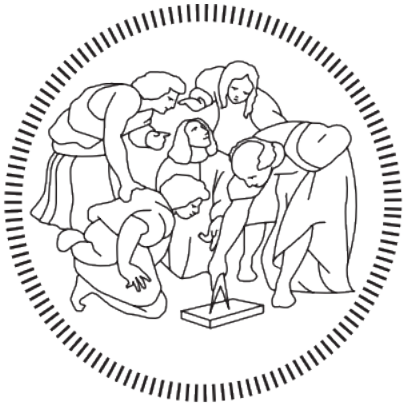


POLITECNICO
MILANO 1863

Financial Engineering

Assignment 2 EPLF, Group 16

2023-2024

Alice Vailati - CP: 10683600 - MAT: 222944
Andrea Tarditi - CP: 10728388 - MAT: 251722
Jacopo Stringara - CP: 10687726 - MAT: 222456
Nicolò Toia - CP: 10628899 - MAT: 247208

Introduction

In this Assignment we perform a random search in the hyper-parameters space of a DNN and comment on the results obtained through Optuna. In particular, the objective is to minimize the loss function by changing the values of two hyper-parameters: the Learning Rate and the size of the Hidden Layers.

Learning rate

The Learning Rate is a hyper-parameter that determines the size of the steps taken when applying the gradient descent algorithm.
In our case, the loss function that we want to minimize is:

$$R(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$$

The following example is a oversimplification of the gradient descent algorithm and it shows how the learning rate is a key ingredient in correctly minimizing the loss function R(θ):

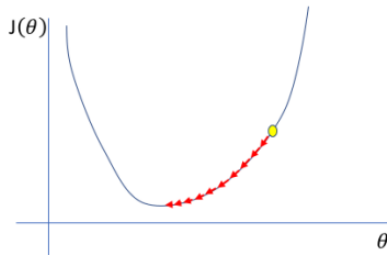


Figure 1: low learning rate

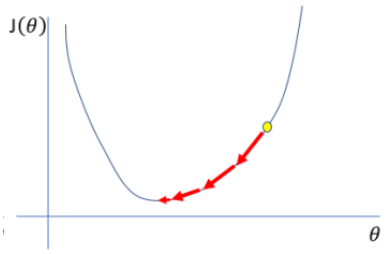


Figure 2: correct learning rate

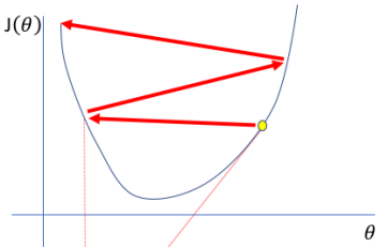


Figure 3: high learning rate

As we can see from these brief examples if the learning rate is too small, it can lead to slow convergence; on the other hand, if it's too big the model learns from it's noise and it doesn't converge.

Hidden size

The Hidden Size is the number of neurons in the hidden layers of the Neural Network: if we choose a size too small for this parameter, we risk under fitting the data (and thus not learning the underlying pattern in the data); on the other hand, a too large size leads to over fitting and thus not only learning from data but also from noise.
Let's now consider a single Neuron, its output will be:

$$l_1 = g(x_i \cdot w_1 + b_1)$$

where g is a simple non linear transformation, l_1 is the output of the Neuron and x_i is the input.
Every Neuron takes a small linear transformation and a non linear one, the power of a Neural network is the concatenation of these transformations through out the hidden size and in each layer.
The next layer of the transformations is:

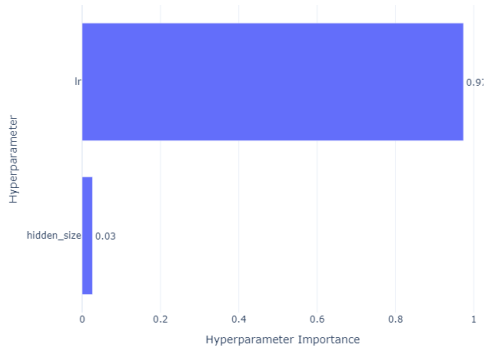
$$l_2 = g(l_1 \cdot w_1 + b_2) \cdot + b_3$$

The number of link's between a set of layers is so dependent from the number of Neuron's (hidden size) in each layer and therefore it's a Key parameter in the characterization of a Neural Network.

For different choices of g we could obtain different solutions: g is commonly called activation function and could differ for every Neuron.

