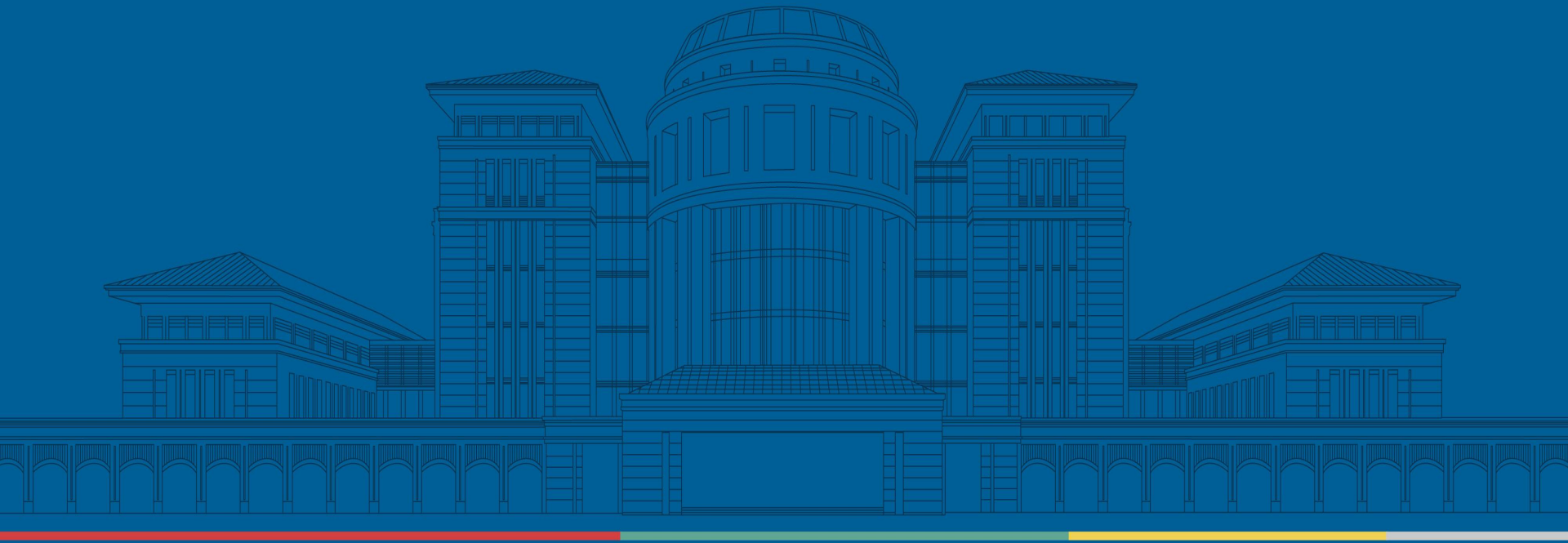
澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Group Report

speaker：Xiongyi Li

17th Nov 2023



Outline

- Deep Reinforcement Learning
  - basics
- Views in Self Direction
  - soliding
  - personal thoughts

Deep Reinforcement Learning

- Basics

First part
- Machine Learning Fundamentals
- Monte Carlo
- Reinforcement Learning Basic Concepts

Deep Reinforcement Learning

- Machine Learning Fundamentals

1.1 Linear models

eg. house purchasing

considering price, distance to work, size, years of construction....

$$\boldsymbol{x} = \left[x_1, x_2, \cdots, x_d\right]^T.$$

weights(权重ω)

offset(偏移量b)

$$f\left(\boldsymbol{x}; \boldsymbol{w}, b\right) \triangleq \boldsymbol{x}^T \boldsymbol{w} + b.$$

$$f\left(\boldsymbol{x}; \boldsymbol{w}, b\right) \triangleq w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b.$$

training data

$$\widehat{y}' = f\left(\boldsymbol{x}'; \widehat{\boldsymbol{w}}, \hat{b}\right),$$

loss(损失函数)

$$L\left(\boldsymbol{w}, b\right) = \frac{1}{2n} \sum_{i=1}^{n} \left[f(\boldsymbol{x}_i; \boldsymbol{w}, b) - y_i\right]^2.$$

Deep Reinforcement Learning

- Monte Carlo

2.1 random variable

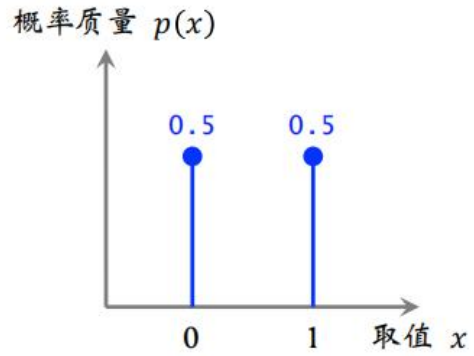
probability mass function(概率质量函数)

probability density function(概率密度函数)

