Lab 5

Dominik Rafacz

06.05.2020

Loading data and splitting into train and test

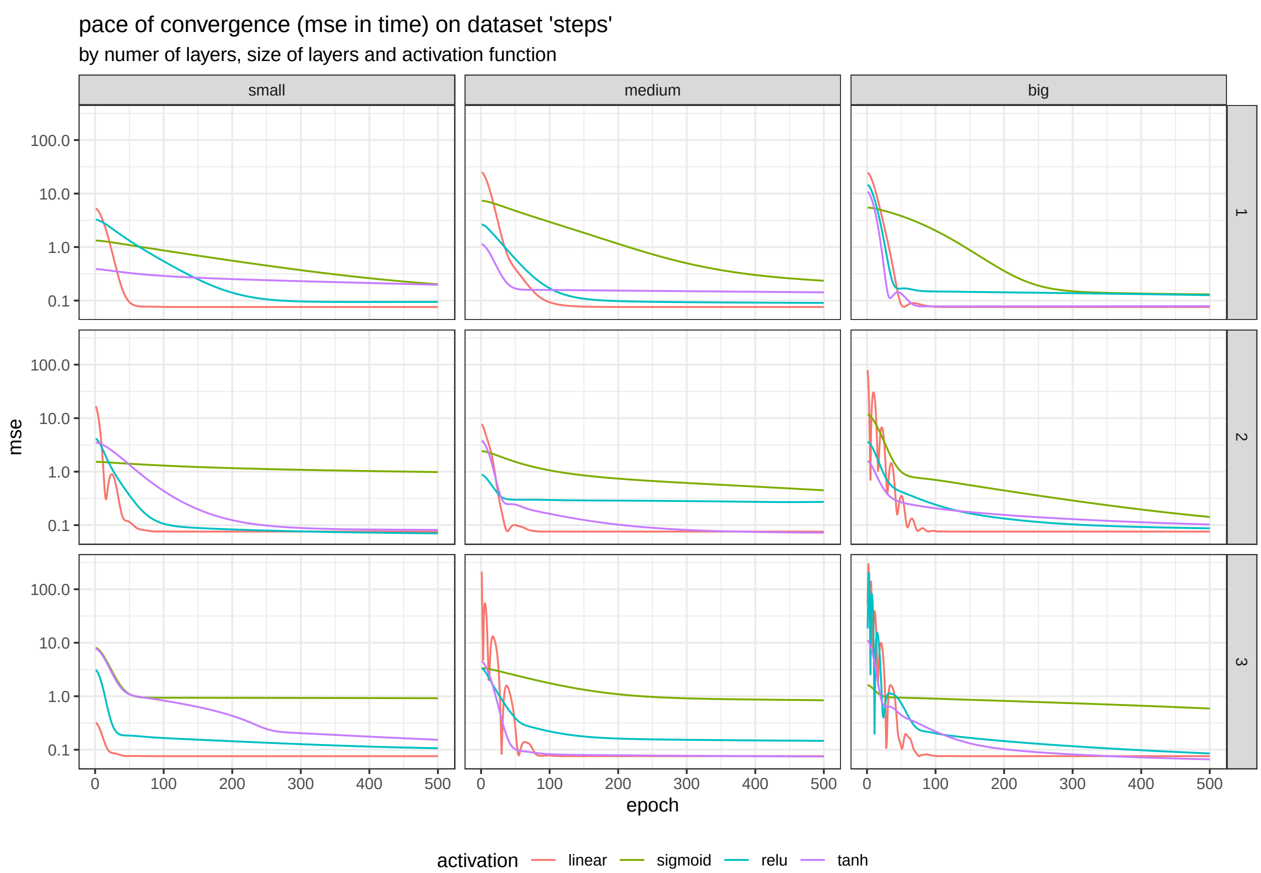
I've loaded four datasets and splitted it into train and test datasets.

