动作来影响环境，并通过观察转移后的状态和返回的奖赏来感知环境**。例如在种西瓜任务中，西瓜的生长状态变化（状态转移）以及最终收获西瓜的好坏对应的奖赏，都是由自然环境决定的，机器无法直接操控。
- 定义：**强化学习**是一种机器学习方法，其中智能体（机器）在环境中采取一系列动作，环境根据智能体的动作反馈相应的状态和奖赏，智能体的目标是学习一个最优策略，以最大化长期累积奖赏。智能体通过感知环境的状态，选择动作，然后环境根据潜在的转移函数和奖赏函数进行状态转移并给予智能体奖赏，**智能体只能观察和适应环境的变化，而不能直接决定环境的状态转移和奖赏返回。**

K –摇臂赌博机题目

- 在 K –摇臂赌博机问题中，“仅探索”法能保证每次都选择到期望奖赏最大的摇臂。

K –摇臂赌博机题目

- 在 K –摇臂赌博机问题中，“仅探索”法能保证每次都选择到期望奖赏最大的摇臂。（错误）

K –摇臂赌博机题目

- 原因：“仅探索”法是将所有尝试机会平均分配给每个摇臂，其目的是估计每个摇臂的奖赏期望，但在这个过程中会失去很多选择当前最优摇臂（即期望奖赏最大的摇臂）的机会，因为它没有优先选择那些在探索过程中表现较好的摇臂，而是平均对待每个摇臂，所以不能保证每次都能选择到期望奖赏最大的摇臂。
- 定义：K –摇臂赌博机是强化学习中的一个理论模型，它有 K 个摇臂，赌徒每次投入硬币后可选择按下其中一个摇臂，每个摇臂以一定概率吐出硬币，但赌徒事先不知道这个概率。赌徒的目标是通过一定策略最大化自己获得的奖赏（硬币数量）。

有模型学习题目

- 在有模型学习中，策略迭代算法在每次改进策略后不需要重新进行策略评估。

有模型学习题目

- 在有模型学习中，策略迭代算法在每次改进策略后不需要重新进行策略评估。（错误）

有模型学习题目

- 原因：策略迭代算法包括策略评估和策略改进两个步骤。在每次改进策略后，由于策略发生了变化，之前对策略的值函数评估就不再适用，需要重新进行策略评估，以确定新策略的累积奖赏情况，从而判断是否达到最优策略或者继续进行改进。如果不重新评估，就无法知道新策略的优劣，也无法继续朝着最优策略的方向迭代。
- 定义：有模型学习是指在强化学习任务中，假设机器已经对环境进行了建模，即已知马尔可夫决策过程四元组（状态空间、动作空间、转移函数、奖赏函数），在这样的已知模型环境中进行学习的方法。策略迭代算法是有模型学习中的一种求解最优策略的方法，它从一个初始策略出发，先进行策略评估（计算该策略下每个状态的累积奖赏），然后根据评估结果改进策略，不断重复这两个步骤，直到策略收敛。

免模型学习题目

- 蒙特卡罗强化学习方法只能用于估计状态值函数，不能用于估计状态 –动作值函数。（错误）

免模型学习题目

- 蒙特卡罗强化学习方法只能用于估计状态值函数，不能用于估计状态 –动作值函数。

免模型学习题目

- 原因：在免模型情形下，由于模型未知导致策略评估困难，蒙特卡罗强化学习通过多次采样求取平均累积奖赏来近似期望累积奖赏。虽然最初策略迭代算法估计的是状态值函数，但在模型未知时，蒙特卡罗方法将估计对象转变为状态 –动作值函数，因为最终的策略是通过状态 –动作值函数来获得的。它通过从起始状态出发，使用策略采样得到轨迹，记录轨迹中状态 –动作对的累积奖赏采样值，多次采样后求平均来估计状态 –动作值函数。
- 定义：免模型学习是指学习算法不依赖于环境建模的强化学习方法，因为在现实中环境的转移概率、奖赏函数往往难以得知。蒙特卡罗强化学习是免模型学习中的一种方法，它通过多次采样来近似期望累积奖赏，适用于 T 步累积奖赏的强化学习任务，通过在环境中执行动作并观察结果，利用采样轨迹来估计值函数。

