On the problem of repeated supervised learning

My first scientific paper

Andrey Veprikov Anton Khritankov Alexander Afanasyev

MIPT
Dolgoprudny, Russia

Spring 2023

Introduction and Related work

In this research paper, we delve into the intricacies of continuous learning artificial intelligence systems as they interact with and influence their environment. Repeated supervised learning appears in many machine learning applications, for example in

1. recommendation systems [5]
2. healthcare [1]
3. predictive policing [3]

Contributions of this paper are as follows

1. Develop a mathematical model to examine the process of repeated and multiple learning, prediction, and dataset updating.
2. Find necessary and sufficient conditions for various effects associated with repeated machine learning that can be tested in experiments
3. Conduct several synthetic experiments based on our findings, hoping to contribute to a better understanding of the behavior of continuous learning AI systems.

Problem statement

The object of our research will be the set **R** of distribution density functions

$$\mathbf{R} := \left\{ f : \mathbb{R}^n \to \mathbb{R}_+ \text{ and } \int\limits_{\mathbb{R}^n} f(x)dx = 1 \right\}$$

and a mapping D as feedback loop mapping.
We consider an autonomous (time-independent explicitly) discrete dynamical system. There is a dedicated variable - the step number, which increases, at step t and $t+1$ of which the ratio is fulfilled

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

where D becomes an evolution operator on the space of the specified functions f and the initial function $f_0(x)$ is known.
According to Conjecture 1 from [4] the positive feedback loop in a system $\mathrm{D} : R \to R$ exists if D is a contraction mapping, but this conjecture was not proved.

Assumptions for D : $\mathbf{R} \to \mathbf{R}$

Theorem 1

If $\|\mathrm{D}\|_1 = 1, \forall f \in \mathbf{R} \hookrightarrow \mathrm{D}(f)(x) \geq 0$ for almost every $x \in \mathbb{R}^n$, and exists D^{-1} such that $\|\mathrm{D}^{-1}\|_1 \leq 1$, then D : $\mathbf{R} \to \mathbf{R}$.

Discussion

In experiments it often difficult to calculate D^{-1} and especially it's norm, so we make a different assumptions. Distribution of our data is approximating by empirical distribution function [2] as follows (for $n = 1$):

$$\hat{F}_N(x) := \frac{\text{number of elements in sample} \leq x}{N} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{X_i \leq x},$$

where X_i are elements of sample. We assume that $(X_1, X_2, X_3, ..., X_N)$ are i.i.d. real random variables.

In this case operator D transoms our data, i.e. translates one empirical distribution function to another. So D : $\mathbf{R} \to \mathbf{R}$ by constructing our experiment.

Inequality on $\|\mathrm{D}\|_q$

Theorem 2

Consider

$$f_A(x) = \frac{1}{\lambda(A)} \cdot \mathbf{1}_A(x),$$

where $A \subset \mathbb{R}^n$ is arbitrary set of a non-path measure, $\lambda(A)$ – the measure of a set A. Then for all $A \subset \mathbb{R}^n : 0 < \lambda(A) < +\infty$ and for all $1 \leq q \leq +\infty$ such that $\mathrm{D}(f_A) \in L_q(\mathbb{R}^n)$ is fulfilled that

$$\|\mathrm{D}\|_q \geq \int\limits_A \mathrm{D}(f_A)(x)dx$$

Discussion

To begin with, note that the result of this Theorem does not depend in any way on whether D converts **R** to **R**.

If you look at it from a probabilistic point of view, then f_A is a distribution density function of vectors uniformly distributed on a set A. If $\|\mathrm{D}\|_q \geq 1$, then D wouldn't be a contraction mapping in $\|\cdot\|_q$, because there always would be function $f \in \mathbf{R}$ such that $\|\mathrm{D}(f)\|_q \geq \|f\|_q$.

