ML 2023/24 Project
Type A

Team Aldra - 14/02/2024

Aru Giacomo 597700 g.aru@studenti.unipi.it
Marzeddu Simone 597134 s.marzeddu@studenti.unipi.it
Raffi Jacopo 598092 j.raffi@studenti.unipi.it

Master Degree in Computer Science (Artificial Intelligence) - University of Pisa

Objectives

- **Research aims:**
  - Explore different neural network topologies;
  - Study different activation functions;
  - Exploit Adamax optimizer.

- **Technical design details:**
  - Single and multi-layer Neural Networks;
  - Standard Backpropagation training algorithm;
  - Ridge Regression and Early Stopping;
  - Adamax optimizer;
  - Ensembling techniques (Bagging).

Software Architecture

- **Main Libraries:**
  - Numpy, Pandas.

- **Main Classes:**
  - **InputNeuron, HiddenNeuron** and **OutputNeuron**, implement units as objects linked by succession relations;
  - **NeuralNetwork**, manages the high-level network of units. Implements training algorithms;
  - **ModelSelection**, performs multi-process cross validation and k-folding with backup management.

Implementation Features (1)

- **Data Preprocessing:** Random Shuffling and Standardisation;
- **Model Selection:** K-Fold Cross-Validation;
- **Topologies and Architectures:** Feed Forward NN:
    - Units: 9-24-32-40 Units;
    - # Layers: 1-2-3.
- **Activation Functions:** Sigmoid (Slope = 1), ReLU, Tanh (Slope = 1), Identity;
- **Weights Initialisation:** Random Uniform Initialisation:
    - Range: - 0.75, + 0.75;
    - Fan-In for Hidden Units.

Implementation Features (2)

- **Training Algorithm:** Standard Backpropagation;
- **Regularization Techniques:** Ridge Regression, Early Stopping;
- **Learning Rate Improvements:** Momentum, Nesterov Momentum, Linear Decay, Adamax;
- **Training Data Consumption:** Stochastic Mode (Online, Mini-batch), Batch Mode;
- **Ensemble Learning Techniques:** Bagging.

Supplementary Approaches

- Design of a flexible modular framework which allows topological combinations (different activation functions for different neurons, different connections between levels, etc.);
- **Learning Rate Decay** between epochs or mini-batches (weight update steps);
- **Adamax**, an algorithm for first-order gradient-based optimization based on the infinity norm [1].

MONK's Results

Task	Hyperparameters*	MSE (TR/TS)	Accuracy (TR/TS) %
MONK1	4 - 0.6 - 0 - 0.6 - 4 - 50 - 150	0.0025 / 0.0052	100% / 100%
MONK2	4 - 0.65 - 0 - 0.9 - 4 - 50 - 150	0.00035 / 0.0022	100% / 100%
MONK3	4 - 0.3 - 0 - 0.3 - 4 - 50 - 600	0.024 / 0.054	98.4% / 92,6%
MONK3 (reg.)	4 - 0.3 - 0.001 - 0.5 - 8 - 50 - 600	0.108 / 0.094	93.4 % / 97.2 %

Hyperparameters*: #Hidden Units, Learning Rate (Internally divided by the batch size), Tikhonov Lambda , Momentum, Batch Size, Min #Epochs, Max #Epochs.

NOTE: All models presented are Feed Forward NNs with a single 4-units Hidden Layer.

Results: MONK1, MONK2

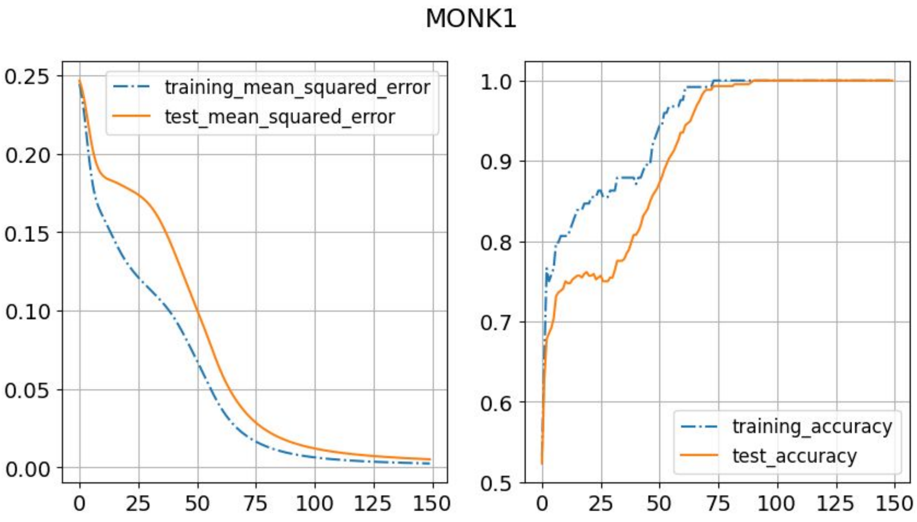


Fig. 1:
MONK1, TR and TS MSE (SX) and Accuracies (DX).

