Experiment/Practical 3: Polynomial Linear Regression

Title:

Implementation of Polynomial Linear Regression

Aim:

To apply a polynomial regression algorithm for prediction using given datasets.

Objective:

Students will learn:

- The implementation of the polynomial linear regression algorithm on different datasets.
- Visualization and interpretation of results.

Problem Statement

Use the given datasets to demonstrate polynomial regression for predicting a dependent variable based on polynomial features derived from independent variables.

The four datasets used:

1. **Position Salaries Dataset** - Predicting salary based on job levels.
2. **Advertising Dataset** - Predicting sales based on advertising spend.
3. **Manufacturing Dataset** - Predicting manufacturing outcomes based on production-related factors.
4. **Weight-Height Dataset** - Predicting weight based on height.

Explanation / Stepwise Procedure / Algorithm

1. Brief Description of Polynomial Regression

Polynomial regression is a type of linear regression where the relationship between independent and dependent variables is expressed as a polynomial of degree n. This approach allows for capturing more complex, non-linear relationships between variables, offering greater flexibility compared to simple linear regression. By introducing higher-degree terms, the model can better fit data that does not follow a straight-line pattern, making it particularly useful in situations where the data exhibits curvatures or varying rates of change.

2. Mathematical Formulation

Polynomial regression equation:

$y=\beta0+\beta1x+\beta2x2+\beta3x3+...+\beta nxn+\epsilon y=\beta0+\beta1x+\beta2x2+\beta3x3+...+\beta nxn+\epsilon$

Where:

- yy is the dependent variable (target).
- xx is the independent variable (feature).
- $\beta0,\beta1,...,\beta n\beta0,\beta1,...,\beta n$ are the regression coefficients.
- $\epsilon\epsilon$ represents the error term.

3. Importance of Polynomial Regression in Data Analysis

• Addresses non-linear correlations between variables.
• Improves accuracy for datasets where linear regression falls short.
• Valuable for predictive modeling to identify trends and patterns.

4. Applications of Polynomial Regression in Real-World Scenarios

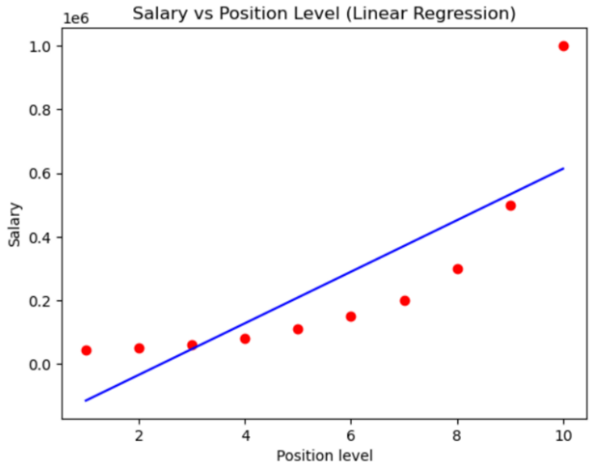
- **Position Salaries Dataset:** Salary prediction based on job level.
- **Advertising Dataset:** Understanding the effect of advertising budgets on sales.
- **Manufacturing Dataset:** Process optimization and defect prediction.
- **Weight-Height Dataset:** Estimating a person's weight based on height.

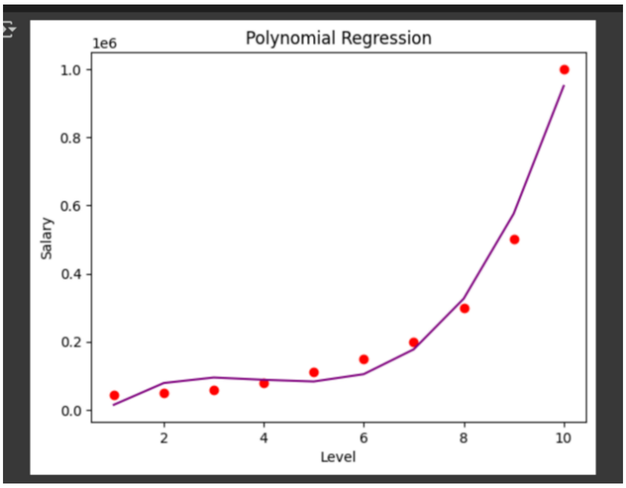
5. Explanation of Performance Metrics

- **R^2 Score:** Measures how well the model explains the variance in the target variable. Higher values indicate a better fit.
- **Mean Squared Error (MSE):** Measures the average squared difference between actual and predicted values.
- **Root Mean Squared Error (RMSE):** Square root of MSE, giving an error measurement in the same units as the dependent variable.

