Лекция 1

Обучение с подкреплением: введение

Дополнительные главы машинного обучения

Андрей Фильченков

19.02.2021

План лекции

- Понятие обучения с подкреплением
- Задача о многоруком бандите
- Обзор вариантов постановок

- В презентации используются материалы курсов
  «Машинное обучение» К.В. Воронцова,
  «Машинное обучение с подкреплением» А.И. Панова
  CS234: Reinforcement Learning, E. Brusnkill
  Reinforcement Learning, D. Silver
- Слайды доступны: **shorturl.at/wGV59**
- Видео доступны: **shorturl.at/ovBTZ**

План лекции

- Понятие обучения с подкреплением
- Задача о многоруком бандите
- Обзор вариантов постановок

Абстрактная постановка задачи

Задана **среда**, в которой действует агент, взаимодействуя с ней через **действия**, за которые он получает **награду.**

Агент максимизирует **суммарную награду** за счет выбора наиболее подходящей **стратегии** взаимодействия со средой: хорошие действия положительно **подкрепляются** больше наградой.

Отдельная область ML

RL исторически развивался отдельно от всех остальных ветвей машинного обучения и сейчас это отдельная область.

Почему это не сводится к обучению с учителем? (по наблюдаемому состоянию среды, объекту, нужно предсказать действие, метку)

Отдельная область ML

Почему это не сводится к обучению с учителем? (по наблюдаемому состоянию среды, объекту, нужно предсказать действие, метку)

1. Нам не дан набор данных, мы собираем данные одновременно с предсказанием
2. Нам не даны метки, мы не знаем, какое действие правильное

