Abalone Dataset: Predict the Ring age in years

Ming Yin, Haoting Qi, Anqi Chen

# Introduction

The concept of deep learning (DL) was first introduced by Geoffrey Hinton, a pioneer in the field of neural networks and machine learning. In 2006, Hinton published a paper titled "Reducing the Dimensionality of Data with Neural Networks" in Science. The paper proposed an unsupervised layer-wise pretraining strategy, which addressed the vanishing gradient problem in training deep neural networks. This method allowed deep neural networks to better learn hierarchical features from data, leading to significant progress in tasks such as image recognition, speech processing, and natural language understanding.

Since then, deep learning has been widely adopted in both academia and industry. Deep neural networks (DNNs) have become the backbone for complex tasks such as computer vision, speech recognition, and natural language processing. The strength of deep learning lies in its ability to automatically extract complex, nonlinear features from high-dimensional data. The pretraining method proposed by Hinton enabled deep networks to handle more complex data structures, surpassing the limitations of traditional shallow models.

In this paper, we explore the similarities and differences between deep learning and traditional neural networks, with a focus on their performance in regression and classification tasks. Using the Abalone dataset, we compare the predictive power of traditional models such as linear regression and logistic regression with deep learning models, and examine the potential of deep learning to improve prediction accuracy.

Despite the significant advancements in both deep learning and traditional machine learning models, there is still insufficient research comparing the performance of these models on specific datasets, particularly in biological applications like age prediction from physical measurements. While traditional models such as linear and logistic regression have been widely used due to their simplicity and interpretability, their predictive power may be limited when dealing with complex, nonlinear relationships in data. On the other hand, deep learning models, especially deep neural networks, are known for their ability to capture such relationships. However, their performance and benefits over simpler models in tasks like age prediction from the Abalone dataset remain underexplored.

This project aims to fill this gap by systematically comparing the predictive performance of traditional models and deep learning models. By investigating how well these models handle both regression and classification tasks using the same dataset, we seek to identify whether the added complexity of deep learning yields significant improvements in predictive accuracy, especially when compared to well-established, simpler models.

In this project, we investigate the predictive performance of traditional machine learning models and deep learning models on the Abalone dataset for age prediction (measured by the number of rings). Specifically, we compare the effectiveness of linear regression, logistic regression, and various deep learning models in both regression and binary classification tasks. For the classification task, we focus on distinguishing between abalones aged below 7 and those aged 7 or above based on their physical measurements.

Our main goals are:

* To evaluate the performance of linear regression and logistic regression, both with and without normalization, in predicting age and classifying age groups.
* To explore the effectiveness of deep learning models, including neural networks implemented in TensorFlow and PyTorch, trained using stochastic gradient descent (SGD) with various configurations (e.g., different numbers of hidden layers and neurons, learning rates).
* To compare the impact of data preprocessing techniques (such as normalization and feature selection) on the accuracy and performance of the models.
* To analyze and report key performance metrics such as RMSE, R-squared, AUC, and accuracy for both regression and classification tasks across 30 experiments, with the results documented and compared.

Through this study, we aim to determine if the added complexity of deep learning models significantly improves predictive accuracy over traditional models when applied to biological data such as the Abalone dataset.

# Methodology

## Data description and pre-processing

The data selected for this paper was used to predict the age of abalone by physical measurements. This data has a total of 4,177 samples with a total of 9 observations, where the age of the abalone (abalone rings) was selected as the explanatory variable for the test. For the data, the details are shown in Table1.

1. Compare two different SVM models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Name | Type | Unit | range | Mean | Description |
| Independent  Variable | Length | continuous | mm | [0.075,0.815] | 0.524 | perpendicular to length |
| Diameter | mm | [0.055,0.650] | 0.4079 | Longest shell measurement |
| Height | mm | (0.000,1.130] | 0.1395 | with meat in shell |
| Whole weight | grams | [0.002,2.8255] | 0.8288 | whole abalone |
| Shucked weight | grams | [0.001,1.488] | 0.3594 | weight of meat |
| Viscera weight | grams | [0.0005,0.7600] | 0.18061 | gut weight (after bleeding) |
| Shell weight | grams | [0.0015,1.0050] | 0.2389 | after being dried |
| Dependent  Variable | Sex | discrete | — | M, F, I (infant) | — | Gender |
| Rings | — | n∈N∩[1.29] | 9.932 | Rings for age |

Table1 shows that this data is numerical except for the variable Sex, so in this paper, gender M (male) is mapped to 0, F (female) to 1, and I (infant) to 2. as their labels in the model. In order to explore the correlation relationship between the variables, a heat map was used to plot the correlation Figure 1.

