MALIS Project 1 Report:
Linear regression for classification

EL BACHA Tonia
SAADE Tarek

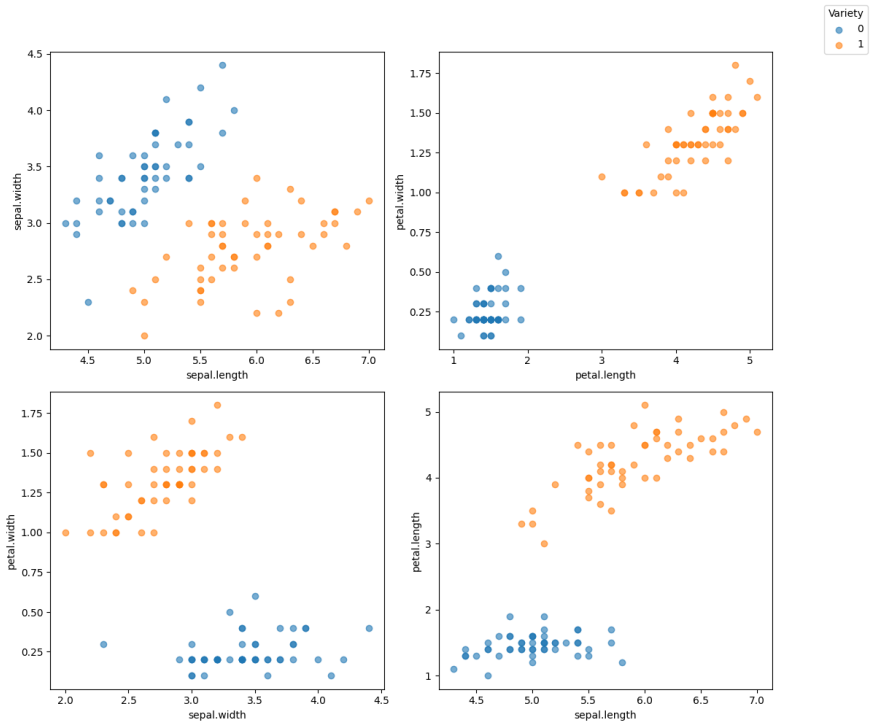
November 11, 2024

1 Introduction

The primary objective of this project is to explore the use of linear regression in classification tasks and examine its limitations in this context. To achieve this, we developed a binary classifier for the first task, followed by a multi-class classifier in the second task, both using linear regression as the foundational model.
The project leverages tools from the **scikit-learn** library, including the **LinearRegression** model, **train-test-split** for dividing the data into training, validation, and testing sets, and **accuracy-score** to evaluate model performance. Additionally, **pandas** was used for handling the dataset as data frames, facilitating data manipulation and processing.

2 Binary Classifier

The original Iris dataset contains three flower classes. To perform binary classification, we filtered the dataset to retain only two classes—Setosa and Versicolor—encoded as binary labels (Setosa as 0 and Versicolor as 1). The data was then split into training (70%), validation (15%), and testing (15%) sets.
A linear regression model was trained on the training set, followed by predictions on the validation set. Since these predictions are continuous, a decision rule was applied to convert them into binary classifications. Given that the target labels are 0 and 1, a natural threshold of 0.5 was set to distinguish the classes: predictions equal to or greater than 0.5 were classified as Versicolor (1), while values below 0.5 were classified as Setosa (0).
After training, the model's performance was validated on the validation set and subsequently tested on the test set, achieving an accuracy score of 1.0 (100%) on the test data.



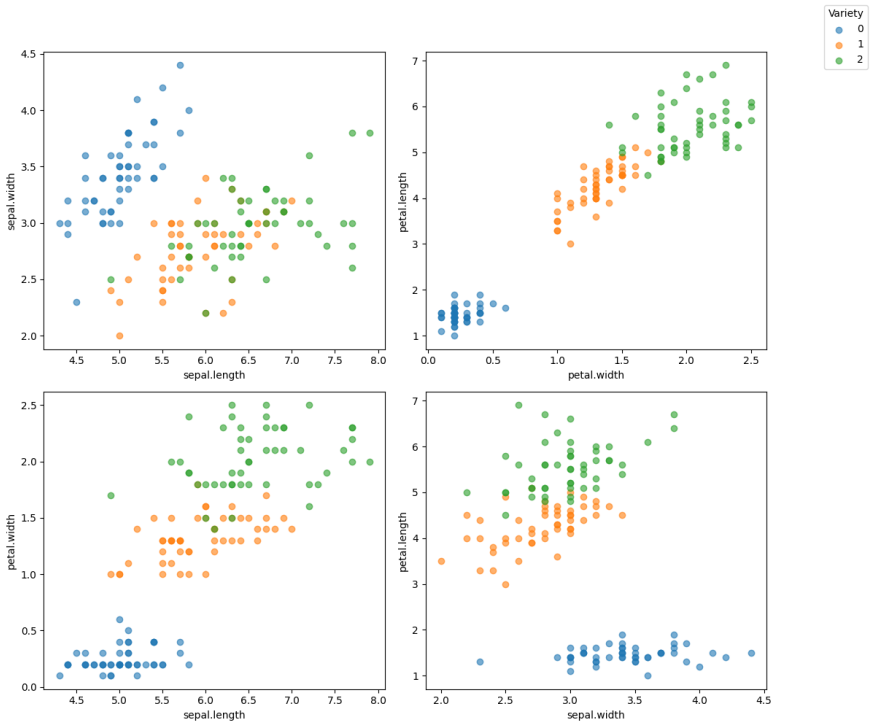
This result was expected, as visualizing various feature pairs for Setosa and Versicolor reveals clear linear separability between the two classes.

3 Multi-Class Classifier

For the multi-class classification task, we encoded the three flower classes in the Iris dataset as labels 0, 1, and 2. The implementation of the multi-class classifier using linear regression followed a similar approach to the binary classifier, but required two decision thresholds to separate the three classes. Given the label values, the first threshold was set between 0 and 1, and the second between 1 and 2.
Initially, we set the thresholds to 0.5 and 1.5, which resulted in an accuracy score of 0.8 on the validation set. After tuning the thresholds to 0.7 and 1.2, the accuracy improved to 1.0. Encouraged by this result, we then tested the model on the test set, where it also achieved an accuracy score of 1.0.

4 Discussion

- In testing the binary model across all combinations of class pairs, with the threshold consistently set at the midpoint between the labels, the model achieved an accuracy of 1.0 for all cases.



- While visualizations of feature pairs for the three classes show some overlap, indicating that individual features alone may not linearly separate the classes, it appears that the classes are linearly separable when considering all four feature dimensions.

In conclusion, linear regression proves to be a simple yet effective tool for classification tasks, and it can be extended to multi-class problems as well. However, its primary limitation is its reliance on linear separability, which restricts its performance on more complex, non-linear data distributions.

Tonia EL BACHA was responsible for the task 1
Tarek SAADE was responsible for task2
Report and conclusions where made together.