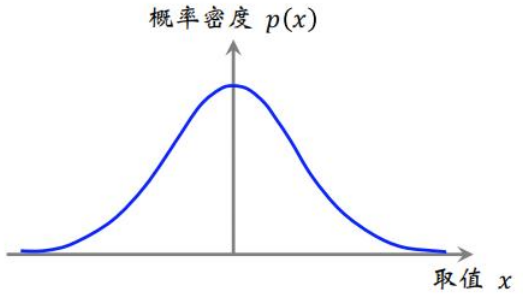
expectation(期望)

$$\mathbb{E}_{X\sim p(\cdot)}[h(X)] = \sum_{x\in\mathcal{X}} p(x)\cdot h(x).$$

$$\mathbb{E}_{X\sim p(\cdot)}[h(X)] = \int_{\mathcal{X}} p(x)\cdot h(x)\,dx.$$



离散概率分布

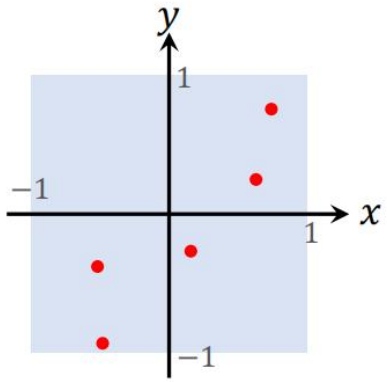


连续概率分布

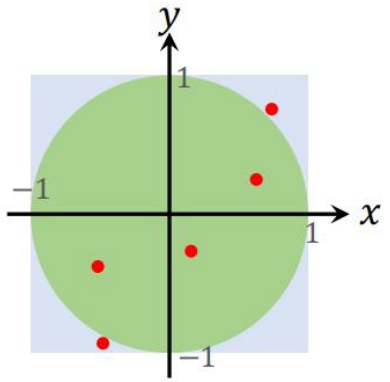
Deep Reinforcement Learning

- Monte Carlo

2.2Monte Carlo estimation



从蓝色正方形中做随机抽样，
得到n个红色的点。



抽到的红色的点可能落在绿色的圆内部，也可能落在外部。

estimation of π

p(green) = $\frac{\pi}{4}$

all dots = n; dots in green = variable M

expectation[M] = $\frac{\pi n}{4}$

In experiment, we find that m dots in green.

when n is big enough, number m is very close to expectation[M]

so m $\approx \frac{\pi n}{4}$, then $\pi \approx \frac{4m}{n}$

Deep Reinforcement Learning

- Reinforcement Learning Basic Concepts

3.1 Terminology

environment(环境)

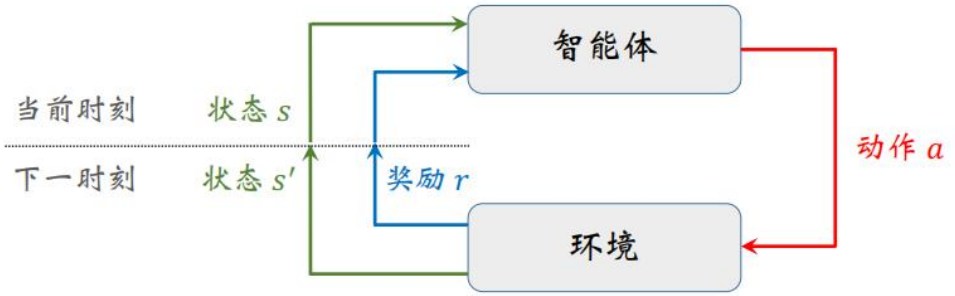
agent(智能体)

state(状态s)

action(动作a)

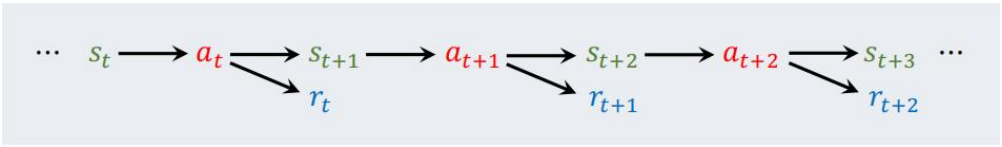
policy(决策π)

reward(奖励r)



trajectory(轨迹)： all states actions rewards

Markov property: $\mathbb{P}(S_{t+1} \mid S_t, A_t) = \mathbb{P}(S_{t+1} \mid S_1, A_1, S_2, A_2, \cdots, S_t, A_t).$



Deep Reinforcement Learning

- Reinforcement Learning Basic Concepts

3.2 Return

return: cumulative future reward (U_t)

$$U_t = R_t + R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_n.$$

optimum policy(最优策略): maximize expectation[return]

discounted return(折扣回报)

$$U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot R_{t+2} + \gamma^3 \cdot R_{t+3} + \cdots$$

Deep Reinforcement Learning

- Reinforcement Learning Basic Concepts

3.3 Value function(价值函数)

U_t is a variable ⟶ unknown

action-value function(动作价值函数):

$$Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}, \cdots, S_n, A_n}\Big[U_t \Big| S_t = s_t, A_t = a_t\Big].$$

related to s_t, a_t,π

optimal action-value function(最优动作价值函数):

$$Q_\star(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t), \qquad \forall\, s_t \in \mathcal{S}, \quad a_t \in \mathcal{A}.$$

state-value function(状态价值函数):

$$V_\pi(s_t) = \sum_{a \in \mathcal{A}} \pi(a|s_t) \cdot Q_\pi(s_t, a).$$

related to s_t & π

Views in Self Direction

- knowledge to solid
- personal thoughts

Views in Self Direction

- knowledge to solid

1. deep learning
2. computer vision
3. robotic control

There is still a lot to learn

Views in Self Direction

- personal thoughts

application:

- DRL for more intelligent NPC
- robots with DPL for like somewhat boxing learning (from video or live, may be applied for other areas)

Thank You!

Avenida da Universidade, Taipa, Macau, China
Email : mc35289@um.edu.mo Website : www.um.edu.mo