一　问题重述

传统的机器学习方法在求解具有连续动作或状态的问题时其效率较低，使用特殊的微分对策问题可以评估一种机器学习算法的算力。

传统的羊-犬博弈问题是一个很好的特殊微分对策问题。羊需要逃出一个半径为R的圆形圈，羊的速率为v且在逃跑的过程中始终保持不变。羊的逃逸路径上的每一点距离圆心的距离始终不会减少，在此条件下羊具有任意的转弯能力。只要羊逃出圆形圈则获得胜利。犬沿着圆周以定速率V围捕羊，任何时刻具有选择圆周两个方向其中之一的能力。

请建立数学模型研究一下问题：

1. 通过运动学精确建模求解犬的最优围堵策略；
2. 假设犬以最优策略围堵，基于精确建模求解羊可以逃逸胜出的条件；
3. 假设羊理解自己的能力、限制和躲避犬围堵而逃逸的目标，但不具备基于运动学的最优化决策知识，假设2中羊可以逃逸的条件被满足，给出一种机器学习方法，使得羊通过学习训练后实现逃逸；
4. 设计一套评价体系，定量评价3中给出的机器学习方法的学习能力；
5. 提出并定量评价更多的羊逃逸机器学习方法。

二　问题分析

2.1　对于问题一的分析

问题一需要通过运动学精确建模求解犬的最优围堵策略。我们已知犬的目标只有一个，预测羊在圆上哪个点逃逸并且往这个点移动。设犬所在位置为A，羊所在半径在圆上的端点为B，犬的最优围堵策略就是沿着劣弧$\overset{\frown}{AB}$运动。羊往接近犬的方向运动时犬有很大概率能在圆上围抓到羊，羊往远离犬的方向运动时犬只有沿着劣弧$\overset{\frown}{AB}$运动才能更快接近羊。

2.2　对于问题二的分析

问题二要基于犬以最优策略围堵，基于精确建模求解羊可以逃逸胜出的条件。我们求出在犬采用最优围堵策略的条件下，羊能够逃逸的极小条件，即求出羊的速率V_s的下限。这是一个典型的微分博弈、追逃对策问题。对于此问题我们需要讨论羊的速度相对于犬只的几种不同情况，分别是羊的速率小于犬的速率；羊的速率等于犬的速率；羊的速率大于犬的速率。同时我们还应当考虑初始时羊与犬的位置。

2.2.1　微分博弈

微分博弈是指在时间连续的系统内，多个参与者进行持续的博弈，力图最优化各自独立、冲突的目标，最终获得各参与者随时间演变的策略并达到纳什均衡，即任何参与者都没有单独改变策略的意愿。首先对微分对策的一般形式进行介绍。设P、E是对策双方，动态系统状态方程为：$x(t) = f[x(t), u(t), v(t), t]$，给定初始状态为$x(t) = x_0$式中$x(t)$为n维P、E双方的状态变量；$u(t)$为$m_u$维P方控制变量；$v(t)$为$m_v$维E方控制变量，$u(t)$和$v(t)$的各分量$u_i(t)$和$v_i(t)$均为$t$的分段连续函数，$[u(t), v(t)]$为容许控制策略或简称控制策略；$f(\cdot)$为n连续可微的向量函数；$t \in [t_0, t_f]$。取性能指标为：

$$H[x(t), u(t), v(t), \lambda(t), t] = L[x(t), u(t), v(t), t] + \lambda^T f[x(t), u(t), v(t), t] \tag{1}$$

2.2.2　纳什-庞特里亚金最大最小原理

定义Hamilton函数如下：

$$H[x(t), u(t), v(t), \lambda(t), t] = L[x(t), u(t), v(t), t] + \lambda^T f[x(t), u(t), v(t), t]$$

若 $u(t)$和$v(t)$是最优控制，则有：

(1) 状态变量$x(t)$与协态变量$\lambda(t)$满足以下协态方程

$$x(t) = \frac{\partial H}{\partial \lambda} \tag{2}$$

$$\dot{\lambda}(t) = -\frac{\partial H}{\partial x} \tag{3}$$

(2) 边界条件

$$x(t_0) = x_0 \tag{4}$$

$$\lambda(t_f) = \frac{\partial \theta[x(t_f), t_f]}{\partial x(t_f)} \tag{5}$$

(3) 对于任意$t \in [t_0, t_f]$，Hamilton函数满足下述条件:

$$\begin{aligned} H[x^*(t), u^*(t), v^*(t), t] = & \min_u \max_v H[x^*(t), u(t), v(t), \lambda^*(t), t] \\ = & \max_v \min_u H[x^*(t), u(t), v(t), \lambda^*(t), t] \end{aligned} \tag{6}$$