Looking at Figure 1.1, the correlation between Rings and Shell weight is 0.63, which is the highest correlation of Rings, indicating that shell weight is a better predictor of age. The correlations between Rings and Length, Diameter and Height are between 0.56 and 0.57, indicating that the size of abalone is moderately correlated with age, the larger the size, the older the age may be, but it is not a very strong linear relationship.The correlation between Rings and Sex is -0.35, and there is a weak negative correlation between them.

Using Figure 1.1, pick the two most relevant variables (one positive and one negative) to Rings, Shell weight and Sex, and create a scatter plot of them against Rings. Due to the large number of samples, there is an overlap of points. The number is indicated by the colour, when the colour is dark there is a high number and when the colour is light there is a low number.

The relationship between Rings and the other two variables can be determined by the colour shades. The first plot in Figure 1.2 shows that as Shell weight increases, Rings also tends to increase, and although there is some dispersion in the data, the overall trend is clear. For the second plot, since Sex is a discrete variable, the data points present three vertical columns corresponding to different sexes (coded 0, 1, 2). For these three sexes, Rings are more numerous in the interval 5-20. However, the negative correlation between Rings and Sex is weak and there is no clear trend.

For the first graph, the distribution of Shell\_weight, it can be seen that the distribution of this data shows a clear positive right-skewed, with the majority of shell weights concentrated in the range of 0.1 to 0.4, and a small number of individuals with shell weights above 0.6. Therefore, the

1. Scatter plot of Rings versus Shell weight and Sex.

values of shell weight are small and concentrated in the low range, with fewer larger shell weights.For the distribution of Diameter, the distribution of the diameter data is close to

1. Scatter plot of Rings versus Shell weight and Sex.

normal, with most individuals having diameters in the range of 0.3 to 0.5. This indicates that the data are roughly symmetrical.Regarding the distribution of Rings, the image shows a clear positive skewed distribution, with the majority of individuals having a Rings number between 5 and 15, and fewer individuals having a Rings number greater than 20. This indicates that the majority of individuals are young or middle-aged, with only a small number of individuals being older.

## Model and optimisation methods

In this paper, two models are used, one model is Liner model and the other model is neural network. the advantages and disadvantages of the models are determined by comparing the final prediction results. The principle of the model is as follows.

### Liner model

Linear models are a class of models for regression and classification tasks that assume a linear relationship between output and input features. In simple linear regression, this relationship is represented by a linear equation. Its mathematical expression is as follows:



A simplified representation in matrix form is:



In linear regression, for parameter estimation, the most commonly used method for estimating parameters is Sum of Squared Errors (SSE), which finds the best-fitting parameter values by least squares:



By taking the partial derivatives of the above objective function with respect to β and setting them to zero, a closed solution mini-mising the sum of squares of the errors can be obtained, which is actually a process of finding the gradient:





This results in an optimal solution to the linear regression, i.e., the solution that minimises the sum of squared errors. However, the matrixis invertible. When there is multicollinearity between the features, may not be invertible and regularised regression is required.

### Logistics model

Logistic regression is actually a classification algorithm for solving binary classification problems, which predicts the probability of a category and maps the output values of a linear combination via the Logistic function to the [0,1]. The sigmoid function is as follows:



At this point the probability is calculated and if classified according to the p(0<p<1), there is the following classification:



Logistic regression uses a Log-Likelihood Function to estimate the model parameters β. The log-likelihood function is of the form:

 The optimal parameters are found by maximising this log-likelihood function β.

### Neural network

For the Multilayer Perceptron (MLP) model, which solves regression problems, it is essentially a feed-forward neural network that is trained by a back-propagation algorithm, and SGD is commonly used to optimise the weights. Although MLP is essentially a non-linear model, MLP can be used as a linear regression model if no activation function is introduced.