Limit in a weak sense to δ function

Theorem 3

If $f_t : \mathbb{R} \to \mathbb{R}$ such that $\forall t \in \mathbb{N} \hookrightarrow \|f_t\|_1 = 1$, $f_t(x) \geq 0$ in almost every point $x \in \mathbb{R}$ and

$$\exists \psi : \mathbb{N} \to \mathbb{R} : \psi(t) \xrightarrow[t \to +\infty]{} +\infty \text{ and } \exists g \in L_1(\mathbb{R}) \text{ such that}$$

$$\forall t \in \mathbb{N}\ \forall y \in \mathbb{R} \hookrightarrow f_t\left(\frac{y}{\psi(t)}\right) \leq \psi(t) \cdot |g(y)|$$

Then $f_t(x) \xrightarrow[t \to \infty]{} \delta(x)$ in a weak sense, i.e.

$$\lim_{t \to +\infty} \left(\int\limits_{-\infty}^{+\infty} f_t(x)\phi(x)dx \right) = \phi(0),$$

where ϕ is continuous function with compact support

Discussion

If in some step t the model $H(x, \theta, t)$ starts to give good predictions , then, in the probabilistic formulation, this means that the density function of the distribution of the object-sign vectors becomes similar to the delta function. Based on these considerations, we can compare each operator D to an operator $\widetilde{\mathrm{D}}$, where $\widetilde{\mathrm{D}}$ transforms $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ at each step, where $\mathbf{y}_{\text{pred}} = H(\mathbf{X}, \theta, t)$.

Consequences of the Theorem 3

Lemma 1

If $\forall x \neq 0 \hookrightarrow f_t(x) \underset{t\to\infty}{\longrightarrow} 0$ and $f_t(0) \underset{t\to\infty}{\longrightarrow} +\infty$, then we can take $\psi(t) = \dfrac{f_t(0)}{|g(0)|}$

In the following we assume that the operator D has the form

$$\mathrm{D}^t(f_0)(x) = \psi(t) \cdot f_0(\psi(t)x) \tag{1}$$

Lemma 2

$\{\mathrm{D}^t\}_{t\in\mathbb{N}}$ is semigroup, i.e. $\mathrm{D}^\tau \circ \mathrm{D}^\kappa = \mathrm{D}^{\tau+\kappa}$, if and only if $\psi(\tau+\kappa) = \psi(\tau)\cdot\psi(\kappa)$

Lemma 3

If D has the form (1), then D becomes an isometry in the Kullback–Leibler divergence, i.e. $\forall f_1, f_2 \in \mathbf{R} \hookrightarrow \rho(f_1, f_2) = \rho(\mathrm{D}(f_1), \mathrm{D}(f_2))$

Lemma 4

If D has the form (1), then $\|\{\nu_k^t\}_0^{+\infty} - e_0\|_1 \underset{t\to\infty}{\longrightarrow} 0$, where ν_k^t – k-th moments of a random variable $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ on step t and $e_0 = (1, 0, 0, 0, ...)$

Lemma 5

If D has the form (1) and we consider a set $\mathbf{R}^* := \left\{ \phi(s) := \int\limits_{-\infty}^{+\infty} e^{isx} f(x)dx \;\middle|\; f \in \mathbf{R} \right\}$ – set of characteristic functions. Then for all $\phi_0 \in \mathbf{R}^*$ is fulfilled that $\mathrm{D}^t(\phi_0)(s) \underset{t\to\infty}{\longrightarrow} \phi_\delta(s) = 1 \; \forall s \in \mathbb{R}$

Important example of ψ and g

Important example of operator D that translates any function from **R** into a δ function is as follows

$$\mathrm{D}^t(f_0)(x) = t \cdot f_0(t \cdot x)$$

Here we take $g(x) = f_0(x)$ and $\psi(t) = t$.

