

Введение в искусственный интеллект

На базе дисциплины «Вычислительные сети,
системы и телекоммуникации»

Технологический университет
Королёв
2020

Введение в искусственный интеллект

Лекция №5 — «Обучение в подкреплении».

- Метод SARSA
- Q-обучение

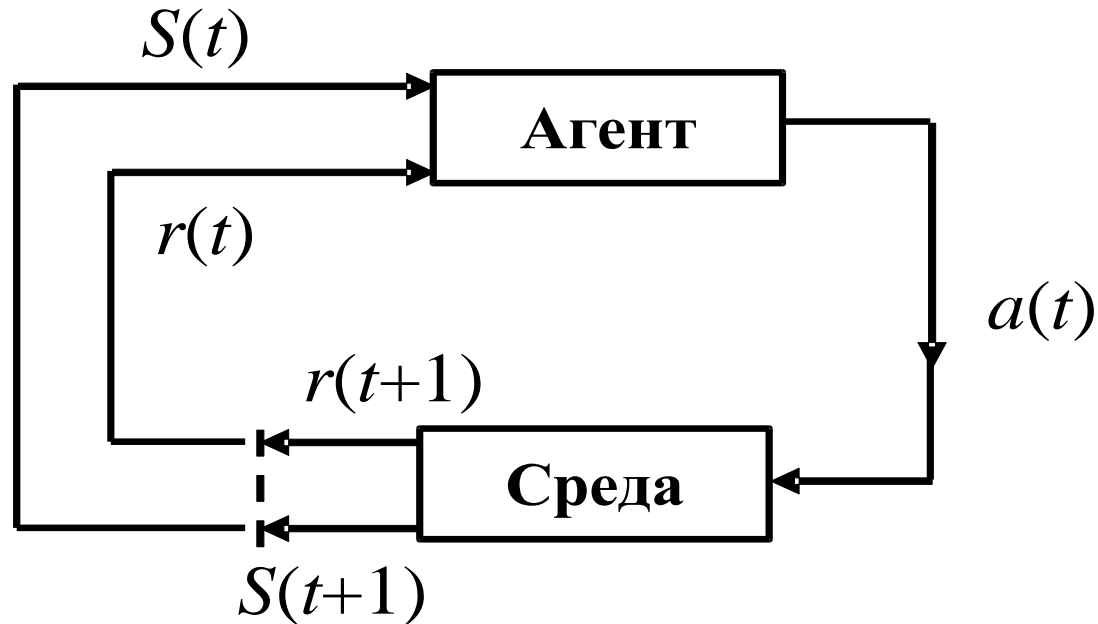
Обучение с подкреплением

- Автономный агент взаимодействует с внешней средой.
- Внешняя среда задается набором состояний.
- Агент может выполнять определённые действия.
- Агент имеет стратегию: функцию преобразования состояния среды в действие.
- В ответ на каждое действие агента внешняя среда формирует подкрепление (награду или наказание).
- Используя подкрепление, агент модифицирует свою стратегию.

SARSA

SARSA = state, action, reward, state, action

$S(t) \rightarrow a(t) \rightarrow r(t) \rightarrow S(t+1) \rightarrow a(t+1)$



SARSA

Глобальная цель агента – максимизировать суммарную награду (сумму подкреплений на каждом шаге):

$$U(t) = \sum_{k=0}^{\infty} \gamma^k r(t+k)$$

где $r(t)$ – подкрепление на шаге t ,
 γ – коэффициент дисконтирования, $\gamma \in [0,1]$.

Q-обучение

Оценка величина суммарной награды:

$$Q(S(t), a(t)) = E(U(t))$$

$$\begin{aligned} Q(S(t), a(t)) &= E(r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) + \dots) = \\ &= E(r(t) + \gamma(r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \dots)) = \\ &= E(r(t) + \gamma Q(S(t+1), a(t+1))) \end{aligned}$$

Ошибка временной разницы:

$$\delta(t) = r(t) + \gamma Q(S(t+1), a(t+1)) - Q(S(t), a(t))$$

Q-обучение

Система управления агента содержит значения $Q(S,a)$ для всех возможных пар (S,a) .

Стратегия π : $a = \pi(S)$

ϵ -жадное правило, $0 < \epsilon \ll 1$:

с вероятностью $1-\epsilon$: $a = \operatorname{argmax}(Q(S,a_i))$,

с вероятностью ϵ : случайное значение a .

