1: introduction to Regression

What is regression in Machine Learning?

is a type of supervised learning used to predict continuous, numerical output values (like house prices, age, or stock market trends) based on input features

it is predict a continuous outcome (y) based on the value of one or more predictor variables (x)

How is Regression different from classification?

The fundamental difference is that classification predicts a discrete label or category spam or not spam, while regression predicts a continuous, real valued number

Difference:

Regression: how many?

Classification: which class?

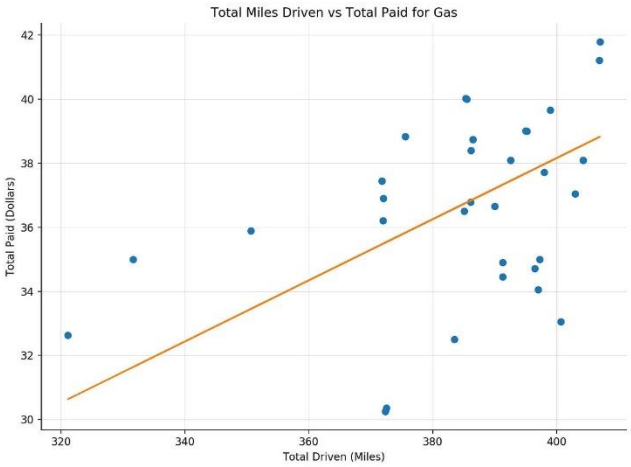
Give one real-life example of regression and one of classification?

Type	Real-life Example	Target Type
Regression	car resale, age, brand, and condition	Continuous value
Classification	classifying transactions as fraudulent or legitimate	Discrete label
Classification	classifying students as pass or fail based on study hours, attendance, and past scores	Discrete label
Regression	electricity in a household based on temperature, and time of day	Continuous value

2. Types of Regression

1: Linear Regression: it provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events to one line

Advantages	Limitations
Simple and easy to implement	not good for nonlinear data
Easy to interpret	Sensitive to outliers
Scalable for large datasets	Cannot capture complex patterns
Suitable for real-time prediction	Multicollinearity can distort coefficients
Good baseline model	Not suitable for categorical outputs



One real world use case:

This picture shows it predict y total paid $ shows x to miles driven

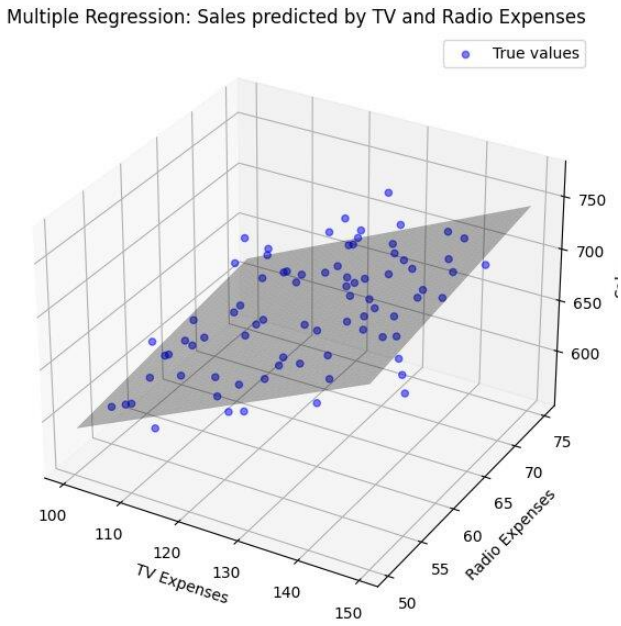
2. Multiple Linear Regression MLR: Models the relationship between two or more independent variables inputs and a continuous dependent variable

Advantages:

1. Handles multiple independent variables simultaneously
2. Easy to interpret coefficients for each predictor
3. Useful for predicting continuous outcomes

Limitations:

1. Assumes linear relationships between inputs and output
2. Sensitive to outliers
3. Multicollinearity (highly correlated predictors) can distort results



One real world use case:

Predicting sales of Tv & Radio based on height, weight, year, electricty

Waxa jira kuwo kale

3. Logistic Regression

4. Ordinal Regression

5. Multinomial Logistic Regression

3. Polynomial regression:

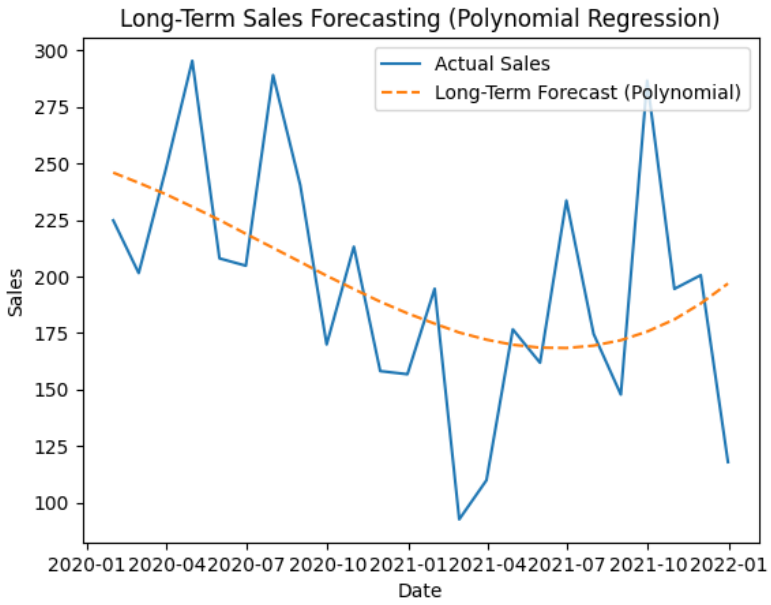
It adapts to nonlinear curves/relationships, such as up/down trends

Advantages

1. Flexibility: Polynomial regression can capture relationships from linear to highly nonlinear which can be crucial when performing real world data set that do not follow simple linear relationships
2. Improved Accuracy: The more complex a model is, the more complex the predictions will be, and thus polynomial regression will outperform simple linear regression
3. Still Linear in Parameters: Although the relationships between the variables are nonlinear, the model can be easily estimated through linear regression
4. Predictive Power: Forecasting with nonlinear relationships, such as growth curves and seasonal patterns, is very effective.

Limitations

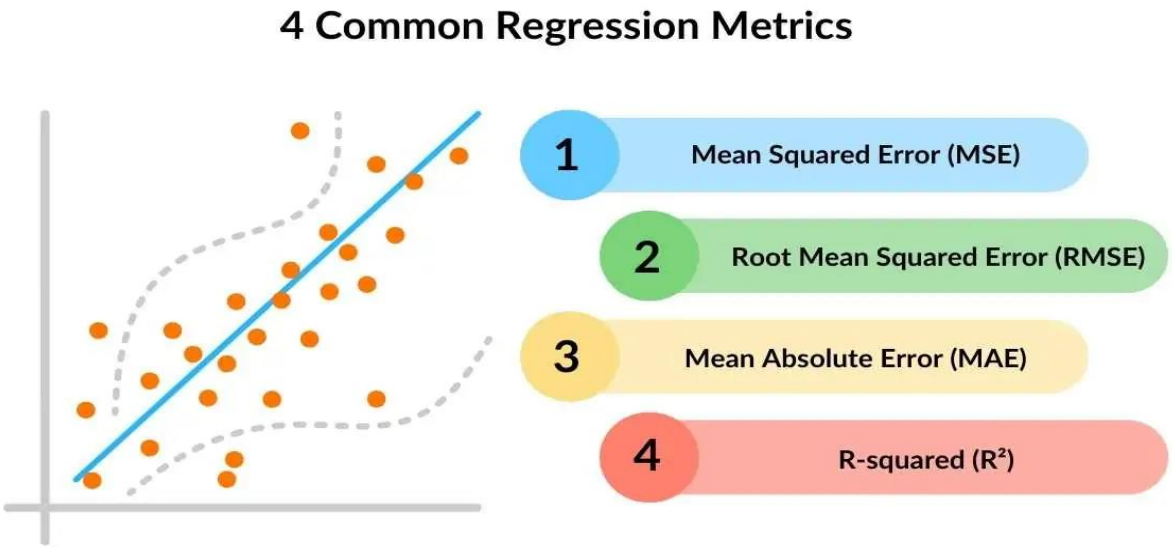
1.Overfitting Risk: Noise could be captured by real trends on high-degree
2.Extrapolation Issues: Inaccurate predictions fall outside the observed range
3.Interpretability: Coefficients in higher degrees are comparatively more difficult to reason than lower degree
4.Sensitive to Outliers: Curve fit can be affected by extreme values



One real world use case:

Researchers aim to understand long sales in polyonominal regression in the sales to date making polynomial regression a suitable modeling approach

3. Regression Metrics



1. MAE_Mean Absolute Error: (Celceliska khaladka toosan) is measures the average absolute difference between the predicted values and the actual values

Example: If MAE = 5, your predictions are off by 5 units on average

2. MSE – Mean Squared Error: (MAE Khaladka la laba jibbaaray) it penalizes larger errors more than smaller ones because errors are squared

Example: If one prediction is very far from the actual value, it will increase the MSE significantly

3. RMSE – Root Mean Squared Error: (Square root of MSE natiijo la mid ah unug yada asalka) it penalizes larger errors, but now the units are the same as the original data, making it easier to interpret

Example: If predicting house prices in $1000s and RMSE = 10, your predictions are off by about $10,000 on average

4. R^2 – Coefficient of Determination: (inta uu modelku xogtaada fahmay) R^2 measures how much of the variation in the dependent variable is explained by your model

1 = perfect prediction, 0 = model predicts no better than the mean, < 0 = model is worse than predicting the mean

Example: If $R^2 = 0.95$ your model explains 95% of the variation in the data

Comparison table showing their differences:

Type	**Relationship Shape**	**Example Features**	**Best For**
Linear Regression	Straight line	Years of Experience Salary	Simple, single-variable predictions
Multiple Linear Regression	Multi-dimensional plane	Age, Education, Experience Salary	Predicting outcomes with multiple factors
Polynomial Regression	Curved line	Time, $Time^2$ Plant Growth	Non-linear growth or trends

4. Underfitting and Overfitting

Overfitting: (waa marka Model aad u complex u noqdo si fiican ayuu u xafidaa training data laakiin wuxuu ku fashilmaa data cusub in isbaro ama u soo saaro) when model gives accurate predictions for training data but not for new data, the model tries to predict outcomes for new data sets

Underfitting: (waa marka Model aad u simple u noqdo wax badan ma fami karo) when the model is too simple to capture the underlying pattern in the data other words, the model is not complex enough to represent the true relationship between the input and output variables

What causes overfitting, especially in polynomial regression?

High degree polynomials very wiggly curve that tries to fit all training points exactly, this causes the curve to work well on training data, but poorly on test data (large generalization error)

1. Model Complexity: Highly complex models with numerous parameters can capture noise in the training data, leading to overfitting
2. Imbalanced Data: When certain user or item categories dominate the dataset, the model may overfit to these categories while neglecting others
3. Overtraining: Training a model for too many epochs can lead to overfitting, as the model starts to memorize the training data
4. Financial Losses: In e-commerce, poor recommendations can lead to lost sales and reduced customer loyalty

Example:

Degree 2 or 3 Usually captures a general trend

Degree 10 Covers all points, and can be an oscillating curve that makes no sense

Give 2–3 methods to prevent overfitting

1. Split dataset: training and testing sets example 80/20 to check generalization
2. Cross-validation: Split data into k folds, train/test multiple times so all data is used
3. Early stopping: Stop training when validation loss starts increasing to prevent memorization
4. Data augmentation & L1/L2 regularization
5. Reduce model complexity: Remove layers or reduce the number of neurons to make the model simpler

Key Results / Insights:

1. it provided accurate forecasts for future sales
2. Important predictors of sales included website traffic, promotions, and product category
3. Business could plan inventory and marketing more efficiently
4. Demonstrated the practical impact of regression analysis in business decision-making

5. Real-World Case Study — Regression in Business

Title: Predicting Retail Store Sales

Type: Multiple Linear Regression

Goal: Predict weekly sales for retail stores to optimize inventory, expensive and staffing

Data:

1. Historical sales data
2. store size
3. promotion schedules
4. holidays
5. purchase inventory

Key Results / Insights:

1. Store size vs promotion schedules were the strongest predictors of sales
2. helped store managers plan inventory & workforce efficiently

References:

1: https://builtin.com/data-science/regression-machine-learning

2: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/

3: https://medium.com/@data-overload/understanding-polynomial-regression-a-powerful-tool-for-complex-relationships-d2394a898fd6

4: https://www.ibm.com/think/topics/classification-vs-regression

5: https://net-informations.com/ml/mla/poly.htm?utm_source=chatgpt.com

6: https://medium.com/data-science/8-simple-techniques-to-prevent-overfitting-4d443da2ef7d

7: https://shorturl.at/9pKxC

By: Mohamed Mohamud