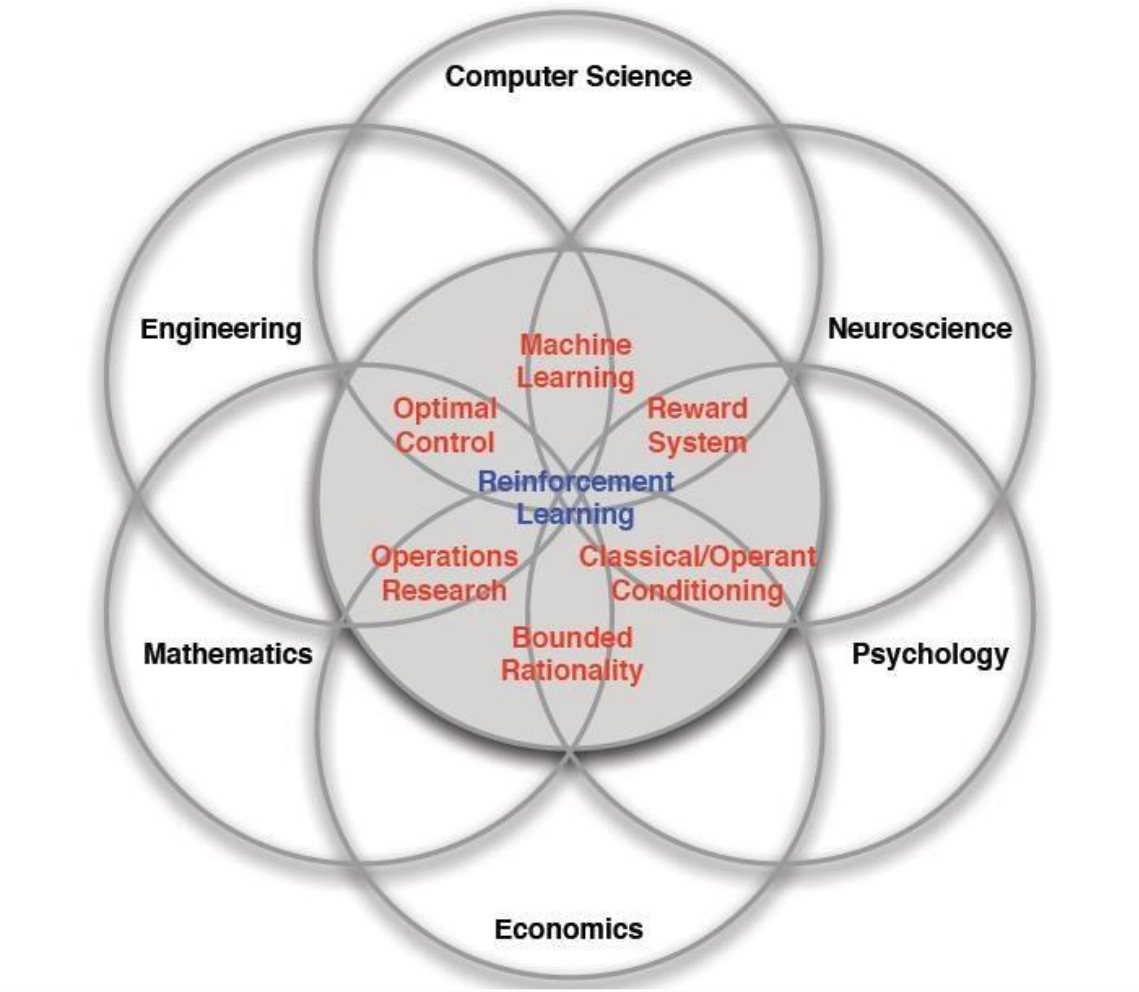
Exploration vs exploitation

Exploration (исследование) это трата ресурсов на понимание того, как устроена среда.

Exploitation (использование) это трата ресурсов на извлечение выгоды из среды.

В RL необходимо находить компромисс (trade-off) между исследованием и использованием.

Междисциплинарные связи



Optimization + neuroscience

Математически, корни RL лежат в **теории оптимального управления** и **оптимального планирования** (выбор действий, планирование и составление расписания).

Идеологически, RL связан с **оперантным обуславливанием** (поведение людей и животных и их стратегии обучения).

Предсказание и контроль

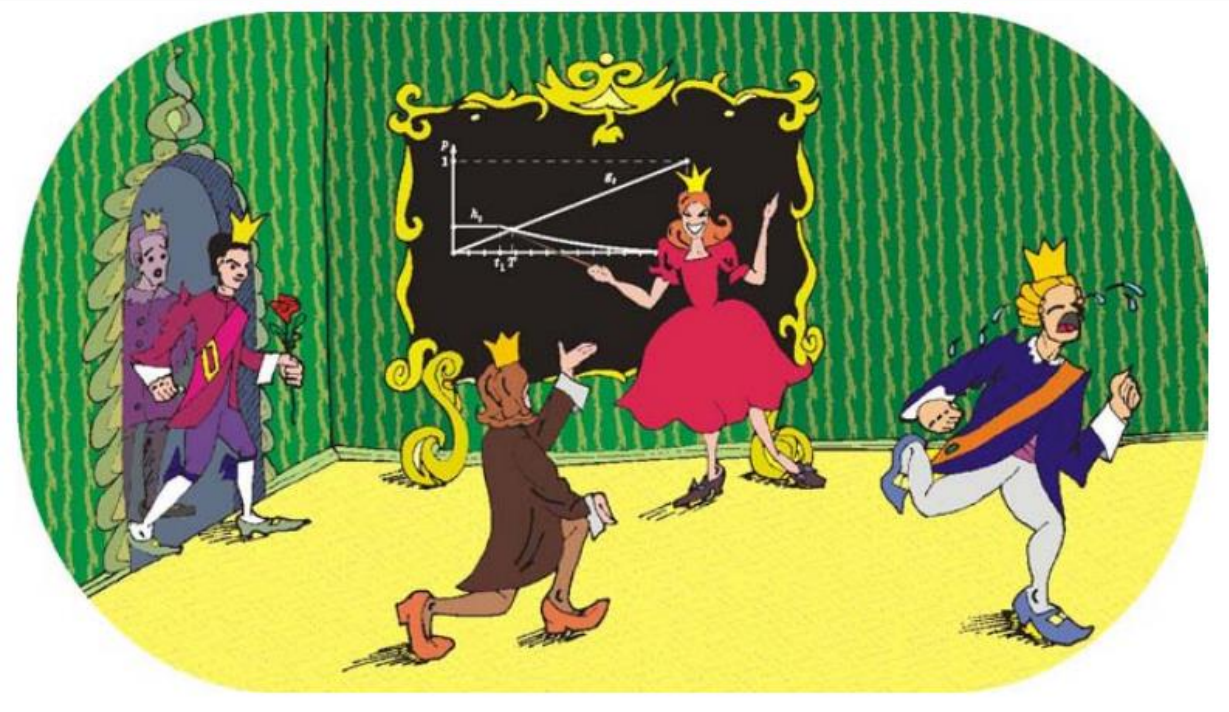
- **Предсказание**: агент должен предсказывать награду для предпринимаемых действий
- **Управление**: агент должен выбирать управляющие действия, максимизирующие награду.