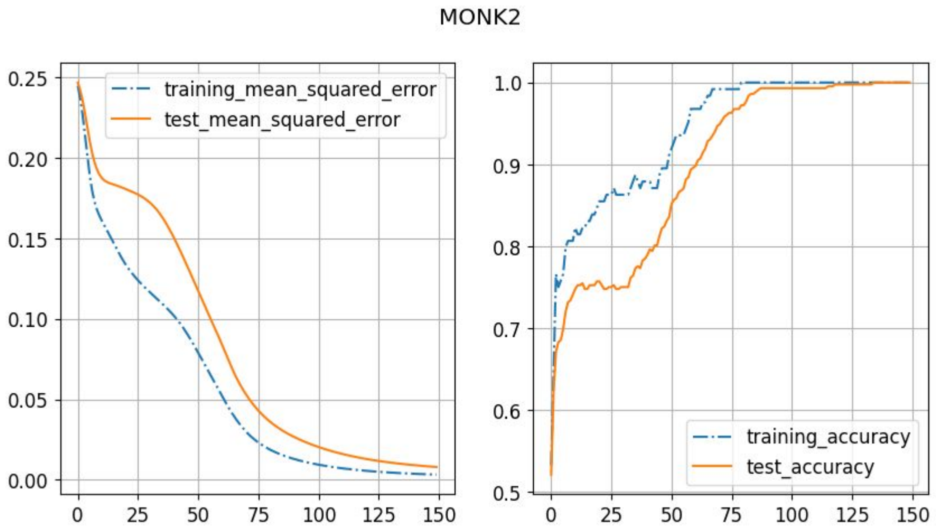


Fig. 2:
MONK2, TR and TS MSE (SX) and Accuracies (DX).

Results: MONK3, MONK3 (reg.)

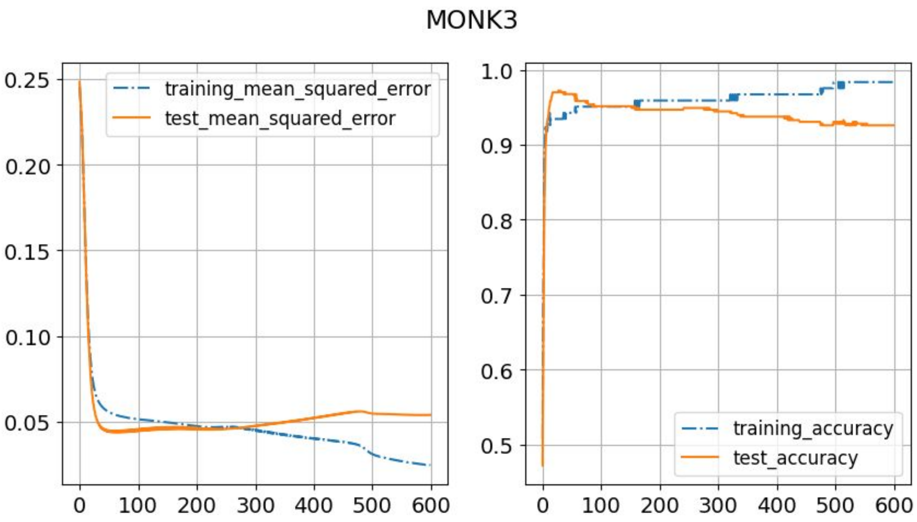


Fig. 3:
MONK3, TR and TS MSE (SX) and Accuracies (DX).

