Neural Network Correlation of Boiling Point with Chemical Properties

1. Introduction

This report presents an analysis of the correlation between the normal boiling point of chemical compounds and their molecular properties, including molecular weight and acentric factor. The study involves:

- Linear regression modeling
- Implementation of a neural network
- Evaluation of training data effects on model accuracy

2. Data Preprocessing

The dataset contains the following columns:

- Common Name
- Molecular Weight (MW)
- Critical Temperature (Tc)
- Acentric Factor (w)
- Normal Boiling Point (Tb)

Steps followed:

1. Data Extraction: The dataset was loaded into a matrix AllData.
2. Feature Selection: MW and w were extracted as independent variables (X), while the reduced boiling point (Tb/Tc) was the dependent variable (y).
3. Random Sampling: 100 random compounds were selected for training.

3. Linear Regression Analysis

A multiple linear regression model was built using:

Theta = (X^T X)^-1 X^T y

Regression Results:

- Equation of Fit: y = 0.5958 + 0.0002 * MW + 0.1546 * w
- Coefficient Values:
  - Theta0 (Bias): 0.5958
  - Theta1 (MW): 0.0002

  - Theta2 (w): 0.1546
- R^2 Value: 0.7861

The moderate R^2 value suggests potential nonlinear relationships.

4. Neural Network Model

To improve prediction accuracy, a neural network was implemented with:
- Input layer: 2 neurons (MW and w)
- Hidden layers: 2 layers with ReLU activation
- Output layer: 1 neuron for prediction

Training Details:
- Data Split:
  - 10% training (~600 samples)
  - 90% validation & testing
- Loss Function: Mean Squared Error (MSE)
- Optimization Algorithm: Adam

Training Results:
- Epoch 1 Loss: 0.4349 (Training), 0.2716 (Validation)
- Mean Absolute Error (MAE): 0.6397

5. Effect of Changing Training Data Proportion

Different training proportions were tested:
- With 10% training data, the model showed moderate generalization.
- Increasing training data improved accuracy but led to overfitting.
- Best performance was observed with 20%-30% training data.

6. Conclusion

- Linear Regression provided a reasonable correlation with an R^2 of 0.7861.
- Neural Network Model showed potential improvements but required careful tuning.
- Future Work:
  - Try additional features such as critical pressure.
  - Experiment with deeper networks and different activation functions.