Математическая модель эффекта обратной связи в системах искусственного интеллекта

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

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 03.04.01 Прикладные математика и физика

2024

Цель исследования

В данной работе решается задача математического моделирования систем с адаптивным управлением. Системе с ИИ ставится в соответствие дискретная динамическая система, по поведению которой можно судить об исходном объекте.

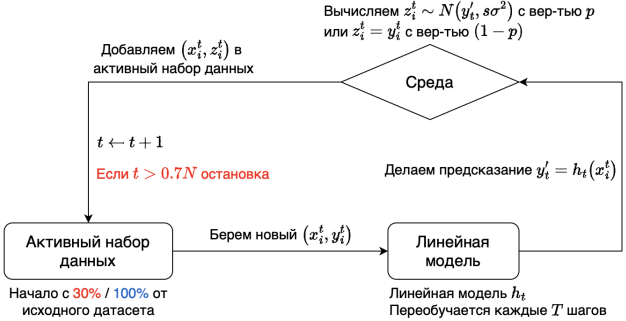

Рис. 1: Две постановки эксперимента. Скользящее окно и обновление выборки.

Примеры процессов многократного машинного обучения:

1. Эффекты петель обратной связи (feedback loop)
2. Усиление ошибок (error amplification)
3. Пузыри фильтров (filter bubbles) и эхо-камеры (echo chambers)

Постановка задачи и теорема о предельном множестве

Определим следующую дискретную динамическую систему:

$$f_{t+1}(x) = \mathrm{D}_t(f_t)(x) \quad \text{для} \quad \forall x \in \mathbb{R}^n, t \in \mathbb{N} \quad \text{и} \quad \mathrm{D}_t \in \mathbb{D}, \tag{1}$$

где D_t обычно называется оператором эволюции, $f_t(x)$ – функции плотности вероятности распределения данных системы, а начальная функция $f_0(x)$ задана.

Теорема 1 (Предельное множество системы (1))

Для любой функции плотности $f_0(x), x \in \mathbb{R}^n$ и дискретной динамической системы (1), пусть существуют $g(x) \in L_1(\mathbb{R}^n)$ и $\psi_t \geq 0$ такие, что $f_t(x) \leq \psi_t^n \cdot |g(\psi_t \cdot x)|$ для всех $t \in \mathbb{N}$ и $x \in \mathbb{R}^n$. Тогда, если ψ_t расходится к ∞, плотности $f_t(x)$ стремятся к дельта-функции, $f_t(x) \underset{t \to +\infty}{\longrightarrow} \delta(x)$ слабо.

Если ψ_t сходится к 0, тогда плотности $f_t(x)$ сходятся к нулевому распределению, $f_t(x) \underset{t \to +\infty}{\longrightarrow} \zeta(x)$ слабо.

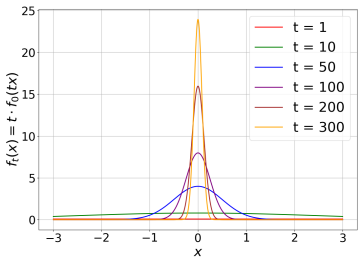

Рис. 2: Пример использования Теоремы 1 с $\psi_t = t$ для $\mathcal{N}(0; 1)$.

Анализ условий существования петель обратной связи и автономности системы (1)

Лемма 1 (Стремление моментов к нулю)

Если система (1) *с* $n = 1$ *удовлетворяет условиям Теоремы 1 и* $\psi_t \to \infty$, *тогда все* $2k$*-тые моменты невязок* $y - h(X)$ *убывают со скоростью как минимум* ψ_t^{-2k}.

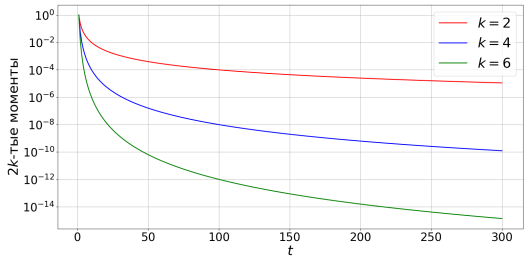

Рис. 3: Пример использования Леммы 1.

Теорема 2 (Критерий автономности)

Если операторы эволюции D_t *динамической системы* (1) *имеют вид* $D_{\overline{1,t}}(f_0)(x) = \psi_t^n \cdot f_0(\psi_t \cdot x)$, *тогда система* (1) *автономна тогда и только когда, когда* $\psi_{\tau+\kappa} = \psi_\tau \cdot \psi_\kappa \ \forall \tau, \kappa \in \mathbb{N}$.

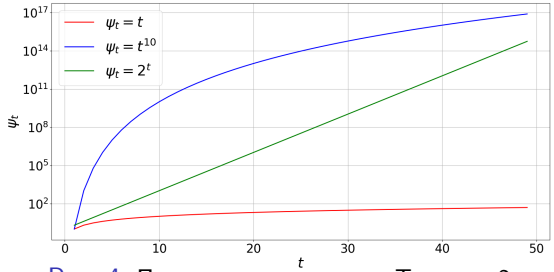

Рис. 4: Пример использования Теоремы 2.

