Activity 1: Prediction with Back-Propagation and Linear Regression

**A2-FaltaosHany-ParadePatilKuldeep**

Github link : https://github.com/hany019/A1.git

**Part 1: Selecting and analyzing the datasets**

**Preprocessing Overview**

In the initial phase of our project, we conducted a thorough preprocessing routine for three datasets. This process included normalization of variables, shuffling data to eliminate bias, and splitting into training, validation, and testing sets. These steps were encapsulated within a custom function named preprocess\_dataset, which streamlined the workflow for efficient data handling.

* Dataset 1(turbine)
* Dataset 2(synthetic)
* Dataset 3(real\_estate): <https://www.kaggle.com/datasets/elakiricoder/gender-classification-dataset>

**Data Preprocessing Steps**

**Normalization**

We applied the MinMaxScaler to all datasets, scaling input and output variables to a [0, 1] range. This normalization is crucial for the following reasons:

* Uniformity: Ensures that all variables have an equal weight in the model's predictions.
* Model Efficiency: Facilitates algorithms, especially those involving gradient descent, by leading to quicker convergence.
* Practicality: MinMax normalization was an optimal choice since the datasets were pre-cleaned.

**Outlier Detection**

For outlier detection, we employed the Z-score method to pinpoint significant deviations from the mean.

This was crucial in maintaining the integrity of our predictive models:

* Dataset 1(turbine): No outliers were detected, suggesting a well-distributed dataset.
* Dataset 2(synthetic): 13 outliers were identified and subsequently handled to prevent distortions in model predictions.
* Dataset 3(real\_estate): Detected 6 outliers, which were addressed to enhance the dataset's robustness for modeling.

**The implementation code for Dataset 3's outlier detection is as follows:**

**z\_scores3 = np.abs(stats.zscore(dataset3[input\_columns3])) outlier\_indices3 = np.where(z\_scores3 > z\_score\_threshold) outliers\_dataset3 = dataset3.iloc[outlier\_indices3[0]]**

**Data Splitting**

Post-normalization and outlier rectification, we split the datasets as follows:

* Dataset 1: Training (191, 5), Validation (192, 5), Test (68, 5)
* Dataset 2: Training (400, 10), Validation (400, 10), Test (200, 10)
* Dataset 3: Training (165, 7), Validation (166, 7), Test (83, 7)

These divisions are integral to effective model training and evaluation.

**Handling Missing Values**

Missing values in Dataset 3 were managed using mean imputation:

**for column in input\_columns3: dataset3[column] dataset3[column].fillna(dataset3[column].mean())**

**Turbine Power Calculation**

**Background and Rationale**

The dataset for the turbine, referred to as A1-turbine.txt, included variables essential for calculating the power output of a hydroelectrical turbine. However, it lacked direct measurements of the turbine's power output. To bridge this gap, we computed the power of the hydroelectrical turbine, a dependent variable crucial for training our predictive model. This computation was not only necessary to complete the dataset but also to enable the model to learn the relationship between the input features and the power output effectively.

**Computation Methodology**

The power output was calculated using the standard hydroelectric power equation, which is a product of several factors:

* Water Density (ρ): A constant representing the density of water, typically at 1000 kg/m³.
* Gravitational Acceleration (g): The acceleration due to gravity, approximately 9.82 m/s².
* Water Flow (Q): The volume flow rate of water, provided in the dataset.
* Height Difference (ΔH): The vertical distance the water falls, also provided in the dataset.
* Efficiency (e): This factor accounts for the efficiency of the turbine system. Since the turbine's specifications were not provided, and to maintain simplicity, the efficiency factor was not considered in our calculation.

By combining these factors into the equation P = ρ \* g \* Q \* ΔH \* e, we were able to estimate the power output for each instance in the dataset. The absence of the efficiency variable is not expected to significantly impact the analysis due to the relative consistency of turbine efficiency and its minor influence on the comparative power output across different instances.

**Categorical Value Representation**

Although Dataset 3 lacked categorical variables, we prepared to use one-hot encoding or label encoding if necessary.

**Data Normalization**

Both input and output variables were normalized using MinMaxScaler:

**# For input variables dataset3[input\_columns3] = scaler.fit\_transform(dataset3[input\_columns3]) # For the output variable output\_scaler3 = MinMaxScaler() dataset3[output\_column3] = output\_scaler3.fit\_transform(dataset3[[output\_column3]])**

# Part 2: Implementation of BP

**Dataset 1:**

* **Mean Squared Error (MSE): 0.0002219075155183274**
* **R-squared (R2) Score: 0.9961163445541038**

**A screen shot of a computer

Description automatically generated**

Analysis: Dataset 1 exhibits exceptional performance in regression analysis. The model achieved an impressively low MSE, indicating that the predicted values are very close to the actual values. The R2 score of approximately 0.996 suggests that nearly 99.6% of the variance in the dependent variable can be explained by the independent variables. This dataset showcases a highly accurate and precise model fit.