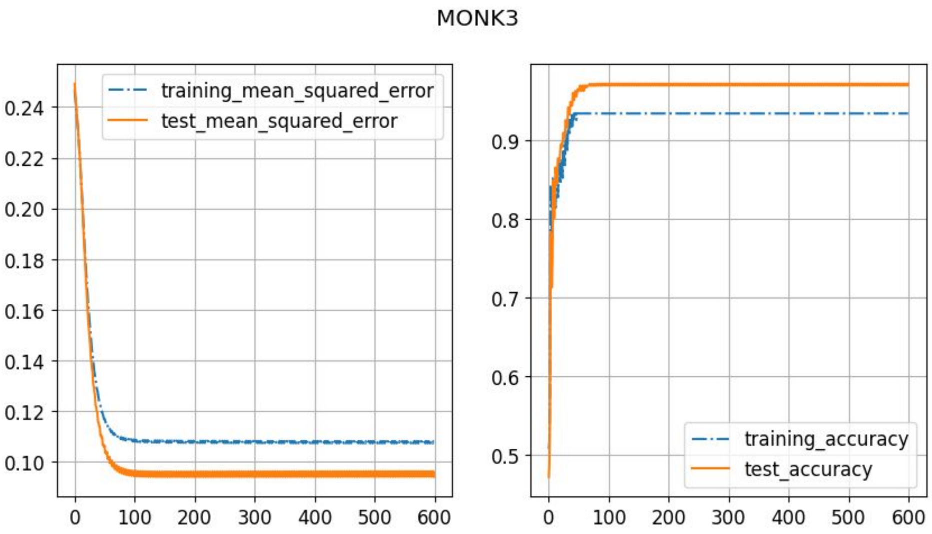


Fig. 4:
MONK3 (reg.), TR and TS MSE (SX) and Accuracies (DX).

ML Cup: Data Splitting

- Preprocessing: Random Shuffling and Standardization;
- Validation Schema:

3-Fold CV TR + VL (80 %)	Hold-out Assessment TS (20 %)

- Retraining on TR + VL;
- Model Assessment on TS;
- Final Retraining on TR + VL + TS.

ML Cup: Our Project Path (1)

1. General tests on our implementation and its functionalities:
    a. A priori exclusion of particular connection structures of units (exclusive use of networks with fully connected layers).
2. Study of different topologies (number of units and activation functions, layers and data consumption modes (batch, mini-batch and online)) focused on models’ stability and approximation capabilities:
    a. we excluded networks with more then 3 layers and/or 40 units;
    b. we excluded topologies combining multiple activation functions for hidden layers;
    c. we excluded Batch Mode and Nesterov Momentum;
    d. we excluded SoftPlus as ReLU performances were generally superior.
3. First coarse-grain 3-Fold CV on a 2 layered (12 Units per layer) ReLU NN, 1 layered Sigmoid NN with 32 Units and a 1 layered Tanh NN with 32 Units, all with both Adamax and standard optimization algorithms.

ML Cup: Coarse-grain CV Configurations

Topology	2 Hidden Layers **ReLU** 12-12 Units
Batch size	10, 20, 30, 50
Min #Epochs	100
Max #Epochs	800
Learning Rate	0.01, 0.03, 0.05, 0.1
Momentum α	0.5, 0.75, 0.9
Tikhonov λ	1e-6

ES Tolerance %	0.0001 %
Patience	5
LR Decay τ	125, 175
Adamax	Yes, No
Adamax LR	0.002, 0.02, 0.008, 0.1
Exp. Decay Rate 1	0.9
Exp. Decay Rate 2	0.99

ML Cup: Coarse-grain CV Configurations

Topology	1 Hidden Layer **Tanh** 32 Units
Batch size	10, 20
Min #Epochs	100
Max #Epochs	800
Learning Rate	0.01, 0.05, 0.1
Momentum α	0.75, 0.9
Tikhonov λ	1e-7, 1e-6

ES Tolerance %	0.0001 %
Patience	5
LR Decay τ	100, 137.5, 175
Adamax	Yes, No
Adamax LR	0.05, 0.1, 0.2
Exp. Decay Rate 1	0.9, 0.8
Exp. Decay Rate 2	0.999, 0.9

ML Cup: Coarse-grain CV Configurations

Topology	1 Hidden Layer **Sigmoid** 32 Units
Batch size	10, 20
Min #Epochs	100
Max #Epochs	800
Learning Rate	0.005, 0.01, 0.05, 0.1
Momentum α	0.75, 0.9
Tikhonov λ	1e-7, 1e-6

ES Tolerance %	0.0001 %
Patience	5
LR Decay τ	100, 137.5, 175
Adamax	Yes, No
Adamax LR	0.005, 0.02, 0.1
Exp. Decay Rate 1	0.9
Exp. Decay Rate 2	0.999

ML Cup: Coarse-grain CV Configurations

Topology	MEE (Standardised)	MSE (Standardised)	MEE Variance	MSE Variance	Appendix Refs.
32 Sigmoid	0.103085	0.015038	3.5e-5	2e-6	slide 30
32 Sigmoid (Adamax)	0.223192	0.065483	7.73e-4	1.78e-4	slide 30
32 Tanh	0.118104	0.019526	7.4e-5	8e-6	slide 29
32 Tanh (Adamax)	0.148201	0.030771	3.70e-4	4.6e-5	slide 29
12-12 ReLU	0.304497	0.122745	3.891e-3	2.549e-3	slide 31
12-12 ReLU (Adamax)	0.343800	0.151048	7.58e-4	7.35e-5	slide 31

ML Cup: Our Project Path (2)

4. We selected the hyperparameter configuration that produced the lowest validation MEE:
   a. we excluded Adamax optimization algorithm;
   b. we excluded ReLU and Tanh as Hidden Units' activation functions.
5. Tests on different topologies with fixed hyperparameters of the best model found (1 layered Sigmoid NN with 32 Units);
6. Attempt to smooth learning curves by balancing LR, LR Decay and Momentum to mitigate the initial learning instability:
   a. we decided to not persevere with this approach as the choice of hyperparameters would have fallen close to the best 3-Fold CV model.