Analyze results

From Optuna's dashboard, we observe that there is a huge difference in importance of hyperparameters:



The learning rate is by far more important than the hidden size.
As confirmation of this fact, let us consider two examples from the trial table: in the first one, we keep the Hidden Size fixed and change the Learning Rate slightly, and vice versa.

Learning Rate	Hidden Size	Value
2.79×10^{-5}	64	0.28223
5.7364×10^{-4}	64	0.24966

In this case, the difference between the two Learning Rates is around 6.87×10^{-4}, resulting in a difference in the value of the loss function of 0.037. Now, let's try to change the Hidden Size significantly while leaving the Learning Rate relatively unchanged.

Learning Rate	Hidden Size	Value
5.24×10^{-5}	576	0.24657
5.27×10^{-5}	128	0.27323

In this case, the difference between the two Hidden Sizes is 448, resulting in a difference in the value of the loss function of 0.027.
From Optuna, we found that the best couple of hyperparameters is the following:

Learning Rate	Hidden Size	Loss Function
0.000708695	384	0.219225362

It's interesting to notice the fact that, as said above, the Learning Rate has as magnitude 10^{-4} and not less as seen in the previous examples: in this way we avoid overfit and obtain the lowest possible value of loss function. The dimension of the hidden layers seems to be quite small with respect to other that have been chose randomly by the algorithm (from 64 to 916 layers).
It's necessary to underline the fact that we are doing considerations only on 50 iterations and so they could be not very precise.

1. Appendix

Trial Number	Value	Hidden layers size	Learning Rate
1	0.26110535860061646	512	1.1604210027950516e-05
2	0.275857537984848	704	0.015722911914775702
3	0.2265419065952301	832	0.001797488075858277
4	0.2294926792383194	960	0.0027036126770796443
5	0.286393940448761	512	0.017054321532784314
6	0.23346757888793945	64	0.0012716563940357362
7	0.2685941457748413	320	2.917550032021065e-05
8	0.23355332016944885	640	0.0003405572171813509
9	0.2676813304424286	512	0.012320424351901687
10	0.25934088230133057	512	2.6269040827336377e-05
11	0.8050574660301208	512	0.07948570976238072
12	0.41429823637008667	512	0.025729870357879776
13	0.2246258705854416	960	0.00020913627820688186
14	0.2369268536567688	192	0.015444210175486596
15	0.22846148908138275	832	0.002929259356536738
16	0.22483374178409576	896	0.001012600799975367
17	0.2567295730113983	640	2.5777695347105532e-05
18	0.25080910325050354	128	0.0001799244299715466
19	0.23048047721385956	192	0.003279973729214162
20	0.2267168164253235	640	0.0008629021150033828
21	0.24464581906795502	896	3.96132731141447e-05
22	0.2375701665878296	256	0.014127550363094817
23	0.23951904475688934	448	0.01087882731316453
24	0.24290457367897034	320	0.0006472410802068822
25	0.24085460603237152	192	0.017308577022592517
26	0.26132330298423767	704	1.548065469376969e-05
27	0.2471289485692978	896	2.7896263819692494e-05
28	0.24834784865379333	512	0.008357507989490285
29	0.7628317475318909	320	0.06499985557797118
30	0.2552589178085327	64	0.00032243829345234286
31	0.22928354144096375	896	0.006242468881893661
32	0.23503044247627258	384	0.002791222881768311
33	0.26153838634490967	192	8.018205933364054e-05
34	0.27065664529800415	128	8.706258694881354e-05
35	0.23656028509140015	704	0.0005575342182714217
36	0.25093314051628113	64	0.0026767367302258337
37	0.22943782806396484	320	0.0025011459987604513
38	0.22757463157176971	640	0.0007630725492723245
39	0.27670252323150635	128	2.674447404127627e-05
40	0.254611074924469	576	0.00014914989959890782
41	0.2654847800731659	320	1.0652301853008555e-05
42	0.24963372945785522	128	0.014260285981320175
43	0.22748921811580658	576	0.0022707414235194298
44	0.22631801664829254	832	0.0003023771721764421
45	0.261157363653183	448	2.6806147754869867e-05
46	0.22790038585662842	64	0.009924563884599608
47	0.22768758237361908	704	0.0004141845506090858
48	0.25117722153663635	960	1.7898214668522725e-05
49	0.22225944697856903	576	0.0006785864131083171
50	0.2633822560310364	576	1.35640538597191e-05

Table 1: Full data for all trials