

Python金融计算（第六讲）： 智能交易与深度强化学习

谢文杰

华东理工大学 金融学系

2022年春

目录

- 1 深度强化学习简介
- 2 深度学习简介
- 3 深度Q神经网络算法
- 4 深度强化学习应用

目录

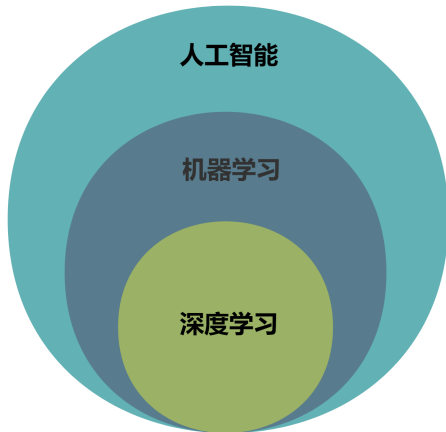
- 1 深度强化学习简介
- 2 深度学习简介
- 3 深度Q神经网络算法
- 4 深度强化学习应用

深度强化学习简介

智能系统有不同类型，也有不同层次，融入了人类生活的方方面面。在计算机科学中可以将智能分成计算智能、感知智能、决策智能、认知智能以及通用智能等。

- 近年来，感知智能某些方面机器也超过了人类，主要得益于深度学习的蓬勃发展，如人脸识别、图像识别等。
- 2016年AlphaGo横空出世，在围棋上打败了人类围棋世界冠军李世石，智能机器的决策智能惊艳全球。AlphaGo用到的关键技术之一就是深度强化学习。

人工智能和机器学习的关系



深度学习与经典机器学习的差异

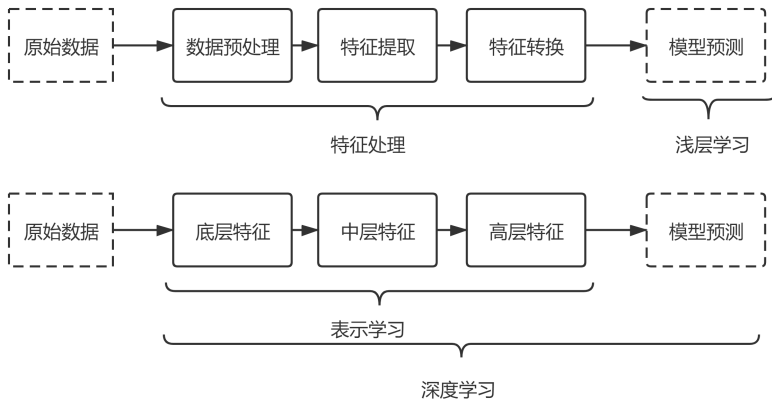
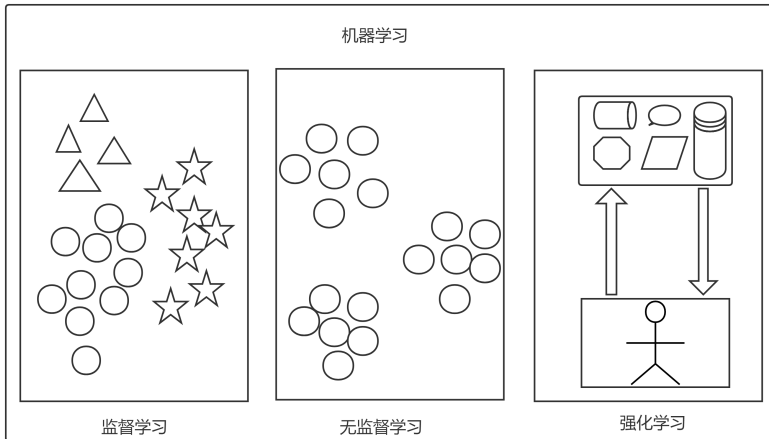
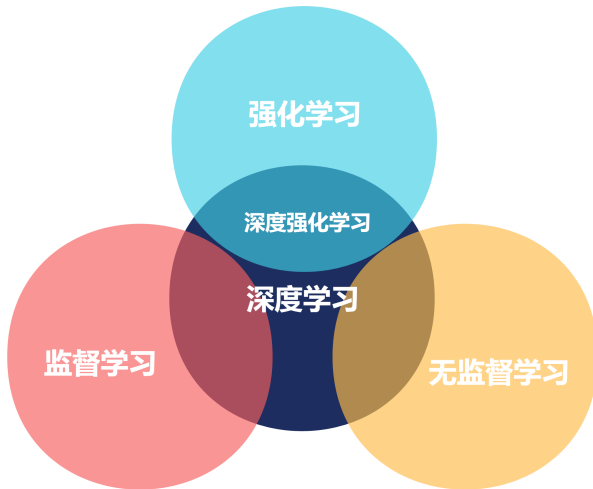