Предел к дельта-функции или нулевому распределению

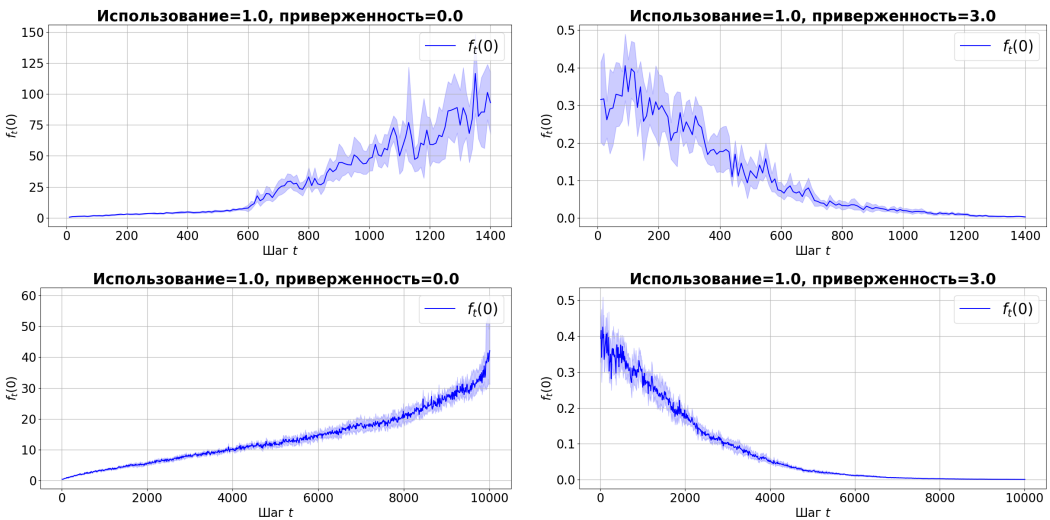

Рис. 5: Постановка скользящее окно (сверху), обновление выборки (снизу).

Стремление моментов к нулю и исследование систем на автономность

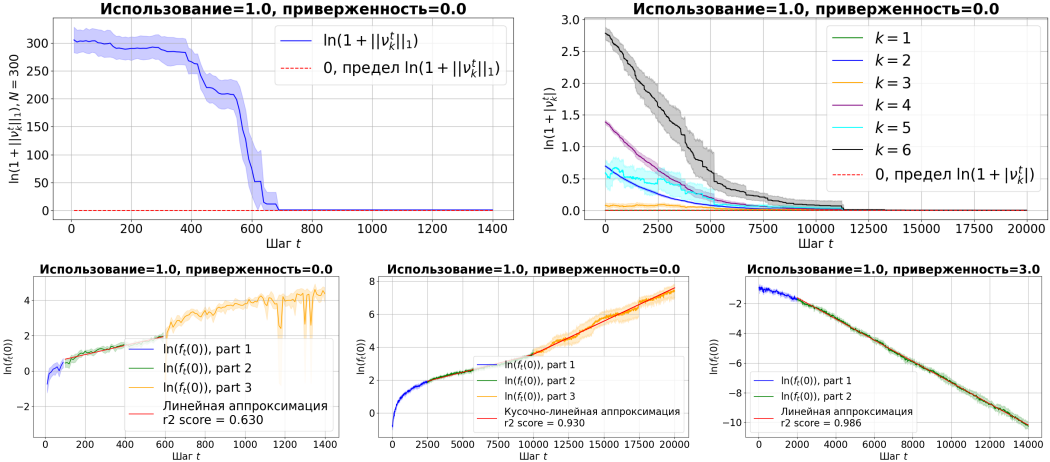

Рис. 6: Стремление моментов к нулю(сверху): скользящее окно(слева), обновление выборки(справа). Проверка автономности(снизу): скользящее окно(слева), обновление выборки(середина и справа).

Выносится на защиту

1. Построена математическая модель эффекта петель обратной связи с использованием дискретных динамических систем
2. Были получены результаты для определения предельного множества динамической системы, достаточных условий существования петли обратной связи и критерий автономности
3. Разработан стенд проведения вычислительных экспериментов, симулирующий процесс многократного машинного обучения

Публикации

1. Veprikov A., Afanasiev A., Khritankov A. A Mathematical Model of the Hidden Feedback Loop Effect in Machine Learning Systems // arXiv preprint `https://arxiv.org/abs/2405.02726`
2. Веприков А. С., Афанасьев А. П., Хританков А.С. Математическая модель эффекта обратной связи в системах искусственного интеллекта // Сборник тезисов 21-й Всероссийской конференции Математические методы распознавания образов (ММРО-21). – Российская академия наук, 2023. – С. 35-37