ELEC-8900-57 Special Topics – Machine Learning

Polynomial Regression Algorithm

Presented by:

Kush Dhuvad

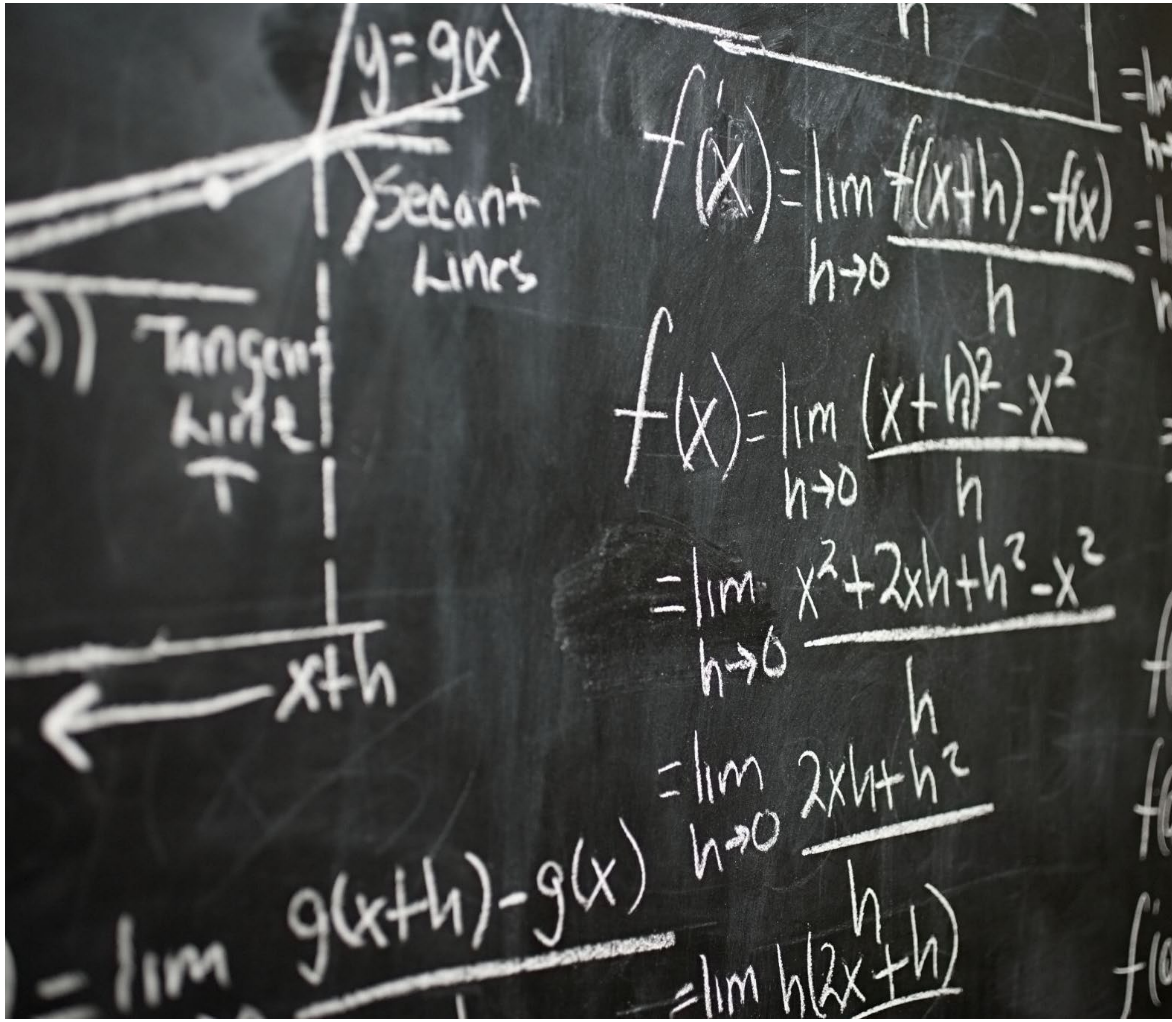
Taksh Shah

Dheeraj Varanasi

Instructor: Dr. Yasser Alginahi

Date: 29 September 2023

Table of contents:



Linear Regression

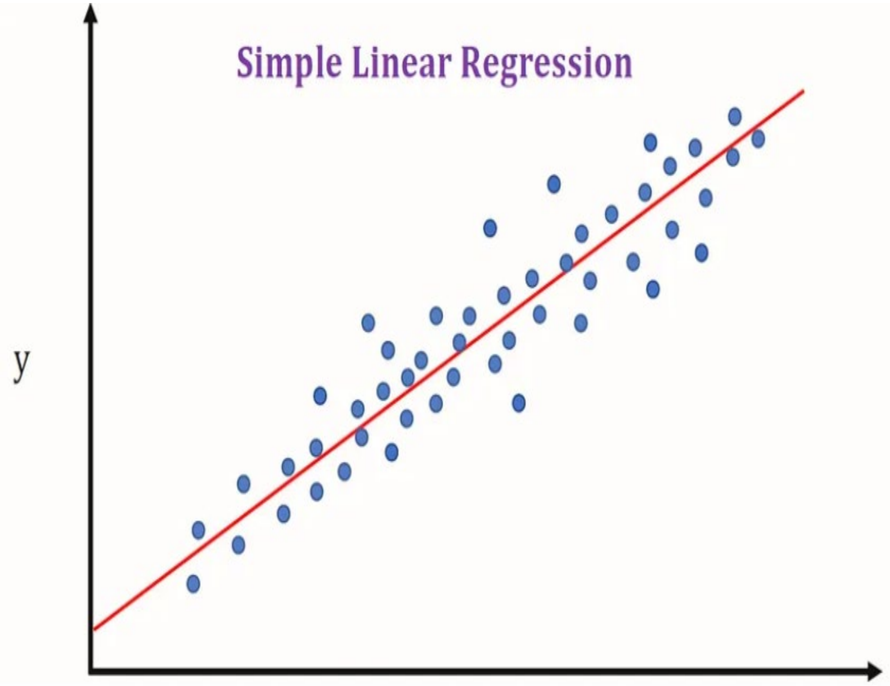
A statistical tool to model the relationship between a single dependent variable and multiple independent variables by fitting a linear equation to the observed data.

This equation then helps to further predict and understand the relationship between variables.

Linear Regression Equation

- To calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

$$Y = b_0 + b_1 X$$



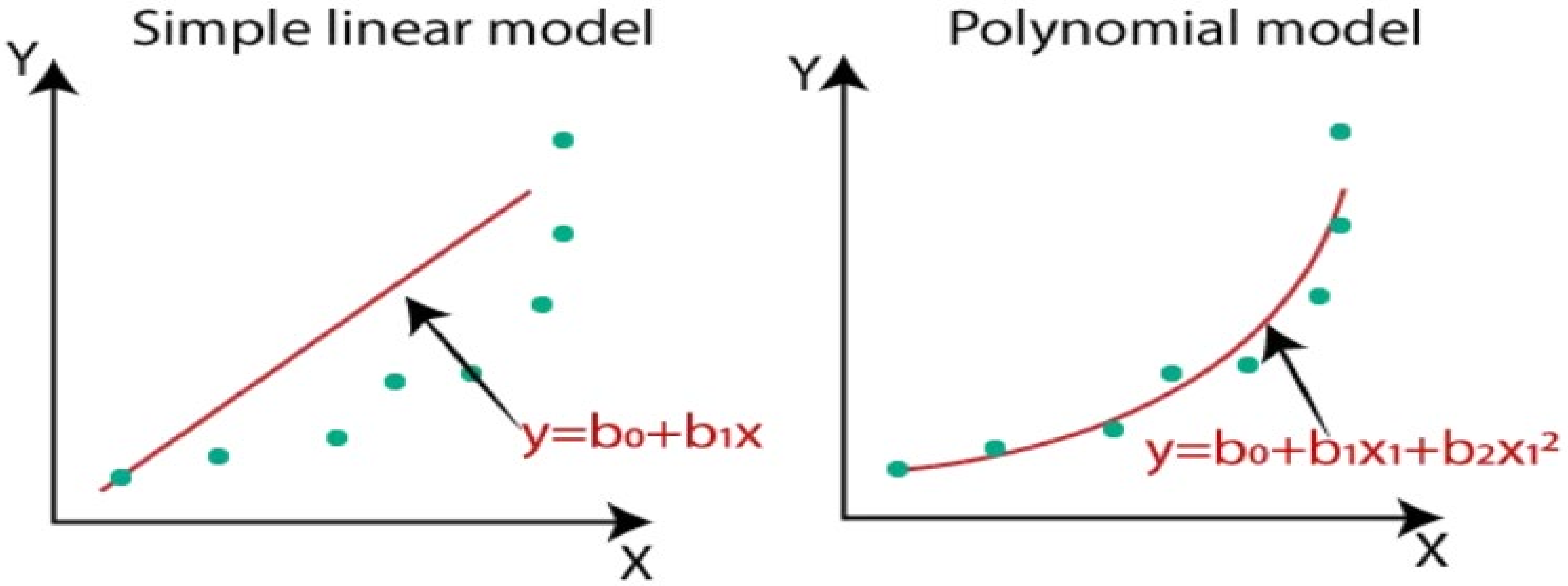
[Source: https://tinyurl.com/3k4kaxap]

