Математическая модель эффекта обратной связи в системах искусственного интеллекта

Выпускная квалификационная работа бакалавра

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

Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

18 мая 2024 г.

Введение

При применении алгоритмов машинного обучения в реальных задачах часто наблюдается явление, когда среда и результаты работы модели начинают влиять друг на друга. Поэтому важно изучать системы искусственного интеллекта в долгосрочной перспективе. Примерами таких эффектов могут служить

1. Эффекты петель обратной связи (feedback loop) (Khritankov, 2023a,b; Taori et al., 2023)
2. Усиления ошибок (error amplification) (Mansoury et al., 2020; Adam et al. 2022)
3. Пузырей фильтров (filter bubbles) и эхо-камер (echo chambers) (Davies et al., 2018; Terren et al., 2021)

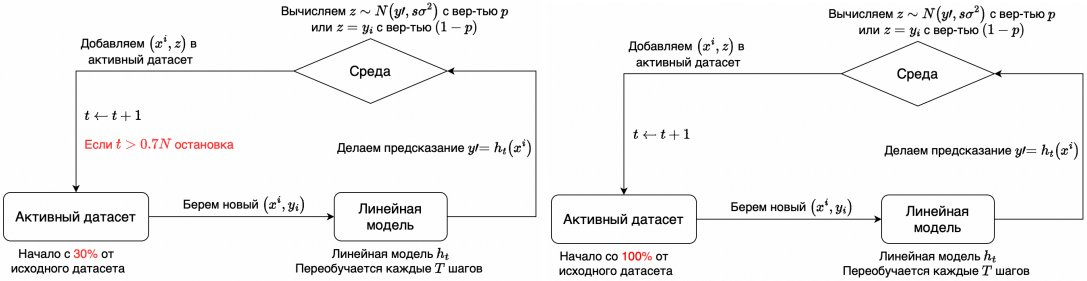


Рис. 1: Две различные постановки эксперимента. Скользящее окно (слева) и обновление выборки (справа).

Постановка задачи и теорема о предельном множестве

Определим следующую дискретную динамическую систему:

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

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 (Veprikov et al., 2024)

Для любой функции плотности $f_0(x), x \in \mathbb{R}^n$ и дискретной динамической системы (1)*, если 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 нулю, тогда плотности $f_t(x)$ сходятся к нулевому распределению, $f_t(x) \underset{t \to +\infty}{\longrightarrow} \zeta(x)$ слабо.

Для задачи регрессии, когда данные имеют вид (X, y) Теорема 1 записывается не для данных в системе, а для случайного вектора невязок модели h, вида $y - h(X)$.

Из Теоремы 1 можно вывести вид огибающих отображений $\mathrm{D}_{\overline{1,t}}(\cdot) := \mathrm{D}_t(\mathrm{D}_{t-1}(...\mathrm{D}_1(\cdot)...))$, которые имеют вид

$$\mathrm{D}_{\overline{1,t}}(f_0)(x) = \psi_t^n \cdot f_0(\psi_t \cdot x) \quad \forall x \in \mathbb{R}^n \text{ и } \forall t \in \mathbb{N}. \tag{2}$$

Если операторы эволюции системы (1) имеют вид (2), можно записать формулу для вычисления ψ_t:

$$\psi_t \simeq f_t(0). \tag{3}$$

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

Предположение 1, сформулированное в [1], гласит, что в системе (1) существует петля положительной обратной связи, если оператор D_t является сжимающим в метрическом пространстве предсказаний модели.

Лемма 1 (Veprikov et al., 2024)

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}.

Если оператор эволюции системы (1) *удовлетворяет* (2), *тогда все (не только четные) моменты* $y - h(X)$ *убывают со скоростью* ψ_t^{-k}.

Если существует $q \in [1; +\infty]$ *такой, что* $\{\nu_k^0\}_{k=1}^{+\infty} \in l_q$ *и оператор эволюции системы* (1) *удовлетворяет* (2), *тогда* $\{\nu_k^t\}_{k=1}^{+\infty} \in l_1$ *и* $\{\nu_k^t\}_{k=1}^{+\infty} \xrightarrow[t\to\infty]{l_1} 0$.

Из Теоремы 1 следует, что существует специальный вид отображений (2), для таких отображений можно вывести критерий автономности системы (1).

Теорема 2 (Veprikov et al., 2024)

Если операторы эволюции D_t *динамической системы* (1) *имеют вид* (2), *тогда система* (1) *автономна тогда и только когда, когда*

$$\psi_{\tau+\kappa} = \psi_\tau \cdot \psi_\kappa \ \forall \tau, \kappa \in \mathbb{N}. \tag{4}$$

Сохранение предельного множества при преобразовании признаков

Анализируется оператор G, осуществляющий преобразование $(X, y) \xrightarrow{G} (X', y')$ пространства признаков X и целевой переменной y для модели машинного обучения в задаче обучения с учителем.

