Задача оптимального управления в системах искусственного интеллекта с обратной связью

Андрей Сергеевич Веприков
Научный руководитель: д.ф.-м.н. А. С. Хританков

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных

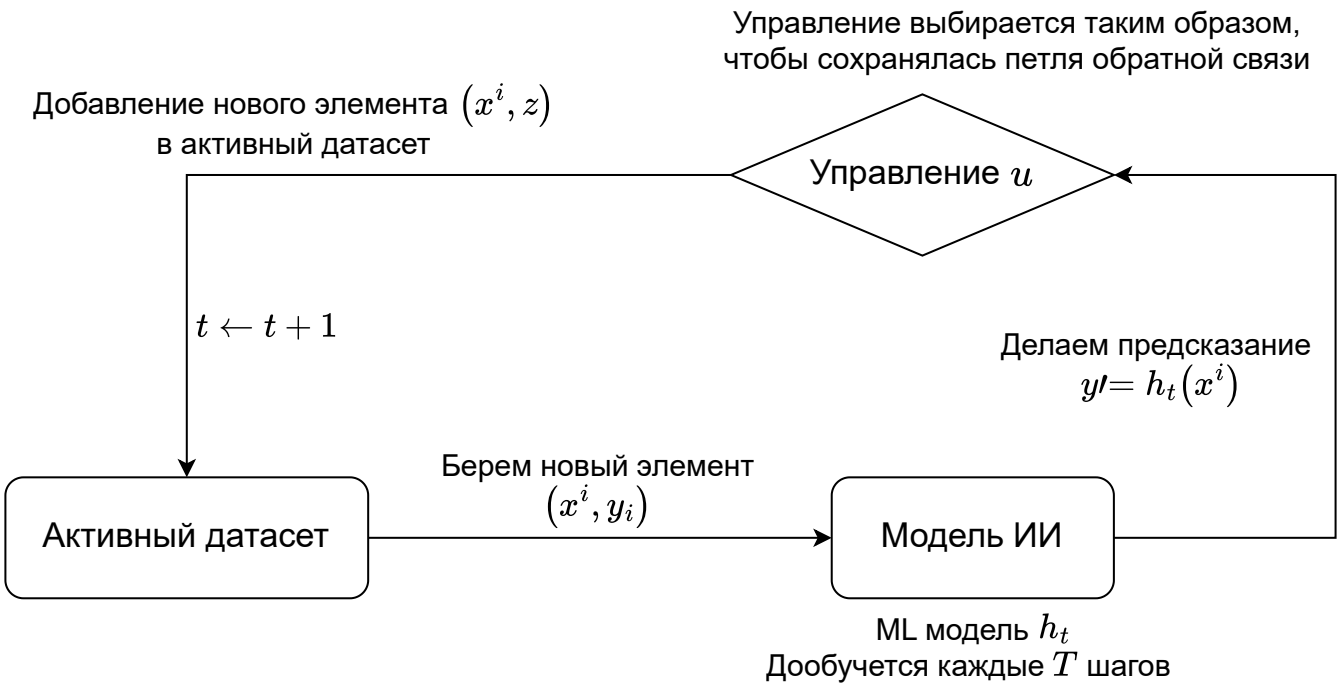
Декабрь 2024

Примеры эффектов петель обратной связи

1. Самоисполняющееся пророчество (self-fulfilling prophecy)
2. Вынужденное смещение данных (data drift) в рекомендательных системах [Khritankov, 2023]
3. Усиление ошибок (error amplification) со временем в задаче медицинского прогнозирования [Adam et al., 2022]
4. Дрейф данных в системах предиктивного полицейского контроля [Ensign et al., 2018]
5. Пузыри фильтров (filter bubbles) [Davies, 2018]

Задача оптимального управления в системах искусственного интеллекта с обратной связью

Пусть дано множество функций плотности состояний системы $\mathcal{F}$, в которое нам необходимо попасть при $t \to \infty$. На каждом шаге t доступна ограниченная выборка данных системы и управление $u_t \in \mathcal{U}$.



Модельная задача оптимального управления в системе ИИ с петлей обратной связи.

Математическая постановка задачи

Определим пятерку: $(\mathcal{S}, \mathcal{U}, \mathbf{D}, \rho, \gamma)$, где

1. $\mathcal{S}$ – функции плотности (состояния)
2. $\mathcal{U}$ – доступные управления (действия)
3. $\mathbf{D}$ – множество всех возможных операторов эволюции (ядро переходов)
4. ρ – функция расстояния до множества $\mathcal{F}$ (награда)
5. $0 < \gamma < 1$ – дисконтирующий коэффициент

Что нам не доступно	Что нам доступно
$f_t \in \mathcal{S}$ $\mathrm{D}_t \in \mathbf{D}$	Ограниченная выборка из $f_t \in \mathcal{S}$ $u_t \in \mathcal{U}$, $\rho(f_t, \mathcal{F})$, γ

Математическая постановка задачи

Оптимизационная задача:

$$\min_{u_1,\ldots,u_\infty} \quad \sum_{t=0}^{+\infty} \gamma^t \rho\left(f_t, \mathcal{F}\right), \tag{1}$$

при условии

$$f_{t+1} = \mathrm{D}_t\left(f_t, u_t\right), \quad \mathrm{D}_t \in \mathbf{D}, \tag{2}$$

$$g(f_t, u_t) \leq 0, \tag{3}$$

где g – функция, показывающая, что при применении управления u_t в системе с данными $\sim f_t$ сохранится петля обратной связи, если всех оставшихся управлений не будет.

Предлагаемые способы решения [TODO]

- Динамическое программирование
- Гауссовские фильтры, PoMDP

Список литературы

Adam, G. A., Chang, C.-H. K., Haibe-Kains, B., and Goldenberg, A. (2022).

Error amplification when updating deployed machine learning models.

In *Machine Learning for Healthcare Conference*, pages 715–740. PMLR.

Davies, H. C. (2018).

Redefining filter bubbles as (escapable) socio-technical recursion.

Sociological Research Online, 23(3):637–654.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018).

Runaway feedback loops in predictive policing.

In *Conference on fairness, accountability and transparency*, pages 160–171. PMLR.

Khritankov, A. (2023).

Positive feedback loops lead to concept drift in machine learning systems.

Applied Intelligence, pages 1–19.