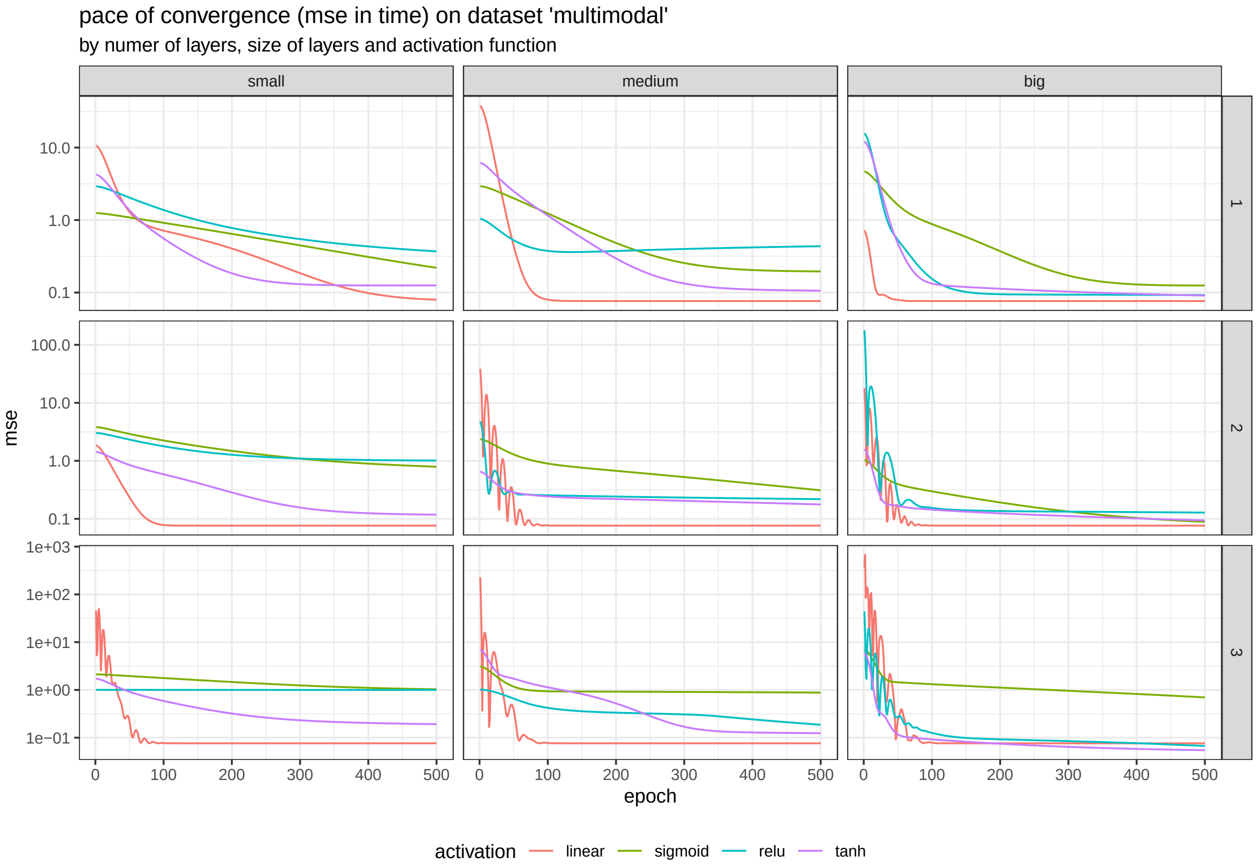
Prepraring experiment

For each dataset I've chosen: - four possible activation functions (linear, sigmoid, relu, tanh), - three possible numbers of hidden layers (one, two, three), - three possible sizes of layers (3, 5, 10).