Решаемые задачи

- Управление технологическим процессом
- Роботы
- Размещение рекламных баннеров
- Управление ценами и ассортиментом
- Торговля на бирже
- Маршрутизация в сетях
- Игры

Задача о разборчивой невесте

Также известна как задача секретаря, задача приданого султана, задача суетливого жениха, проблема остановки выбора



Постановка задачи

Вы — невеста (принцесса), желающая выйти замуж (за принца). n принцев уже выстроились в очередь перед вашей комнатой.

Каждый принц последовательно заходит. Либо вы говорите «да» и тут же играете свадьбу, либо говорите «нет», и он навсегда уходит (в слезах).

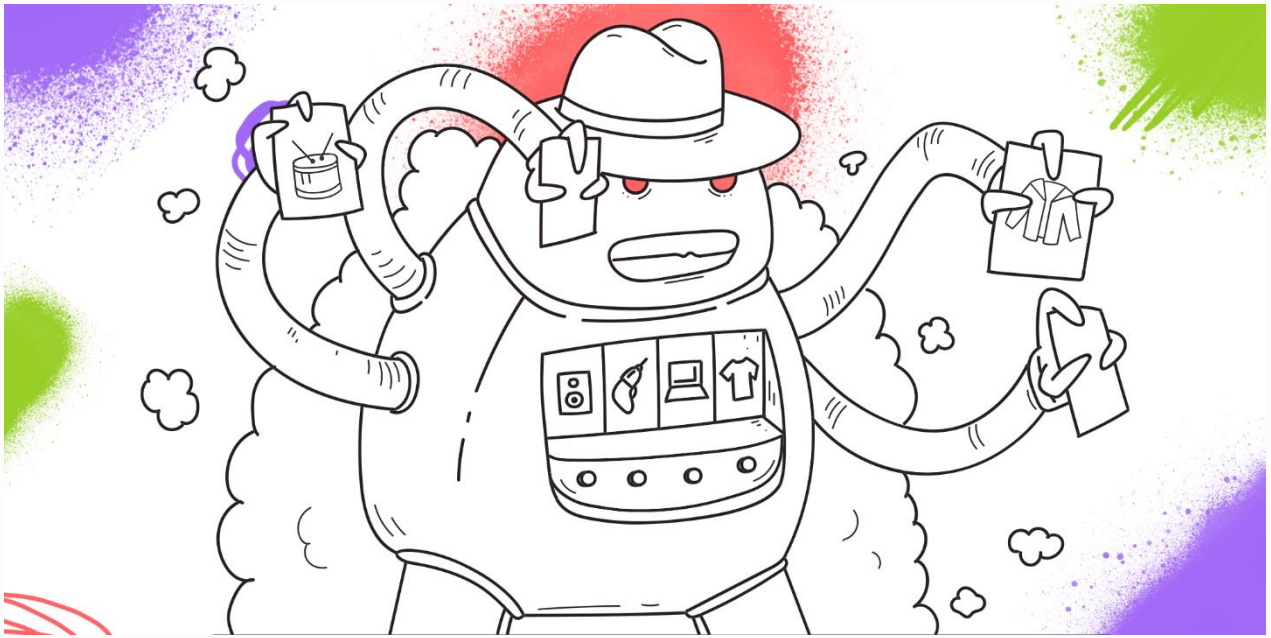
Какая для вас лучшая стратегия, чтобы максимизировать матожидание качества жениха?

План лекции

- Понятие обучения с подкреплением
- Задача о многоруком бандите
- Обзор вариантов постановок

Задача о многоруком бандите

Многорукий бандит (multi-armed bandit): среда не меняется, награда для каждого действия задается независимо



Постановка задачи

$a \in A$ — **действие**, $r \in \mathbb{R}$ — **награда**

$\$(a)$ — награда за действие, задаваемая неизвестным распределением $p_a(r)$

$\pi(t)$ это **стратегия** агента в момент времени t, которую можно представить как распределение над A: $\pi_t(a)$

Агент инициализирует $\pi(1)$ и на каждом шаге:

- предпринимает действие $a_t \sim \pi(t)$;
- среда возвращает награду $r \sim \$(a_t)$;
- агент улучшает стратегию до $\pi(t+1)$.

