MACHINE LEARNING AND AI

ЛЕКЦИЯ 10

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ (REINFORCEMENT LEARNING)

Обучение с подкреплением - это отображение множество ситуаций в действия с целью максимизировать числовой сигнал – вознаграждение. Учащемуся не говорят, какие действия следует предпринять, вместо этого он должен выяснить, какие действия приносят наибольшее вознаграждение (методом проб и ошибок). В самых интересных и сложных случаях действия могут повлиять не только на немедленную награду, но и на следующую ситуацию, а через нее и на все последующие награды. Эти две характеристики – ***поиск методом проб и ошибок*** и ***отложенное вознаграждение*** – являются двумя наиболее важными отличительными чертами обучения с подкреплением.

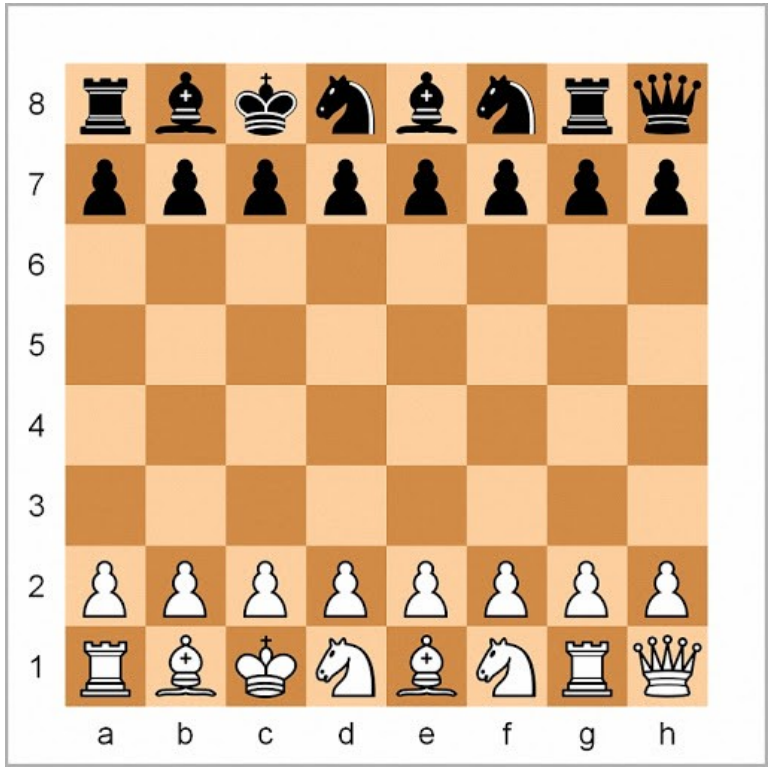
НЕКОТОРЫЕ ОТЛИЧИТЕЛЬНЫЕ ОСОБЕННОСТИ

- **RL** не является ни обучением с учителем, ни обучением без учителя
- Одной из проблем считают поиск компромисса между исследованием и эксплуатацией
- **RL** рассматривает проблему целенаправленного взаимодействия агента с неопределенной средой.

ПРИМЕРЫ ЗАДАЧ

- Шахматист делает ход. Выбор основывается как на планировании – предвидении возможных ответов и встречных ответов – так и на немедленных интуитивных суждениях о желательности определенных позиций и ходов
- Детеныш антилопы с трудом поднимается на ноги через несколько минут после рождения. Через полчаса он уже бегает со скоростью 40 км/ч
- Мобильный робот решает, следует ли ему войти в новую комнату в поисках мусора, который нужно собрать, или попвтаться вернуться к своей станции зарядки аккумулятора. Он принимает решение на основе текущего уровня заряда аккумулятора и того, насколько быстро и легко ему удавалось найти зарядное устройство в прошлом

СТРАТЕГИЯ



СРАВНИТЕЛЬНАЯ СИЛА ШАХМАТНЫХ ПРОГРАММ

Магнус Карлсен
ЭЛО 2847

Stockfish 13
ЭЛО 3516

МАТЧ AlphaZero - Stockfish

VS

Матч из 1000 партий

Счёт +155 – 6 = 839

ЭЛЕМЕНТЫ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

- ***Policy*** (политика) определяет способ поведения обучающегося агента в данный момент времени. Грубо говоря, политика - это отображение воспринимаемых состояний окружающей среды и действий, которые необходимо предпринять в этих состояниях. Это соответствует тому, что в психологии можно было бы назвать набором правил или ассоциаций "стимул-реакция"
- ***Reward signal*** *(*сигнал вознаграждения) определяет цель задачи обучения с подкреплением. На каждом временном шаге среда отправляет агенту обучения с подкреплением одно число, называемое вознаграждением. Единственная цель агента - максимизировать общее вознаграждение, которое он получает в долгосрочной перспективе.