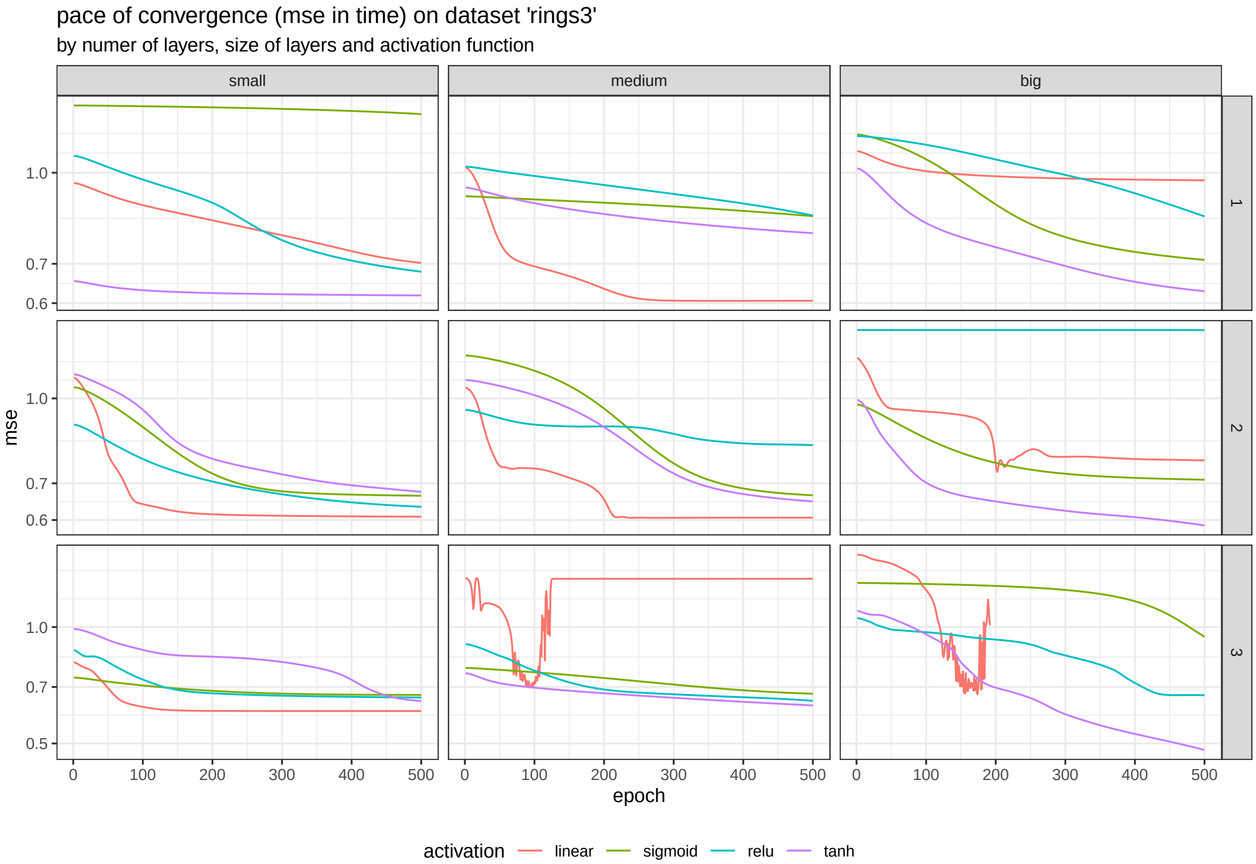
Next, I've trained one network for each dataset for each combination of those parameters (totalling $4 \times 4 \times 3 \times 3 =$ 144 networks). I've used momentum training with some default values (linear activation of last layer for regression, softmax for classification, using momentum optimizer, $\eta = 10^{-5}$ for regression and $\eta = 10^{-7}$ for classification).