Why Polynomial Regression?

Complex Data Patterns: Polynomial regression is essential when data exhibits nonlinear relationships and intricate patterns that simple linear models cannot capture.

Improved Predictive Accuracy: It enhances predictive accuracy by allowing the model to fit real-world data more closely, reducing bias in predictions.

Linear Regression vs Polynomial Regression



[Source: https://tinyurl.com/52t36unu]

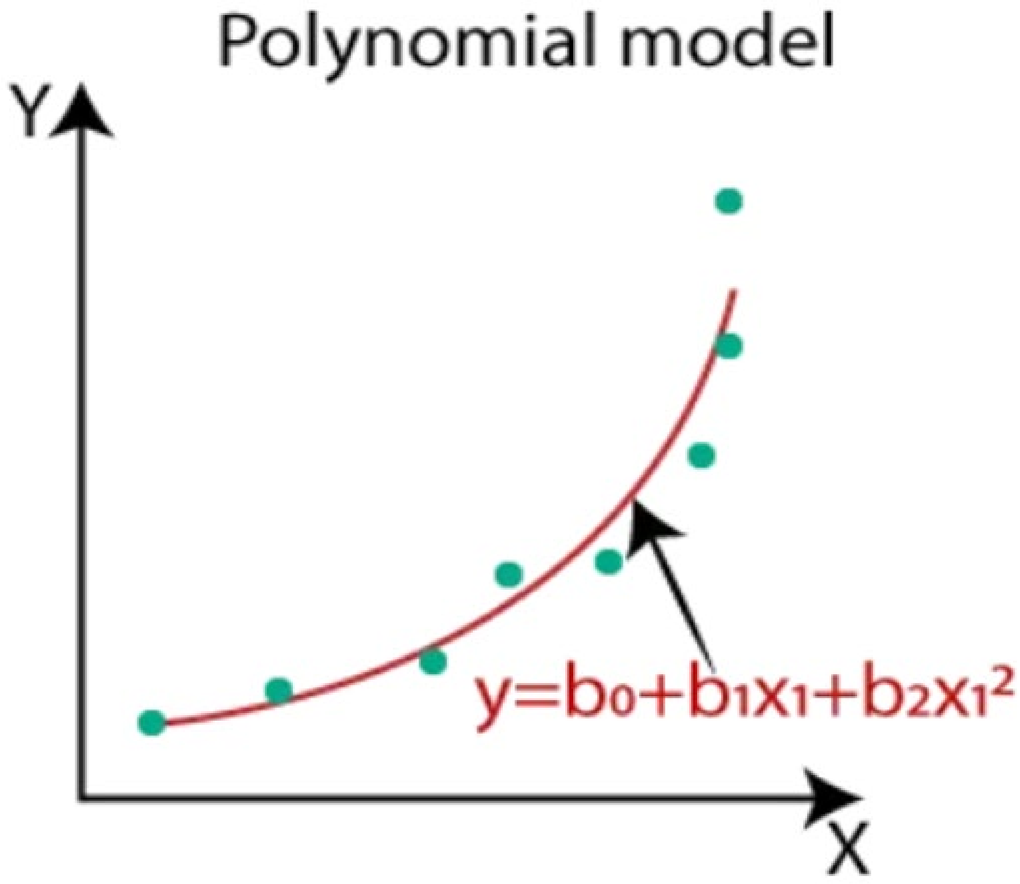
Polynomial Regression

It helps to analyze the relationship between the independent variable x and the dependent variable y, modelled as an nth degree polynomial in x.

