Impact of Activation Functions and Learning Rates on Non-Linear Pattern Learning in Neural Networks

Group 12: Sohbat Sandhu & Yilin Long

Abstract

This study investigates the effects of different ***activation functions*** and ***learning rates*** on the ability of a neural network to learn non-linear data patterns. The experiment involves generation of synthetic datasets featuring various geometric and non-linear ***data patterns*** for binary classification tasks using neural networks. The testing loss, representing the mean squared error, was assessed across 36 treatments and blocking combinations. The experiment employs a factorial design and uses ANOVA with contrasts to analyze the main effects and interactions to identify significant factors influencing performance.

Introduction

Neural networks depend heavily on hyperparameters like activation functions and learning rates to model non-linear patterns effectively. Activation functions introduce non-linearity, allowing networks to capture complex relationships, while learning rates control weight updates during optimization. A fixed neural network architecture with the same activation function and learning rate at each neuron will be employed to ensure consistency across experiments. This experiment explores how different activation functions and learning rates affect a neural network's ability to accurately classify such patterns.

Outline

The **Detail of the experimental design** section will include information about Experiment Factors and Levels, Experimental plans, and response variable.

The **Statistical Analysis** section will include ANOVA modelling and approach, model checking, exploration of effect , contrast analysis.

The **Summary and Conclusions** section will include Summary of the experiment, potential improvements and future uses for this experiment.

Detail of Experimental Design

Experiment Factors and Levels

The experimental factors and levels including the reasons behind the importance for inclusion in the experiment are mentioned in Table 3.

Experiment Plan

This study follows a **full-factorial design** with blocking to examine all combinations of activation functions and learning rates across different data patterns. A total of **36 treatment combinations** (4 activation functions × 3 learning rates × 3 patterns) are tested.

Data Generation

The data was generated using a synthetic generator with the variables X_1 and X_2 signifying the two dimensional cartesian coordinates x and y. The data points are labelled 1 or -1 for positive (blue points)) and negative classes (yellow points). See Figure 6 for spiral data pattern, Figure 7 for Gaussian Mixture Data Pattern and Figure 8 for concentric data patterns.

Fixed Neural Network Setup

The neural network can be seen in Figure 9.

- **Architecture:** Fixed with 3 hidden layers containing 4, 8, and 4 neurons, respectively.
- **Output Layer:** Sigmoid activation for binary classification.
- **Optimizer:** Gradient Descent optimizer, fixed across all configurations, adjusts weights and biases by using the accumulated gradients.
- **Epochs:** Each treatment combination runs for 100 epochs, and the testing loss at the final epoch is recorded.
- **Features:** The input features include X_1, X_2, ${X_1}^2$, ${X_2}^2$, X_1X_2 for the neural network.

Blocking and Randomization

1. **Blocking:** Data patterns act as the blocking factor to control for variations introduced by pattern complexity.
2. **Randomization:** Treatment combinations are randomized within each block (data pattern). The order of experiments is randomized independently across blocks.

The data after running the Neural Network was collected and stored in a file. See Table 4 for preview.

Response Variable

Testing loss or error at the last epoch serves as the response variable, indicating the model's generalization capability.

Statistical analysis

ANOVA Modelling and Approach

The model includes interaction terms between these factors: pattern * activation, activation * learning.rate, and learning.rate * pattern. This allows the model to investigate whether the effect of one factor depends on the level of another factor. The resulting anova table is table1. Table 1 reveals that both pattern and activation function are significant factors affecting the test error, with p-values less than 0.001, indicating a strong influence on model performance. Learning rate is also significant at the 5% level ($p = 0.0477$), suggesting it has a smaller yet notable impact on the test error.

The interaction between pattern and activation is highly significant ($p < 0.001$), meaning the effect of the activation function on test error depends on the data pattern. However, the interactions between activation and learning rate, and pattern and learning rate are not significant, with p-values greater than 0.05, indicating that learning rate does not substantially affect the relationship between these factors.

