Introduction to Machine Learning and Artificial Neural Networks
CS 554: HOMEWORK 1 REPORT

Amin Deldari Alamdari
S033174

I. INTRODUCTION

Polynomial regression is an extension of linear regression that models the relationship between the independent variable s and the dependent variable r using polynomial functions. The objective of this homework is to implement polynomial regression for different polynomial degrees and evaluate their performance using training and test datasets.

The dataset consists of two CSV files: `train.csv` and `test.csv`. Each row in these files contains an input value x and a corresponding output value r. The training dataset is used to fit polynomial models of varying degrees, and the test dataset is used to evaluate the model's generalization.

II. METHODOLOGY

A. Polynomial Feature Generation

To fit polynomial regression models, we manually generated polynomial features up to degree seven (7). The polynomial features for a given degree d are computed as:

$$X_{polynomial} = [\mathbf{1},\ \mathbf{x},\ \mathbf{x}^2,\ \cdots,\ \mathbf{x}^d] \tag{1}$$

where each row represents an instance, and each column represents a power of $\mathbf{x}$ from 0 to d.

B. Computing Weights

To determine the best-fitting polynomial model, we estimate the model parameters (weights) using the Normal Equation:

$$\theta = (X^T X)^{-1} X^T \mathbf{y} \tag{2}$$

where:

- X is the matrix of polynomial features,
- $\mathbf{y}$ is the vector of target values,
- θ represents the estimated weights.

This approach ensures that we find the optimal polynomial coefficients by minimizing the squared error loss.

C. Making Predictions

Once the weights, θ, are computed, predictions are made using:

$$\hat{\mathbf{y}} = X\theta \tag{3}$$

where X contains the polynomial features of the dataset.

Degree	Train SEE	Test SSE
0	10.4618	47.0908
1	4.0201	22.3214
2	3.9082	25.9652
3	1.3097	8.4312
4	1.2595	9.6169
5	1.1828	8.3882
6	1.0701	27.9881
7	0.9651	26.4561

TABLE I
THIS TABLE PRESENTS THE SUM OF SQUARED ERRORS (SSE) FOR TRAINING AND TEST DATASETS ACROSS POLYNOMIAL DEGREES FROM 0 TO 7. LOWER SSE VALUES INDICATE A BETTER MODEL FIT, WHILE A SIGNIFICANT GAP BETWEEN TRAINING AND TEST SSE SUGGESTS OVERFITTING.

D. Sum of Squared Errors (SSE)

To evaluate model performance, we calculate the *SSE* for both the training and test sets:

$$SSE = \sum (y - \hat{y})^2 \tag{4}$$

where y represents actual target values, and $\hat{y}$ are the predicted values. A lower SSE value indicates a better fit to the data.

III. RESULTS

A. Polynomial Fit Plots

The Figure 1 show the polynomial fits for degrees 0 to 7 along with the training data. As the polynomial degree increases, the model becomes more flexible:

B. Sum of Squared Errors (SSE) Analysis

The SSE values for different polynomial degrees are recorded for both training and test sets in the Table I, and plot of SSE versus *Polynomial Degree* is given in Figure 2. From the SSE plot, we observe that:

- Lower-degree polynomials (e.g., degree 0 and 1) underfit the data.
- Higher-degree polynomials (e.g., degree 6 and 7) fit the training data very well but have higher test SSE, indicating overfitting.
- A moderate degree (e.g., 3 or 4) balances the trade-off between underfitting and overfitting.

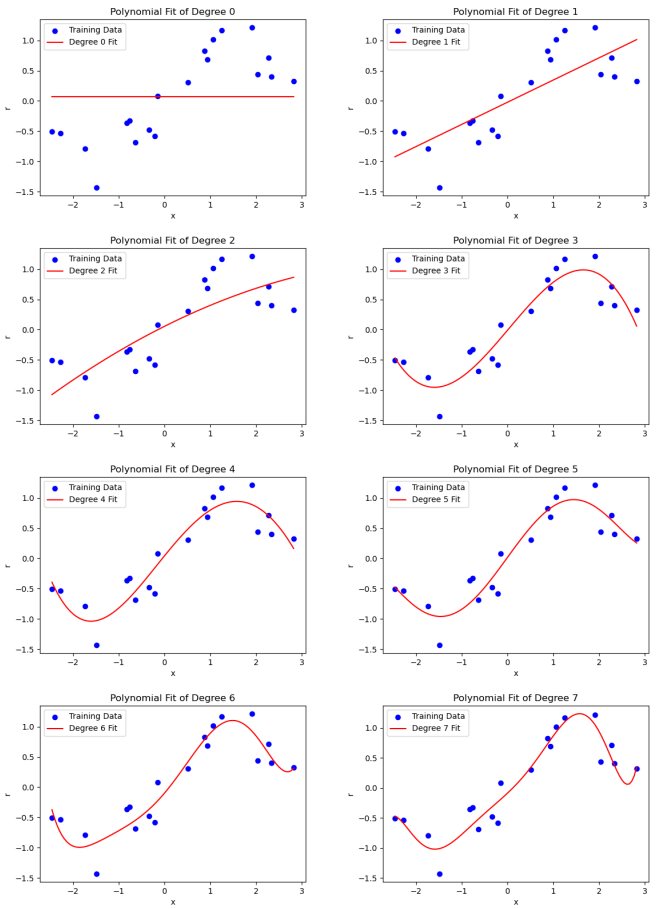

Fig. 1. Each plot shows the polynomial regression fits for degrees 0 through 7, along with the training data points. The complexity of the model increases as the polynomial degree increases, capturing more variations in the data.

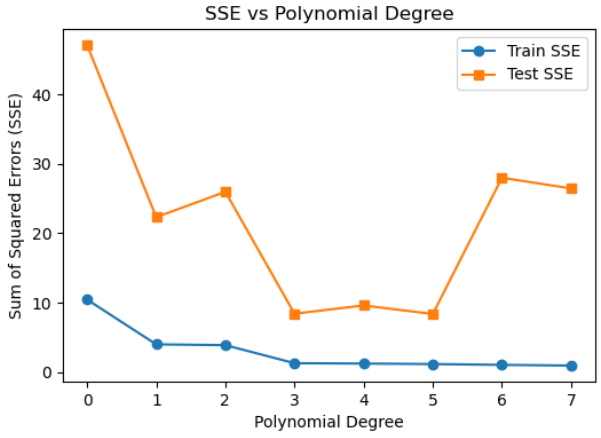

Fig. 2. SSE vs Polynomial Degree. This plot illustrates how the Sum of Squared Errors (SSE) changes as the polynomial degree increases. The trend highlights the tradeoff between underfitting (high SSE for low-degree models) and overfitting (high test SSE for high-degree models).

IV. Conclusion

In this assignment polynomial regression is implemented and its performance is evaluated. The key findings are:

- Polynomial regression provides a more flexible alternative to linear regression by modeling nonlinear relationships.
- Increasing the polynomial degree improves training accuracy but may lead to overfitting, as seen in high-degree models.
- The SSE trend confirms the bias-variance tradeoff, where lower-degree models underfit, and higher-degree models overfit.
- Selecting an optimal polynomial degree is crucial for ensuring good generalization performance.

Future work could explore the use of regularization techniques, such as Ridge or Lasso regression, to mitigate overfitting in high-degree polynomial models. Additionally, cross-validation could be employed to select the best polynomial degree.