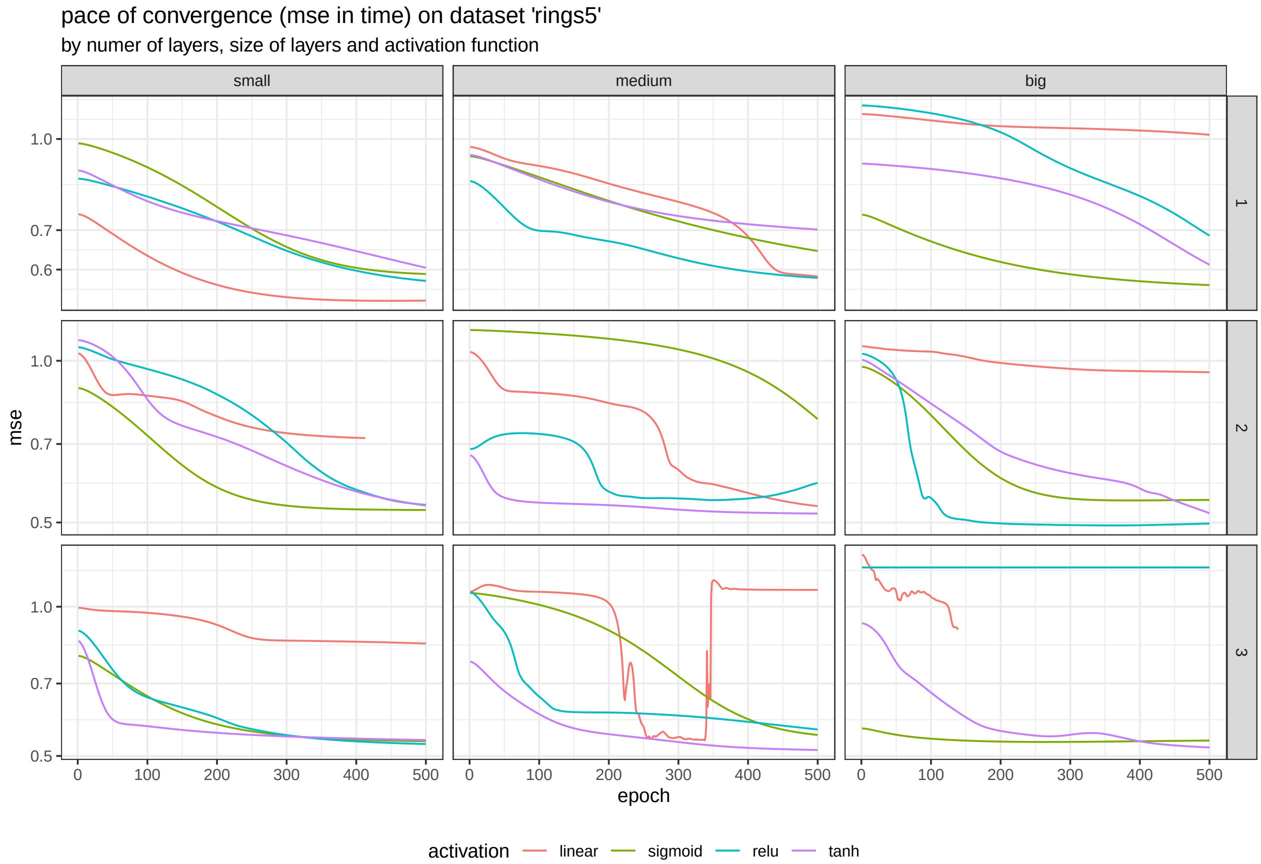
Convergence pace





pace of convergence (mse in time) on dataset 'multimodal'
by numer of layers, size of layers and activation function
small
medium
big
1
2
3
mse
10.0
1.0
0.1
100.0
1e+03
1e+02
1e+01
1e+00
1e−01
0
100
200
300
400
500
epoch
activation
linear
sigmoid
relu
tanh