**Dataset 2:**

* **Mean Squared Error (MSE): 0.001559855351424817**
* **R-squared (R2) Score: 0.971593333802795**

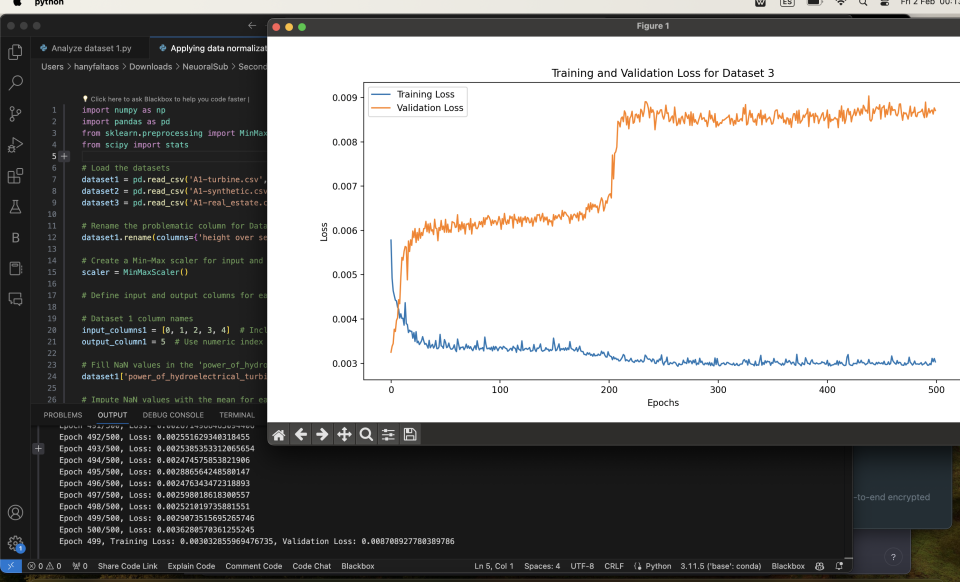
**A screenshot of a computer

Description automatically generated**

Analysis: Dataset 2 also demonstrates strong performance in regression analysis. The MSE, though slightly higher than in Dataset 1, is still quite low, indicating a good fit. The R2 score of approximately 0.972 indicates that about 97.2% of the variance in the dependent variable can be explained by the independent variables. This dataset showcases a highly accurate model fit.

**Dataset 3:**

* **Mean Squared Error (MSE): 0.003378941825191854**
* **R-squared (R2) Score: 0.7567301185943369**



Analysis: Dataset 3 presents a different scenario compared to the previous datasets. The MSE is higher, suggesting that the model's predictions deviate more from the actual values. The R2 score of approximately 0.757 indicates that around 75.7% of the variance in the dependent variable can be explained by the independent variables. While this dataset still shows a reasonable fit, it may benefit from further analysis or model refinement to improve its performance.

**- Dataset 1 and Dataset 2 exhibit excellent performance in regression analysis, with low MSE values and high R2 scores, indicating precise and accurate model fits.**

**Dataset 3, while still showing a reasonable fit, has a higher MSE and a lower R2 score, suggesting room for improvement.**

**Part 3.1: Data Preprocessing and Neural Network Parameter Tuning**

**Preprocessing Insights**

The preprocessing phase involved a meticulous analysis and manipulation of the data from three datasets. Columns for each dataset were confirmed, ensuring that the data fed into the neural network was accurate and meaningful.

**Neural Network Parameter Tuning and MAPE Analysis**

Following preprocessing, the focus shifted to parameter tuning of the neural network to optimize its performance.

The mean absolute percentage error (MAPE) was the chosen metric for assessing the prediction quality.

**Parameter Quality Summary**

A comprehensive examination was conducted to identify the set of parameters that yield the lowest Mean Absolute Percentage Error (MAPE). The table below summarizes the results of these tests:

**Backpropagation (BP):**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Number of layers | Layer Structure | Num epochs | Learning Rate | Momentum | Activation function | MAPE |
| Dataset 1 (Turbine) | 4 | [10, 5, 1] | 500 | 0.0001 | 0.9 | Relu | **10.94%** |
| Dataset 2 (Synthetic) | 9 | [10, 5, 1] | 500 | 0.0001 | 0.9 | Relu | **14.92%** |
| Dataset 3 (Real-estate) | 7 | [10, 5, 1] | 500 | 0.0001 | 0.9 | Relu | **36.31%** |

