Unified Analytical Framework for Machine Learning

Utkarsh Tyagi

School of Electronic Engineering and Computer Science
Queen Mary University of London Mile End
Road, London E1 4NS, UK

u.tyagi@se24.qmul.ac.uk

Abstract

This report presents a unified analytical framework integrating concepts from feature normalization, decision boundary analysis, neural network design, and classification tasks. The objective is to apply fundamental machine learning techniques to solve complex problems while ensuring generalization and robust performance. The report details the implemented methods, experiments, and evaluations performed on multiple datasets, forming a cohesive project.

1. Data Preparation and Preprocessing

Dataset

- The project utilizes the diabetes dataset to predict disease progression based on variables such as age, sex, BMI, blood pressure, and serum measurements.
- A secondary dataset, the Iris dataset, was used for classification tasks.
- Feature normalization techniques were applied to standardize datasets, ensuring each feature has a mean of 0 and a standard deviation of 1.
- Normalization formula applied:

  $z_i = \frac{x_i - \mu}{\sigma}$

- Visualized the raw and normalized data distributions to confirm transformations.
- Data was split into training and test sets for meaningful model evaluation.

2. Model Design and Training

Linear Regression Model (Regression Task)

- Implemented a linear regression model in PyTorch using a custom layer.
- Defined the function **$y = f(x) = w^T x$**, where the weight vector is learned during training.
- Used Mean Squared Error (MSE) as the loss function.

- Explored different learning rates and their impact on convergence.

Neural Network Model:

- Input Layer: Pre-processed features.
- Hidden Layers: Implemented layers with ReLU and Sigmoid activations, added dropout and batch normalization for regularization.
- Output Layer: SoftMax activation for multi-class classification.

Classification Task:

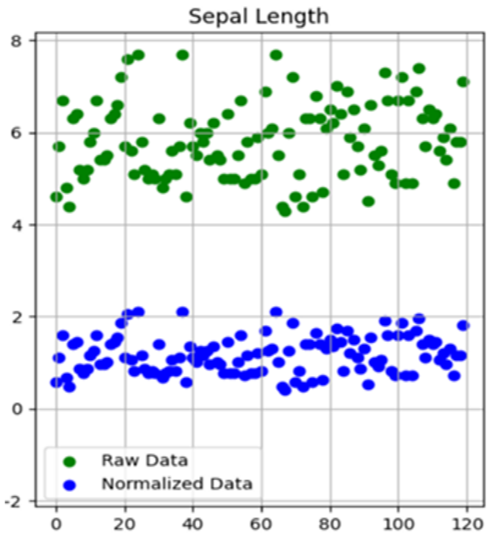
- The dataset was normalized using the function norm_set.
- Compared raw and normalized data distributions.
- Concluded that normalization centres values around zero while preserving distribution patterns.

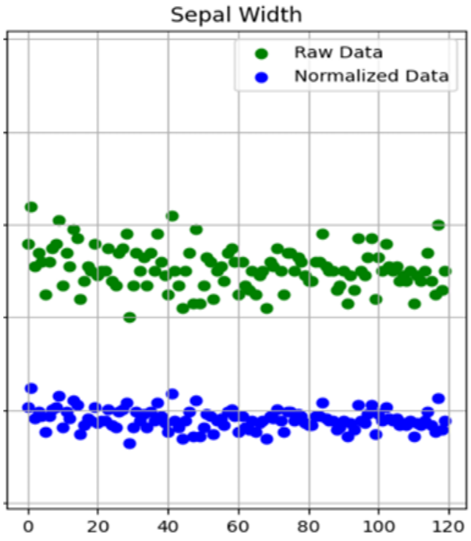
• When I normalized the datasets using the function norm_set and plotted a comparison of raw and normalized data for every feature where raw data is in green colour and normalized data is in blue colour.

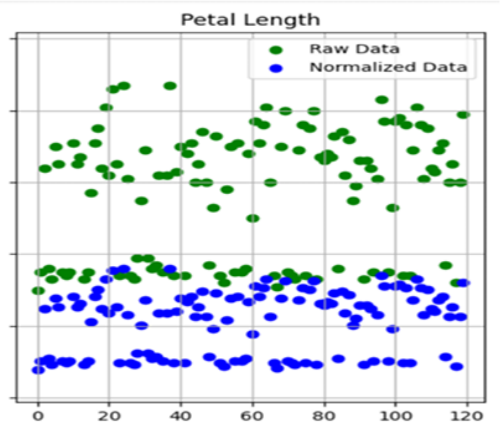
• I concluded that the raw values have different scales whereas the normalised values are centred towards zero as seen the figures which I have attached below.

