Regression vs Classification

Supervised Learning

Daniel E. Acuna
Associate Professor, University of Colorado Boulder

Contents of This Video

In this video, we will cover:

- Definition of regression and its characteristics
- Real-world regression examples and evaluation metrics
- Definition of classification and its characteristics
- Real-world classification examples and evaluation metrics
- Key differences between regression and classification
- How to identify whether a problem is regression or classification

What is Regression?

Regression predicts **continuous numerical values**:

- House prices
- Temperature forecasts
- Exam scores
- Income predictions
- Stock prices

Regression Examples

Let's look at concrete examples:

- **Housing Price Prediction**: Predicting a house will sell for $325,500 based on square footage, location, etc.
- **Wage Forecasting**: Predicting someone will earn $72,150 per year based on their experience and education
- **Stock Market**: Predicting a stock will be worth $142.75 tomorrow

All are **continuous values** (can be any number in a range)

Evaluating Regression Models

Regression models are typically evaluated using:

- **Mean Squared Error (MSE)**: Average of squared differences between predictions and actual values
- **Mean Absolute Error (MAE)**: Average of absolute differences between predictions and actual values
- **Root Mean Squared Error (RMSE)**: Square root of MSE (in same units as prediction)

Lower values indicate better performance

Regression Error: A Simple Example

House Price Prediction

House	Actual	Predicted	Error	Error²
1	$200K	$220K		
2	$150K	$130K		
3	$300K	$310K		
4	$250K	$270K		
5	$180K	$160K		

Regression Error: Calculated

House Price Prediction

House	Actual	Predicted	Error	Error²
1	$200K	$220K	+$20K	400M
2	$150K	$130K	-$20K	400M
3	$300K	$310K	+$10K	100M
4	$250K	$270K	+$20K	400M
5	$180K	$160K	-$20K	400M

Mean Absolute Error (MAE) = (|20|+|20|+|10|+|20|+|20|)/5 = $18K

Mean Squared Error (MSE) = (400+400+100+400+400)/5 = 340M

Root MSE (RMSE) = $\sqrt{340M} \approx \$18.4K$

What is Classification?

Classification predicts **discrete categories/classes**:

- Email: Spam or Not Spam
- Medical: Disease or Healthy
- Image: Cat, Dog, or Bird
- Customer: Will Subscribe or Won't Subscribe

Classification Examples

Let's look at concrete examples:

- **Spam Detection**: Classifying an email as "Spam" or "Not Spam"
- **Medical Diagnosis**: Classifying a tumor as "Malignant" or "Benign"
- **Sentiment Analysis**: Classifying reviews as "Positive", "Neutral", or "Negative"

All are **discrete categories** (only specific values allowed)

Binary vs. Multi-class Classification

Binary Classification

- Two possible classes
- Examples:
  - Spam/Not Spam
  - Pass/Fail
  - Fraud/Not Fraud

Multi-class Classification

- More than two classes
- Examples:
  - Animal type (cat/dog/bird)
  - Movie genre (action/comedy/drama)
  - Digit recognition (0-9)

Evaluating Classification Models

Classification models are typically evaluated using:

- **Accuracy**: Percentage of correctly classified instances
- **Precision**: Proportion of positive identifications that were actually correct
- **Recall**: Proportion of actual positives that were correctly identified
- **F1 Score**: Harmonic mean of precision and recall

Higher values indicate better performance

Classification Error: A Simple Example

Email Spam Classification

Consider 5 emails in our test set:

Email	Content	Actual Class	Predicted Class	Correct?
1	“Win a free vacation now!”	Spam	Spam	
2	“Meeting at 3pm tomorrow”	Not Spam	Not Spam	
3	“Claim your prize money”	Spam	Not Spam	
4	“Project report attached”	Not Spam	Not Spam	
5	“Increase your followers”	Spam	Spam	

Classification Error: Evaluated

Email Spam Classification

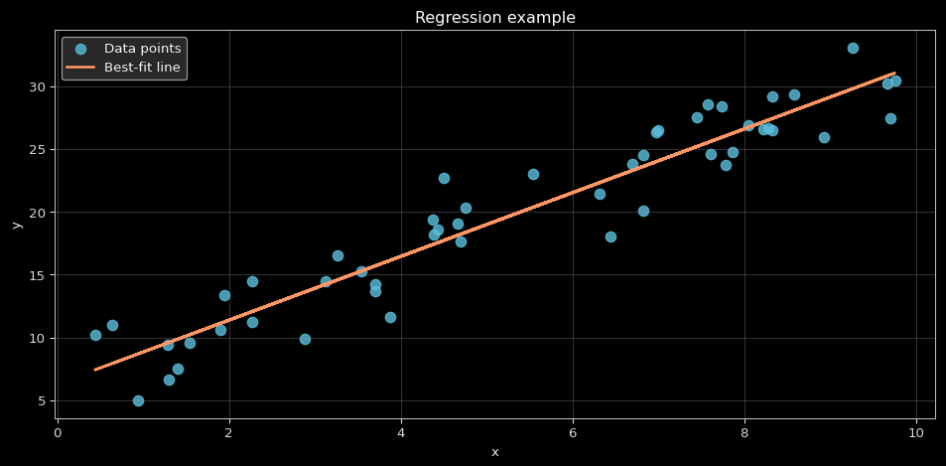
Email	Content	Actual Class	Predicted Class	Correct?
1	“Win a free vacation now!”	Spam	Spam	✓
2	“Meeting at 3pm tomorrow”	Not Spam	Not Spam	✓
3	“Claim your prize money”	Spam	Not Spam	✗
4	“Project report attached”	Not Spam	Not Spam	✓
5	“Increase your followers”	Spam	Spam	✓