We use Box-Cox transformation to determine if a transformation of the response variable (the test error) is needed to stabilize variance and make the data more normally distributed. And according to Figure 1 and Figure 2, no transformation is needed, because the Box-Cox plot suggests that the response variable (test error) is already approximately normally distributed, with the transformation parameter close to 1. This indicates that the data's variance is stable.

Model Checking

We use the QQ plot to check the normality of the residuals, and according to Figure 3, the residuals roughly follow the diagonal line, indicating that they are approximately normally distributed. This supports the assumption of normality, suggesting that no transformation of the response variable is necessary.

We use the residual vs. fitted values plot to check for constant variance, and according to Figure 4, the residuals exhibit a random scatter around zero with no clear patterns, indicating that the variance of the residuals is constant. This suggests that the

assumption of homoscedasticity holds, and no further adjustments to the model are needed.

Exploration of Effects

We got Figure 5, (interaction plot between Pattern and Test Error by Activation) because the ANOVA results show a significant interaction between these two factors (p-value = 1.08e-08). This suggests that the effect of the activation function on test error is not independent of the data pattern. Sigmoid performs worse, showing higher test errors, and remains consistent across all patterns. Linear and Tanh are similar, with Linear slightly outperforming Tanh on the concentric pattern. This plot highlights how the choice of activation function should be influenced by the data pattern to optimize performance.

Contrast Analysis

We perform the contrast to evaluate the relative effect of each activation function on the test error. The contrast allows us to quantify how the mean test errors for different activation functions (Linear, ReLU, Sigmoid, and Tanh) compare against each other.

Table 2 is the test error means for each activation level, and we calculate the contrast for the activation factor based on the following formula:

$$k = c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 \ldots + c_j\alpha_j$$

$\sum_{i=1}^{j} c_i = 0$, in this case, $c_1 = 1,\ c_2 = -1,\ c_3 = 1,\ c_4 = -1$

$$k = \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4$$

where: α_1 = 0.25 (Linear), α_2 = 0.27 (ReLU), α_3 = 0.51 (Sigmoid), α_4= 0.27(Tanh) So, K = 0.22

A positive contrast value suggests that Sigmoid and Linear activation functions lead to worse performance compared to ReLU and Tanh.

Conclusions and Discussion

Summary

We built a model that includes interaction terms between pattern, activation function, and learning rate to assess how these factors interact and influence the model's performance. The ANOVA results highlighted significant interactions, especially between pattern and activation function, indicating that the effect of the activation

function varies depending on the data pattern. To check if the response variable (test error) required transformation, we performed a Box-Cox transformation analysis, and the results showed that no transformation was necessary. We include diagnostic checks, including the QQ plot and residual vs. fitted values plot, supporting the assumptions of normality and homoscedasticity.

Finally, we calculated the contrast to compare the performance of different activation functions. The positive contrast value suggests that Sigmoid and Linear activation functions lead to lower test errors compared to ReLU and Tanh, providing a guide for selecting the most effective activation functions.

Scope for Improvement

We can incorporate other activation functions or experimenting with non-linear transformations for the input data could potentially improve performance. Secondly, Cross-validation could also be employed to assess the model's generalizability and reduce the risk of overfitting.

Future Use

The findings from this analysis can guide future model development by helping to select the most effective activation function for different data patterns. Additionally, the contrast method used here can be applied in future research to compare the performance of various hyperparameter configurations, further optimizing neural network models for a range of tasks.