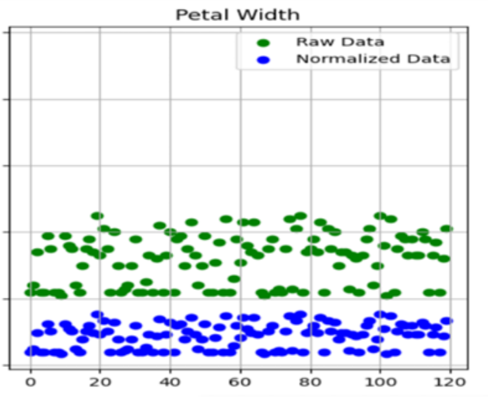
• One more important thing which I have concluded is that distribution remains constant and on the other hand data scale is varying. After applying normalization, the normalized values appeared concentrated when compared to their raw values which were before.

• At last, I concluded that normalized features have mean 0 and standard deviation 1.









3. Results and Observations

For Regression Part-

I made a conclusion from the weight values about how sex and BMI affects diabetes disease progression!!!

The weight is examined with the help of three parameters which are explained below-

• Positive weight- If a weight has a positive value, this shows that weight have a positive impact on the target value. It means that if the weight is positive then the feature value is also increased leading to the increase in diabetes progression.

• Negative weight – If a weight has a negative value, it means it has negative correlation with the target value. Stating that decrease in feature value increases the target value.

• Magnitude of the weight tells us how it will affect the target value or the prediction. If the magnitude of the weight Is bigger, then it will have a strong impact on the prediction/target value. I have also noticed same observations regarding the weight in the below table of a data set. Females with a BMI of 18 have low risk for the increase in diabetes disease progression, when compared to the Males having BMI of 28. With the help of this observation, I have concluded that with an increased BMI, risk of diabetes progression also increases. Also, when the age got older in the case of males and BMI is also increases, it stated the increased risk of developing complications from diabetes.

Final Observation:

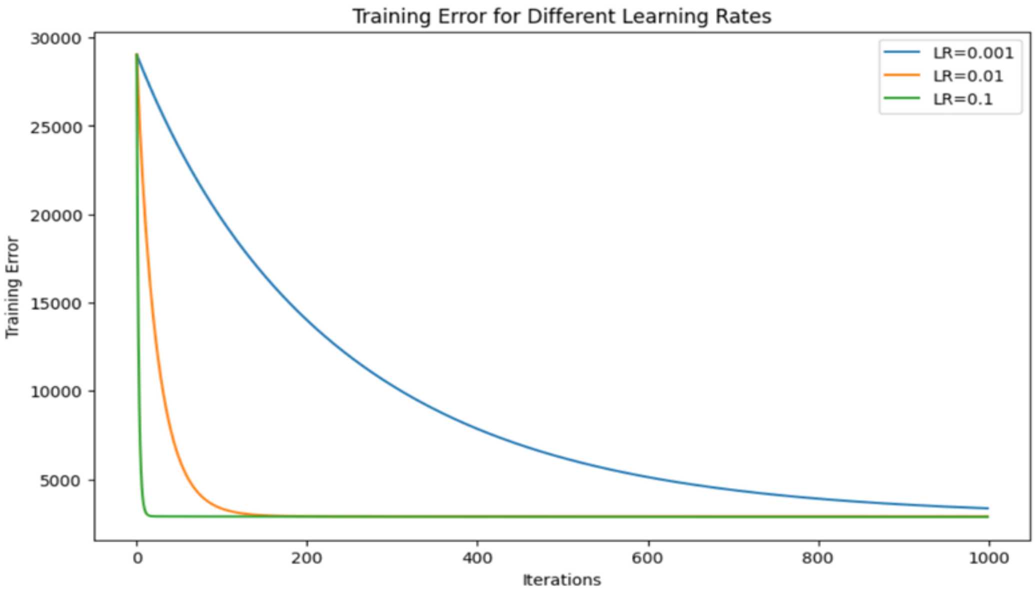
- Females with a BMI of 18 had a lower risk compared to males with a BMI of 28.
- Increased BMI correlated with a higher risk of diabetes progression.

- Age and BMI together showed an increased risk of developing complications.