Меры оценки

Средняя награда за t шагов:

$$Q_t(a) = \frac{\sum_{i=1}^{t} r_i \,[a_i = a]}{\sum_{i=1}^{t} [a_i = a]}.$$

Ценность действия a:

$$Q^*(a) = \lim_{t\to\infty} Q_t(a) \to \max_{a\in A}.$$

Сожаления (regrets):

$$R(\pi) = \sum_{i=1}^{t} \left(r_{\pi(t)} - \mathrm{E} r_{\mathrm{best}(t)}\right) \to \min_{\pi}.$$

Жадная стратегия

Множество действий с наибольшее наградой:

$$A_t = \arg\max_{a \in A} Q_t(a)\,.$$

Жадная стратегия — выбираем произвольно из A_t:

$$\pi_{t+1}(a) = \frac{1}{|A_t|}[a \in A_t].$$

Нет исследования!

Для компромисса **ε-жадная стратегия**:

$$\pi_{t+1}(a) = \frac{1-\varepsilon}{|A_t|}[a \in A_t] + \frac{\varepsilon}{|A|}.$$

Стратегия Softmax

Мягкий поиск компромисса expl-expl .

Softmax strategy:

$$\pi_t(a) = \frac{\exp(Q_t(a)/\tau)}{\sum_{a' \in A} \exp(Q_t(a')/\tau)},$$

где τ — температура:

при $\tau \to 0$ сдвигаемся к использованию;

при $\tau \to \infty$ сдвигаемся к исследованию.

UCB-1

Одна из лучших полужадных стратегий

UCB-1 (upper confidence bound):

$$A_t = \arg\max_{a \in A} \left(Q_t(a) + \sqrt{\frac{2 \ln t}{k_t(a)}} \right),$$

где $k_t(a) = \sum_{i=1}^{t} [a_i = a]$.

Экспоненциальное сглаживание

Можно адаптировать, когда среда нестационарна.

Рекуррентный пересчет Q_t:

$$Q_{t+1}(a) = (1 - \alpha_t)Q_t(a) + \alpha_t r_t(a_t) = \\ = Q_t(a) + \alpha_t\big(r_t(a_t) - Q_t(a)\big).$$

При $\alpha_t = \frac{1}{k_t(a)+1}$ Q_t равно среднему.

При $\alpha_t = \text{const}$ Q_t экспоненциально сглажено.

Сравнение с подкреплением

Средняя награда: $\bar{r}_{t+1} = \alpha(r_t - \bar{r}_t)$

Преимущество (advantage) действия:

$$\mathrm{ad}_{t+1}(a_t) = \mathrm{ad}_t(a_t) + \beta(r_t - \bar{r}_t)$$

Сравнение с подкреплением (reinforcement comparison):

$$\pi_{t+1}(a) = \frac{\exp(\mathrm{ad}_{t+1}(a))}{\sum_{a' \in A} \exp(\mathrm{ad}_{t+1}(a'))}.$$

Почему здесь нет параметра температуры?

Сравнение с подкреплением

Средняя награда: $\bar{r}_{t+1} = \alpha(r_t - \bar{r}_t)$

Преимущество (advantage) действия:

$$\mathrm{ad}_{t+1}(a_t) = \mathrm{ad}_t(a_t) + \beta(r_t - \bar{r}_t)$$

Сравнение с подкреплением (reinforcement comparison):

$$\pi_{t+1}(a) = \frac{\exp(\mathrm{ad}_{t+1}(a))}{\sum_{a' \in A} \exp(\mathrm{ad}_{t+1}(a'))}.$$

Мы выбираем компромисс при помощи β.

Стратегия преследования

Вместо жадной стратегии

$$\pi_{t+1}(a) = \frac{[a \in A_t]}{|A_t|}$$

можно использовать **стратегию преследования (pursuit strategy)**:

$$\pi_{t+1}(a) = \pi_t(a) + \beta\left(\frac{[a \in A_t]}{|A_t|} - \pi_t(a)\right).$$

План лекции

- Понятие обучения с подкреплением
- Задача о многоруком бандите
- Обзор вариантов постановок задач

Награда как цель

Какой бы не была цель, она должна выражаться через награды.