MLP is consisting of an input layer, a hidden layer and an output layer. Assuming that the model has L layers of neural networks, the output of layer l can be expressed as:



Among them:

is the input of layer l (linear combination). is the l layer of the weight matrix. is the output of the previous layer. is the l layer of the bias vector.

For the regression problem, the MLP uses the Mean Squared Error (MSE) as a loss function of the form:



Next, Stochastic Gradient Descent (SGD) is used to optimise the model parameters. At each iteration, the weights and biases are updated using a small batch or a sample of the dataset, gradually reducing the loss function values. The formula for updating the parameters each time is:





Through continuous iteration, SGD adjusts the weights W and bias b until the loss function is minimised and the optimal linear mapping relation is found.

### Normalization

The goal of normalisation is to scale the data to a specified range (usually between [0, 1]) with the following formula:



Normalisation is important when the distribution of the data has no obvious shape characteristics, or when the data is within a specific range. Normalisation is also often used when neural networks are involved, for instance, which require the input data to be within a certain range.

## Software suite

For this experiment, this article mainly who used these libraries in python, its main functions are as follows.

In addition, this paper sets the seed through the random.seed function of the random library, and the value of the random seed is related to the number of runs. Thus, the repeatability of the results of this experiment is achieved.

# Results

|  |  |  |  |
| --- | --- | --- | --- |
| Library | Goal | Function | Description |
| Pandas | preprocessing | — | provides a DataFrame data structure for storing and manipulating |
| NumPy | computing | — | providing support for multi-dimensional arrays and matrix |
| Matplotlib | drawing | — | suitable for a variety of data visualisation needs |
| Scikit-learn | Classification and regression | train\_test\_split | split the dataset into training and test sets |
| LinearRegression | for linear regression model construction to predict variables. |
| LogisticRegression | logistic regression models, for binary classification tasks. |
| MLPRegressor | multilayer perceptual machine regression through neural networks |
| evaluate models | mean\_squared\_error | calculates the mean square error |
| r2\_score | calculates the R2 score, values closer to 1 representing a better |
| roc\_auc\_score | calculates the area under the ROC curve (AUC) |
| roc\_curve | generate ROC curve data for performance of the classification |
| preprocessing | MinMaxScaler | for data normalisation to scale the data to [0, 1] |
| torch | modelling | nn.Module | base class of a neural network model ( a trainable module) |
| nn.Linear | fully connected neuron layer |
| nn.ReLU | common activation function, Rectified Linear Unit (ReLU) |
| nn.MSELoss | loss function (MSE) |
| optim.SGD | stochastic Gradient Descent (SGD) |
| Keras | modelling | models.Sequential | stacking the layers of a neural network in order |
| layers.Dense | is the fully connected layer in Keras |
| layers.Input | layers that define the shape of the model input |
| optimizers.SGD | stochastic Gradient Descent (SGD) |

1. Functions used in python and their libraries

# Discussion

## Results Summary

This project evaluates the performance of various methods for age prediction using linear regression, logistic regression models, and neural networks trained with stochastic gradient descent (SGD) on the abalone dataset. The results indicate that the models have achieved our project goals.

The following two figures summarize the evaluation results of the different methods:

Key observations from the data:

The standard linear regression model achieved a root mean square error (RMSE) of 2.11 and an r-squared value of 0.51, indicating moderate predictive ability. Normalizing the features kept the RMSE constant but increased the R-squared to 0.54, indicating that the model fit was enhanced. In contrast, the RMSE of the linear regression model using only two features increased to 2.34, showing a decrease in performance due to the loss of potentially critical information. The neural network model trained using stochastic gradient descent showed a significant advantage with a minimum RMSE of 2.05 and a maximum R-squared of 0.57, highlighting its effectiveness in capturing the complex relationships in the data.

## Methodology Limitations

### Over-fitting

Neural networks typically consist of a large number of parameters, which increases their ability to adapt to complex data models. However, all three training methods showed better performance on the training set compared to the test set, suggesting that over-fitting may be present. Over-fitting occurs when the model captures patterns specific to the training data, resulting in reduced performance on the unseen test data. This suggests that the model may be difficult to effectively generalize to new data, thus affecting its predictive accuracy and robustness in real-world applications.