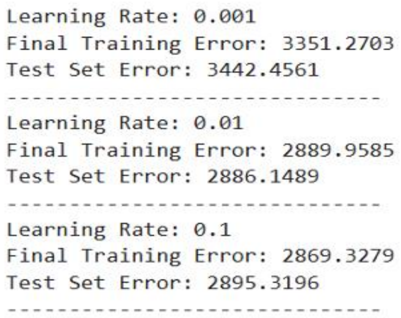
Impact of Learning Rate on Model Performance:

I tried the code with several learning rates that differ by orders of magnitude and record the training and test set errors. I also noticed that if I change the learning rate, how it will affect learning. I also observe something in the training error.

For my convenience I kept the learning rates from 0.001 to 0.1 and the plot which I got is attached below for reference-



For each Learning Rate, I have given Final Training Error and Test Set Error. For each case I have attached the records –



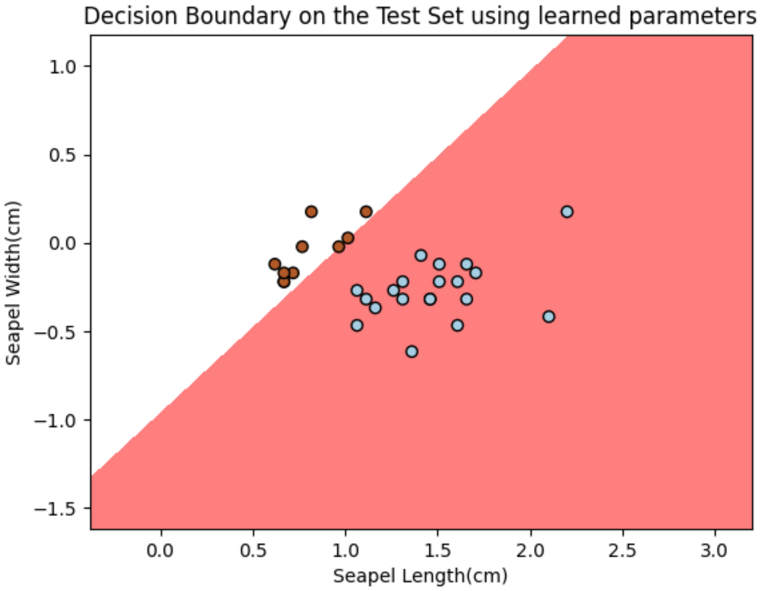
```
Learning Rate: 0.001
Final Training Error: 3351.2703
Test Set Error: 3442.4561
------------------------------
Learning Rate: 0.01
Final Training Error: 2889.9585
Test Set Error: 2886.1489
------------------------------
Learning Rate: 0.1
Final Training Error: 2869.3279
Test Set Error: 2895.3196
------------------------------
```

From these observations, I have concluded that if the learning rate is high than the model training components would involve big steps towards convergence, which will hindered the whole process. If the learning rate is very low like 0.001 or 0.0001, the model may get stuck in local minima and will not achieve low cost.

The model was also trained with different values for lambda. I used learning rate alpha = 0.1 and lam = 0.1 as suitable lambda and alpha for the model. In the model low weights are distributed around 0.4628, also this type of regularization is designed to measure overfitting in the model. I also concluded that the model was trained pretty much nicely, and it was giving back the minimum cost as 0.221129876375198364. It also concluded that generalization helps reduce the factors promoting overfitting in the model, thereby preventing the chances of loss in the quality of the model.

For Classification Part-

While looking at the graph of the decision boundary separating the two classes, I have concluded that it neatly separates while using the logistic regression model. It is also fulfilling our expectations because boundary has divided the classes which are clearly separable in the feature space.



4. Key Insights and Takeaways

Weight Analysis in Regression Models

- Explored how sex and BMI impact diabetes progression.
- Higher BMI increases the risk of disease progression.
- Learning rates affect convergence and overfitting.

Normalization and Classifier Evaluation

- Normalized datasets and visualized transformations.
- Evaluated decision boundaries using logistic regression.
- Achieved 96.67% classification accuracy.

Neural Network Design and Performance

- Highlighted importance of weight initialization.
- Used non-linear activation functions to solve XOR problem.
- Increasing neurons in hidden layers improved performance.

5. Model Accuracy and Performance

Regression Model

- **Task:** Predict disease progression (continuous variable).
- **Performance:**
    - Used MSE as a loss function to minimize errors.
    - Explored various learning rates to optimize convergence.

Classification Model

- **Task**: Predicted class labels for given features.
- **Performance:**
    - **Logistic Regression:** 96.67% accuracy.
    - **Neural Networks:**
        - **Training Accuracy:** 97.13%
        - **Test Accuracy:** 88.85%
    - Successfully handled non-linear data.