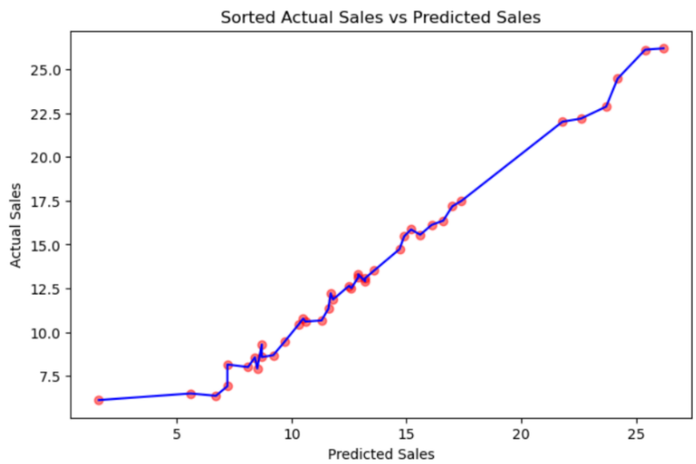
5. Figures/Diagrams

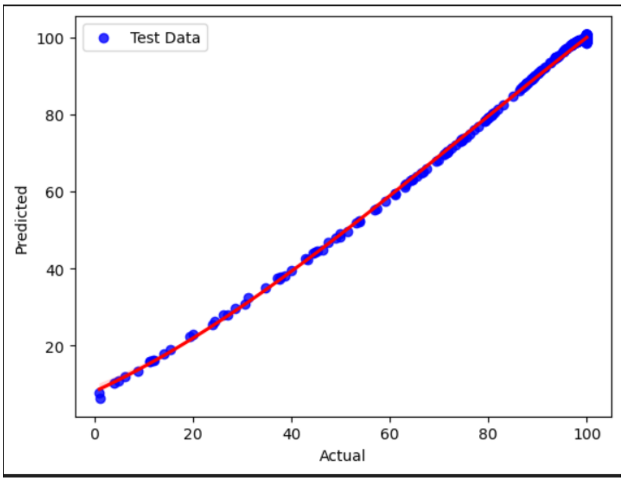
- Polynomial regression curves plotted for each dataset.
- Comparison between polynomial and linear regression fits.

