esade

Sequential Copies

11th of October 2022

ESADE & UB

Index

01

Introduction to ML & Copies

Machine learning

- It studies algorithms and statistical models to perform a specific task in the absence of explicit rules and by relying on individual patterns and inference instead
- Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.

Machine learning minimum definition

- A task *t* to solve
- Given these examples: {data set}
- And an error metric: *M*
- Learn a function *f* that minimizes M on {data set}

Supervised Learning I

Let's provide the minimum notation:

- Dependent variable is denoted as Y
- Independent variables, are denoted as $X_1, \ldots, X_p$ respectively. For short, we write $X = (X_1, \ldots, X_p)$
- It is assumed that Y and X are related as follows:

$$Y = f(X) + \epsilon$$

where f is a function that relates the variables in X with the dependent variable Y.

Supervised Learning II

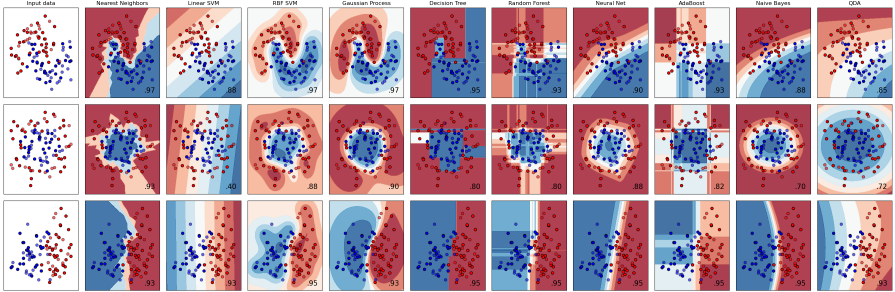
Supervised learning task

Find a function $\hat{f}$ that approximates f as well as possible, based on the examples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ that we are given

- Then $\hat{f}$ can be used to predict the dependent variable on any given input
- $\hat{f}$ is what we call the model
- The process by which $\hat{f}$ is built is called training.

Decision boundary

Supervised machine learning models define a decision boundary that separates the data space into different regions, according to the different values of the dependent variable *y*.



Single-pass copy

$$f_C^*(\theta^*, x) = f_O(x)$$

- Copying problem: $\theta^* = \arg\max_\theta \int_{z \in S} P(\theta | f_O(z))\ dz$
- Need for unlabeled data: $S = \{s_j\}_{j=1}^N | s_j \in X$
- Copying problem reformulation: $\theta^* = \arg\max_\theta \sum_{z \in S} P(\theta | f_O(z))$

02

Sequential Copies

From a set to a sequence of subsets

The problem is we cannot generate a N large enough to solve the copying problem in a single step. Therefore, we must generate a sequence of smaller sets as follows

$$S_i \subseteq S_{i+1} \subseteq \cdots \subseteq S$$

Then, the copying problem is reformulated as

$$\theta_i^* = \arg\max_{\theta} \sum_{z \in S_i} \mathcal{P}(\theta | f_O(z), f_C(\theta_{i-1}^*, z))$$

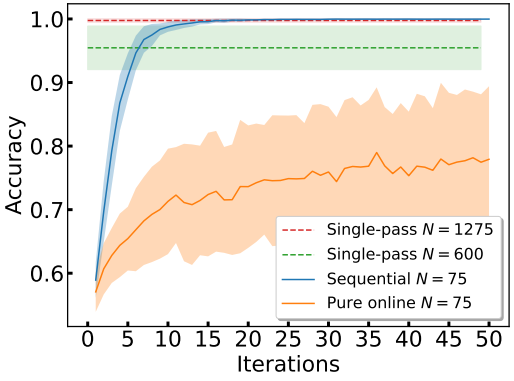
It can be proof that a sequence of parameters $\left\{\theta_i^*\right\}_i$ defined as $\theta_i^* = \arg\max_{\theta \in \Theta} F_i(\theta)$, converges to $\theta^* = \arg\max_{\theta \in \Theta} F(\theta)$, where Θ is the complete parameter set.

Sequential versus single-pass/online I

	Single-pass	Sequential	Online
N	Large and constant	Increase monotonically	Small and constant
Accuracy	High N-dependent	Increase per iteration	Increase per iteration Upper bound N-dependent
Consumption	$T \propto O(t \cdot N)$	$T \propto O(t^2 \cdot N)$	$T \propto O(t \cdot N)$

Sequential versus single-pass/online II

$f_O(SVC_{rbf}\ ,\ X)$



30 runs of $f_C(NN\ ,\ Z)$

Takeaways

- The sequential approach is an intermediate and flexible solution to mitigate the single pass and pure online problems.
- It has a square computational time.
- It needs the same memory as the single pass.
- Solution → Reduce the number of data points per iteration.

03

Uncertainty

Model compression

- ML aims to compress the relevant information contained in the data.
- It is not trivial to define a compression measure.

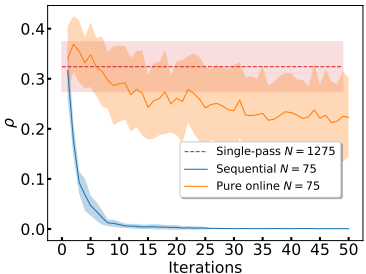
Model compression in the copy framework

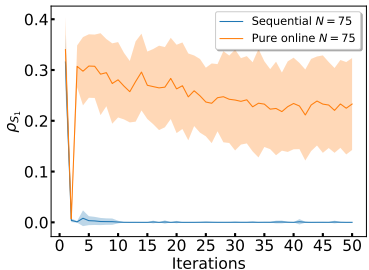
Assuming the original is perfect, i.e., the original model has zero uncertainty when processing new synthetic data, we can build a compression estimator based on the uncertainty that the copy has concerning the original model.