Лемма 2 (Вериков и др., 2023)

1. *Если* $\exists\, T \in \mathbb{N}$: *для* $\forall t \geq T$ *выполнено* $L(y_t, h_t, X_t) \geq L(y'_t, h'_t, X'_t)$ *и функции плотности* $y_t - h_t(X_t) \to \delta(x)$, *то функции плотности* $y'_t - h'_t(X'_t) \to \delta(x)$.
2. *Если же* $\exists\, T \in \mathbb{N}$: *для* $\forall t \geq T$ *выполнено* $L(y_t, h_t, X_t) \leq L(y'_t, h'_t, X'_t)$ *и функции плотности* $y_t - h_t(X_t) \to \zeta(x)$, *то функции плотности* $y'_t - h'_t(X'_t) \to \zeta(x)$.

Квадратичная (MSE) и другие функции потерь, часто применяемые в задаче регрессии, могут быть разложены на составляющие смещения, разброса и неустранимого шума в данных:

$$L(y, h, X) = \text{Bias}^2(y, h, X) + \text{Var}(y, h, X) + \sigma^2.$$

Лемма 3 (Вериков и др., 2023)

1. *Если* $\exists\, T \in \mathbb{N}$: *для* $\forall t \geq T$ *преобразование* G *уменьшает сумму смещения и разброса на шаге* t, *тогда* $\forall t \geq T$ *выполнено* $L(y_t, h_t, X_t) \geq L(y'_t, h'_t, X'_t)$.
2. *Если же* $\exists\, T \in \mathbb{N}$: *для* $\forall t \geq T$ *преобразование* G *увеличивает сумму смещения и разброса на шаге* t, *тогда* $\forall t \geq T$ *выполненно* $L(y_t, h_t, X_t) \leq L(y'_t, h'_t, X'_t)$.

Предел к дельта-функции или нулевому распределению

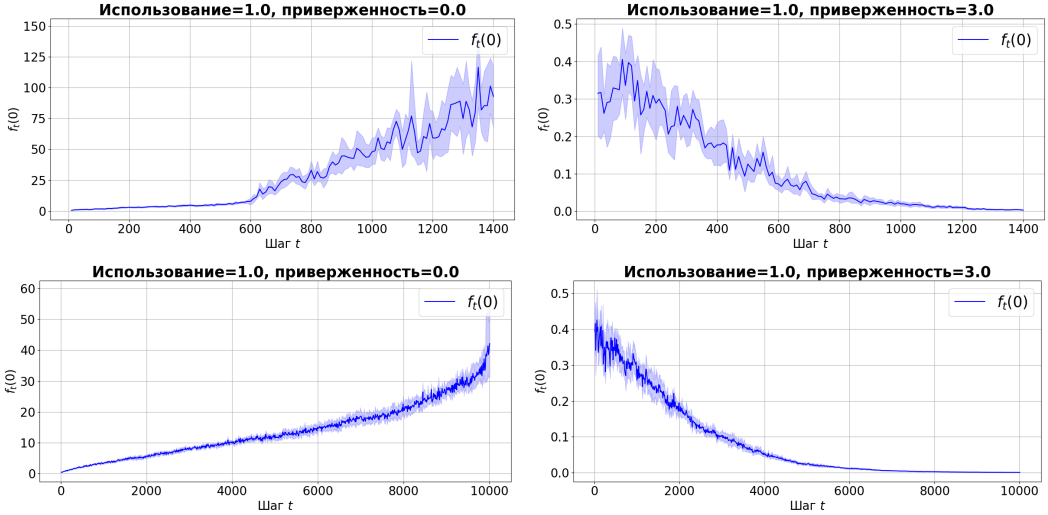


Рис. 2: Постановка скользящее окно (сверху), обновление выборки (снизу).

