INTRO TO REG-2

By Rishabh Dubey

AGENDA

- ✅overfitting and underfitting (Bias-Variance tradeoff)
- ✅ Polynomial Regression
- ✅Lasso & Ridge Regression
- ✅ Classification vs Regression
- ✅ Implement Logistic Regression to classify binary outcomes

Quiz Time

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

Question 1: What does Linear Regression primarily do?

- A) Classifies data into categories
- B) Finds relationships between variables
- C) Clusters similar data points
- D) Encrypts data

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

Question 2: What is the equation for simple linear regression?

- A) $y = mx + b$
- B) $y = ax^2 + bx + c$
- C) $y = m_1x_1 + m_2x_2 + c$
- D) $y = \frac{x}{m} + b$

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

Question 3: How does Multiple Linear Regression differ from Simple Linear Regression?

- A) Uses more than one independent variable
- B) Uses non-linear relationships
- C) Only works for categorical data
- D) Requires deep learning

QUICK REVISION QUIZ (MCQS & TRUE/FALSE)

Question 4: Which metric is commonly used to evaluate a linear regression model?

- A) Accuracy
- B) Mean Squared Error (MSE)
- C) Recall
- D) Entropy

Advanced Regression Techniques

RECAP OF REGRESSION

Linear Regression predicts continuous values

Equation of Linear Regression:

$$y=b0+b1x+b2x2+...+bnxn$$

Evaluating a regression model:

- R-squared (R^2) Score: Measures how well the model explains the variance in data
- Mean Squared Error (MSE), Root Mean Squared Error (RMSE)

Why do we need Advanced Regression?

- Sometimes, relationships between variables are non-linear, and Linear Regression doesn't work well.

POLYNOMIAL REGRESSION

📌Problem with Linear Regression::

- If data has curves, a straight line cannot fit well.
- Example: Predicting temperature changes over time (hourly temperature fluctuations).

📌Solution:

- Polynomial Regression: Introduce higher-degree terms to capture curves.
- Equation **y=b0+b1x+b2x2+b3x3+...+bnxn**

📌How it works?

- Transform the original X values into polynomial features
- Fit a Linear Regression model on transformed features

(BIAS-VARIANCE TRADEOFF)

📌**Overfitting (High Variance):**

- Model captures noise, performs well on training data but poorly on test data
- Happens when model is too complex

📌**Underfitting (High Bias)**

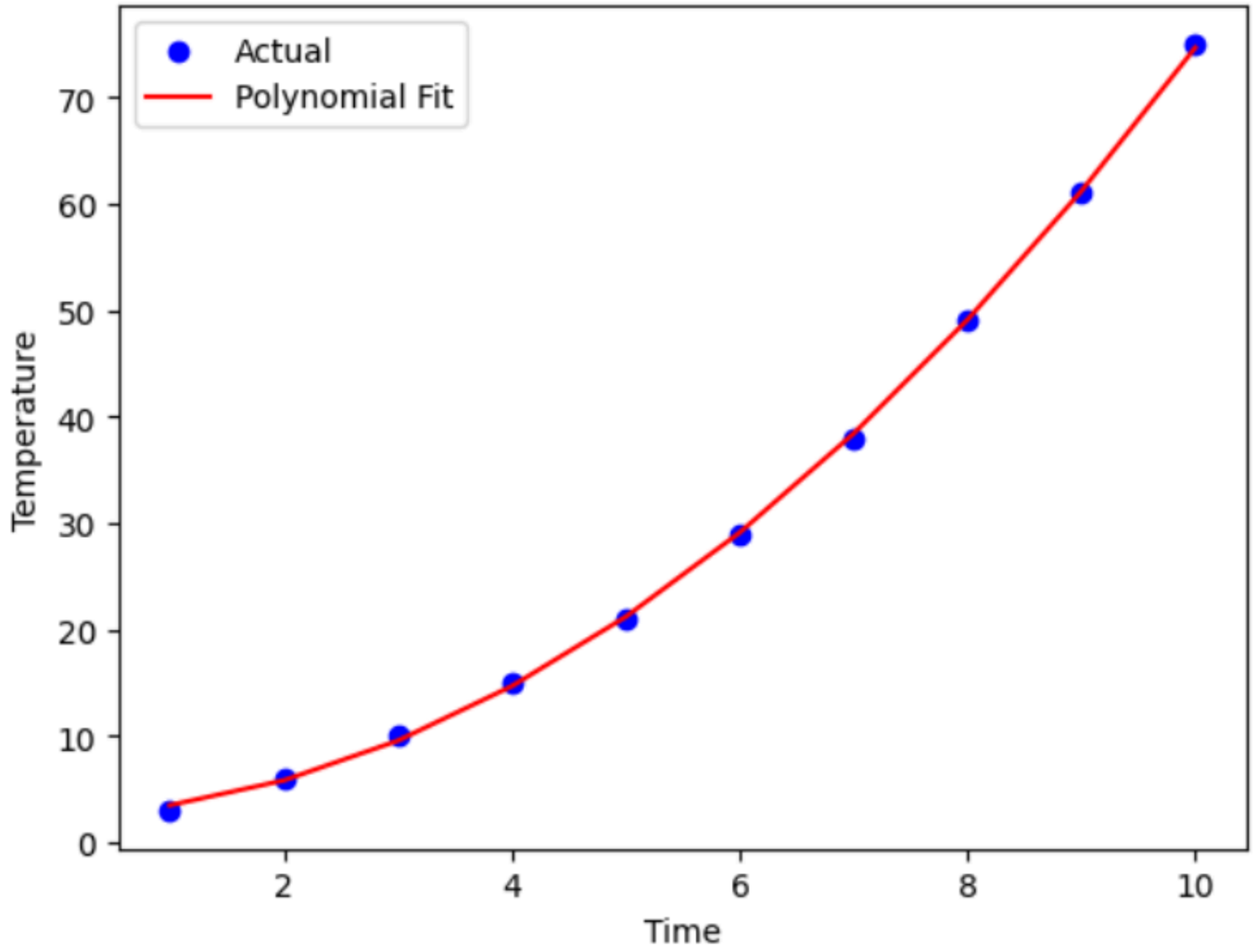
- Model is too simple and fails to capture patterns
- Happens when model is too simplistic