### Insufficient Hyper-parameter Search

Despite the initial exploration of neural network models with specific combinations of hyper-parameters (e.g., number of layers, learning rate, etc.). However, the potential capabilities of the models may not have been fully explored due to the more limited scope of configuration search. The restricted space for hyper-parameter tuning may result in the model failing to achieve its theoretically optimal performance, especially when dealing with complex high-dimensional data, making it difficult to fully capture its generalization ability and optimal solution. Further hyper-parameter optimization processes (e.g., grid searches or stochastic searches) are especially necessary to more fully evaluate the model's performance, thus fully exploring the performance limits of neural networks.

### Insufficient Performance

In this project, the highest R-squared value obtained by the model is about 0.57, indicating that it has significant limitations in explaining the variance of the dependent variable, and that the currently adopted neural network architecture may be too simple to effectively capture the underlying complex relationships in the data. Specifically, the simplified model design may lead to insufficient generalization of the model when dealing with high-dimensional data, further affecting its prediction performance on unknown samples. Therefore, in order to improve the explanatory ability and prediction accuracy of the model, it will be necessary to consider introducing more complex network structures or deeper feature engineering.

## Improved Modeling

### Regularization Techniques

In order to reduce the risk of over-fitting, various regularization techniques can be employed. For example, methods such as L1 regularization (Lasso) and L2 regularization (Ridge) effectively limit the complexity of the model by introducing a penalty term, thus enhancing the model's generalization ability on unseen data. In addition, the introduction of a dropout layer in the neural network architecture can further reduce the over-fitting phenomenon and improve the robustness and reliability of the model by randomly disabling a certain percentage of neurons during the training process. These regularization methods are widely used in modern machine learning to ensure that the model maintains good predictive performance when dealing with complex data.

### Hyper-Parameter Optimization

Implement a more comprehensive hyper-parameter optimization strategy for the model. For example, Bayesian optimization is an effective method that can outperform traditional grid search or random search in the identification of hyper-parameter combinations. This approach evaluates the hyper-parameter space by constructing a probabilistic model, which enables broader exploration and more efficient search. This optimization strategy can significantly improve the performance of the model and better tune the model parameters to specific data characteristics and task requirements in complex tasks.

### Complex Model Architecture

To enhance the model's ability to capture complex relationships in the data, more advanced neural network architectures should be considered. Specifically, utilizing deep networks with additional hidden layers, or depending on the characteristics of the data, architectures such as Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) may provide significant performance gains. These architectures have a greater ability to effectively learn the complex patterns and dependencies present in high-dimensional datasets, leading to more accurate predictions and more reliable model generalization capabilities.

## Conclusions

### Major Contributions

This project provides a comprehensive comparison between traditional machine learning models (e.g., linear regression and logistic regression) and deep learning models, highlighting the strengths and weaknesses of each approach in predicting age and categorizing age groups in abalone datasets. By systematically evaluating the performance of regression models with and without normalization, this project highlights the importance of data cleaning techniques in machine learning tasks, especially in biological datasets where there can be significant variation in feature scales, and this contribution helps to reinforce best practices in data preparation.

In addition, this project explores deep neural network (NN) methods with a focus on hyper-parameter optimization, including the number of hidden layers, the number of neurons, and the learning rate. By systematically tuning these parameters, this research identifies optimal configurations in terms of improving prediction accuracy and minimizing error metrics. This analysis emphasizes the importance of balancing model complexity to prevent over-fitting and points out the significant impact of activation function and learning rate on convergence speed. These insights provide actionable guidance for researchers aiming to optimize neural network design for biological data applications.

Finally, this analysis reports in detail key performance metrics such as root mean square error (RMSE) and R-squared value. By recording and comparing these metrics across 30 experiments, the study establishes a robust framework for evaluating model effectiveness, providing an important reference point for future predictive modeling research.

### Directions for Future Research

Future research could extend the findings of this project by investigating several key areas. These include the application of more advanced machine learning techniques, such as integrated methods and hybrid models, to improve prediction accuracy. Additionally, exploration of feature selection and dimension reduction techniques could improve model performance by reducing complexity. Finally, examining the applicability of these models across different datasets and domains may further validate their robustness and generalization.

1. Dou J, Yunus A P, Bui D T, et al. Dou, Jie, et al. "Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan." pp.641-658, 2020.