值函数近似题目

- 值函数近似方法只能用于连续状态空间的值函数学习，不能用于有限状态空间。

值函数近似题目

- 值函数近似方法只能用于连续状态空间的值函数学习，不能用于有限状态空间。 （错误）

值函数近似题目

- 原因：虽然值函数近似方法主要是为了解决连续状态空间无法用表格值函数记录状态值的问题，但在某些情况下，对于有限状态空间，如果状态数量非常大，直接使用表格值函数可能会面临存储和计算效率的问题，此时也可以考虑使用值函数近似方法来提高效率和泛化能力。例如，当状态空间维度很高时，即使是有限状态，使用表格值函数也可能不现实，值函数近似可以通过参数化的方式来表示值函数，减少存储空间和计算量。
- 定义：值函数近似是针对现实强化学习任务中状态空间可能连续或状态数量巨大而无法用表格值函数记录状态值的情况，直接对连续状态空间（或复杂有限状态空间）的值函数进行学习的方法。它通过假设值函数的某种形式（如线性函数），利用样本数据学习参数，使得学得的值函数尽可能近似真实值函数，常用最小二乘误差度量近似程度，并通过梯度下降等方法优化参数。

模仿学习题目

- 直接模仿学习中，从人类专家决策轨迹数据中学得的策略模型可以直接作为最优策略使用，无需再进行改进。

模仿学习题目

- 直接模仿学习中，从人类专家决策轨迹数据中学得的策略模型可以直接作为最优策略使用，无需再进行改进。（错误）

模仿学习题目

- 原因：直接模仿学习中，利用人类专家决策轨迹数据通过监督学习学得的策略模型只是一个初始策略。虽然它借鉴了人类专家的经验，但可能存在局限性，例如数据不完整、环境变化等因素。通过强化学习方法基于环境反馈对初始策略进行改进，可以使其适应更复杂的环境情况，进一步优化策略，从而获得更好的性能。所以不能直接将学得的策略模型作为最优策略，而需要进一步改进。
- 定义：模仿学习是指在强化学习中，机器从人类专家的决策过程范例中学习的方法。直接模仿学习是其中一种方式，它通过获取人类专家的决策轨迹数据，将状态作为特征，动作作为标记，利用监督学习方法（如分类或回归算法）学得策略模型，该模型可作为机器强化学习的初始策略，后续还可通过强化学习基于环境反馈进行改进。

任务与奖赏题目

- 强化学习任务中，学习的目的是找到能使（）最大化的策略。
- A. 单步奖赏
- B. T 步累积奖赏
- C. γ 折扣累积奖赏
- D. 长期累积奖赏

任务与奖赏题目

- A. 单步奖赏
- 错误
- A 选项单步奖赏是在某个特定时刻采取一个动作后立即获得的奖赏反馈。然而，仅仅追求单步奖赏的最大化可能会导致短视行为，无法保证在整个学习过程中获得最优结果。例如，在一个迷宫探索任务中，如果智能体只关注当前步骤获取的小奖赏（如捡起一个小道具获得的即时奖励），而忽略了探索更有潜力通向目标但当前奖赏较低的路径，可能最终无法到达目标获得更大的累积奖赏。所以强化学习的目的不是单纯使单步奖赏最大。

任务与奖赏题目

- B. T 步累积奖赏
- 错误
- B 选项 T 步累积奖赏是考虑有限的 T 步内的累积奖赏，它在一定程度上考虑了短期的累积效果。但在很多实际情况中，强化学习任务是一个持续的过程，T 步的限制可能无法涵盖整个任务的全貌。比如一个长期投资决策任务，仅考虑前 T 步的累积收益可能无法把握长期的投资趋势和潜在的更大回报，而且 T 的选择往往比较困难且具有局限性，所以不能仅仅以 T 步累积奖赏最大化为目标。

任务与奖赏题目

- C. γ 折扣累积奖赏
- 错误
- C 选项γ折扣累积奖赏是一种考虑未来奖赏衰减的累积方式，它通过折扣因子γ来平衡当前奖赏和未来奖赏的重要性。虽然它在一定程度上解决了无限期累积奖赏计算的问题，但它仍然只是长期累积奖赏的一种计算方式，不能完全代表强化学习的最终目标。强化学习更强调在整个长期的交互过程中获得最优的累积奖赏，而不局限于某一种特定计算方式下的累积奖赏最大化。