图: 深度学习与经典机器学习的差异

机器学习三大方向



深度学习和强化学习融合



深度强化学习经典论文

- 2013年, NIPS, DeepMind “Playing Atari with Deep Reinforcement Learning”
- 2015年, Nature, Google DeepMind “Human-level control through deep reinforcement learning”
- 2016年, Nature, Google DeepMind “Mastering the game of Go with deep neural networks and tree search”
- 2017年, Nature, Google DeepMind “Mastering the game of Go without human knowledge”
- 2018年, Deepmind 提出了 Alpha Zero 和 AlphaFold。
2020年, DeepMind的第二代AlphaFold 在国际蛋白质结构预测竞赛（CASP）获得了冠军。

目录

- 1 深度强化学习简介
- 2 深度学习简介
- 3 深度Q神经网络算法
- 4 深度强化学习应用

深度学习实例

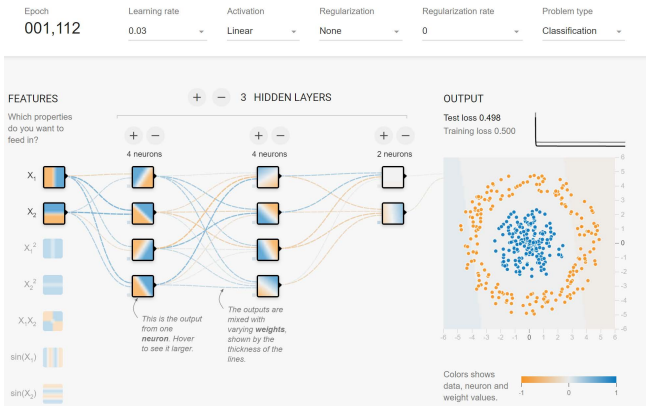


图: 线性不可分情况下监督学习分类问题示例。画图工具网
址<https://playground.tensorflow.org>

目录

- 1 深度强化学习简介
- 2 深度学习简介
- 3 深度Q神经网络算法
- 4 深度强化学习应用

金融市场中马尔科夫决策过程建模

定义

基于离散马尔科夫决策过程 (Markov Decision Process , MDP) 的金融市场环境模型可用一个六元组 $(S, \mathcal{A}, P, R, \gamma, H)$ 表示, 其中:

- S 表示金融市场环境状态集合,
- \mathcal{A} 表示智能体动作集合,
- $P : S \times \mathcal{A} \times S \rightarrow [0, 1]$ 是金融市场环境状态转移函数。
- $R : S \times \mathcal{A} \times S \rightarrow \mathcal{R}$ 是金融市场环境奖赏函数, \mathcal{R} 为连续的区间, $R(s_t, a_t, s_{t+1}) \in \mathcal{R}$, $R_{\max} \in \mathbb{R}^+$ (e.g., $[0, R_{\max}]$),
- $\gamma \in [0, 1)$ 是折扣系数。
- H 是投资期限。

Q-learning

下面简单介绍时序差分Q-learning算法的伪代码：

Algorithm 11: 时序差分 Q-learning 算法的伪代码

Input: 状态空间 S , 动作空间 \mathcal{A} , 折扣系数 γ , 以及环境 E , 初始化的状态-动作值函数 $Q(s, a) = 0$, 初始采样策略采用 $\pi(a | s) = \frac{1}{|\mathcal{A}|}$ 。