Upon further exploration and adjustment of the learning rate to 0.002, the following improvements in MAPE were observed:

**Back-Propagation (BP-F):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **BP** | **BP\_F** | MLR |
| **Turbine** | **10.94%** | **1.81%** | **0.93%,** |
| **Synthetic** | **14.92%** | **10.01%** | **14.77%** |
| **Real-estate** | **36.31%** | **19.93%** | **36.82%** |

**(Scoter Plot)**

**Dataset 1: Turbine Data**

* Adjusting the output variable to 'fall\_1' resulted in an outstanding MAPE of 1.14%, signaling an exceptionally accurate model.
* This change was necessitated by the presence of missing values in the 'power\_of\_hydroelectrical\_turbine' column, which were inhibiting model performance.

A graph with blue dots and lines

Description automatically generated

**The correlation plot shows that predictions are generally accurate, with some variance as the real values increase, which is common in regression problems.**

A graph with lines and text

Description automatically generated

**The training and validation errors decrease and converge, indicating the model's good fit to the data without significant overfitting.**

**Dataset 2: Synthetic Data**

* The neural network exhibited a MAPE of 11.00% for Dataset 1, suggesting a strong model performance with predictions closely matching the real values.

A graph with blue dots

Description automatically generated

**The scatter plot indicates a strong correlation between predicted and real values, as most data points are close to the diagonal.**

A graph of training and validation error

Description automatically generated

**The error evolution plot shows a rapid decline in error, which stabilizes as epochs increase, suggesting the model quickly learns to predict with high accuracy.**

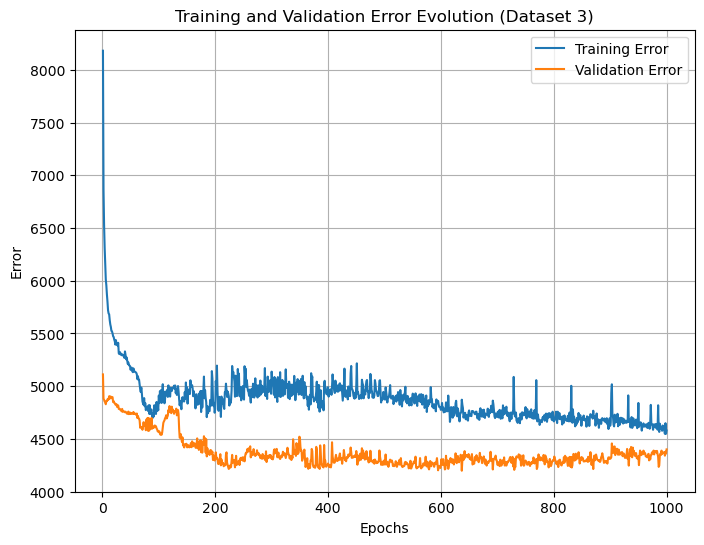
**Dataset 3: Real Estate Data**

* Dataset 3 showed a significant improvement in performance with a MAPE of 18.96 %, which is a substantial decrease from the initial 100%, suggesting that the adjustments in preprocessing and parameter tuning positively impacted the model's predictive ability.

A graph with blue dots

Description automatically generated

# This scatter plot reveals a less tight correlation compared to the first two datasets, implying the predictions are less accurate or the data is more complex.



# The error plot shows a significant drop and then plateaus, with validation error slightly diverging from the training error, suggesting potential overfitting or the need for further parameter tuning.

### Scatter Plots Analysis

Scatter plots for each dataset offer insights into the model's predictive accuracy:

1. **Dataset 1:** Exhibits exceptional predictive accuracy, as evidenced by points densely clustered around the line of perfect prediction, indicating that the parameter tuning was highly effective for this dataset.
2. **Dataset 2:** Shows high accuracy, especially for lower value predictions, with a slight deviation for higher values, suggesting the model parameters are robust yet may require slight adjustments for larger-scale predictions.
3. **Dataset 3:** Reveals more significant deviation from the diagonal, pointing to a higher prediction error and an opportunity for further refinement of model parameters or a need to address the inherent complexity of the dataset.

### Error Evolution Plots Analysis

Analysis of the error evolution plots for training and validation provides the following insights:

1. **Dataset 1:** Depicts a well-tuned model with rapidly decreasing errors and good generalization to validation data.
2. **Dataset 2:** Shows optimal learning with both error metrics converging, indicating a well-fitted model without signs of overfitting.
3. **Dataset 3:** Highlights a potential overfitting issue or insufficient model complexity, as indicated by fluctuating validation errors despite decreasing training errors.