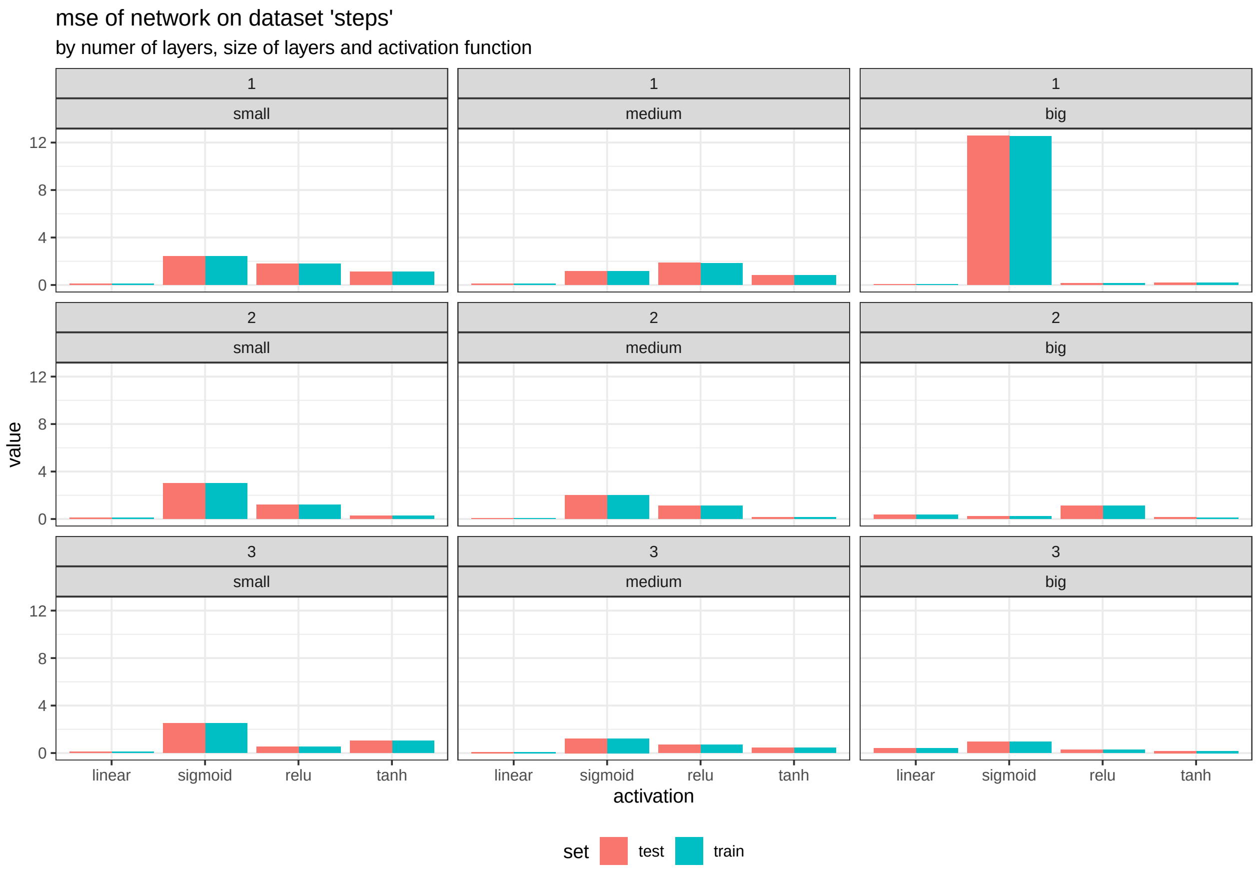

pace of convergence (mse in time) on dataset 'rings3'
by numer of layers, size of layers and activation function
small
medium
big
1
2
3
mse
1.0
0.7
0.6
0.5
epoch
0
100
200
300
400
500
activation
linear
sigmoid
relu
tanh