任务与奖赏题目

- D. 长期累积奖赏
- 正确
- D 选项长期累积奖赏涵盖了从智能体开始与环境交互到任务结束或趋于稳定的整个过程中所获得的所有奖赏之和。它综合考虑了智能体在各个阶段的决策对最终结果的影响，体现了智能体在整个学习过程中的全局最优性能。无论是单步奖赏、T 步累积奖赏还是折扣累积奖赏等，都是为了更好地近似或实现长期累积奖赏最大化这个最终目标。因此，强化学习任务的学习目的是找到能使长期累积奖赏最大化的策略，D 选项正确

K –摇臂赌博机题目

- 以下关于ϵ–贪心算法在 K –摇臂赌博机中的描述，正确的是（）
- A. 每次尝试时，总是选择当前平均奖赏最高的摇臂
- B. 每次尝试时，以固定的概率选择当前平均奖赏最高的摇臂，以1-ϵ概率随机选择一个摇臂
- C. 每次尝试时，以ϵ的概率随机选择一个摇臂，以1-ϵ概率选择当前平均奖赏最高的摇臂
- D. 每次尝试时，完全随机选择一个摇臂

K –摇臂赌博机题目

- 以下关于ϵ–贪心算法在 K –摇臂赌博机中的描述，正确的是（C）
- A. 每次尝试时，总是选择当前平均奖赏最高的摇臂
- B. 每次尝试时，以固定的概率选择当前平均奖赏最高的摇臂，以1–ϵ概率随机选择一个摇臂
- C. 每次尝试时，以ϵ的概率随机选择一个摇臂，以1–ϵ概率选择当前平均奖赏最高的摇臂
- D. 每次尝试时，完全随机选择一个摇臂

K –摇臂赌博机题目

- A. 每次尝试时，总是选择当前平均奖赏最高的摇臂
- 错误
- A 选项描述的是“仅利用”法的策略，在这种方法下，智能体每次都选择当前已知的平均奖赏最高的摇臂，完全不进行探索新摇臂的操作。这种策略在摇臂奖赏分布稳定且已知的情况下可能表现较好，但在实际情况中，由于初始时对摇臂奖赏情况了解有限，可能会错过其他摇臂潜在的更高奖赏，容易陷入局部最优解。例如，如果一个新的摇臂在初始几次尝试中表现不佳，但实际上其长期平均奖赏可能很高，“仅利用”法就会因为过早放弃对该摇臂的探索而无法发现其真正价值。所以 A 选项不符合ϵ–贪心算法。

K –摇臂赌博机题目

- B. 每次尝试时，以固定的概率选择当前平均奖赏最高的摇臂，以1–ϵ概率随机选择一个摇臂
- 错误
- B 选项中概率选择的方式错误。ϵ –贪心算法是以的概率进行探索，即随机选择一个摇臂，而不是选择当前平均奖赏最高的摇臂；以1–ϵ的概率进行利用，即选择当前平均奖赏最高的摇臂（若有多个最优摇臂，则随机选取一个）。如果按照 B 选项的方式，会导致探索和利用的操作颠倒，无法实现–贪心算法在探索新摇臂和利用已有信息之间的平衡，不利于找到最优摇臂。所以 B 选项错误。

K –摇臂赌博机题目

C. 每次尝试时，以ϵ的概率随机选择一个摇臂，以1-ϵ概率选择当前平均奖赏最高的摇臂

正确

C 选项正确描述了ϵ–贪心算法的操作方式。以ϵ的概率随机选择摇臂，可以保证在一定程度上对所有摇臂进行探索，即使是那些尚未充分了解的摇臂也有机会被选中，从而有可能发现更高奖赏的摇臂；以1-ϵ的概率选择当前平均奖赏最高的摇臂，则是在利用已有的信息，选择当前看起来最优的摇臂，以获取相对较高的奖赏。通过合理调整的值，可以在探索和利用之间进行权衡，适应不同的摇臂奖赏分布情况，提高找到最优摇臂的概率。

K –摇臂赌博机题目

D. 每次尝试时，完全随机选择一个摇臂

错误

D 选项完全随机选择一个摇臂，这是一种纯粹的探索策略，没有利用已有的关于摇臂平均奖赏的信息。在 K –摇臂赌博机问题中，完全随机选择会使智能体花费大量时间在尝试各种摇臂上，而无法有效地利用已有的经验来提高获取奖赏的效率。虽然探索是必要的，但没有利用的探索效率低下，很难在有限的尝试次数内找到最优摇臂，所以 D 选项不符合ϵ–贪心算法的特点。