ML Cup: Our Project Path (3)

7. Fine-grained 3-Fold CV setting the best topology of the previous iteration (NN with a single 32-unit hidden layer, Sigmoid activation function):
    a. we selected the hyperparameter configuration with lowest validation MEE.
8. Application of an ensemble learning technique (**Bagging**) to control the variance of the best model resulting from the last grid search:
    a. 32 models, all with the same hyperparameters configuration, were trained on random samplings with replacement from the training set;
    b. the resulting model uses as output the arithmetic mean of the internal models' predictions.
9. The test set (TS) obtained from the initial 20% Hold-out has now been exploited to assess the final ensemble model generalisation capabilities;
10. Final Retraining on TR + VL + TS.

ML Cup: Fine-grain CV Configurations

Topology	1 Hidden Layer **Sigmoid** 32 Units
Batch size	6, 8, 10, 11, 20
Min #Epochs	100, 150
Max #Epochs	500, 800
Learning Rate	0.005, 0.01, 0.05, 0.07, 0.09, 0.1, 0.11, 0.13
Momentum α	0.75, 0.85, 0.9, 0.92, 0.95
Tikhonov λ	1e-9, 1e-8, 1e-7, 1e-6

ES Tolerance %	0.0001 %,
Patience	5
LR Decay τ	100, 137.5, 145, 165, 175, 185, 200
Adamax	No

Color Legend:

- Original Coarse-grain 3-Fold CV;
- Fine-grain 3-Fold CV.

ML Cup: Final Model - Hyperparameters

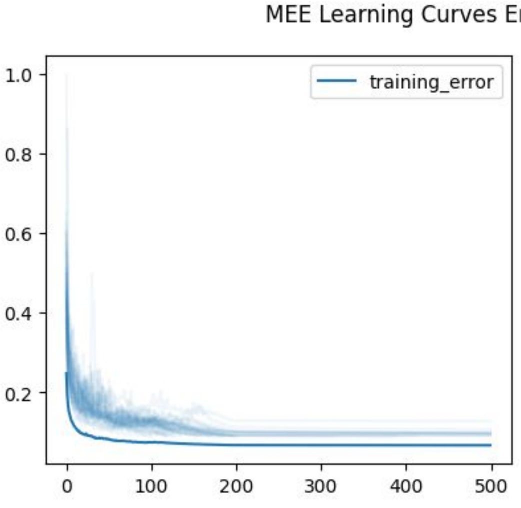
Selecting the configuration with the lowest validation MEE after the 3-fold CV resulted in the following final model:

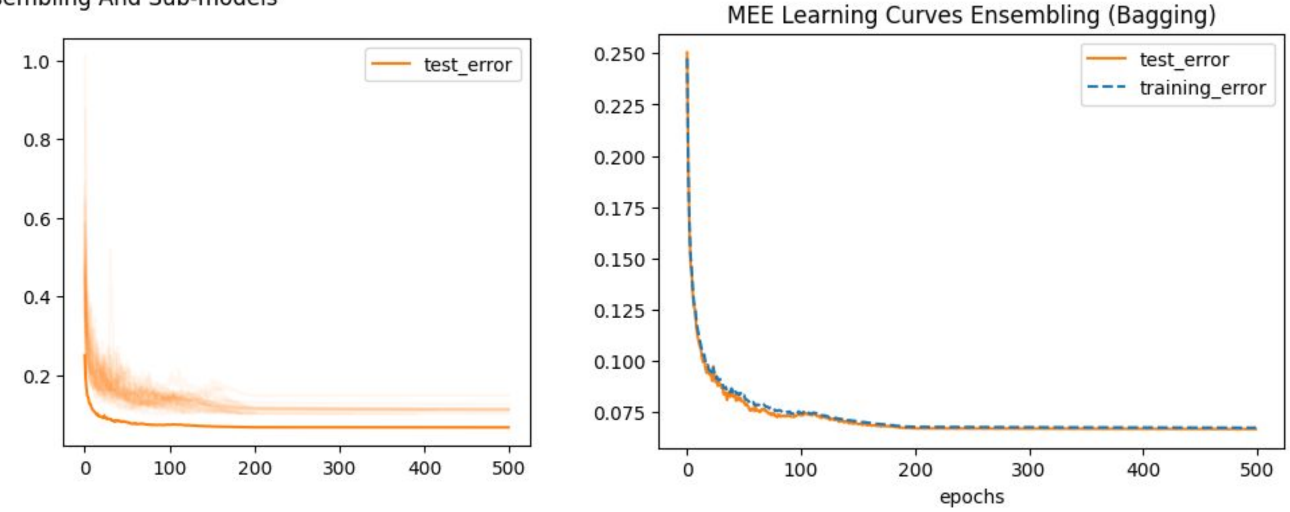
- **Topology:** 1 Hidden Layer with 32 Hidden Units (Sigmoid Activation Function);
- **MEE Validation:** 2.43;
- **MEE Validation Variance:** 0.018.