Примеры:

AlphaGo — награда за выигрыш партии

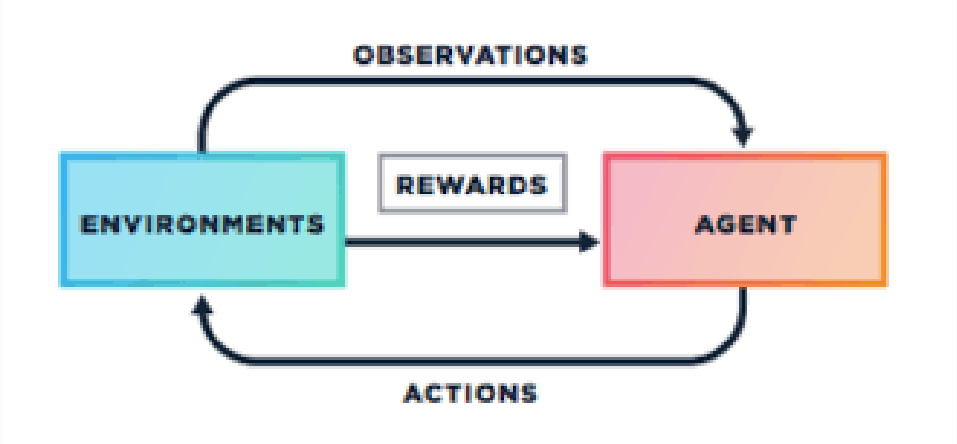
Atari — награда за получение очков от хода

Беспилотный автомобиль — награда за скорость прибытия и потребление энергии, штрафы за нарушение правил и повреждения автомобиля

Среда сложнее

Среда в общем случае меняет состояния под воздействием агента.

Состояния среды в общем случае лишь частично наблюдаемы.



Некоторые упрощения

1. Среда обладает свойством марковости:

$$\Pr(s_{t+1}|s_t, s_{t-1}, \dots) = \Pr(s_{t+1}|s_t)$$

2. Среда полностью наблюдаема:

$$o_t = s_t$$

Формализация наблюдаемой среды

Среда находится в одном из **состояний** $s \in S$. Переход между состояниями обусловлен $\tau(t) \sim p(a_t, s_t)$

Агент инициализирует $\pi(1|s) = \{s_1(a|s)\}$, и далее на каждом шаге:

- агент делает действие $a_t \sim \pi(t|s_t)$;
- среда возвращает награду $r_a \sim \$(a_t|s_t)$ и переходит в состояние $s_{t+1} \sim \tau(t)$.
- агент меняет стратегию на $\pi(t+1|s)$.

Марковский процесс принятия решений:

$$\Pr(s_{t+1} = s'; r_{t+1} = r' | s_t, a_t, r_t, \dots) = \\ = \Pr(s_{t+1} = s'; r_{t+1} = r' | s_t, a_t).$$

Особенности

Выборка $\{(s_t, a_t, r_t)\}$ не является независимой

Распределение $p(s_t, a_t, r_t)$ может меняться во времени и не зависеть от стратегии агента

Награды могут быть получены с задержкой

Награды могут быть разреженными и зашумленными

Что можно оценивать

Будущая награда: $R_t = r_{t+1} + r_{t+2} + \cdots$

Дисконтированная награда: $R_t = \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$

где γ коэффициент дисконтирования

Ценность состояния s при стратегии π:

$$V^{\pi}(s) = \mathrm{E}_{\pi}(R_t|\pi(t) = \pi) = \mathrm{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|\, \pi(t) = \pi\right)$$

Ценность действия в состоянии s при стратегии π

$$Q^{\pi}(s, a) = \mathrm{E}_{\pi}(R_t|\pi(t) = \pi, a_t = a) =$$

$$= \mathrm{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|\, \pi(t) = \pi, a_t = a\right)$$

Уравнение Беллмана

Пусть нам известно **распределение перехода**

$$\mathcal{P}^{a}_{ss'} = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$$

и **ожидание награды**

$$\mathcal{R}^{a}_{ss'} = \mathrm{E}(r_{t+1} | s_t = s, a_t = a, s_{t+1} = s').$$

Уравнение Беллмана:

$$V^{\pi(t)}(s) = \sum_{a} s_t(a|s) \sum_{s} \mathcal{P}^{a}_{ss'} \left(\mathcal{R}^{a}_{ss'} + \gamma V^{\pi}(s') \right).$$

Но это предположение никогда не выполняется.

