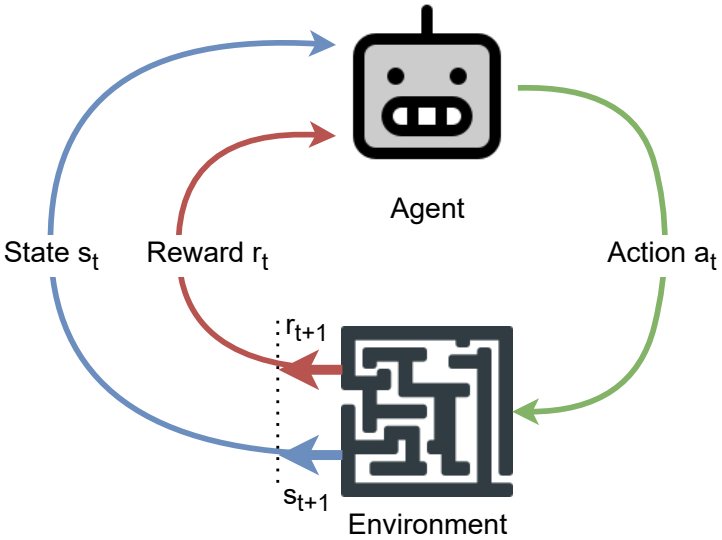
Обзор приложений RL в задачах кредитования

Сергей Самсонов[1], Алексей Масютин[2], Щербаков Игорь[3], Васильева Наталья[3]

[1] HDI Lab, НИУ ВШЭ, https://cs.hse.ru/hdilab/
[2] Центр ИИ НИУ ВШЭ, https://cs.hse.ru/aicenter/
[3] Управление инструментов и моделей, Сбер

Обучение с подкреплением: общая парадигма



Марковский процесс принятия решений (MDP)

MDP: S - пространство состояний, A - пространство действий, $\gamma \in (0; 1]$ - дисконтирующий фактор.

Процесс взимодействия агента с MDP: на шаге номер t,

- агент находится в состоянии $s_t \in \mathcal{S}$;
- агент выбирает действие $a_t \in \mathcal{A}$;
- агент переходит в следующее состояние $s_{t+1} \sim P(\cdot|s_t, a_t)$ и получает награду $r(s_t, a_t)$.

Задача агента: найти политику $\pi \colon \mathcal{S} \to \mathcal{A}$, максимизирующую функцию ценности

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right].$$

Q-функция, соответствующая политике π:

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right].$$

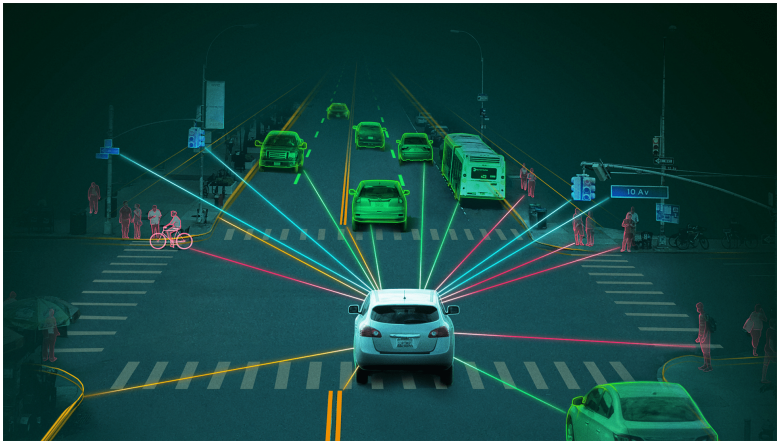
Q-обучение

- Рассмотрим процесс online-обучения агента с помощью $Q-$обучения:

$$\begin{cases} \delta_t = r(s_t, a_t) + \gamma \max_{a'}\{Q^t(s_{t+1}, a')\} - Q^t(s_t, a_t), \\ Q^{t+1}(s_t, a_t) = Q^t(s_t, a_t) + \alpha_t \delta_t, \\ \pi_{t+1}(s) \in \arg\max_a \{Q^{t+1}(s, a) + \underbrace{b_{t+1}(s, a)}_{\text{Бонус}}\}; \end{cases}$$

- Политика π_t, в соответствии с которой действует агент, связана с текущей оценкой Q-функции .