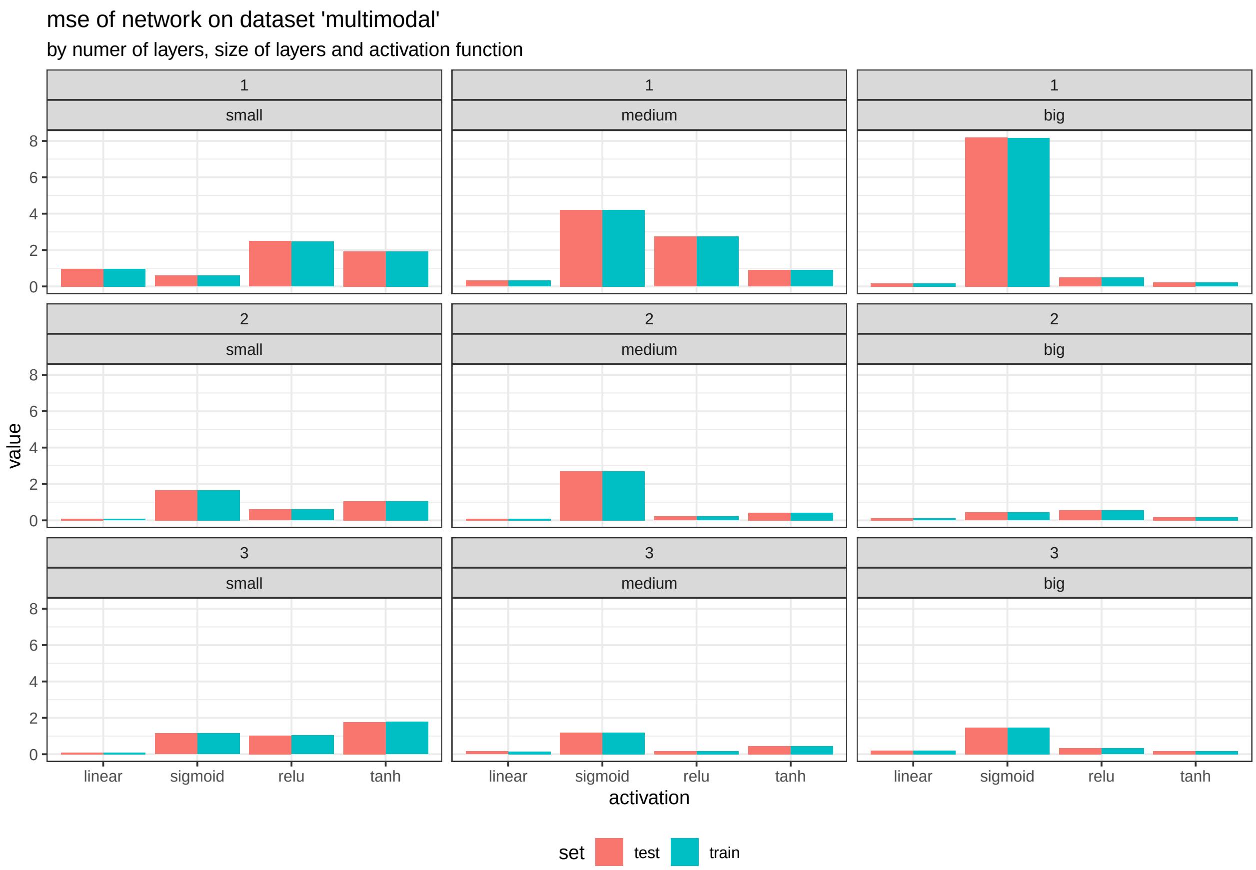


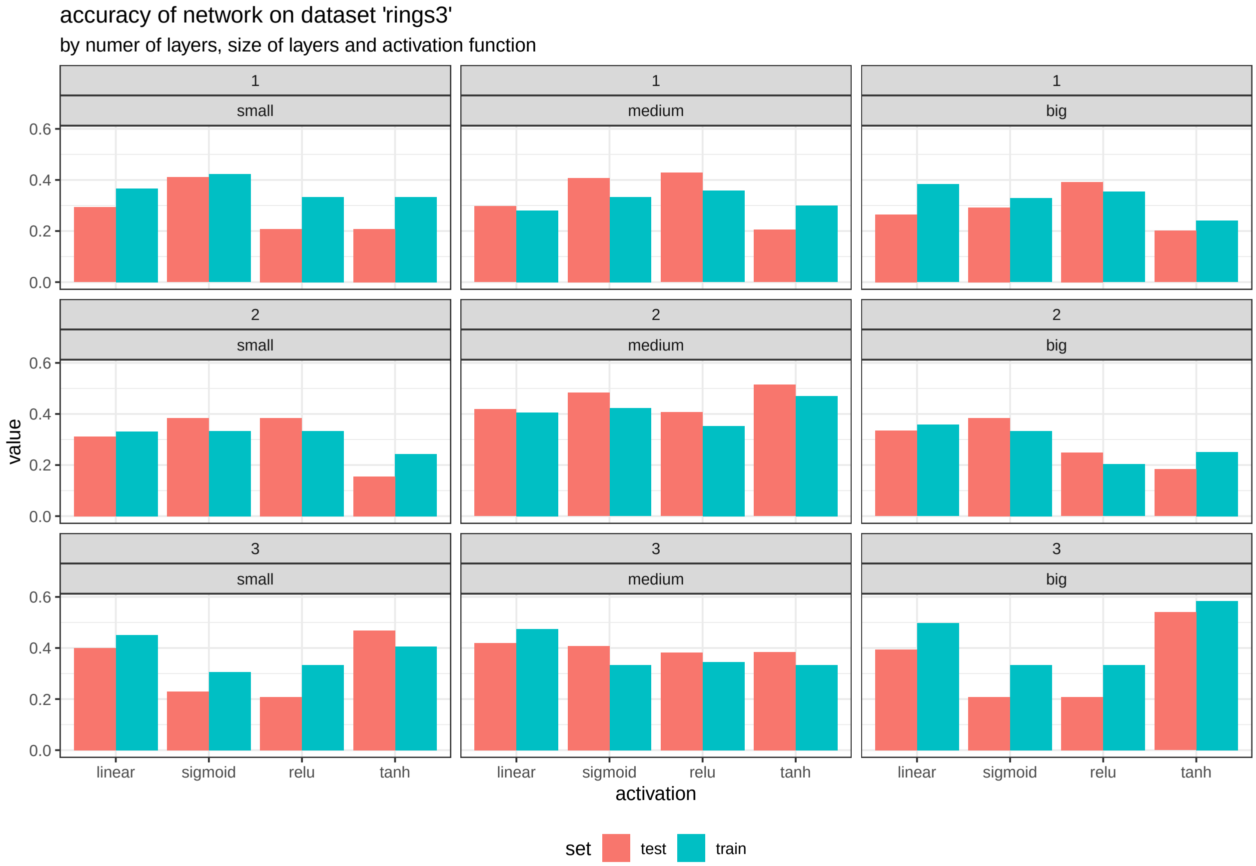
We can draw some conclusions from those plots:

- Linear activation is not once the fastest one, however, it's the most unstable of all – especially in classification problems.
- ReLU easily drops into local optimum (it's caused by it's derivative form) and do not reach the global optimum; however, it can be quite fast
- Sigmoid is slow. Sometimes very slow.
- Tahn converges faster than ReLU and sigmoid and sometimes reaches better values than other activation functions.
- The bigger layers the slower convergence – not necessarily better values.
- The more layers the greater instability.