It is a form of regression model, used especially when the relationship between variables is non-linear and the straight line obtained by linear regression is not adequate to accurately represent the data.

Polynomial Regression Equation

$$y = b_0 + b_1x + b_2x^2 + \dots + b_nx^n$$



[Source: https://tinyurl.com/52t36unu]

Degrees of Polynomial

The highest power raised to x in the polynomial equation is determined as the degree of polynomial.

Higher degrees result in more complex models. Each additional degree adds a new term to the equation, making the model more intricate.

A higher degree allows the model to fit the training data more closely.

Polynomial features

Features are the input variables that are used to predict the dependent variable y using a polynomial equation.

These polynomial features are generated by raising the original features to various powers, allowing the model to capture nonlinear relationships.

Over Fitting

Overfitting is a common issue in polynomial regression model that occurs when the degree of the equation is excessively high.

An overfit model works exceptionally well on training data but while on new data, it leads to poor predictive performance since it fits the training data too closely including the noise.

Benefits

Flexibility: It can capture non-linear patterns and curves in the data, making it suitable for a wider range of data distributions.

Higher Accuracy: Polynomial Regression can offer higher accuracy especially in case of problems with curved or oscillating patterns.

Visual Representation: Representation with polynomial curves makes data visualization and communication of results more efficient.

Limitations

Overfitting: High degree polynomials can lead to overfitting, where the model fits well to the training data, but fails to fit well to unseen data.

Complexity: The complexity of the model increases with the increasing degree of the polynomial.

Data availability: Polynomial Regression requires a sufficient amount of data, especially for higher-degree polynomials.

Bias – variance tradeoff

Mean Squared Error (MSE) - a commonly used performance metric for regression models.

Test MSE - the MSE when a model is applied to unknown data, which is what we are interested in.

Test MSE is decomposed into variance and bias.

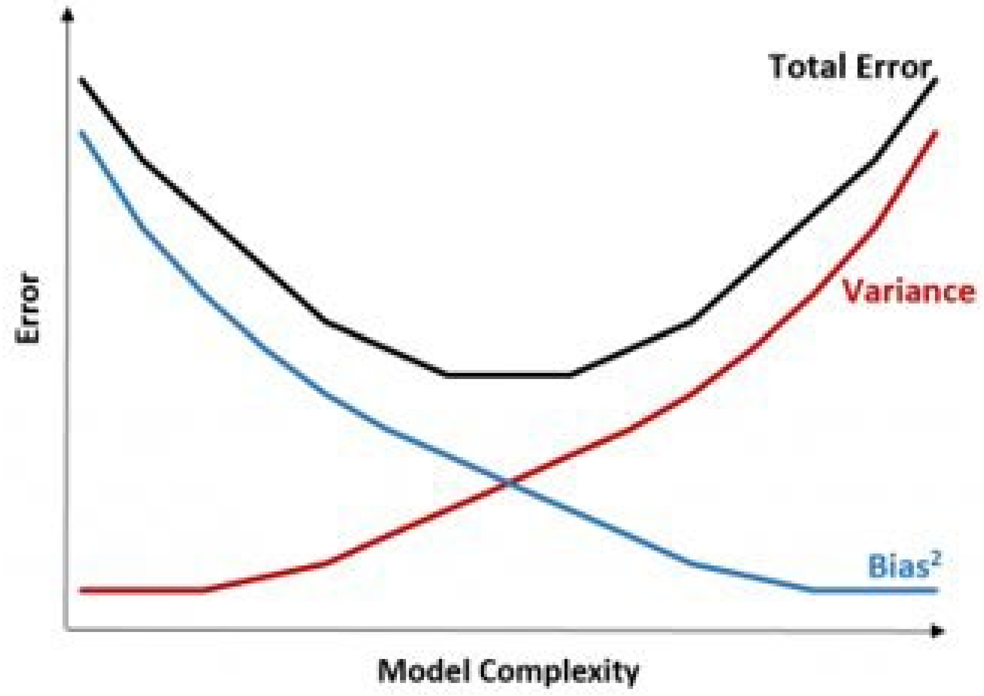
Variance - refers to the amount by which our function would change if we trained our model using a different training set.