Оффлайн RL: постановка проблемы



- Основная проблема онлайн-подхода: исследование новых стратегий может быть слишком дорогим;
- Одно из решений - алгоритм CQL (conservative Q-learning, [Kumar et al., 2020]);

Risk-based pricing с помощью RL: данные и спецификация модели

- Рассмотрим постановку задачи [Khraishi and Okhrati, 2022];
- https://www8.gsb.columbia.edu/cprm/research/datasets: 200000 заявок на автомобильное кредитование, собранных с июля 2002-го по август 2004-го года;
- Состояние агента:

$$s_t = [Term_t, Amount_t, FICO_t, PD_t, PreviousRate_t, LoanType_t, \ldots];$$

- Действие: $a_t > 0$ - годовая процентная ставка;

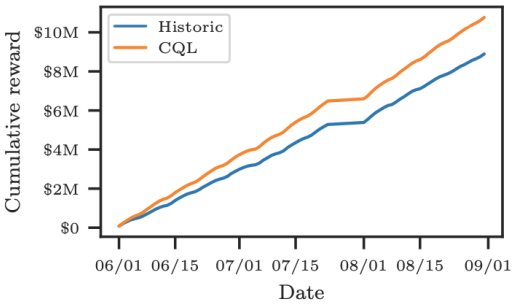
[Khraishi and Okhrati, 2022]: моделирование наград

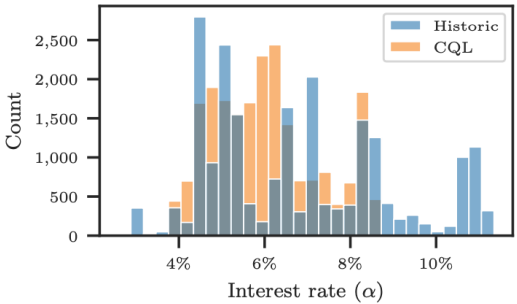
- Награда агента:

$$r(s_t, a_t) = p(+|a_t, s_t)[(1 - PD_t) * (TP_t - Cost_t) - PD_t * LGD_t * Cost_t],$$

- $p(+|a_t, s_t)$ - вероятность для заемщика принять предложение о кредите по текущей цене;
- $p(+|a_t, s_t)$ моделируется с помощью логистической регрессии;
- PD_t - вероятность дефолта заемщика (оценивается с помощью $FICO_t$);
- $LGD_t = 0.5$ для всех потенциальных заемщиков

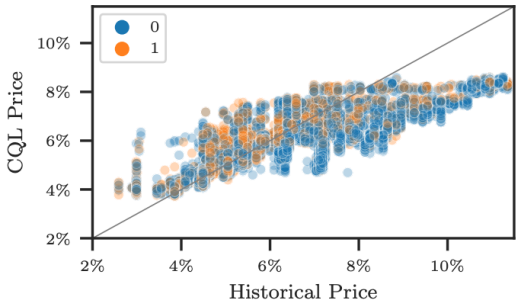
[Khraishi and Okhrati, 2022]: полученные результаты





- Кумулятивная награда выше на $\approx 21\%$, по сравнению с историческими данными;
- Средний interest rate снижается с 6.8% до 5.9%;

[Khraishi and Okhrati, 2022]: полученные результаты



- CQL неявно оценивает чувствительность заемщиков к измнению ставки;
- Достоинство RL подхода: устойчивость к миспецификации моделей;
- В различных моделях price-response CQL демонстрирует прирост профита $+5\%$ до $+24\%$;
- Прямая оптимизация награды $r(s_t, a_t)$ отличается большей диперсией: прирост профита от -7% до $+34\%$;

Constrained MDP: введение

- Введем функции затрат $C_i(s, a)$ для $i = 1, \ldots, n$;
- Ограничим множество рассматриваемых политик Π

$$\mathbb{E}_{s\sim\mu, a\sim\pi(\cdot|s)}[C_i(s, a)] \leq B_i, \quad i = 1, \ldots, n,$$

  где μ - стационарное распределение. Иными словами, мы хотим, чтобы в среднем затраты в каждый момент времени не превышали порог B_i для всех видов затрат $i = 1, \ldots, n$.
- Задача: найти политику, максимизирующую $V^\pi(s)$ при заданных ограничениях $\pi \in \Pi$.