Using model compression to evaluate copy uncertainty

We can evaluate uncertainty using a norm. For example, the normalized euclidean norm of the distance between the outputs of the original and copy at iteration *i* is as follows

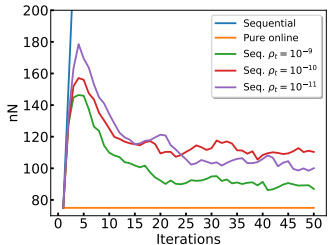
$$\rho_i(\theta^*_{i-1}, z^k_j) = \frac{\left| f_C(\theta^*_{i-1}, z^k_j) - f_O(z^k_j) \right|}{\sqrt{n_c}} \in [0, 1]$$

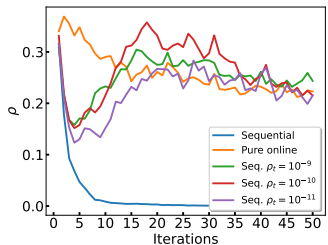


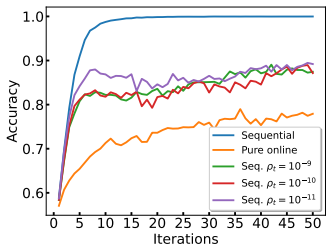


Uncertainty as a dropping points measure

- As observed, the sequential approach has small uncertainty.
- We can fix a threshold for ρ and remove those data points below the threshold.
- Dropping reduces time and memory costs.







Takeaways

- The sequential approach without dropping has no clear advantages regarding computational costs.
- Uncertainty management reduces time and memory costs.
- We observe a trade-off between accuracy and dropping due to forgetting.

04

Model forgetting

Catastrophic forgetting

Problem description

When training on new tasks or categories, online ML models tend to forget the previously learned information, which means new data will override the previously learned model parameters degrading model performance for the past data points.

Without fixing this problem, online ML models cannot have long-term memory because they forget the existing knowledge when learning new data points.

Forcing the model to remember

Since we proved the parameter convergence in the sequential approach, we can force model long-term memory limiting its capacity to update its parameter from two consecutive iterations as

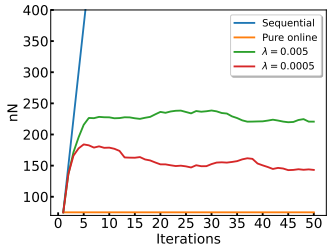
$$\left|\left|\theta^*_{i+1} - \theta^*_i\right|\right| \longrightarrow 0$$

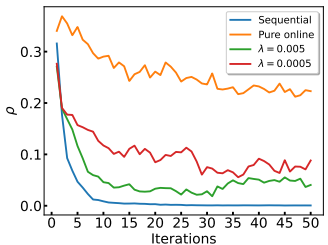
We can include this restriction in the parameter optimization process (loss function or regularization term).

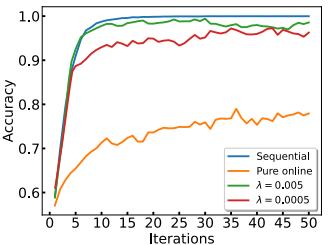
Combining uncertainty and long-term memory

A possible loos function including uncertainty and memorizing can be as follows

$$\mathcal{L} = \sum_{j \in S_i^*} \rho_i(\theta_{i-1}^*, z_j) + \lambda \cdot \left|\left|\theta_i - \theta_{i-1}^*\right|\right|$$







$\rho = 1e^{-10}$ for all runs

Takeaways

- Forcing the sequential approach to remember improves accuracy.
- Using λ we can control computational costs (memory and time).
- Then the sequential approach is a suitable alternative to the single-pass and purely online policies.

05

Conclusions

Conclusions

- Single-pass and online methodologies have good properties and drawbacks.
- Sequential approach is a time-consuming intermediate approach.
- To mitigate its computational cost, we introduce an uncertainty measure for dropping unnecessary data points.
- Dropping reduces model accuracy, so we force the sequential approach to remember limiting parameter updates.
- Combining uncertainty and memory accuracy increases.

Next Steps

- Complete experiments with all UCI databases
- Create a clean GitHub project
- Study incremental learning scenario
- Interpretable sequential copies

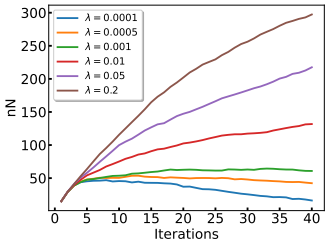
esade

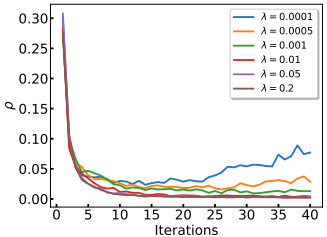
Questions?

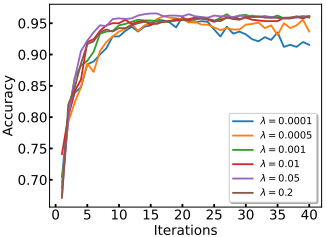
UCI: iris

The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (setosa, virginica and versicolor).

parameter	value	parameter	value
n_samples_iter	15	max_iter	40
learning_rate	0.0005	n_epochs	1000
n_runs	30	batch_size	32
thresh	1e-10		







UCI: wine

The wine dataset contains the results of a chemical analysis of wines grown in a specific area of Italy. Three types of wine are represented in the 178 samples, with the results of 13 chemical analyses recorded for each sample.