Batch size	Min #Epochs	Max #Epochs	Patience	ES Tolerance %	Tikhonov λ	Momentum α	LR Decay τ	LR
8	150	500	5	0.0001 %	1e-9	0.85	200	0.11

ML Cup: Final Model - Bagging

# Internal Models	Ensembling Function	MEE Training (TR + VL)	MEE Test (TS)
32	Arithmetic Mean	1.3334	1.3324





ML Cup: Discussion (1)

Proposed techniques results:

- **Learning Rate:** The learning rate and associated hyperparameters were found to be the lynchpin of our model selection. Small variations in these values can radically change the resulting model;
- **Momentum:** High momentum values were definitely favoured by model selection, but we found that values above 0.9 were characterised by high instability;
- **Tikhonov Regularization:** Low values of λ always allowed good regularization of the models;
- **Batch Size:** Large mini-batch sizes showed lower convergence speed but higher learning curve stability, while small mini-batch sizes performed better but were more unstable.

ML Cup: Discussion (2)

- **Flexible and Modular Implementation:** It allowed for several experimentation in terms of variety of case studies, but resulted also in high temporal costs for exhaustive researches. Despite that, it was educationally relevant;
- **Learning Rate Decay:** Finding the right trade-off for τ allowed us to balance the convergence speed and the stability of the learning curve;
- **Adamax:** Despite not passing the model selection in terms of validation MEE, the Adamax models consistently showed great stability of the learning curves, at the expense of a slow convergence speed (graphs on slides 29, 30, 31);
- **Ensembling:** Following the 3-Fold CV, the MEE validation variance of the best model was still problematic. Taking advantage of ensemble techniques (Bagging) allowed us to obtain a final result with more reliable performance.

Conclusions

What we drew and what we learned:

- **Practical Approach:** Working on a practical Machine Learning project allowed us to delve deeper into the theory, and to resolve doubts about certain nuances that could arise from theoretical study alone;
- **Working with Hyperparameters:** Although we were confident in the theory behind Model Selection, approaching this technique on a practical level allowed us to really understand the influence of hyperparameters and their tuning;
- **Working with long computations:** The Machine Learning project required a lot of time and computational resources, more than any other project we had faced in the past. This practical test led us to organise a work plan for our equipment, studying the division and timing of the various processes.

Blind Test Results File: Aldra_ML-CUP23-TS.csv
Our Nickname: Aldra

Bibliography

1. Diederik P. Kingma e Jimmy Ba. Adam: A Method for Stochastic Optimization. 2017. arXiv: 1412.6980 [cs.LG].

Appendix

The appendices provide additional data and graphs to support the arguments in this report:

- **Best Discarded Models** (Slides 26, 27, 28, 29, 30)**:** Collection of graphs, configurations and performances of the best discarded models for each type of activation function, with and without the use of Adamax;
- **Learning curve after the smoothing attempt** (Slides 31, 32)**:** Implementation details and demonstration graph concerning the attempt to smooth the learning curves starting from the best output of model selection;
- **Learning curve after the final Retraining phase** (Slide 33);
- **Training Speed and Hardware Data** (Slide 34).

ML Cup: Model Selection - Best Discarded Configurations (No Adamax)

Topology	Batch size	Min #Epochs	Max #Epochs	Patience	ES Tolerance %	Tikhonov λ	Momentum α	LR Decay τ	LR
32 Tanh	10	100	800	5	0.0001 %	1e-7	0.75	175	0.1
12-12 ReLU	10	100	800	5	0.0001 %	1e-6	0.5	175	0.01

ML Cup: Model Selection - Best Discarded Configurations (Adamax)

Topology	Batch size	Min #Epochs	Max #Epochs	Patience	ES Tolerance %	Tikhonov λ	Exp. Decay Rate 1	Exp Decay Rate 2	LR
32 Sigmoid	10	100	800	5	0.0001 %	1e-6	0.9	0.999	0.1
32 Tanh	10	100	800	5	0.0001 %	1e-7	0.9	0.999	0.1
12-12 ReLU	10	100	800	5	0.0001 %	1e-6	0.9	0.999	0.1