Output: 最优策略 π^*

```
1 for  $k = 0, 1, 2, 3, 4, 5, \dots$  do
2   % 每次循环针对一条轨迹
3   初始化状态  $s$ 
4   for  $t = 0, 1, 2, 3, 4, \dots, T$  do
5     % 采用  $\epsilon$ -贪心策略生成轨迹中的一步
6     
$$\pi(s, a) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}|}, & a = \arg \max_a Q(s, a) \\ \frac{\epsilon}{|\mathcal{A}|}, & a \neq \arg \max_a Q(s, a) \end{cases}, \quad (4.50)$$

7     产生一步轨迹  $(s, a, r, s')$ 
8     其中  $a$  和  $r$  是基于  $\epsilon$ -贪心策略产生的动作和即时奖励。
9      $Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a') - Q(s, a))$ 
10    智能体进入下一个状态  $s = s'$ 
11    if  $s$  为终止状态 then
12      | 开始下一条轨迹采样
13  % 计算最优策略
14  for  $s \in S$  do
15    |  $\pi^*(s) = \arg \max_a Q(s, a)$ 
```

策略学习示例

策略背后，个体考虑时都会有一个收益矩阵，如下所示

表: 基于不同天气动作的收益矩阵

	下雨	不下雨
带伞	1	-1
不带伞	-2	1

策略学习示例

最优的策略就是下雨带伞，不下雨就不带伞，可以简单表示为矩阵形式

表: 基于不同天气最优策略

天气	下雨	不下雨
动作	带伞	不带伞

策略学习示例

状态-动作值矩阵(Q-table)可以表示如下:

$$\begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ \vdots \\ s_{n-1} \\ s_n \end{array} \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_{m-1} & a_m \\ Q_{11} & Q_{12} & Q_{13} & \cdots & Q_{1,m-1} & Q_{1m} \\ Q_{21} & Q_{22} & Q_{23} & \cdots & Q_{2,m-1} & Q_{2m} \\ Q_{31} & Q_{32} & Q_{33} & \cdots & Q_{3,m-1} & Q_{3m} \\ Q_{41} & Q_{42} & Q_{43} & \cdots & Q_{4,m-1} & Q_{4m} \\ Q_{51} & Q_{52} & Q_{53} & \cdots & Q_{5,m-1} & Q_{5m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ Q_{n-1,1} & Q_{n-1,2} & Q_{n-1,3} & \cdots & Q_{n-1,m-1} & Q_{n-1,m} \\ Q_{n1} & Q_{n2} & Q_{n3} & \cdots & Q_{n,m-1} & Q_{nm} \end{pmatrix}$$

其中 s_i 表示状态 i , a_j 表示动作 j , Q_{ij} 表示智能体在状态 s_i 情况下执行动作 a_j 的价值。

Deep Q Network

DQN全称是Deep Q Network，基于Q-learning演化而来，Q-learning作为强化学习核心算法，有着悠久的历史。Q-learning算法核心是学习状态-动作值函数 $Q(s, a)$ ，然后在给定的状态下选择最优动作。在学习过程中 $Q(s, a)$ 更新公式如下

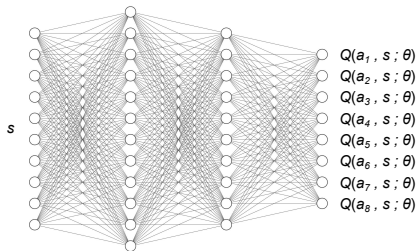
$$Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a') - Q(s, a)) \quad (1)$$

深度强化学习和经典强化学习的最大区别在于，值函数和策略函数都使用深度神经网络进行逼近，而算法更新过程中更新状态-动作值函数 $Q(s, a; \theta)$ 的参数 θ ，因此将状态-动作值函数 $Q(s, a)$ 参数化