K –摇臂赌博机题目

- 以下关于ϵ–贪心算法在 K –摇臂赌博机中的描述，正确的是（C）
- A. 每次尝试时，总是选择当前平均奖赏最高的摇臂
- B. 每次尝试时，以固定的概率选择当前平均奖赏最高的摇臂，以1–ϵ概率随机选择一个摇臂
- C. 每次尝试时，以ϵ的概率随机选择一个摇臂，以1–ϵ概率选择当前平均奖赏最高的摇臂
- D. 每次尝试时，完全随机选择一个摇臂

有模型学习题目

- 在基于 T 步累积奖赏的策略评估算法中，值函数的计算是通过（）进行迭代更新的。
- A. 从值函数的初始值出发，直接计算每个状态的 T 步累积奖赏
- B. 利用 Bellman 等式，根据前一步的状态值函数和当前的转移概率、奖赏函数来计算当前状态的值函数
- C. 随机选择状态，根据该状态下执行动作后的奖赏和下一个状态的值函数来更新当前状态的值函数
- D. 以上都不对

有模型学习题目

- 在基于 T 步累积奖赏的策略评估算法中，值函数的计算是通过（B）进行迭代更新的。
- A. 从值函数的初始值出发，直接计算每个状态的 T 步累积奖赏
- B. 利用 Bellman 等式，根据前一步的状态值函数和当前的转移概率、奖赏函数来计算当前状态的值函数
- C. 随机选择状态，根据该状态下执行动作后的奖赏和下一个状态的值函数来更新当前状态的值函数
- D. 以上都不对

有模型学习题目

- A. 从值函数的初始值出发，直接计算每个状态的 T 步累积奖赏
- 错误
- A 选项从值函数的初始值出发直接计算每个状态的 T 步累积奖赏，这种方式没有考虑到状态之间的转移关系和动态变化。在基于 T 步累积奖赏的策略评估算法中，仅仅计算每个状态单独的 T 步累积奖赏是不够的，因为状态之间是相互关联的，一个状态的价值不仅取决于自身的直接奖赏，还与后续状态以及状态转移的概率有关。例如，在一个迷宫游戏中，一个状态的价值不仅仅取决于当前位置获得的即时奖赏，还与从该位置通过不同动作转移到其他位置后的后续奖赏有关。所以 A 选项不符合策略评估算法的值函数计算方式。

有模型学习题目

- B. 利用 Bellman 等式，根据前一步的状态值函数和当前的转移概率、奖赏函数来计算当前状态的值函数
- 正确
- B 选项利用 Bellman 等式进行迭代更新是正确的。Bellman 等式基于马尔可夫决策过程（MDP）的特性，它将当前状态的值函数表示为当前状态下采取不同动作转移到下一个状态的期望值，这个期望值是通过前一步的状态值函数、当前的转移概率和奖赏函数来计算的。在基于 T 步累积奖赏的策略评估算法中，通过不断地利用 Bellman 等式进行迭代，可以逐步更新每个状态的值函数，使其更准确地反映在给定策略下的累积奖赏期望。例如，在每一次迭代中，根据当前对每个状态的值函数估计，结合转移概率和奖赏函数，重新计算每个状态的新值函数，经过多次迭代后，值函数会收敛到一个稳定的状态，准确表示在 T 步内的累积奖赏期望。所以 B 选项正确。

有模型学习题目

- C. 随机选择状态，根据该状态下执行动作后的奖赏和下一个状态的值函数来更新当前状态的值函数
- 错误
- C 选项随机选择状态并根据该状态下执行动作后的奖赏和下一个状态的值函数来更新当前状态的值函数，这种方式缺乏系统性和完整性。在策略评估算法中，需要对所有状态进行全面的考虑和更新，而不是随机选择状态进行局部更新。如果只是随机选择状态进行更新，可能会导致某些状态的值函数更新不及时或者不准确，无法准确评估整个策略下的累积奖赏。而且这种方式没有充分利用 MDP 的结构和已知的转移概率信息，无法保证值函数的收敛性和准确性。所以 C 选项不符合基于 T 步累积奖赏的策略评估算法要求。

免模型学习题目

- 时序差分学习与蒙特卡罗强化学习相比，其优势在于（）
- A. 时序差分学习不需要进行采样
- B. 时序差分学习可以在每执行一步策略后就更新值函数估计，效率更高
- C. 时序差分学习能更准确地估计值函数
- D. 时序差分学习适用于所有类型的强化学习任务