2.2.3 追逃对策

追逃对策(pursuit-evasion game)是双方对追及与否的对策，一类典型的定性微分对策。参与对策的追、逃两方，分别记为P和E，P方欲迫近E，E方相反，欲远离P，若P方接近到E方的一定范围，并用$\psi(x) \leq 0$(目标集)表示，则称“追及”或“捕获”，这时对策结束。目标集(termination set)是一种集合，是定性微分对策结束时所要求实现的集合。在定性微分对策中，对策结束时所要求实现的条件界定的集合，如用导弹拦截飞机，要求在拦截过程结束时，飞机处于导弹爆炸时能击毁飞机的有效范围内，亦即导弹与飞机间的距离不超过某一给定值。一般地，目标集可写成$\psi(x) \leq 0$。要素：

局中人：此时局中人就是追者与逃者，例如在空战中是一方的歼击机和另一方的轰炸机，在空防中是攻方的轰炸机和守方的高射炮，在导弹战中是一方的导弹和另一方的反导弹等等。

策略：追逃双方都有自己的可选择的行动方案，例如轰炸机可选择飞行路线和投弹方式，本题的羊的逃跑方向等。因为在追踪问题中，局中人例如犬必须每时每刻都掌握双方的相对位置和某些情况以便跟踪追击。同样羊也得每时每刻都能掌握双方的相对位置和某些情况以便躲过。所以用数学描述追踪对策中的策略，也必须要反映出这种连续动态的决策过程。这就得借助微分方程的理论，因此这类对策理论称为微分对策。

于是局中人的一个策略就是决定一个连续控制的规律,即对任意状况变量$X(t)$给出自己应选的控制量。所以控制规律就是状况变量的函数，记为$\phi(X)$(对追者)，$\psi(X)$(对逃者)。一般，当控制规律 $\phi(X)$,$\psi(X)$为各方选定以后，双方的运动情况可由某个微分方程组描述：

$$\frac{dX}{dt} = f(x, \phi, \psi) \tag{7}$$

此方程组也称为运动方程。从理论上说，当给定了初始状况X_0和控制规律 $\phi(X)$,$\psi(X)$后，由运动方程可解出双方在这局对策中自始至终的运动路线。

一局对策的得失:对于追踪对策，当追着了或逃掉了，即到达运动路线的终点时，这局对策也就结束了。但在终点上的得失不光是追着或逃掉，因为有时逃者并非单纯逃，而是企图在被追到以前达到某种目的。那么，追者必须在逃者达到目的之前追着它才是最好的。

2.3 对于问题三的分析

问题三犬采用最优围堵策略，羊满足逃逸条件，通过机器学习训练使羊逃逸。此题没有样本数据，使用机器学习中的强化学习进行处理。强化学习是智能体以“试错”的方式进行学习，通过与环境进行交互获得的奖赏指导行为，目标是使智能体获得最大的奖赏，强化学习不同于连接主义学习中的监督学习，主要表现在强化信号上，强化学习中由环境提供的强化信号是对产生动作的好坏作一种评价(通常为标量信号)，而不是告诉强化学习系统如何去产生正确的动作。

这里面我们选用Double DQN算法来希望取得更好的效果。首先使用
gym构建环境，智能体是羊，环境是羊、犬和圆形草地，犬采用最优围堵策略围堵羊，若羊在一段时间内逃出圈则胜利，这段时间内没逃出或者被犬抓到则失败；状态空间是整个圆组成的点集，是二维的；动作空间是羊每一步可采取的动作的集合；回报的设计主要涉及是否达到目标，采用的时间和犬羊之间的角度等等，后面进行说明。然后随机好犬羊的位置，先为犬的速率V赋初值，然后给羊的速率v赋满足逃逸条件的初值，便可以开始训练。

2.4 对于问题四的分析

问题四要设计一套评价体系，定量评价3中给出的机器学习方法的学习能力。这里主要设计了四个评价指标。

- 胜率：即羊获得胜利次数与总次数之比。
- 完美率：即羊跑出去需要的时间和实际花费的时间之比。
- 切向距离：即羊在切向方向上移动的距离之和。
- 回报值：即羊在最终时刻得到的回报值。

2.5 对于问题五的分析

问题五需要提出并定量评价更多的羊逃逸机器学习方法。对于问题四提出的一些评教指标，我们将对强化学习的其他方法，比如Q-Learning、DQN和Policy Gradients 等算法进行评价。