ML Cup: Model Selection - Best Discarded Configurations (Tanh)

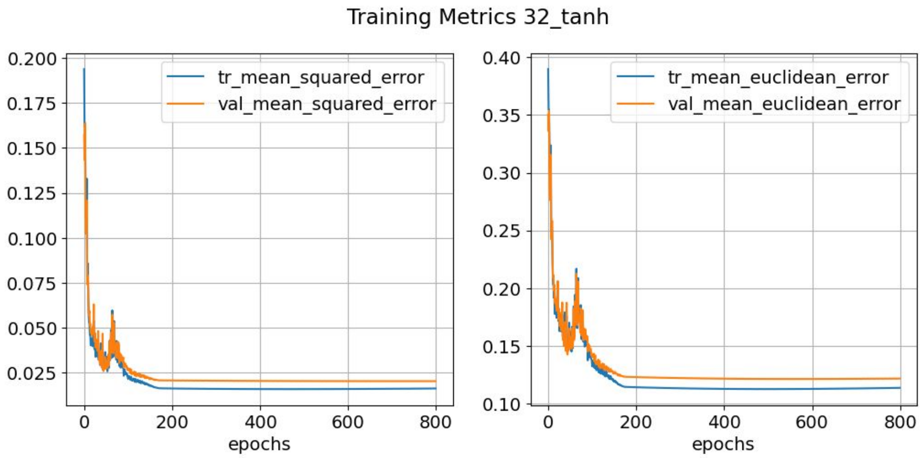


Fig. 7
Tanh no Adamax

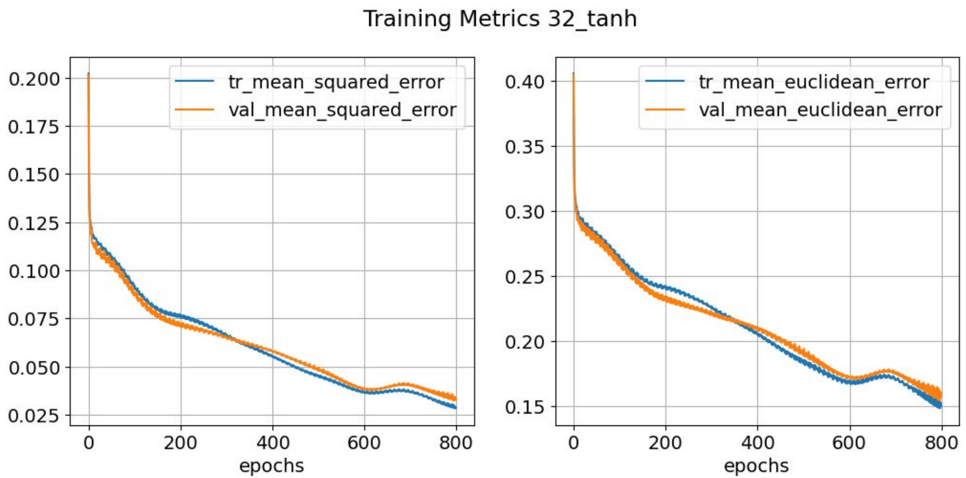


Fig. 8
Tanh with Adamax

ML Cup: Model Selection - Best Discarded Configurations (Sigmoid)

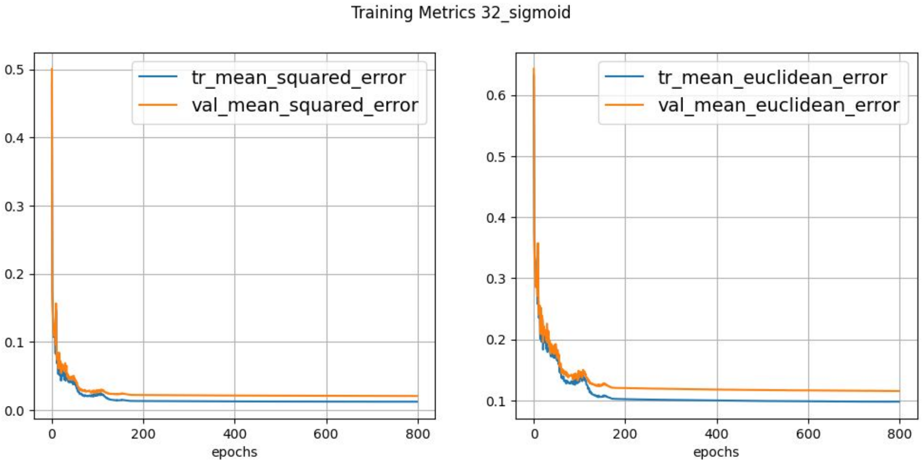


Fig. 9
Sigmoid no Adamax (Best Model, Not Discarded)