免模型学习题目

- 时序差分学习与蒙特卡罗强化学习相比，其优势在于 (B)
- A. 时序差分学习不需要进行采样
- B. 时序差分学习可以在每执行一步策略后就更新值函数估计，效率更高
- C. 时序差分学习能更准确地估计值函数
- D. 时序差分学习适用于所有类型的强化学习任务

免模型学习题目

- A. 时序差分学习不需要进行采样
- 错误
- 时序差分学习同样需要进行采样。它通过智能体在环境中执行动作并观察状态转移和奖赏来获取样本信息，只是与蒙特卡罗强化学习的采样方式和利用样本的时机有所不同。例如，在一个迷宫探索任务中，时序差分学习的智能体在每走一步后，会根据这一步的经验（如当前状态、采取的动作、获得的奖赏和到达的新状态）进行学习，这个过程中就涉及到了采样。所以 A 选项错误。

免模型学习题目

- B. 时序差分学习可以在每执行一步策略后就更新值函数估计，效率更高
- 正确
- 时序差分学习结合了动态规划与蒙特卡罗方法的思想，可以在每执行一步策略后就更新值函数估计。这是因为它利用了马尔可夫决策过程（MDP）的特性，在每个时间步，根据当前状态、动作、奖赏以及下一个状态的值函数估计来更新当前状态 –动作对的值函数。相比之下，蒙特卡罗强化学习需要完成一个完整的采样轨迹（从初始状态到终止状态的一系列状态、动作和奖赏）后才更新值函数估计。在一个复杂的环境中，例如机器人在多个房间中执行任务，时序差分学习能够更快地利用新获取的信息调整策略，而不需要等待整个任务完成，因此效率更高。所以 B 选项正确。

免模型学习题目

- C. 时序差分学习能更准确地估计值函数
- 错误
- 时序差分学习和蒙特卡罗强化学习在不同的环境和任务条件下各有优劣，不能简单地说时序差分学习能更准确地估计值函数。在某些情况下，蒙特卡罗强化学习通过对完整轨迹的采样和平均，可以获得较为准确的长期累积奖赏估计，从而准确估计值函数；而在另一些情况下，时序差分学习由于能够更快地利用新信息，可能在动态环境中表现更好，但并不意味着它总是能更准确地估计值函数。例如，在一个奖赏分布非常稀疏且状态转移具有随机性的环境中，蒙特卡罗强化学习可能需要大量的采样才能准确估计值函数，但一旦估计准确，其结果可能非常可靠；而时序差分学习可能会因为过早地利用不完整信息而导致估计偏差。所以 C 选项错误。

免模型学习题目

- D. 时序差分学习适用于所有类型的强化学习任务
- 错误
- 虽然时序差分学习在很多常见的强化学习任务中表现出色，但它并不适用于所有类型的强化学习任务。例如，在一些具有特殊结构或约束的环境中，可能需要其他专门设计的学习方法。同时，某些任务可能对采样的完整性或精确性有特殊要求，蒙特卡罗强化学习或其他方法可能更合适。所以 D 选项错误。

值函数近似题目

- 在值函数近似中，采用线性函数近似值函数时，参数更新规则是基于（）来最小化误差的。
- A. 随机梯度下降法
- B. 批量梯度下降法
- C. 梯度下降法
- D. 以上都不对

值函数近似题目

- 在值函数近似中，采用线性函数近似值函数时，参数更新规则是基于（C）来最小化误差的。
- A. 随机梯度下降法
- B. 批量梯度下降法
- C. 梯度下降法
- D. 以上都不对

值函数近似题目

- A. 随机梯度下降法
- 错误
- 随机梯度下降法每次从训练数据中随机选取一个样本计算梯度并更新参数。在值函数近似中，虽然随机梯度下降法也是基于梯度的优化方法，但它每次只使用一个样本，可能会导致更新方向不稳定，收敛速度较慢且可能无法收敛到全局最优解。例如，在处理大规模数据时，随机梯度下降法可能会因为个别异常样本而使参数更新偏离最优方向。而在值函数近似采用线性函数近似值函数时，通常需要考虑所有样本或一个样本子集的整体趋势来更稳定地更新参数，所以一般不是基于随机梯度下降法来最小化误差，A选项错误。

