Обучение с подкреплением

Драгунов Никита

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

12 апреля 2018 г.

Постановка задачи

- До сих пор задача ставилась так: есть обучающая выборка, даны значения целевых переменных и нужно их продолжить на всё пространство (обучение с учителем), или есть неразмеченная выборка, и надо выявить ее структуру (обучение без учителя).
- Как работает обучение в реальной жизни? Мы далеко не всегда знаем набор правильных ответов, мы просто делаем то или иное действие и получаем результат.

Марковский процесс принятия решений

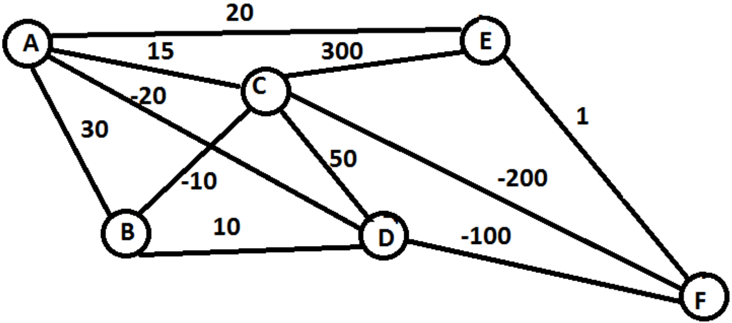
Марковский процесс принятия решений задается пятью параметрами $(S, A, P_\cdot(\cdot,\cdot), R_\cdot(\cdot,\cdot), \gamma)$, где:

- S — конечное множество состояний
- A — конечное множество действий (A_s — конечное множество действий, возможных в состоянии s)
- $P_a(s, s') = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ — вероятность того, что действие a поменяет состояние среды s в момент времени t на состояние s' в момент времени $t+1$
- $R_a(s, s')$ — вознаграждение, получаемое после перехода в состояние s' из состояния s под действием a
- $\gamma \in [0, 1]$ — дисконтирующий множитель, показывающий насколько вознаграждение в краткосрочном периоде важнее вознаграждения в долгосрочном периоде

Марковский процесс принятия решений

Основной задачей MDP является поиск стратегии — функции $\pi : S \times A \to [0, 1]$, задающей вероятность $\pi(a|s) = P(a_t = a|s_t = s)$, при которой суммарная награда будет максимальной:

$$\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1}), \quad a_t = \pi(s_t)$$



Состояния — вершины графа
Действия — переход по ребру
На ребрах указана стоимость действия, отрицательная стоимость соответствует прибыли

Методы решения:

- Наивный подход
- Функции полезности (Value функции)

Марковский процесс принятия решений

Наивный подход:

1. Опробовать все возможные стратегии
2. Выбрать стратегию с наибольшим ожидаемым выигрышем

Проблемы:

1. Количество доступных стратегий может быть очень велико или же бесконечно
2. Чтобы точно оценить выигрыш от каждой стратегии потребуется многократно применить каждую из них

Value функции

Функции полезности:

- $V^\pi(s) = E[R|s,\pi]$ — функция полезности состояния s при стратегии π, показывает ожидаемый выигрыш с начальным состоянием s при дальнейшем следовании стратегии π
- $Q^\pi(s,a) = E[R|s,a,\pi]$ — функция полезности действия a при стратегии π, показывает ожидаемый выигрыш при принятии решения a в состоянии s при дальнейшем следовании стратегии π

Равенства Беллмана

Равенства Беллмана:

$$V^\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} P_a(s, s') \left(R_a(s, s') + \gamma V^\pi(s')\right)$$

$$Q^\pi(s, a) = \sum_{s'} P_a(s, s') \left(R_a(s, s') + \gamma V^\pi(s')\right)$$

Поиск оптимальной стратегии:

$$\pi(s) := \arg\max_a \left\{ \sum_{s'} P_a(s, s') \left(R_a(s, s') + \gamma V^\pi(s')\right) \right\}$$

$$V^\pi(s) := \sum_{s'} P_{\pi(s)}(s, s') \left(R_{\pi(s)}(s, s') + \gamma V^\pi(s')\right)$$

Марковский процесс принятия решений

1. Изучить среду и создать модель
2. Метод временных разностей (Temporal difference)

Метод временных разностей TD(0)

Равенство Беллмана:

$$V^{\pi}(s) = E\left[R | s, \pi\right] = E\left[R_{a_0}(s_0, s_1) + \gamma V^{\pi}(s_1) | s_0 = s, \pi\right]$$

Дано: оцениваемая стратегия π и параметр α
Найти: $V^{\pi}(s)$
TD(0):

1. Произвольная инициализация $V^{\pi}(s)$
2. Повторять до сходимости:
    1. Выполняем действие a согласно стратегии π
    2. Узнаем s' и $R_a(s, s')$
    3. $V^{\pi}(s) := V^{\pi}(s) + \alpha\left(R_a(s, s') + \gamma V^{\pi}(s') - V^{\pi}(s)\right)$
    4. Переходим в состояние s'

Список литературы

Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto
http://incompleteideas.net/book/bookdraft2017nov5.pdf

Markov decision process, Wikipedia
https://en.wikipedia.org/wiki/Markov_decision_process

Reinforcement Learning. Searching for optimal policies I: Bellman equations and optimal policies, Mario Martin
http://www.cs.upc.edu/~mmartin/Ag4-4x.pdf

Temporal difference learning, Wikipedia
https://en.wikipedia.org/wiki/Temporal_difference_learning