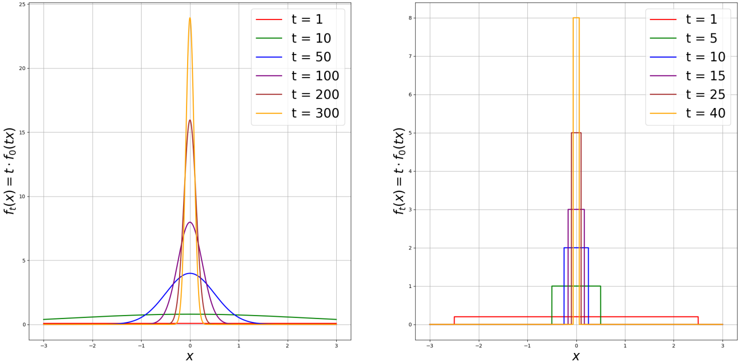
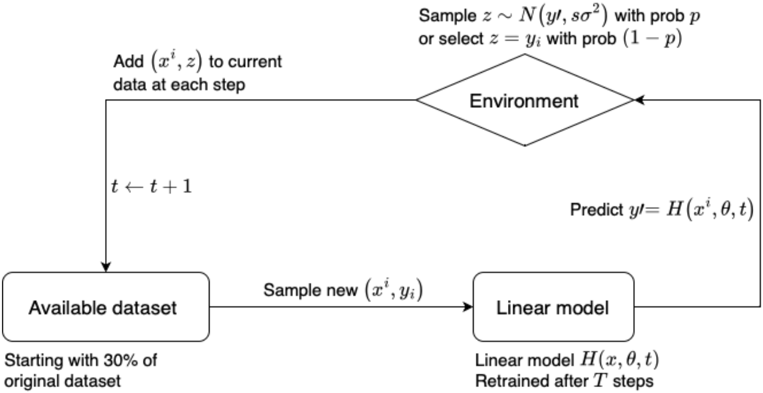


Illustration of weak limit to δ function. $\mathcal{N}(0, 5)$ left, $\mathcal{U}[-2.5, 2.5]$ right.

Experiment Design

In this paper we consider linear model, that is solved as Ridge regression with mean squared error loss function. In the formulation of our experiment, we consider $\mathbf{R}$ as space of density functions of random vectors $(\mathbf{x^i}, y_i)$, where $\mathbf{x^i} \in \mathbf{X}$ and $y_i \in \mathbf{y}$ – the features and target variable of the i-th object.



Sample $z \sim N(y\prime, s\sigma^2)$ with prob p
or select $z = y_i$ with prob $(1-p)$

Environment

Add (x^i, z) to current
data at each step

$t \leftarrow t+1$

Predict $y\prime = H(x^i, \theta, t)$

Available dataset

Sample new (x^i, y_i)

Linear model

Starting with 30% of
original dataset

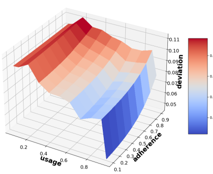
Linear model $H(x, \theta, t)$
Retrained after T steps

Experiment setup.

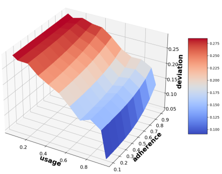
Dispersion decrease

We measured the standard deviation of $\mathbf{y} - \mathbf{y}_{\text{pred}}$ at different *usage* – the probability with which we take $(\mathbf{x}^{\mathbf{i}}, z_i)$ into the active dataset and *adherence* – the parameter by which we multiply σ^2 when sampling z_i.

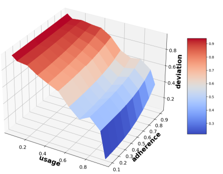
Deviation in original data = 0.1



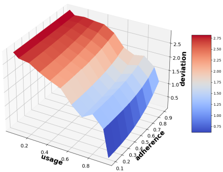
Deviation in original data = 0.3



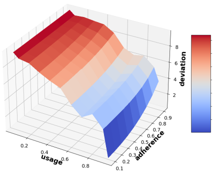
Deviation in original data = 1



Deviation in original data = 3



Deviation in original data = 10



Conclusion

1. In this paper we develop a mathematical model to examine the process of repeated learning.
2. We provide several Theorems and Lemmas based on our mathematical model for better understanding how multiple learning can improve our models.
3. We conduct some synthetic experiments based on our findings, in which we tested the significance of our theoretical results.

References

[1] George Alexandru Adam, Chun-Hao Kingsley Chang, Benjamin Haibe-Kains, and Anna Goldenberg. Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. In *Machine Learning for Healthcare Conference*, pages 710–731. PMLR, 2020.

[2] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[3] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, pages 160–171. PMLR, 2018.

[4] Anton Khritankov. Hidden feedback loops in machine learning systems: A simulation model and preliminary results. In *Software Quality: Future Perspectives on Software Engineering Quality: 13th International Conference, SWQD 2021, Vienna, Austria, January 19–21, 2021, Proceedings 13*, pages 54–65. Springer, 2021.

[5] Anton Khritankov and Anton Pilkevich. Existence conditions for hidden feedback loops in online recommender systems. In *Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part II 22*, pages 267–274. Springer, 2021.