值函数近似题目

- B. 批量梯度下降法
- 错误
- 批量梯度下降法在每次更新参数时需要计算整个训练数据集的梯度。在值函数近似的情境下，如果采用批量梯度下降法，计算所有样本的梯度可能在计算资源和时间上消耗过大，尤其当数据集规模非常大时。而且，在实际应用中，数据可能是动态生成或不断更新的，批量梯度下降法需要重新计算整个数据集的梯度，灵活性较差。相比之下，通常会采用类似梯度下降法的思想，但不是严格意义上的批量梯度下降法，B 选项错误。

值函数近似题目

- C. 梯度下降法
- 正确
- 在值函数近似中采用线性函数近似值函数时，为了使值函数近似的误差最小化（常用最小二乘误差度量），采用梯度下降法对误差求负导数，得到参数更新规则。梯度下降法通过计算目标函数（这里是值函数近似的误差函数）关于参数的梯度，然后沿着梯度的反方向更新参数，以逐步减小误差。这种方法在每次更新时可以考虑到一定数量的样本（可以是整个数据集，也可以是一个小批量样本，类似随机梯度下降法和批量梯度下降法的折衷），在计算效率和稳定性之间取得较好的平衡。例如，在处理连续状态空间的值函数近似时，可以根据采样得到的状态 –动作对数据，计算误差函数关于参数的梯度，然后按照一定的步长更新参数，使值函数逐渐逼近真实值函数，所以 C 选项正确。

模仿学习题目

- 逆强化学习的基本思想是（）
- A. 直接从人类专家范例数据中学习策略，不涉及奖赏函数
- B. 通过调整策略，使机器在给定环境中获得的累积奖赏最大化
- C. 寻找一种奖赏函数，使得范例数据在该奖赏函数环境下对应的策略是最优策略
- D. 利用强化学习算法直接优化人类专家范例数据中的动作选择

模仿学习题目

- 逆强化学习的基本思想是（C）
- A. 直接从人类专家范例数据中学习策略，不涉及奖赏函数
- B. 通过调整策略，使机器在给定环境中获得的累积奖赏最大化
- C. 寻找一种奖赏函数，使得范例数据在该奖赏函数环境下对应的策略是最优策略
- D. 利用强化学习算法直接优化人类专家范例数据中的动作选择

模仿学习题目

- A. 直接从人类专家范例数据中学习策略，不涉及奖赏函数
- 错误
- 直接从人类专家范例数据中学习策略且不涉及奖赏函数，这描述的是直接模仿学习的方式，而非逆强化学习。直接模仿学习是将人类专家决策轨迹中的状态 –动作对抽取出来，构造数据集后利用监督学习（如分类或回归算法）来学习策略模型，重点在于直接模仿专家的动作选择，不关注奖赏函数的构建。而逆强化学习的核心是围绕奖赏函数展开的，所以 A 选项错误。

模仿学习题目

- B. 通过调整策略，使机器在给定环境中获得的累积奖赏最大化
- 错误
- 通过调整策略使机器在给定环境中获得的累积奖赏最大化，这是强化学习的一般目标，但没有体现逆强化学习的独特之处。强化学习通常是在已知或未知环境模型下，通过不断尝试和优化策略来实现累积奖赏最大化。逆强化学习则是从范例数据反推奖赏函数，再基于此训练策略，与该选项表述的过程不同，所以 B 选项错误。

模仿学习题目

- C. 寻找一种奖赏函数，使得范例数据在该奖赏函数环境下对应的策略是最优策略
- 正确
- 逆强化学习的基本思想正是寻找一种奖赏函数，使得范例数据在该奖赏函数环境下对应的策略是最优策略。在逆强化学习中，已知状态空间、动作空间以及人类专家的决策轨迹数据集，其目的是找到一个合适的奖赏函数，使得基于这个奖赏函数在环境中求解出的最优策略所产生的轨迹能够与范例数据一致。然后，就可以使用这个推导出的奖赏函数来训练强化学习策略，让机器学习到与人类专家相似的行为模式。所以 C 选项正确。

模仿学习题目

- D. 利用强化学习算法直接优化人类专家范例数据中的动作选择
- 错误
- 利用强化学习算法直接优化人类专家范例数据中的动作选择，这种说法没有抓住逆强化学习的关键。逆强化学习不是直接优化动作选择，而是先确定合适的奖赏函数，再根据奖赏函数来优化策略进而影响动作选择。它是从一个更高层次的奖赏函数构建出发，来引导策略学习和动作选择优化，所以 D 选项错误。

End