$$Q(s, a) = Q(s, a; \theta) \quad (2)$$

Deep Q Network

DQN算法基于经验数据来拟合状态-动作值函数：



在DQN算法中，通过对值函数进行遍历，找到价值最大的动作作为输出动作，最终得了策略函数：

$$a^* = \arg \max_{a'} Q(s, a'; \theta) \quad (3)$$

Q-learning

下面简单介绍DQN算法的伪代码：

Algorithm 14: DQN 算法的伪代码

Input: 状态空间 S , 动作空间 \mathcal{A} , 折扣系数 γ , 以及环境 E 。

初始化状态-动作值函数 $Q(s, a; \theta)$ 参数 θ 。

初始化目标网络 $Q(s, a; \theta^-)$ 参数 $\theta^- = \theta$ 。

Output: 最优策略 π^*

```
1 for  $k = 0, 1, 2, 3, 4, 5, \dots$  do
2   % 每次循环针对一条轨迹
3   初始化状态  $s$ 
4   for  $t = 0, 1, 2, 3, 4, \dots, T$  do
5     采用  $\epsilon$ -贪心策略 Eq. 10.1 产生一步轨迹  $(s, a, r, s')$ , 并存入经验池
6     if 到了需要更新参数的时候 then
7       从经验池中随机采样小批量  $n$  个状态转移序列对  $(s, a, r, s')$ , 针对每个
          序列  $i$  计算 TD 目标值
8       if  $s'$  是终止状态 then
9          $y_i = r_t$ 
10      else
11         $y_i = r_t + \gamma \max_{a'} Q(s', a'; \theta^-)$ 
12      计算小批量  $n$  个状态转移序列的损失函数  $J(\theta)$  及其梯度:
          
$$\nabla J(\theta) = -\frac{2}{n} \sum_{i=1}^n (y_i - Q(s_i, a_i; \theta)) \nabla Q(s_i, a_i; \theta) \quad (5.13)$$

          更新网络参数  $\theta = \theta + \alpha \nabla \theta$ 
13      % 参数  $\theta$  更新  $C$  次后更新一次  $\theta^-$ 
14      if 间隔  $C$  步 then
15         $\theta^- = \theta$ 
16 返回最优参数  $\theta$ , 并得到最优策略  $\pi^*$ 。
```

目录

- 1 深度强化学习简介
- 2 深度学习简介
- 3 深度Q神经网络算法
- 4 深度强化学习应用**

基于深度强化学习的智能投资决策系统

图展示了基于深度强化学习的智能投资决策系统结构示意图。

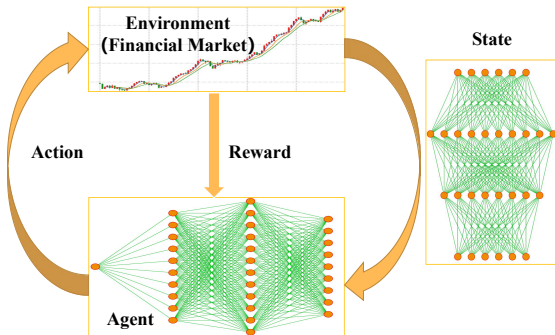





图: 基于深度强化学习的智能投资决策系统结构示意图

基于深度强化学习的智能投资决策系统

在深度强化学习算法实践应用过程中，为了提高对算法的理解可以深入开源代码库进行源码解析。理解从算法原理到编程实现的诸多细节问题和编程技巧。本示例只提供了算法基础原理和入门实践基础，因此对深度强化学习算法编程实现不做要求。下面简单介绍训练交易智能体的DQN核心代码¹。开源代码库中存在很多优秀的深度强化学习算法实现，本实例主要使用stable-baselines代码库。

¹[https:](https://stable-baselines.readthedocs.io/en/master/modules/dqn.html)

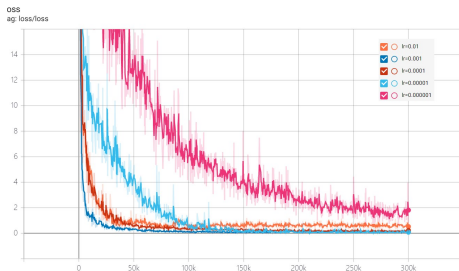
[//stable-baselines.readthedocs.io/en/master/modules/dqn.html](https://stable-baselines.readthedocs.io/en/master/modules/dqn.html)   

基于深度强化学习的智能投资决策系统

- 核心代码中设置深度强化学习算法DQN参数是模型学习和训练的关键。
- DQN算法输入参数MlpPolicy设定了DQN值函数的神经网络模型结构，面对不同的环境状态变量，可采用不同的深度神经网络模型结构，如深度卷积神经网络CnnPolicy、LnMlpPolicy等。
- 参数env_train是深度强化学习算法DQN训练的环境模型，是智能体能否进行实际应用的关键。learning_rate为学习率，是机器学习过程中需要调节的首要参数。