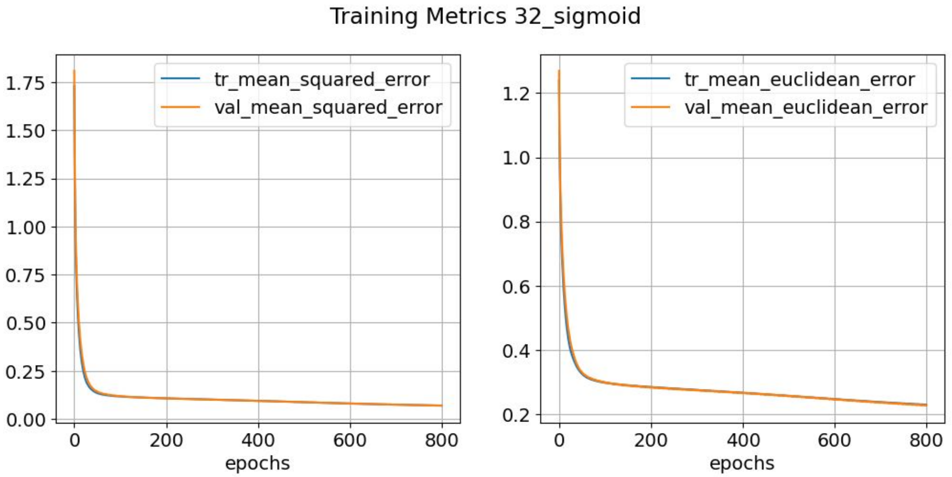


Fig. 10
Sigmoid with Adamax

ML Cup: Model Selection - Best Discarded Configurations (ReLU)

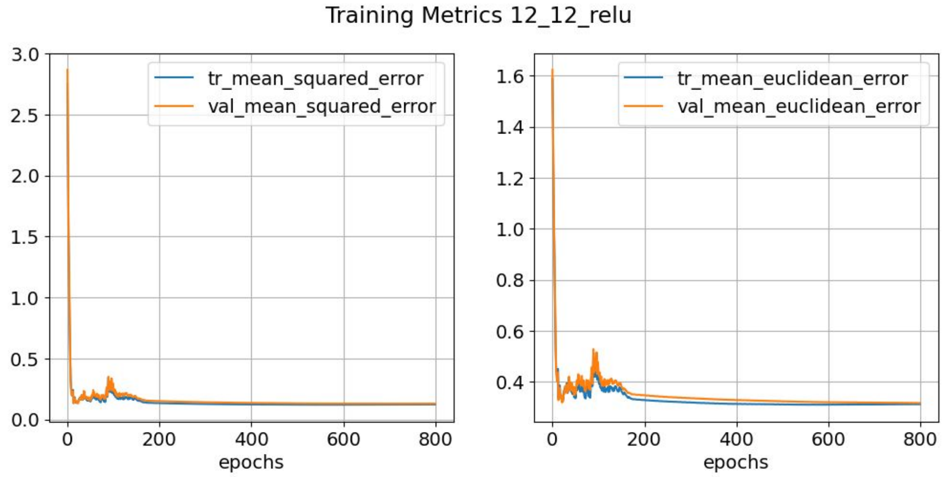


Fig. 11
ReLU no Adamax

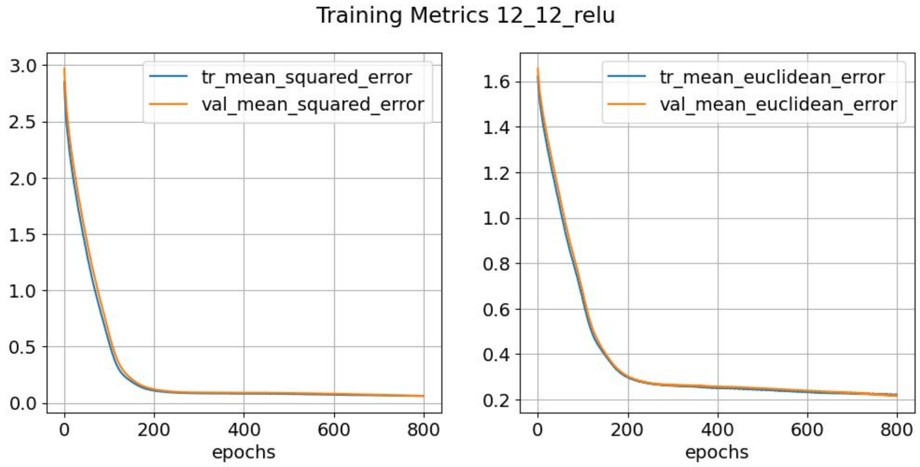


Fig. 12
ReLU with Adamax

ML Cup: Model Selection - Smoothing

Topology	1 Hidden Layer **Sigmoid** 32 Units
Batch size	10
Min #Epochs	100
Max #Epochs	400 - Reduced to limit execution time
Learning Rate	0.0005, 0.002. 0.005, 0.01 - Balancing LR with Decay and Momentum
α Momentum	0.75 - Lowered to reduce initial instability
λ Tikhonov	0.000001