Bias - corresponds to the difference between the expected and the true value for a particular prediction.

Typically, models with low variance have high bias and vice versa.

Aim - minimize the total error as much as possible.

How - find the right balance between variance and bias for a model.



[Source: https://www.statology.org/bias-variance-tradeoff/]

Understanding Polynomial Regression using Mathematics

- Consider the polynomial quadratic equation

$$y=a_0+a_1*x+a_2*x^2+\dots\dots\dots+ +a_n*x^n$$

- We will understand the polynomial regression using below dataset[2]

X	Y
3	2.5
4	3.2
5	3.8
6	6.5
7	11.5

- We will calculate the values of a_0, a_1, and a_2 can be found using below formulae[2]:

$$\sum y_i = na_0 + a_1\left(\sum x_i\right) + a_2\left(\sum x_i^2\right)$$
$$\sum y_i x_i = a_0\left(\sum x_i\right) + a_1\left(\sum x_i^2\right) + a_2\left(\sum x_i^3\right)$$
$$\sum y_i x_i^2 = a_0\left(\sum x_i^2\right) + a_1\left(\sum x_i^3\right) + a_2\left(\sum x_i^4\right)$$

- Calculating the required values we obtain

	x	y	x^2	x^3	x^4	y*x	y*x^2
	3	2.5	9	27	81	7.5	22.5
	4	3.2	16	64	256	12.8	51.2
	5	3.8	25	125	625	19	95
	6	6.5	36	216	1296	39	234
	7	12	49	343	2401	80.5	563.5
Σ	**25**	**27.5**	**135**	**775**	**4659**	**158.8**	**966.2**

- Using the above data in the given formulae:[2]

$$27.5 = 5a_0 + 25a_1 + 135a_2$$

$$158.8 = 25a_0 + 135a_1 + 775a_2$$

$$966.2 = 135a_0 + 775a_1 + 4659a_2$$

- Solving this system of equations we get

$$a_0=12.4285714$$

$$a_1=-5.5128571$$

$$a_2=0.7642857$$

- The required quadratic polynomial model is

$$\mathbf{y=12.4285714 - 5.5128571*x + 0.7642857*x^2}$$

- Using values of x in above equation, we will obtain the values of y for the polynomial regression model.[2]
- Refer the given YouTube video to understand the concept better:

https://www.youtube.com/watch?v=xLZs9xwdHqY

Crunching numbers - Polynomial Regression

Position	Level	Salary
Business Analyst	1	45000
Junior Consultant	2	50000
Senior Consultant	3	60000
Manager	4	80000
Country Manager	5	110000
Region Manager	6	150000
Partner	7	200000
Senior Partner	8	300000
C-level	9	500000
CEO	10	1000000

Dataset: Polynomial Regression Dataset [1]

Position	Level	Salary
Business Analyst	1	45000
Junior Consultant	2	50000
Senior Consultant	3	60000
Manager	4	80000
Country Manager	5	110000
Region Manager	6	150000
Partner	7	200000
Senior Partner	8	300000
C-level	9	500000
CEO	10	1000000

Data Visualization

- **Select the Independent and Dependent variables**
  - X (Independent variable) = Level of the position
  - Y (Dependent variable) = Salary
- **Choosing the polynomial degree**
  - The degree represents the highest power to which you will raise the features.
- **Generate polynomial features**
  - Based on chosen degree you can generate the polynomial features,
  - poly = PolynomialFeatures(degree = 4)