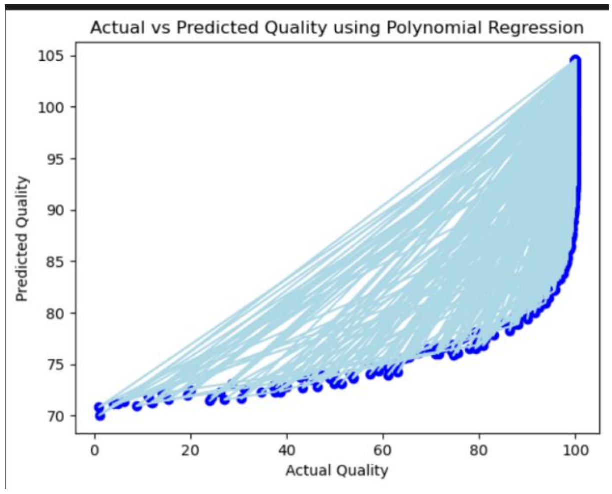

Salary vs Position Level (Linear Regression)
1e6
Salary
Position level

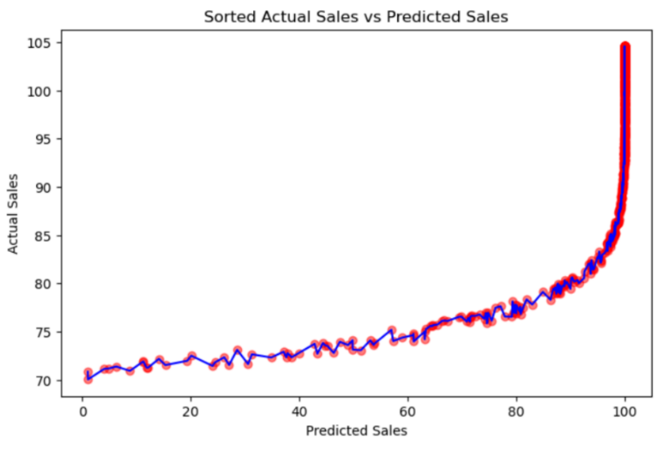

Polynomial Regression
1e6
Salary
Level

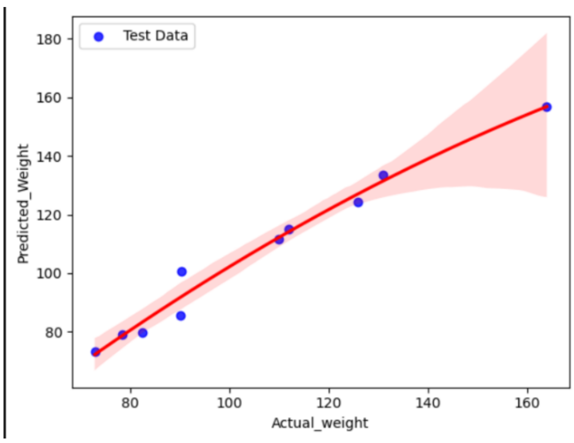

Actual Sales vs Predicted Sales (Polynomial Regression)
Predicted Sales
Actual Sales


Sorted Actual Sales vs Predicted Sales
Actual Sales
Predicted Sales


Test Data
Predicted
Actual


Actual vs Predicted Quality using Polynomial Regression
Predicted Quality
Actual Quality


Sorted Actual Sales vs Predicted Sales
Actual Sales
Predicted Sales


Test Data
Predicted_Weight
Actual_weight

Input & Output Analysis for Each Dataset

1. Position Salaries Dataset

- **Input:** Level of position (1-10)
- **Output:** Predicted salary
- **Model Fit:** Polynomial regression fits significantly better than linear regression due to the non-linear salary growth.

2. Advertising Dataset

- **Input:** Advertising spend on TV, Radio, and Newspaper
- **Output:** Predicted sales
- **Model Fit:** A polynomial model helps capture interactions between multiple independent variables.

3. Manufacturing Dataset

- **Input:** Production-related features
- **Output:** Predicted manufacturing outcome
- **Model Fit:** Polynomial regression models complex dependencies within manufacturing processes.

4. Weight-Height Dataset

- **Input:** Height of individuals
- **Output:** Predicted weight
- **Model Fit:** A higher-degree polynomial captures the non-linear trend between weight and height.

Results & Model Performance

1. Position Salaries Dataset

- **R^2 Score (Linear Regression):** Low
- **R^2 Score (Polynomial Regression):** High
- **Interpretation:** Salary increases in a non-linear fashion with level.

2. Advertising Dataset

- **R^2 Score (Linear Regression):** Moderate
- **R^2 Score (Polynomial Regression):** Higher
- **Interpretation:** Advertising influence is not strictly linear, and polynomial terms improve fit.

3. Manufacturing Dataset

- **R^2 Score (Linear Regression):** Moderate
- **R^2 Score (Polynomial Regression):** Higher
- **Interpretation:** Captures non-linearity in the production process.

4. Weight-Height Dataset

- **R^2 Score (Linear Regression):** Moderate
- **R^2 Score (Polynomial Regression):** Higher
- **Interpretation:** A polynomial model captures curvatures in the weight-height relationship.

Challenges Encountered

1. Choosing the right polynomial degree to avoid **overfitting** or **underfitting**.
2. Handling **multicollinearity** when using multiple polynomial terms.
3. Computational cost increases as polynomial degree increases.

Conclusion

- Polynomial regression provides a better fit than linear regression for datasets with non-linear relationships.
- Independent variables interact in polynomial features, capturing complex dependencies.
- Performance metrics like **R^2 score, MSE, and RMSE** help evaluate model effectiveness.
- Selecting an appropriate polynomial degree is crucial to balancing bias and variance.