Tables and Figures

Tables

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
pattern	2	0.5006	0.25029	611.225	8.44e-13	***
activation	3	0.4036	0.13452	328.509	8.84e-12	***
learning.rate	2	0.0032	0.00162	3.962	0.0477	*
pattern:activation	6	0.1757	0.02928	71.499	1.08e-08	***
activation:learning.rate	6	0.0040	0.00066	1.613	0.2262	
pattern:learning.rate	4	0.0010	0.00025	0.599	0.6708	
Residuals	12	0.0049	0.00041			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 1: ANOVA table

activation <fctr>	test.error <dbl>
Linear	0.25
ReLU	0.27
Sigmoid	0.51
Tanh	0.27

Table 2: Test error means for activation at each level

Experimental Factors	Description	Levels
Activation Function (Treatment)	Critical in directly influencing model performance by introducing non-linear transformations in neural networks.	**ReLU:** Widely used due to its simplicity and computational efficiency, with a low likelihood of vanishing gradients. **Tanh:** Similar to Sigmoid but is better with vanishing gradient due to zero-centered **Sigmoid:** Provide contrasting properties with smoother gradients for comparative analysis. **Linear:** Acts as a baseline to evaluate the necessity of non-linear transformations.
Learning Rate (Treatment)	Learning rates determine the magnitude of weight updates to optimize model performance. Including different levels ensures the exploration of their effects on convergence.	Low (0.001), Medium and (0.003), High (0.01) are chosen to cover a broad spectrum of convergence speeds while avoiding excessively high values that could cause divergence.
Data Patterns (Blocking)	Patterns like are chosen to introduce a variety of non-linear classification challenges, allowing insights into generalizability.	Spiral, Gaussian Mixture, and Concentric Circles represent varying degrees of non-linearity, allowing a

		robust assessment.

Table 3: Description of Experimental Factors and Levels

Description: df [6 × 4]

	pattern <chr>	activation <chr>	learning.rate <dbl>	test.error <dbl>
1	spiral	ReLU	0.001	0.496
2	spiral	ReLU	0.003	0.496
3	spiral	ReLU	0.010	0.465
4	spiral	Tanh	0.001	0.493
5	spiral	Tanh	0.003	0.487
6	spiral	Tanh	0.010	0.421

6 rows

Table 4: Data Preview

Figures

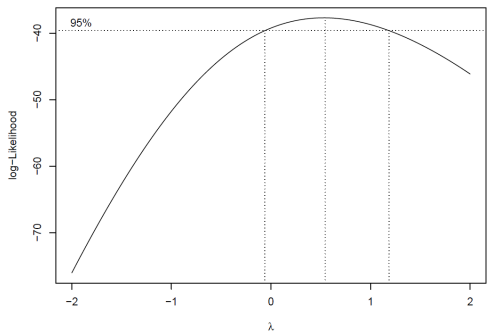


Figure 1: Boxcox plot of test error by pattern

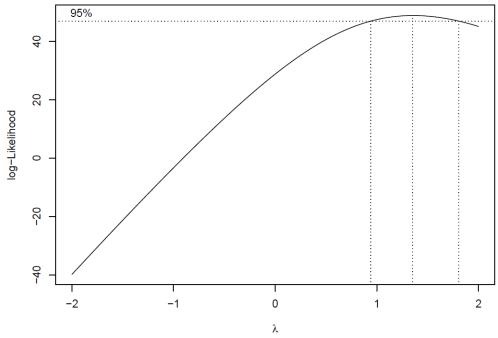
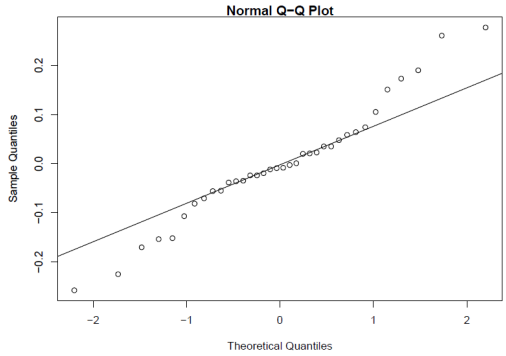