Q-обучение

Алгоритм обучения:

$$\Delta Q(S(t), a(t)) = \alpha \delta(t)$$

$$\begin{aligned} Q'(S(t), a(t)) - Q(S(t), a(t)) &= \\ &= \alpha r(t) + \alpha \gamma Q(S(t+1), a(t+1)) - \alpha Q(S(t), a(t)) \end{aligned}$$

$$\begin{aligned} Q'(S(t), a(t)) &= \\ &= (1 - \alpha) Q(S(t), a(t)) + \alpha (r(t) + \gamma Q(S(t+1), a(t+1))) \end{aligned}$$

Q-обучение

Простой способ реализации

Матрица $N \times M$ со значениями $Q(S, a)$,
где N – число возможных состояний S ,
 M – число возможных действий a .

Значение Q для конкретных значений S и a
определяется как значение в соответствующей
ячейке матрицы.

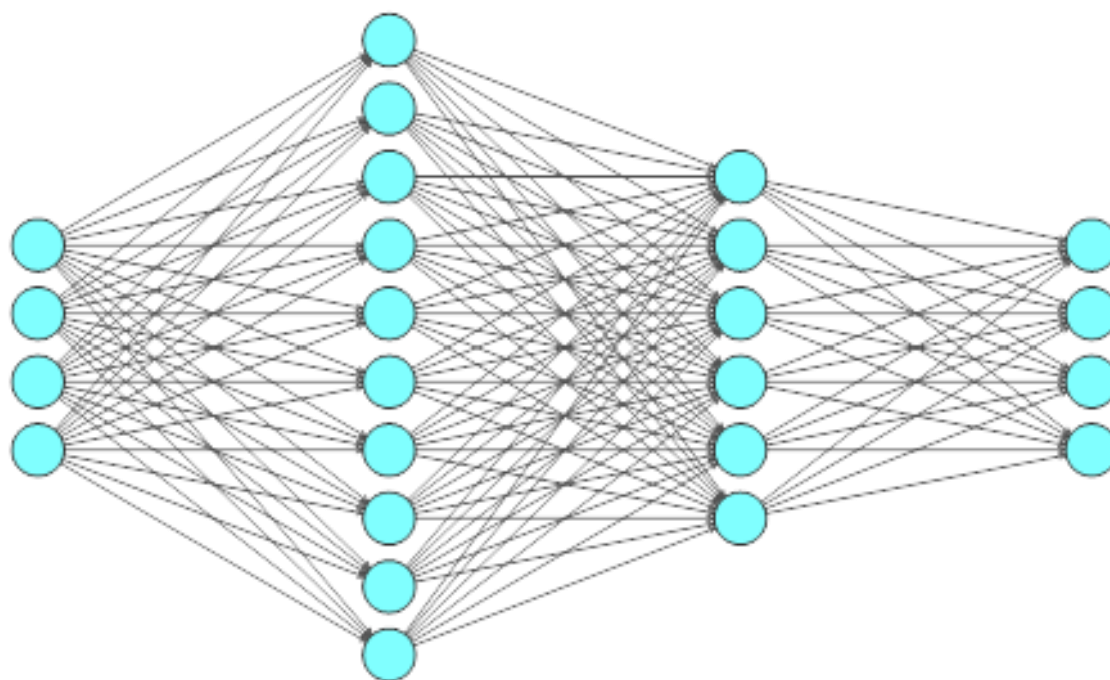
Q-обучение

Более сложный способ реализации

Нейронная сеть:

- число нейронов во входном слое равно размерности состояния S , т. е. достаточно для его однозначного описания,
- число нейронов в выходном слое равно числу возможных действий a .

Q-обучение



Q-обучение

Значение S подаётся на вход нейронной сети, значения на каждом из выходных нейронов трактуются как $Q(S,a)$.

Выбор действия производится в соответствии с принятой стратегией, например, с помощью ϵ -жадного правила.

Обучение нейронной сети производится с помощью метода обратного распространения ошибки. В качестве значения функции ошибки используется ошибка временной разности.

Q-обучение

Если обучать сеть на каждом шаге на основе выбранного действия, то

- 1) сеть будет слишком чувствительная к небольшим, локальным изменениям,
- 2) сеть будет всё время настраиваться на последние по времени события.

Решение: использование случайно выбранных батчей из заранее сохраненной памяти.

Q-обучение

Изначально память пустая, обучение не происходит, но каждый набор данных $(S(t), a(t), r(t), S(t+1))$ записывается в память как отдельный элемент.

Когда память содержит достаточное количество элементов, на каждом шаге случайным образом извлекается батч (некоторое количество элементов памяти) и сеть обучается на каждом из этих элементов.

Q-обучение

Нейронная сеть используется для двух целей:

- 1) выбор действия,
- 2) вычисление ошибки временной разности.

Иногда используют две копии нейронной сети для каждой из этих задач, периодически синхронизируя их значения связей между нейронами.

Материалы

1. <https://habr.com/ru/post/439674/>
2. <https://habr.com/ru/post/308094/>

Спасибо за внимание!