Что можно обучать

Стратегию $\pi(a|s;\theta)$

Функцию ценности состояния $V(s;\theta)$

Функцию ценности действия $Q(s,a;\theta)$

Модель среды $(r_t, s_{t+1}) = \mu(s_t, a_t;\theta)$

Типизация агентов

Оценивающие ценность (value based): обучают только функцию ценности

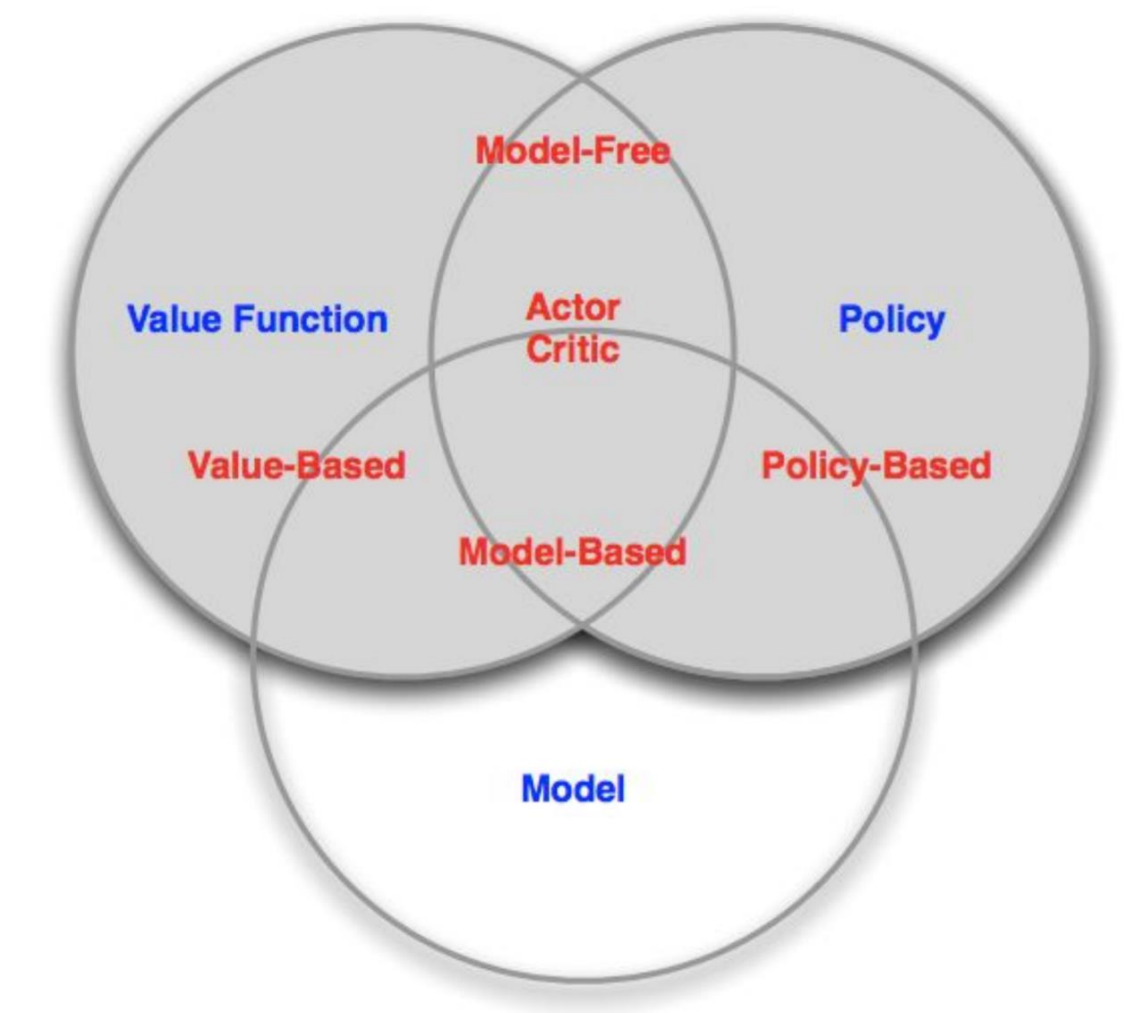
Оценивающие стратегию (policy based): обучают только стратегию

Актор-критик (actor-critic): обучают функцию ценности и стратегию

Безмодельные (model free): обучают только стратегию и/или функцию ценности

Основанные на модели (model-based): в дополнение к предыдущему обучают модель среды

Схема


Model-Free
Value Function
Actor
Critic
Policy
Value-Based
Policy-Based
Model-Based
Model