ES Tolerance %	0.0001 %
Patience	5
LR Decay τ	250, 300 - Increased to slow the decay

ML Cup: Model Selection - Smoothing

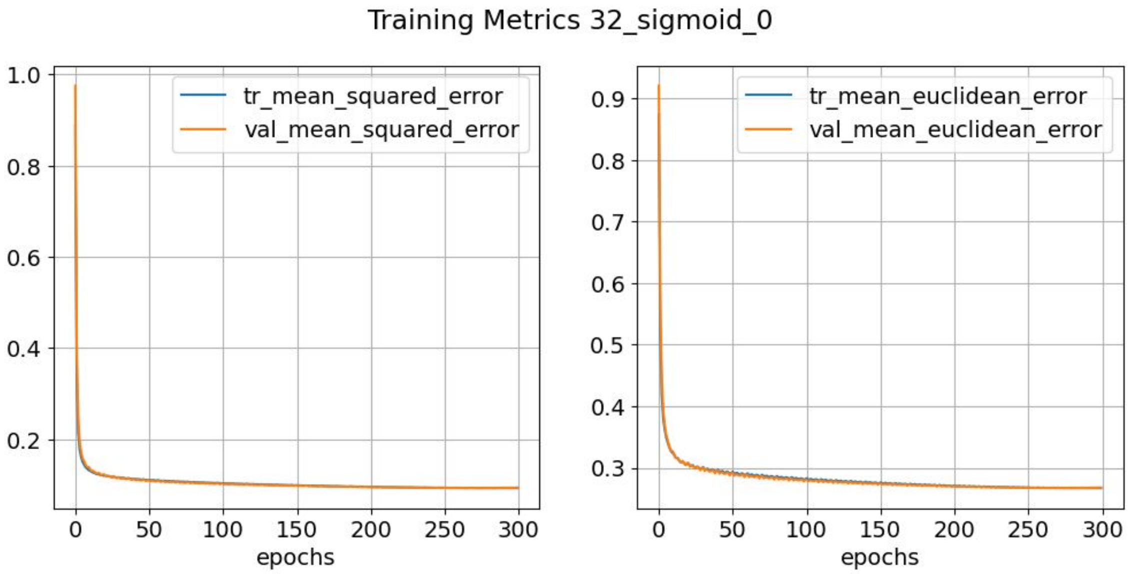


Fig. 13

Smoothed graphs: worse approximation but better stability

ML Cup: Final Retraining Learning Curve

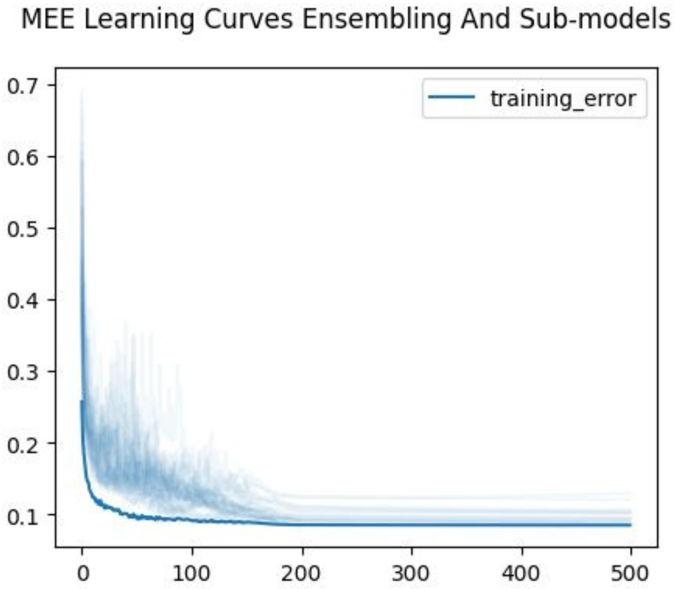


Fig. 14
Learning Curve on TR + VL + TS
(Bagging of 32 Learners)

Training Speed and Hardware Data

	Machine 1	Machine 2	Machine 3
CPU	Intel(R) Core(TM) i7-8750H	Intel(R) Core(TM) i7-10750H	Intel(R) Core(TM) i5-1035G1
Cores	6	6	4
Base Speed	2.2 GHz	2.6 GHz	1.0 GHz
Logical Processes	12	12	8

Training Speed Estimation: around 1.00 Seconds for an Epoch of 533 data.