Стремление моментов к нулю и исследование систем на автономность

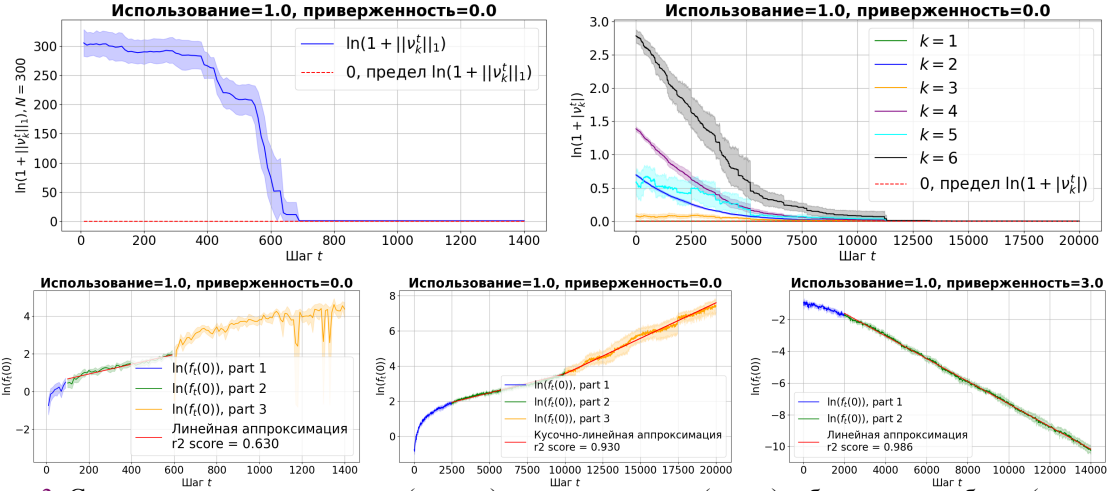
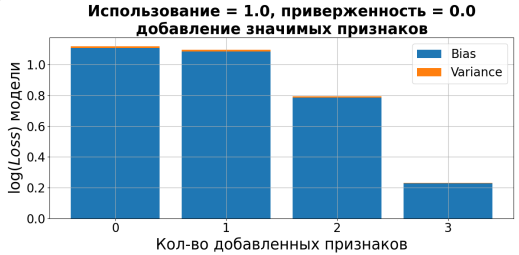
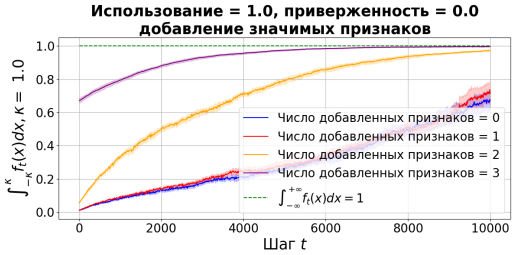
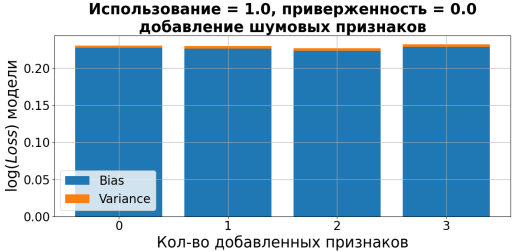


Рис. 3: Стремление моментов к нулю(сверху): скользящее окно(слева), обновление выборки(справа). Проверка автономности(снизу): скользящее окно(слева), обновление выборки(середина и справа).

Сохранение предельного множества при преобразовании признаков







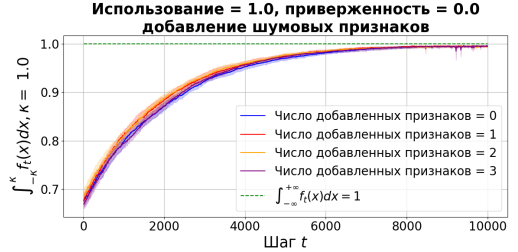


Рис. 4: Добавление информативных (сверху) и неинформативных (снизу) признаков.

Выносится на защиту

1. В данной работе 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, критерий автономности и достаточные условия сохранения предельного множества при преобразовании признаков
3. В данной работе на языке Python разработан стенд проведения вычислительных экспериментов, симулирующий процесс многократного машинного обучения
4. Полученные результаты применимы для контроля качества систем ИИ в медицине, персонализации образования, широком классе СППР для выполнения требований кодекса этики в сфере ИИ и требований ГОСТ
5. Мой вклад в данное исследование состоит в доказательстве теоретических исследований, участие в постановке экспериментов, сборе, обработке и анализе их результатов
6. По результатам работы была подана статья в Q1 журнал (Veprikov et al., 2024), рассказано выступление на конференции ММРО-2023 (Веприков и др., 2023), материалы которой были поданы в журнал ИИПР

Список литературы

1 (Veprikov et al., 2024) 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 и др., 2023) 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

3 (Khritankov, 2023) Khritankov A. Hidden feedback loops in machine learning systems: A simulation model and preliminary results // Software Quality: Future Perspectives on Software Engineering Quality: 13th International Conference, SWQD 2021, Vienna, Austria, January 19–21, 2021, Proceedings 13. –Springer International Publishing, 2021. –С. 54-65.

4 (Taori et al., 2023) Taori R., Hashimoto T. Data feedback loops: Model-driven amplification of dataset biases // International Conference on Machine Learning. –PMLR, 2023. –С. 33883-33920.

5 (Adam et al., 2022) Adam G. A. et al. Error amplification when updating deployed machine learning models // Machine Learning for Healthcare Conference. –PMLR, 2022. –С. 715-740.

6 (Terren et al., 2021) Terren L. T. L., Borge-Bravo R. B. B. R. Echo chambers on social media: A systematic review of the literature // Review of Communication Research. –2021. –Т. 9.

7 (Mansoury et al., 2020) Mansoury M. et al. Feedback loop and bias amplification in recommender systems //Proceedings of the 29th ACM international conference on information and knowledge management. –2020. –С. 2145-2148.

8 (Davies et al., 2018) Davies H. C. Redefining filter bubbles as (escapable) socio-technical recursion // Sociological Research Online. –2018. –Т. 23. –№. 3. –С. 637-654.