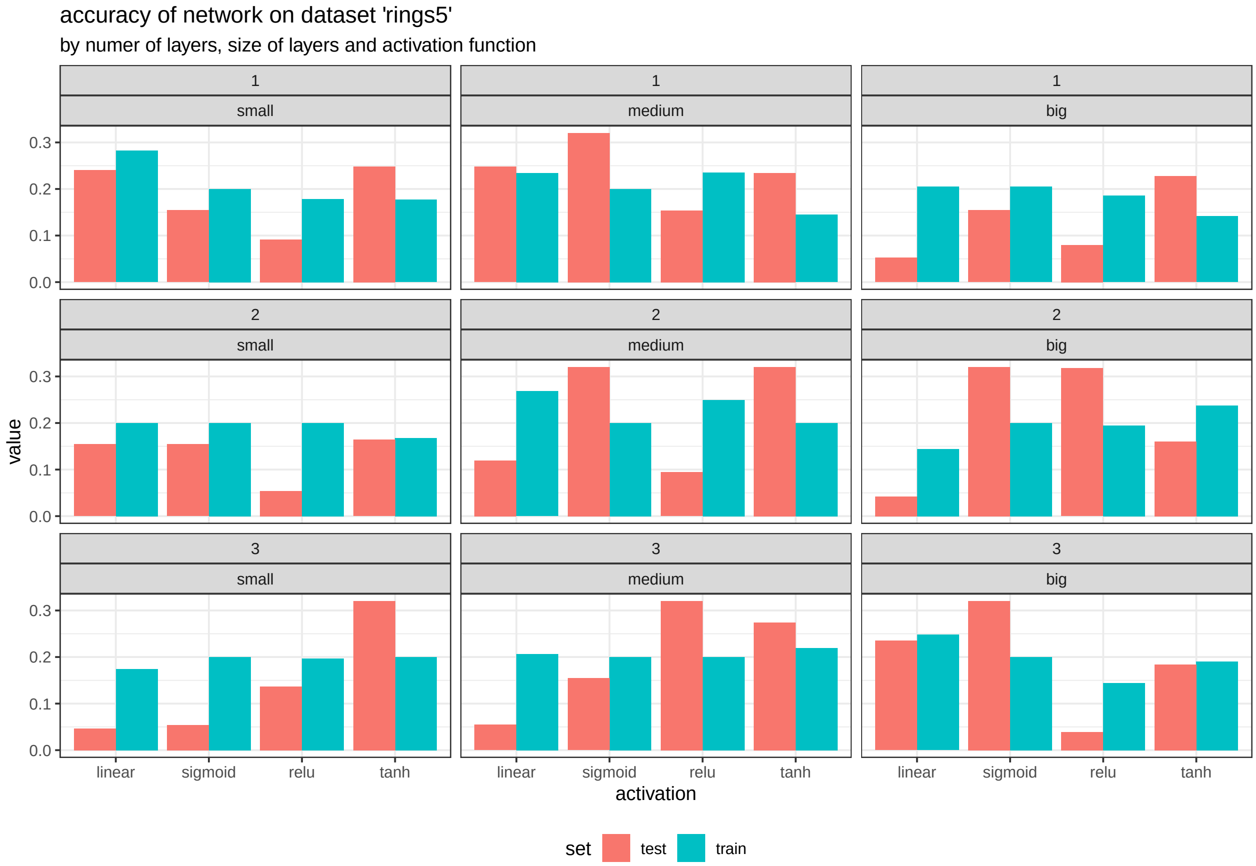
Goodness of fit




mse of network on dataset 'multimodal'
by numer of layers, size of layers and activation function
1
small
1
medium
1
big
2
small
2
medium
2
big
3
small
3
medium
3
big
value
0
2
4
6
8
linear
sigmoid
relu
tanh
activation
set
test
train



accuracy of network on dataset 'rings3'
by numer of layers, size of layers and activation function
1
small
1
medium
1
big
2
small
2
medium
2
big
3
small
3
medium
3
big
value
0.0
0.2
0.4
0.6
linear
sigmoid
relu
tanh
activation
set
test
train



Caution! In plots regarding classification, results on the test set are sometimes better than on the training set – this may be caused by the fact that networks are way undertrained.

The main conclusion is: result of the training heavilly depends on the dataset and parameters.