Polynomial feature generation

University of Windsor

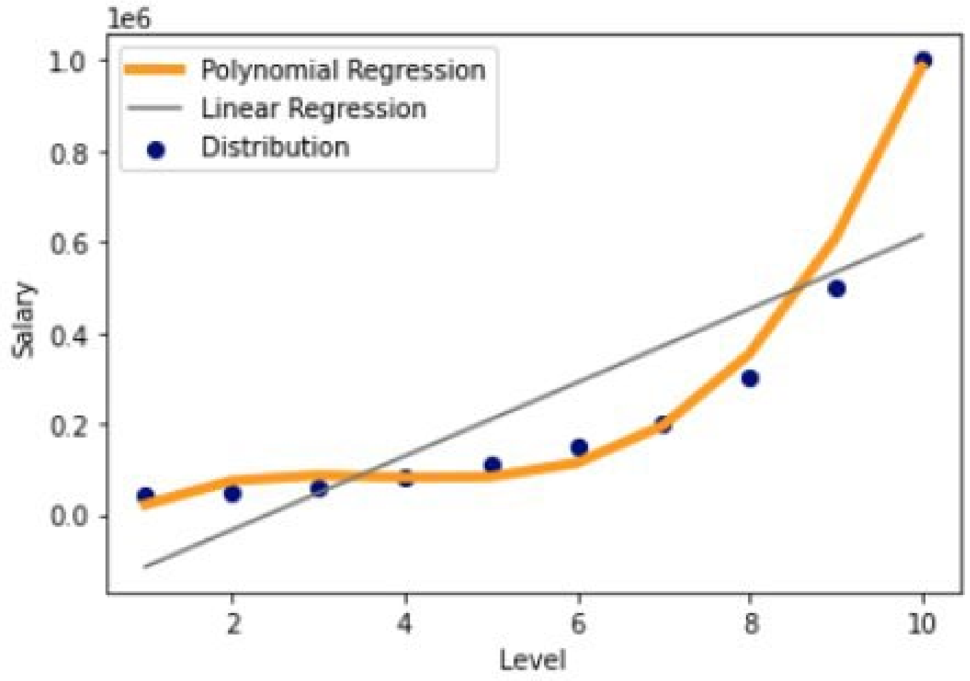
- **Fit the Polynomial Regression Model**
  - Once you've generated the polynomial features, you can use them to fit a polynomial regression model.
  - x_poly = poly.fit_transform(X)
- **Evaluate the Model**
  - After fitting the model, you should evaluate its performance using appropriate evaluation metrics like Mean Squared Error (MSE), R-Squared (R^2) or others depending on the problem.

Model Training and Evaluation

Model Evaluation - Mean Squared Error (MSE)

- **Data split**
  - x_train, x_test, y_train, y_test = train_test_split(x_poly, y, test_size = 0.10, random_state = 42)
- **Prediction**
  - model = LinearRegression()
  - model = model.fit(x_train, y_train)
  - y_pred = model.predict(x_test)
- **Mean Squared Error (MSE)**
  - MSE = mean_squared_error(y_test, y_pred)

Model Visualization



[Source: https://tinyurl.com/59nkn93c]

Real World Applications

Stock Market Analysis: Polynomial regression can be used to model stock price movements and predict trends by capturing the nonlinear patterns that often occur in financial markets. [3]

Climate Modeling: Polynomial regression can be applied to analyze climate data and model complex relationships between factors like temperature, precipitation, and greenhouse gas levels to make predictions about climate change. [4]

Sales Forecasting: Polynomial regression can be employed to forecast sales based on historical data and various marketing strategies, considering nonlinear influences on consumer behavior. [3]

Energy Consumption Forecasting: In energy management, polynomial regression can be used to forecast energy demand patterns, considering various factors like weather conditions and economic indicators. [4]

Conclusion

In conclusion, polynomial regression is a powerful tool for modeling complex relationships between variables in real-world applications.

By using higher-degree polynomial functions, we can capture nonlinear patterns that would be missed by simpler linear models.

However, accurate model evaluation is crucial to ensure that our models are truly predictive and not simply overfitting to noise in the data.

Conclusion

To this end, it is important to use appropriate metrics such as R-squared, adjusted R-squared, and root mean squared error, and to interpret them correctly.

Only by rigorously testing our models on independent data sets can we be confident in their ability to make accurate predictions.

I encourage everyone to consider the power of polynomial regression and to use it wisely and responsibly. By doing so, we can unlock new insights and drive innovation in fields ranging from finance to engineering to medicine.

References

1. Polynomial Regression," *kaggle.com*. https://www.kaggle.com/code/fehimenuruysal/polynomial-regression/input (accessed Sep. 29, 2023).

2. V. M. H, "Quadratic polynomial regression model solved example," VTUPulse, https://www.vtupulse.com/machine-learning/quadratic-polynomial-regression-model-solved-example/ (accessed Oct. 5, 2023).

3. Applications of Polynomial Regression, " *sparkbyexamples.com. https://sparkbyexamples.com/machine-learning/polynomial-regression-with-examples (accessed Sep. 29, 2023)*

4. *Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.*

University of Windsor