Influence of hyperparameters on aggregating predictions of infinite number of experts

Kunin-Bogoiavlenskii Sergey
Expert: R. D. Zukhba
Consultant: A. V. Zukhba

Moscow Institute of Physics and Technology

kunin-bogoiavlenskii.sm@phystech.edu

May 17, 2024

Contents

What can be forecast?

> Tell us what the future holds, so we may know that you are gods.
>
> *Isaiah 41:23*

- Weather conditions
- Economic trends
- Technology advancements
- Consumer behavior
- Population growth
- Political elections outcomes

Targets

Prediction is very difficult,
especially if it's about the future.

Niels Bohr

1. Time series generator implementation
2. Aggregating algorithm implementation
3. Experiments with various hyperparameters

Problem statement

> There are two kinds of forecasters: those who don't know, and those who don't know they don't know.
>
> *John Kenneth Galbraith*

Data

It is assumed that there are multiple generators, whose structure is unknown to the predictors. These generators switch, producing a time series that is subdivided into a sequence of segments - areas of stationarity, which can be studied using machine learning methods.

Gerators implemented:

- Linear
- ARMA

Problem statement

Terms

- X — signals space
- Y — responses space
- $\mathcal{N}$ — set of experts, indexed by natural numbers
- D — desicion space, to which predictions belong
- $\lambda : \mathrm{D} \times \mathrm{Y} \to \mathbb{R}_+$ — nonnegative loss function
- $L^i_T = \sum\limits_{t=1}^{T} l^i_t$ — cumulative loss of expert i during the first T steps
- $H_T = \sum\limits_{t=1}^{T} h_t$ — master's cumulative loss during the first T steps
- $R_T = H_T - L_T$ — master's regret relative to the best partition, where L_T is the cumulative loss of the best partition.

Problem statement

Algorithm

FOR $t = 1, 2, \ldots$:

1. Expert f^t initialization
2. Experts' predictions $f_t^i = f_t^i(x_t),\ 1 \le i \le t$
3. Master's prediction evaluation $\gamma_t = \mathrm{Subst}(\mathbf{f_t}, \widehat{\mathbf{w}}_\mathbf{t})$
4. Computation of master's loss $h_t = \lambda(p_t, y_t)$ and experts' losses l_t^i
5. **Loss Update** weights modification
6. **Mixing Update** weights modification

ENDFOR

Experiments

Metric — R_T, the regret

Initialization weights

Default weights: $w_1^i = \frac{1}{(i+1)\ln^2(i+1)}$

Experimental: $\frac{1}{i^\alpha}$, $\frac{1}{c}$, $\frac{1}{(i+4)\ln(i+4)\ln^2\ln(i+4)}$, etc.

Noise

Different noise variance leads to diverse ability of experts to train, which opens curious quialities of the master algorithm

Window size

As the algorithm does not know the locations of generator switches, finding an optimal training window is also a challenge.

Experiments

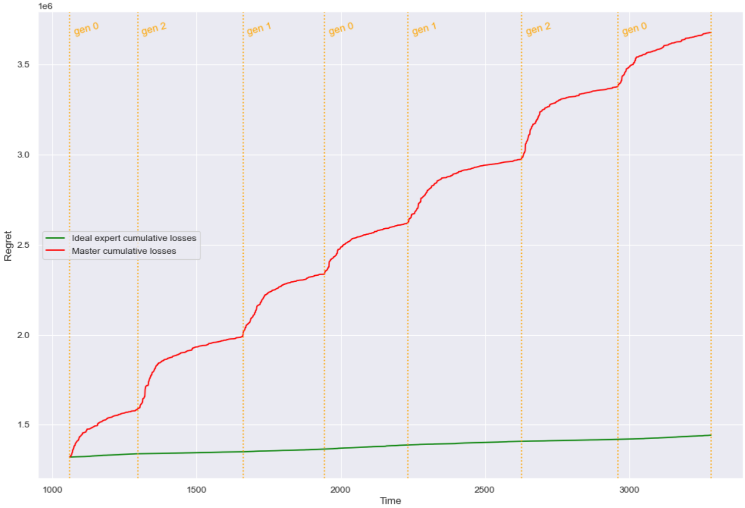
Mixing update scheme

- Start Vector Share - default scheme in GMPP
- Uniform Past Share
- Decaying Past Share
- Increasing Past Share - new proposed scheme

Mixing update coefficients

Default coefficient: $\alpha_t = \frac{1}{t+1}$
Experimental: $\frac{1}{(t+1)^\beta}$, $\frac{1}{c}$, $\frac{1}{(t+c)}$, $\frac{1}{e^{t/3}}$, etc.

Losses plot


1e6
gen 0
gen 2
gen 1
gen 0
gen 1
gen 2
gen 0
3.5
3.0
2.5
2.0
1.5
Regret
Ideal expert cumulative losses
Master cumulative losses
1000
1500
2000
2500
3000
Time

The End