📌 How to Fix It?

✅ Use Polynomial Regression for non-linear data

✅ Use Regularization (Lasso & Ridge) to control model complexity

POLYNOMIAL REGRESSION - CODING DEMO



REGULARIZATION

📌What is Regularization?:

- A technique to prevent overfitting by penalizing large coefficients.
- Forces the model to select only the most important features.

📌Types of Regularization:

- 1️⃣ Lasso Regression (L1 Regularization)

  Shrinks some coefficients to zero, performing feature selection.

- 2️⃣ Ridge Regression (L2 Regularization)

  Shrinks all coefficients equally but doesn’t remove any features.

Introduction to Classification

INTRODUCTION TO CLASSIFICATION

Regression vs. Classification

Regression	Classification
Predicts continuous values	Predicts categories
Example: Predict house price	Example: Spam vs. Not Spam

Common Use Cases:

- ✅ Disease Diagnosis (Diabetic vs. Non-Diabetic)
- ✅ Fraud Detection (Genuine vs. Fraudulent Transaction)
- ✅ Spam Detection (Spam vs. Not Spam)

LOGISTIC REGRESSION (PREDICTING CATEGORIES)

📌What is Logistic Regression?

- A classification algorithm used to predict categorical outcomes.
- Outputs probabilities (values between 0 and 1).
- Uses the sigmoid function to transform linear predictions into probabilities.

📝 Example:

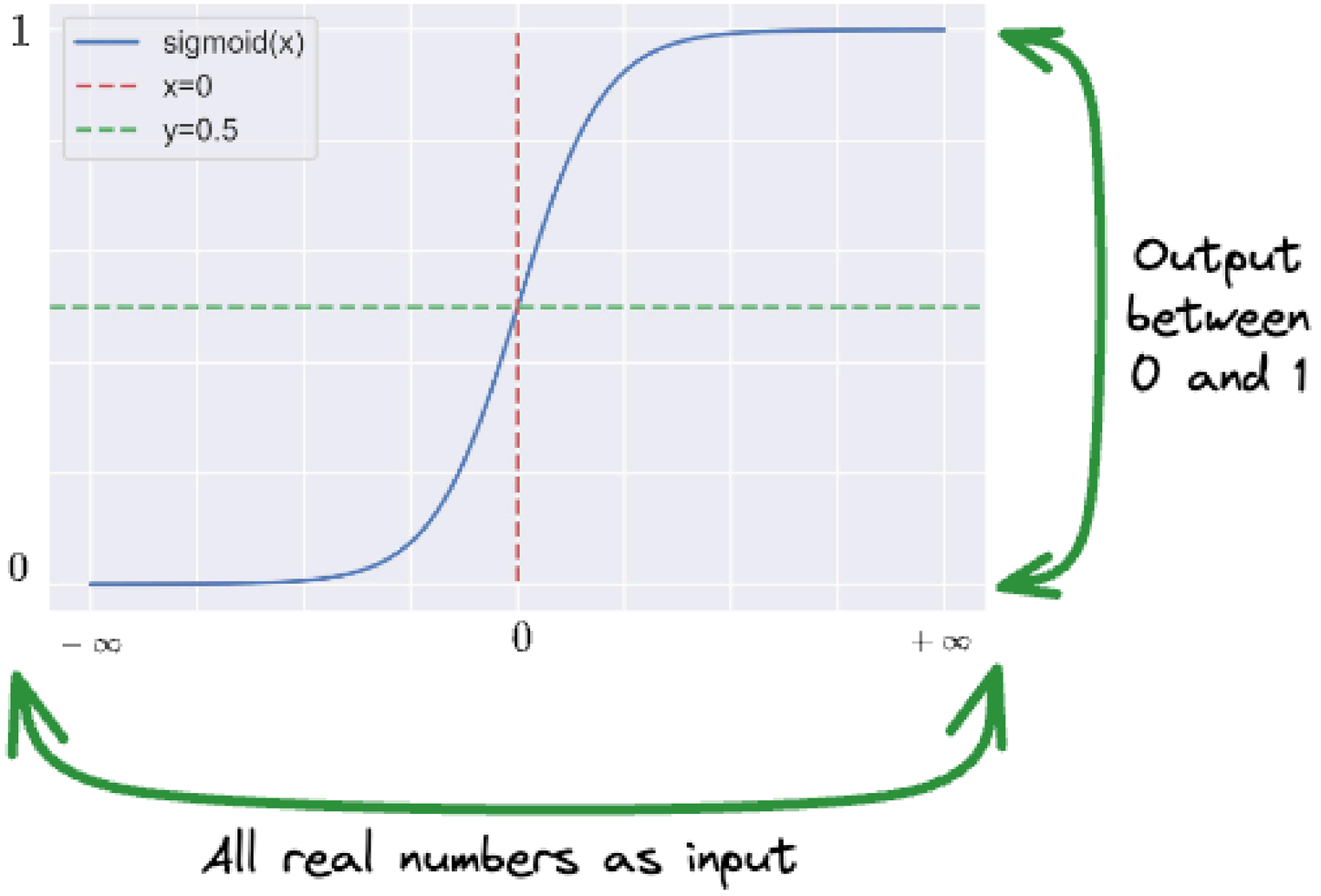
Predicting if an email is Spam (1) or Not Spam (0).

LINEAR VS LOGISTIC REGRESSION

Feature	Linear Regression	Logistic Regression
Output Type	Continuous (any number)	Discrete (0 or 1)
Function Used	Straight Line Equation	Sigmoid Function
Use Case	Predicting house prices	Predicting spam emails

SIGMOID FUNCTION

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$



DECISION BOUNDARY

📌 **Decision Boundary:**

A threshold (e.g., 0.5) is used to classify outcomes.

- If sigmoid output > threshold, classify as 1.
- If sigmoid output < threshold, classify as 0.

- Example: Classifying whether a tumor is benign (0) or malignant (1).

COST FUNCTION & OPTIMIZATION

- Unlike linear regression (which uses MSE), logistic regression uses Log Loss (Binary Cross-Entropy)

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1-y_i) \cdot log(1-p(y_i))$$

SYNTAX

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

- Uses fit() to train the model
- predict() classifies new data

MODEL EVALUATION METRICS

✅ **Accuracy Score:**

- Measures the percentage of correctly classified instances.
- Formula:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

Works well for balanced datasets but may be misleading for imbalanced data.

MODEL EVALUATION METRICS

✅ Confusion Matrix:

- A table that compares actual vs. predicted classes.
- Format:

Actual \ Predicted	Positive (1)	Negative (0)
Positive (1)	TP (True Positives)	FN (False Negatives)
Negative (0)	FP (False Positives)	TN (True Negatives)

Helps identify false positives and false negatives.

MODEL EVALUATION METRICS

✅ **Precision, Recall, F1-score:**

- Precision: Out of predicted positives, how many were actually positive?

$$\text{Precision} = \frac{TP}{TP + FP}$$

MODEL EVALUATION METRICS

- **Recall (Sensitivity)**:Out of actual positives, how many were correctly predicted?

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score**: Balances precision & recall (harmonic mean)

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

SUMMARY

- Logistic Regression is used for binary classification.
- Uses the sigmoid function to output probabilities.
- Decision boundary helps classify outcomes.
- Evaluated using accuracy, precision, recall, F1-score.

THANK YOU