Callealta Andres | A01763701

August 2024

Getting abalone’s age with linear regression

Introduction

In this project, I worked on codding a linear regression model to predict the number of rings in abalones, which are marine mollusks. The goal was to create a model that can make accurate predictions based on various physical measurements of abalones.

Note: the point of getting the numbers of rings is because the age in years of an abalone is equal to the number of rings plus 1.5.

Why this project?

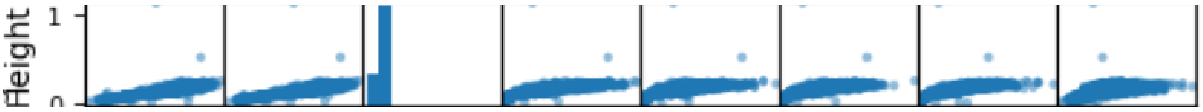
Abalone rings are used to estimate the age of the mollusks, which is important for understanding their growth and managing fisheries. By predicting the number of rings from measurements such as length and weight, we can help in research and management of abalone populations. Otherwise, some people have to count the rings manually, which is a penible job.

What We Did

1. **Data Preparation:**
    - We started with a dataset containing various measurements of abalones, such as length, diameter, and weight.
    - We removed columns that were not useful for our model, specifically "Sex" and "Height", to focus on the features most relevant for predicting the number of rings.

- The dataset was then shuffled to ensure random distribution of data points before splitting it into training and validation sets. This helps in getting a fair evaluation of the model's performance.

  Note 1: Here is the plotting scatter of the "Height" variable, I decided to not use it because we do not see a strong correlation with any other variable, specially with rings, the last one.



  Note 2: I also decided to delete the "Sex" variable because it the only one that is not a continuous/ quantitative variable but a categorical one.

2. **Model development:**
   - We implemented a linear regression algorithm from scratch. Linear regression is a method used to model the relationship between a dependent variable (number of rings) and one or more independent variables (physical measurements).
   - We trained the model using gradient descent, which iteratively adjusts the model's parameters to minimize the error in predictions.

3. **Evaluation:**
   - After training, we evaluated the model's performance by measuring how well it predicted the number of rings on both the training and validation datasets.
   - We calculated the R-squared value, which indicates how well the model explains the variation in the number of rings. A higher R-squared value means better performance.

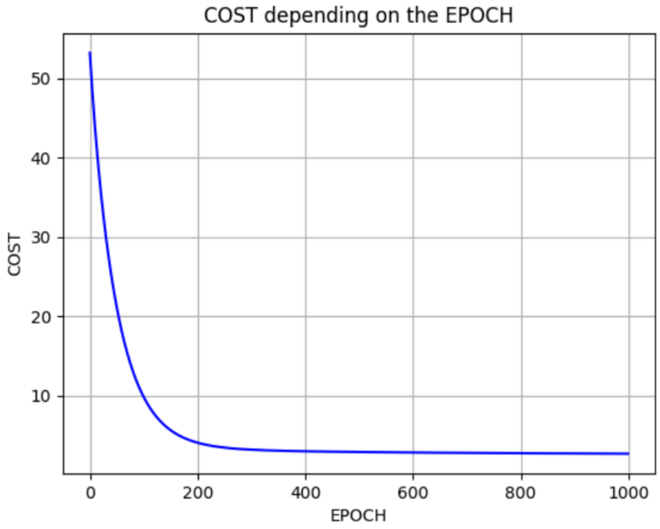
4. **Visualization:**
   - We created a graph to visualize how the model's cost (a measure of prediction error) decreased over time as the model was trained. This helps us understand if the model was learning effectively

Results

- **Final Model Coefficients:** The values of the model parameters that were learned during training.

  Coefficients: [9.91719636 0.43354534 0.86623476 0.35954118 -1.68803941 -0.24365662 2.12608417]

- **Cost History:** The progression of the model's prediction error over training epochs:



- **Prediction Accuracy:** We calculated the percentage of correctly predicted rings and compared the model's performance on training and validation data.

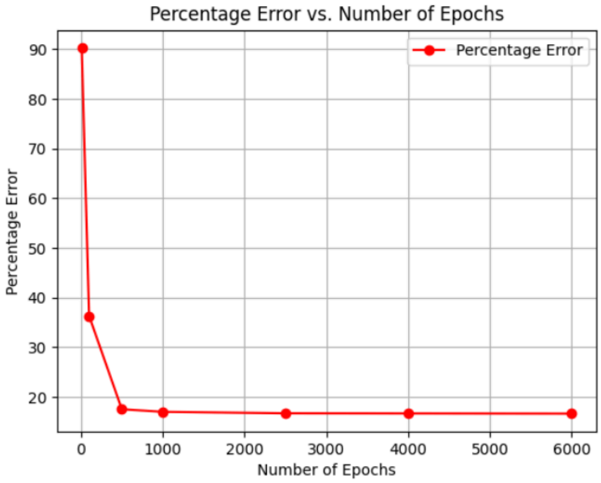
  Here with the parameters of 1000 epochs and a learning rate of 0.01 we have a percentage of correct predictions of 83.01%.

- **R-squared Values:** This metric showed how well our model's predictions matched the actual values.

  Here we have R-SQUARED training of 0.4809 and a R-SQUARED validation of 0.4896.

  A good model should have a R-Squared value of 1, between zero and 1 is a normal value.

- Finally, here is a graph of the progression of the errors in the predictions, related to the number of epochs. The code used for this one in *#comments*, in the same python file.



In summary, this project aimed to build a basic linear regression model to predict the number of rings in abalones, evaluate its performance, and understand its learning process through visualizations and metrics.