[Abe et al., 2010]: RL для взыскания просроченной задолженности

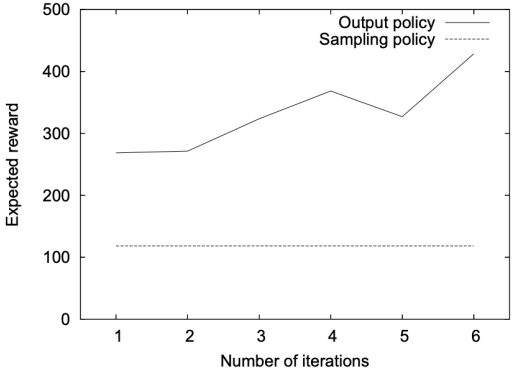
feature	description
Taxpayer features	
num_non_rstrctd_fin_srcs	num non-restricted financial sources
st_inactv_ind	sales tax inactive indicator
num_bnkrptcy_flngs	number of bankruptcy filings
Liability features	
ttl_liability_blnc	total liability balance
sum_cllct_asmts	sum of collectible assessments
sum_asmts_avail_to_wrrnt	sum assessments available to warrant
Transactional features	
tax_pd_lst_yr	tax paid last year
num_pymnts_snc_lst_actn	num of payments since last action
sum_pymnts	sum of payments to date
sum_pymnts_lst_yr	sum of payments last year
Collections features	
num_opn_pfrctd_wrrnts	number of open perfected warrants
dys_snc_lst_wrrnt_pfrctd	days since last warrant perfected

(a) Состояния налогоплательщика

action	description	hours	bound
Collection Actions			
cntct_tp_ml	contact taxpayer by mail	0.01	5000
cntct_tp_phn	contact taxpayer by phone	0.14	2000
crt_wrrnt	create warrant	0.01	5000
crt_ie	create income execution	0.01	10
crt_lvy	create levy	0.09	5000
Movement Actions			
mv_to_do	move to district office	0	5910
mv_to_hivl	move to high value team	0	330
mv_to_cvs	move to collection vendors	0	340
mv_to_ice	move to indiv. case enf.	0	100

(b) Действия агента (налогового управления)

[Abe et al., 2010]: результаты



- Налоговое управление штата Нью-Йорк инвестировало 4 миллиона USD в разработку системы в 2009 году;
- Система в $2009 - 2010$ году позволила увеличить сумму успешно взысканных долгов на 8% (на 83 миллиона USD в аболютных цифрах);

Спасибо за внимание!

Bibliography I

Abe, N., Melville, P., Pendus, C., Reddy, C. K., Jensen, D. L., Thomas, V. P., Bennett, J. J., Anderson, G. F., Cooley, B. R., Kowalczyk, M., Domick, M., and Gardinier, T. (2010). Optimizing debt collections using constrained reinforcement learning.
In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, page 75–84, New York, NY, USA. Association for Computing Machinery.

Khraishi, R. and Okhrati, R. (2022).
Offline deep reinforcement learning for dynamic pricing of consumer credit.
https://arxiv.org/abs/2203.03003.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020).
Conservative q-learning for offline reinforcement learning.
In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Приложения: Уравнения Беллмана

Уравнения Беллмана для политики π:

$$Q^\pi(s,a) = r(s,a) + PV^\pi(s,a)\,,$$
$$V^\pi(s) = Q^\pi(s,\pi(s))\,,$$

где $PV^\pi(s,a) = \sum_{s'} P(s'|s,a)V^\pi(s') = \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\pi(s')]$.

Приложения: Оптимальные уравнения Беллмана

Оптимальная политика $\pi^\star$ - политика, максимизирующая $V^\pi(s)$ для всех состояний $s \in \mathcal{S}$.

Оптимальные V- и Q-функции:

$$V_h^\star(s) = V_h^{\pi^\star}(s), \quad Q_h^\star(s,a) = Q_h^{\pi^\star}(s,a).$$

Оптимальные уравнения Беллмана:

$$\begin{aligned} Q^\star(s,a) &= r(s,a) + PV^\star(s,a)\,, \\ V^\star(s) &= \max_a Q^\star(s,a)\,, \end{aligned}$$

где $PV^\pi(s,a) = \sum_{s'} P(s'|s,a)V^\pi(s') = \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\pi(s')]$.

Оптимальная политика: $\pi^\star(s) = \arg\max_a Q^\star(s,a)$.

Приложения: [Kumar et al., 2020] - консервативное Q-обучение

Что делать, если исторические данные получены с помощью фиксированной политики π_{hist}?

- Пусть $\mathcal{D}$ - распределение исторических данных;
- Функция потерь нового алгоритма:

$$L(\theta) = \alpha \mathbb{E}_{s_t \sim \mathcal{D}} \left[\log \sum_a \exp Q_\theta(s_t, a) - \mathbb{E}_{s,a \sim \mathcal{D}} [Q_\theta(s, a)] - \varkappa \right] + L_{SAC}(\theta),$$

где $L_{SAC}(\theta)$ - функция потерь алгоритма Soft Actor-Critic [Kumar et al., 2020];