ЭЛЕМЕНТЫ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

- В то время как сигнал вознаграждения указывает на то, что хорошо в данный конкретный момент, ***value function*** (функция ценности) указывает, что хорошо в долгосрочной перспективе. Грубо говоря, ***ценность*** состояния – это общая сумма вознаграждения, которую агент может ожидать накопить в будущем, начиная с этого состояния.
- ***Model of the environment*** (модель окружающей среды). Это то, что имитирует поведение среды или, в более общем плане, позволяет делать выводы о том, как среда будет себя вести. Например, с учетом состояния и действия модель может предсказать следующее в результате состояние и следующую награду.

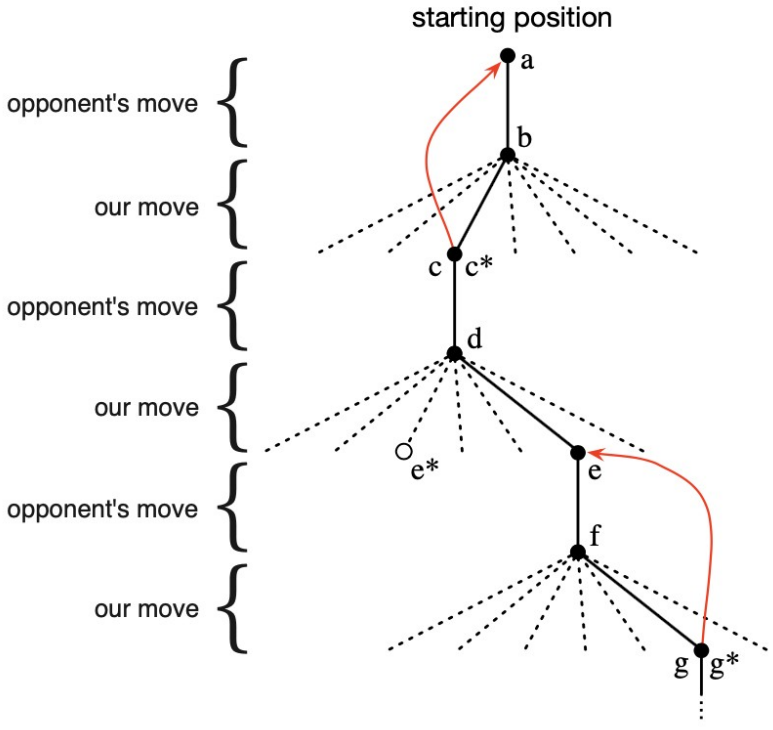
ПРИМЕР – КРЕСТИКИ-НОЛИКИ

X	O	O
O	X	X
		X

Стандартный алгоритм – “minimax”

Здесь **политика** - это правило, которое диктует игроку, какой ход сделать для каждого состояния игры – каждой возможной конфигурации X и O на доске 3x3.

ВЫБОР ПРОДОЛЖЕНИЯ



$$V(S_t) \leftarrow V(S_t) + \alpha\Big[V(S_{t+1}) - V(S_t)\Big]$$

Функция ценности состояния

МНОГОРУКИЕ БАНДИТЫ (k-armed Bandit Problem)

Рассмотрим следующую проблему обучения:

Вы постоянно сталкиваетесь с выбором из k различных вариантов или действий. После каждого выбора вы получаете числовую награду, выбранную из стационарного распределения вероятностей, которое зависит от выбранного вами действия. Ваша цель - максимизировать ожидаемую общую награду за некоторый период времени, например, за выбор более 1000 действий или временных шагов.