基于深度强化学习的智能投资决策系统

Tensorboard与TensorFlow融合，记录和分析模型在训练过程中参数变化情况。Tensorboard记录的日志数据对模型的训练、程序的调整、参数的调优都非常有帮助。图给出模型迭代了30万次后，模型训练过程中损失函数变化曲线。



图：模型训练过程中损失函数变化曲线

基于深度强化学习的智能投资决策系统

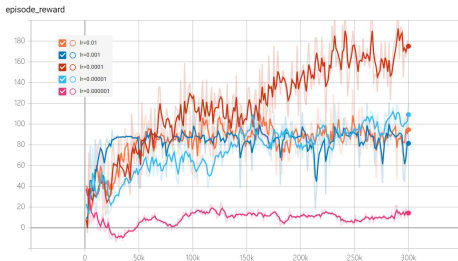
随着模型训练迭代次数的增加，损失函数显著下降，后期下降不明显，且存在一定波动。损失函数的波动也代表了模型收敛的稳定性。图不同实线对应着学习率learning_rate取不同值情况。图中可以发现五种学习率learning_rate情况下损失函数的衰减规律类似。



图：模型训练过程中TD误差变化曲线

基于深度强化学习的智能投资决策系统

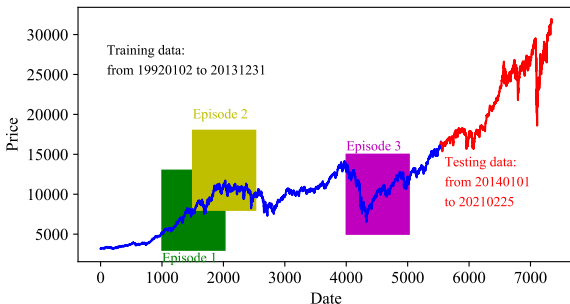
图展示了模型训练收敛情况，模型收敛是深度强化学习训练的第一步，模型训练的最终目的是收敛的决策网络模型能够在投资决策过程中获得较高的累积收益。因此查看智能体单个训练周期内的累积收益情况能够衡量智能体智能决策的绩效。



图：模型训练过程中智能体单个训练周期内累积收益变化曲线

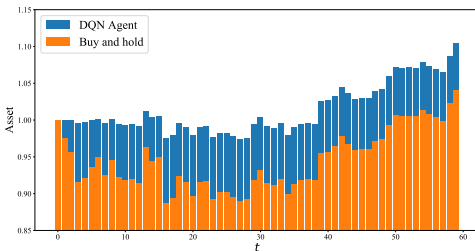
基于深度强化学习的智能投资决策系统

基于离散马尔科夫决策过程 (Markov Decision Process , MDP) 的金融市场环境模型六元组中, 变量 S 表示金融市场状态空间。如何刻画金融市场状态, 一般采用价格时间序列数据。决策过程中智能体不可能考虑全部的金融市场信息, 人类投资者也不可能知晓和处理所有市场信息。



基于深度强化学习的智能投资决策系统

对于泛化能力不强的智能算法而言，测试模型在实际应用中表现至关重要。在测试集中进行测试，查看智能体投资收益情况，如图所示。为了保证模型测试的有效性，需要将测试集与训练集进行严格分离，或者为了能够测试模型的泛化性能，可以应用于不同的市场进行测试。



图：模型测试结果。

基于深度强化学习的智能投资决策系统

为了保证模型测试的有效性，训练集为2010年01月01日至2015年12月31日价格时间序。测试集为2016年01月01日至2016年03月31日价格时间序列。

图中深色直方图表示深度强化学习智能体（DQN Agent）投资策略在不同时刻的资产价值情况。浅色直方图表示买入持有策略（Buy and hold）在不同时刻的资产价值变化情况。图中可以发现深度强化学习DQN智能体投资策略显著好于买入持有策略。在实际应用和分析中，需要进行更多的测试和指标分析，如年化收益、夏普率、最大回撤等。