Figure 2: Boxcox plot of test error by model



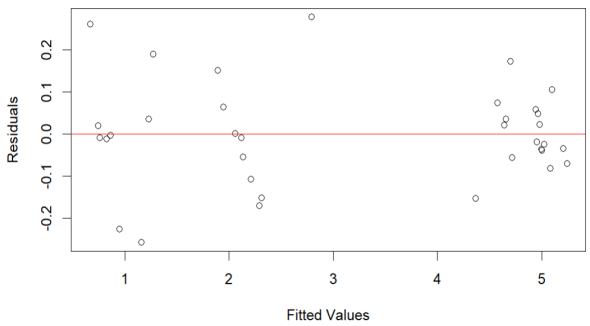


Figure 3: QQ plot of residuals

Figure 4: residuals vs. fitted values

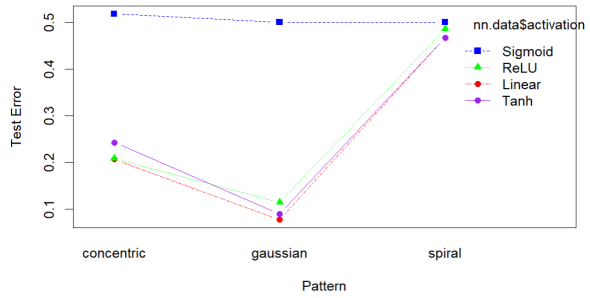


Figure 5: Interaction plot: Pattern vs Test error by activation

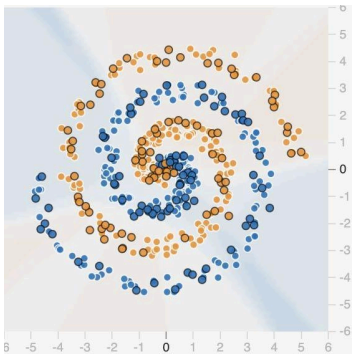


Figure 6: Spiral Data pattern

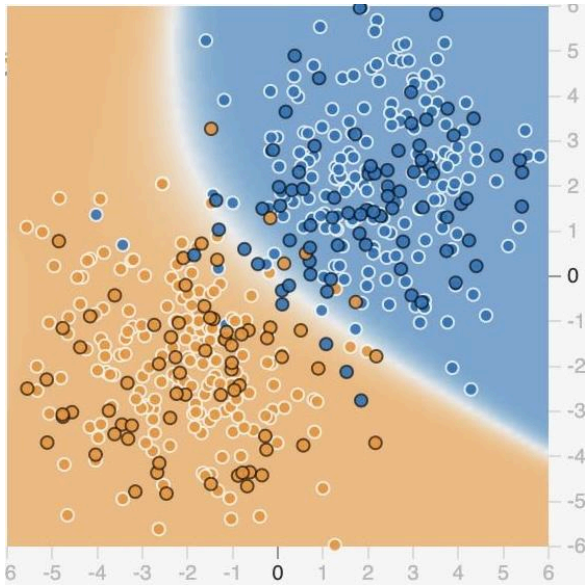


Figure 7: Gaussian Mixture Data pattern

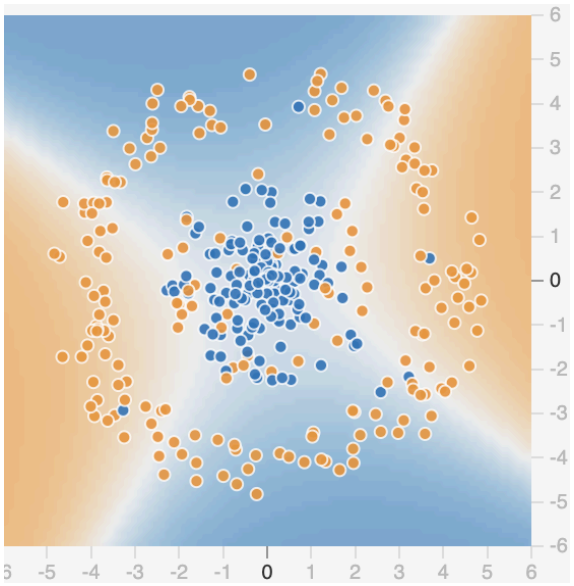


Figure 8: Concentric Circles Data pattern

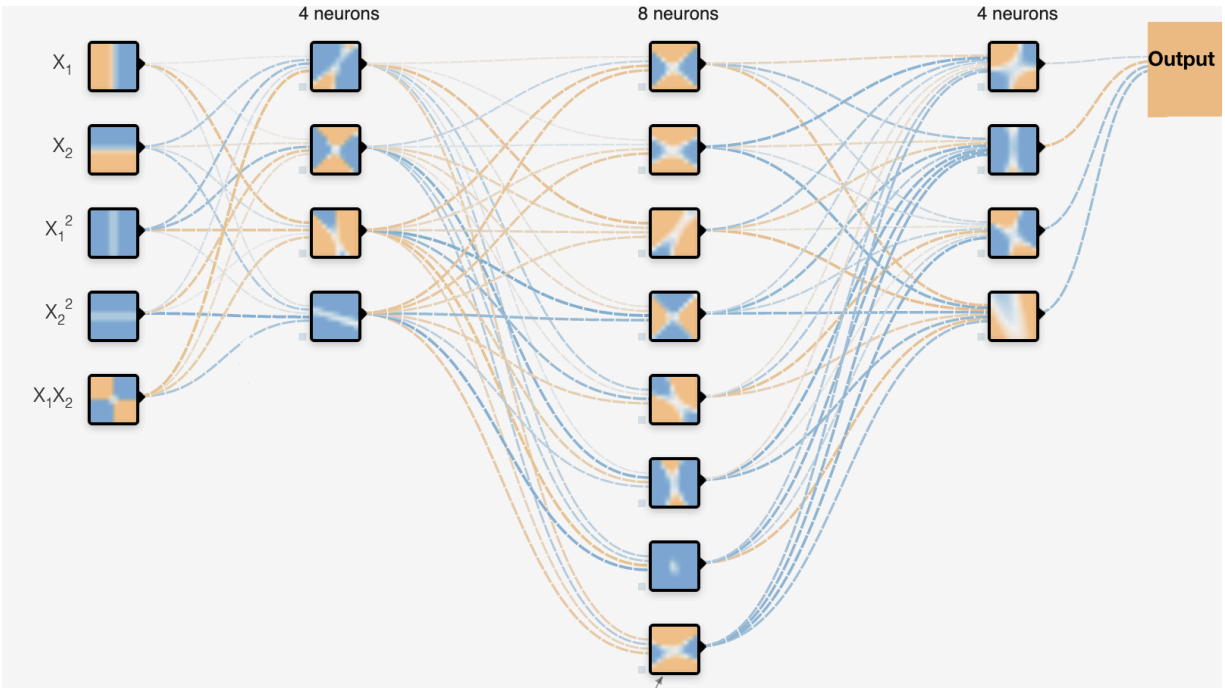


Figure 9: Fixed Neural Network Architecture

Units

The test error is dimensionless, while the learning rate is also dimensionless. The activation function has no units as it represents categorical factors, while data patterns (spiral, concentric, gaussian) are also dimensionless.

Data Appendix

Variable	Description	Variable Type
pattern	The type of data pattern used in the experiment with levels spiral, concentric, and gaussian.	Categorical
activation	The activation function used in the neural network with levels Linear, ReLU, Sigmoid, and Tanh	Categorical

learning.rate	The step size used for weight updates in the neural network during training (0.001, 0.003, 0.01)	Categorical
test.error	The error rate measured using the mean square error recorded at the final epoch of training, indicating the model's performance.	Continuous
X_1	The first feature representing the x-axis Cartesian coordinate for each data point.	Continuous
X_2	The second feature representing the y-axis Cartesian coordinate for each data point.	Continuous