Ценность действия $A_t = a$ с предполагаемой наградой R_t

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

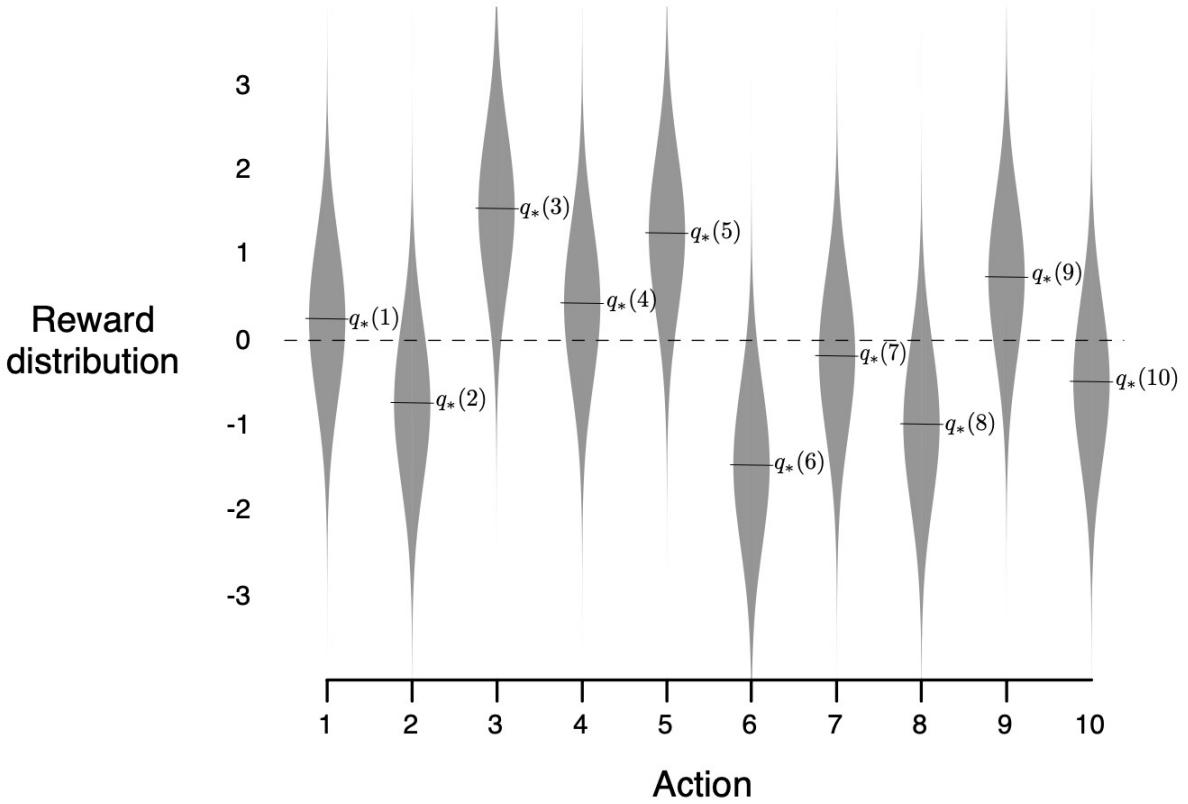
ACTION-VALUE METHOD

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

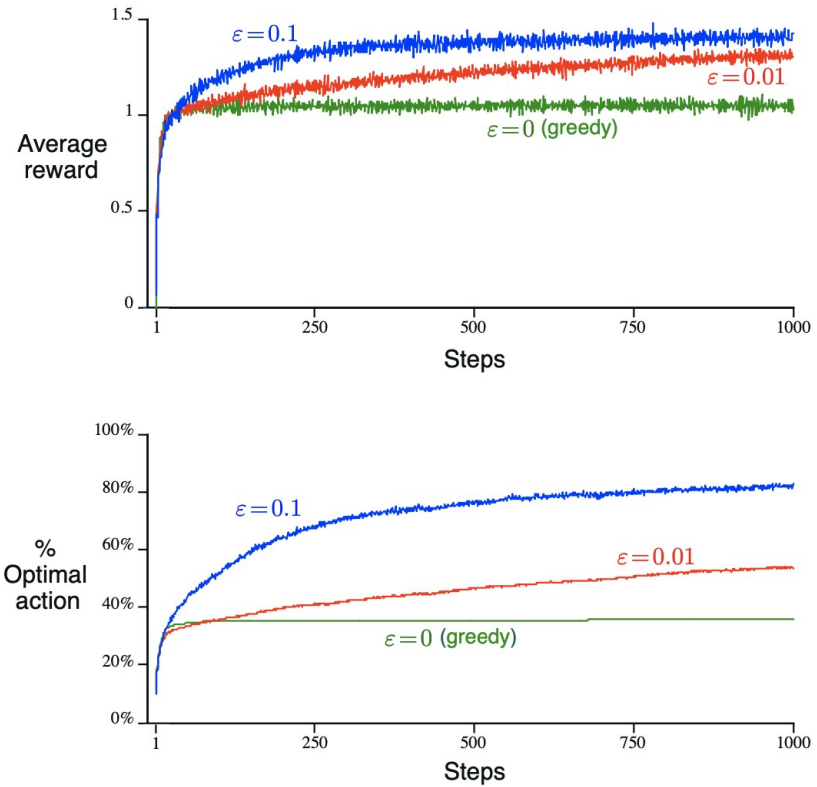
Жадный (greedy) выбор действия предполагает

$$A_t \doteq \arg\max_a Q_t(a)$$

ПРИМЕР 10-РУКИЕ БАНДИТЫ



ПРИМЕР 10-РУКИЕ БАНДИТЫ



РЕАЛИЗАЦИЯ

Ценность выбора действия

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