Evaluation Metrics:

- **Accuracy**: 4/5 = 80% (4 correct predictions out of 5)
- **Precision**: 2/2 = 100% (all predicted spam were actually spam)
- **Recall**: 2/3 = 67% (only found 2 of the 3 actual spam emails)
- **F1 Score**: $2 \times (\text{precision} \times \text{recall})/(\text{precision}+\text{recall}) = 80\%$

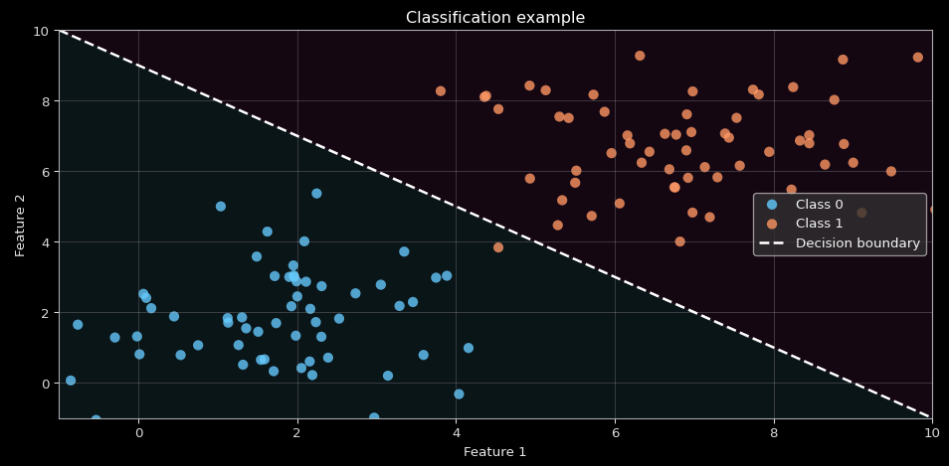
Visualization: Regression vs. Classification

Regression



Finding a line/curve that best fits points

Classification



Finding boundaries that separate classes

Algorithms for Both Tasks

Some algorithms can be used for either task (with modifications):

- **K-Nearest Neighbors (KNN)**
  - Regression: average the values of K nearest neighbors
  - Classification: vote on class based on K nearest neighbors
- **Decision Trees**
  - Regression: predict a number at each leaf
  - Classification: predict a class at each leaf

Output Differences

Regression Output

```
price = 200000 + 150 * sqft +
        50000 * (bedrooms) -
        10000 * (age_of_house)
```

Produces a number (e.g., $325,500)

Classification Output

```
if income > $50k AND age < 30:
    classify as "Likely to Subscribe"
else:
    classify as "Unlikely to Subscribe"
```

Produces a category

Blurred Lines

Sometimes the boundary between regression and classification can blur:

- **Probabilities**: Predicting “80% chance of rain” is regression, but often converted to classification (“Will rain” vs “Won’t rain”)
- **Logistic Regression**: Actually a classification algorithm despite the name! Outputs probabilities that are converted to classes

The key question: Are we predicting a number or a category?

Let's Practice!

Determine if each scenario is regression or classification:

1. Predict a student's exam score (0-100) based on hours studied.
2. Predict whether a student will pass or fail an exam based on hours studied.
3. Predict how many minutes late / early a flight will arrive.
4. Predict which team will win a basketball game.

Answers

- Scenario 1: **Regression** — output is a continuous numerical value.
- Scenario 2: **Classification** — output is one of two categories (pass/fail).
- Scenario 3: **Regression** — output is a numerical value (minutes).
- Scenario 4: **Classification** — output is one of two categories (Team A or Team B wins).

Quick Identification Tip

How to quickly identify the task type:

- Can you list all possible output values? → **Classification**
- Is the output a number on a continuum? → **Regression**
- Is the output naturally expressed as a category? → **Classification**
- Is the output naturally expressed as a quantity? → **Regression**

Summary

Regression

- Predicts continuous numerical values
- Examples: price, temperature, score
- Metrics: MSE, MAE, RMSE
- Focus: How close to actual value

Classification

- Predicts discrete categories
- Examples: spam/not spam, disease type
- Metrics: accuracy, precision, recall
- Focus: Correct or incorrect class

What We've Covered

In this video, we've discussed:

- Regression problems that predict continuous numerical values
- Classification problems that predict discrete categories
- Evaluation metrics specific to each problem type
- Examples showing the difference between predicting "how much" vs. "which category"
- Practical tips for identifying regression vs. classification tasks
- How the same algorithm can be adapted for either task type