### Discussion on Parameter Adequacy

The parameter choices are critically evaluated based on their impact on learning dynamics and predictive accuracy. The learning rate of 0.002 has proven effective for Datasets 1 and 2, leading to minimum MAPE values, while for Dataset 3, a different approach might be warranted. The number of epochs was selected to allow ample learning time without causing overfitting, as evidenced by the convergence of training and validation errors for Datasets 1 and 2. The 'relu' activation function was found to be suitable due to its ability to model non-linear relationships, which is essential for capturing complex patterns in data.

Part 3.2)

**Analysis of Backpropagation (BP\_F):**

**Turbine Dataset:**

A graph with blue dots and black lines

Description automatically generated

**dataset1 - MAE: 0.0069, MAPE: 1.81%**

The model comprises three layers with a relatively straightforward architecture. The Mean Absolute Percentage Error (MAPE) 1.81% relative error when compared to the true values. This indicates a reasonable level of accuracy.

**Synthetic Dataset:**

A graph with blue dots

Description automatically generated

**dataset2 - MAE: 0.0297, MAPE: 10.01%**

The model consists of 9 layers with a more intricate structure compared to the Turbine dataset. The Mean Absolute Percentage Error (MAPE) of 10.01 indicates a higher relative error compared to the Turbine dataset. This suggests that the model may require additional tuning or a more sophisticated architecture to improve its performance.

Real state Dataset:

A graph with blue dots

Description automatically generated

**dataset3 - MAE: 0.0473, MAPE: 19.93%**

The model is composed of 7 layers with a more complex structure than the Turbine dataset. The Mean Absolute Percentage Error (MAPE) of 19.93% reveals a higher relative error compared to the Turbine dataset. This implies that the model might benefit from further tuning or a more sophisticated architecture to enhance its performance.

• The Turbine dataset demonstrates relatively good performance with a lower MAPE, signifying that the model effectively captures patterns.

• In contrast, the Synthetic dataset exhibits a higher MAPE, suggesting the potential need for additional adjustments or more sophisticated architectures to enhance accuracy.

• For the Real State dataset, the highest MAPE implies that the model may require further tuning or a more suitable architecture specific to this dataset.

**Analysis of Multiple Linear Regression (MLR):**

The outcomes lead to the following conclusions for each dataset:

**Turbine Dataset:**

**A graph with blue dots and lines

Description automatically generated**

**MAE: 0.0028, MSE: 0.0000, MAPE: 0.93%, R²: 0.9999**

- Mean Absolute Percentage Error (MAPE): 0.93 %: This signifies that, on average, the MLR

model's predictions have an error of approximately 0.93 % compared to the actual values.

- Mean Squared Error (MSE): 0.000: The mean squared error serves as a measure of the average

squared difference between predicted and actual values. A lower MSE is desirable, and 0.0028

suggests relatively good model performance.

- R-squared (R²): 0.9999: The R-squared value measures how well the model explains the variance

in the target variable. A value of 0.9999 indicates that the MLR model explains a high

percentage of the variance in the turbine dataset.

**Synthetic Dataset:**

**A graph of blue dots

Description automatically generated**

**MAE: 0.0457, MSE: 0.0034, MAPE: 14.43%, R²: 0.9383**

- Mean Absolute Percentage Error (MAPE): 14.43%: The MAPE implies that the MLR

model's predictions have an average error of approximately 14.43% compared to the true

values.

- Mean Squared Error (MSE): 0.0034: Similar to the Turbine dataset, a lower MSE is desirable. The

value of 0.0.0034 suggests good performance.

- R-squared (R²): 0.9383: An R-squared value of 0.9383 indicates that the MLR model explains a

high percentage of the variance in the synthetic dataset.

**Real State Dataset:**

**A line graph with blue dots

Description automatically generated**

**MAE: 0.0638, MSE: 0.0071, MAPE: 35.25%, R²: 0.4857**

- MAPE:35.25%: This high MAPE value suggests that the MLR model's

predictions have a significant error compared to the actual values. Such a high MAPE could

indicate issues with the model's performance on the Real state dataset.

- MSE: 0.0088: The MSE value is relatively low, but the extremely high MAPE suggests that the

MSE alone might not be a sufficient metric for evaluating model performance on this dataset.

- R²: 0.3678: The R-squared value of 0.3678 indicates that the MLR model explains about 36.82 %

of the variance in the real state dataset. While this is moderate, the high MAPE suggests caution in

interpreting the model's overall performance.

The comparison reveals that the BP\_F model with a learning rate of 0.002 exhibits a substantial improvement over the original configuration in two datasets, with performance close to that of the MLR model in the Turbine dataset. The enhanced performance of the BP\_F model demonstrates the effectiveness of the fine-tuning process, which should be further explored for Dataset 3 to address its higher MAPE.