$$\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big],
\end{aligned}$$

$$NewEstimate \leftarrow OldEstimate \; + \; StepSize \; \Big[Target - OldEstimate\Big]$$

Q-LEARNING

FROZEN LAKE

S F F F

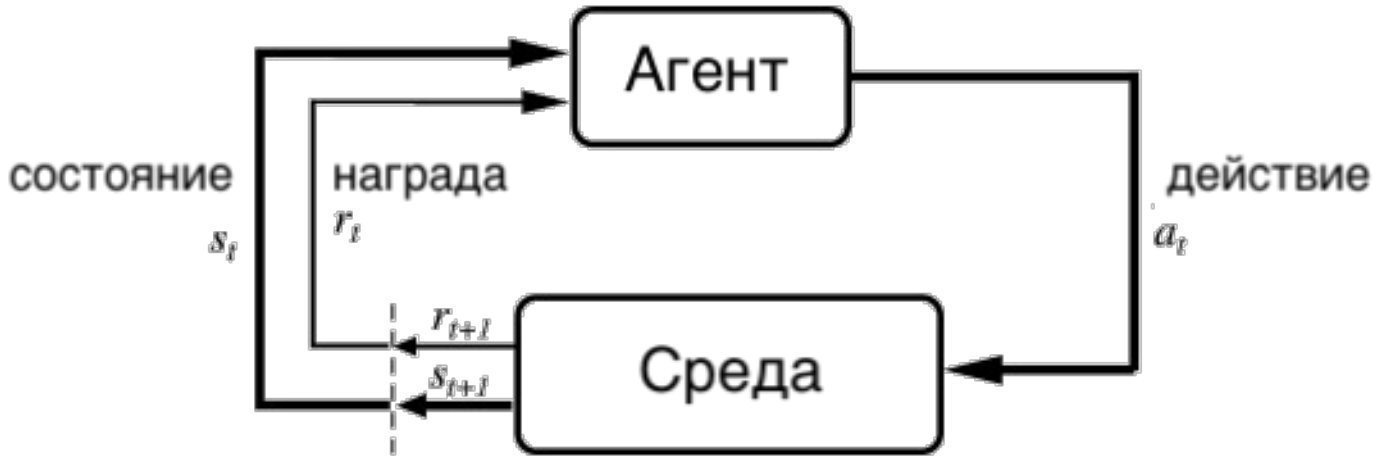
F H F H

F F F H

H F F G

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R(s,a) + \gamma\ max\ Q'(s',a') - Q(s,a)]$$

СХЕМА ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ



ТИПОВАЯ ЗАДАЧА – СТОЛБ НА ТЕЛЕГЕ (CART-POLE)