

1

1  What are Overfitting and Underfitting?

- ◆ Underfitting

Model is too simple

Cannot learn the pattern in the data

 Performance

Poor on training data

Poor on testing data

 Example:

Using a straight line to fit a curved relationship

Cause:

Very simple model

Too few features

Insufficient training

- ◆ Overfitting

Model is too complex

Learns noise + memorizes data

 Performance

Very good on training data

Poor on testing/unseen data

 Example:

Decision tree grown too deep memorizing every data point

Cause:

Too complex model

Too many features

Too little data

You said:

1. What are Overfitting and Underfitting?
2. What is the difference between Supervised and Unsupervised Learning?
3. What is a training dataset and a testing dataset? Why is data splitting important?
4. What is feature scaling and why is it needed in some algorithms?
5. How does a Linear Regression model work?

ChatGPT said:

Nice set of questions — these are classic interview + exam favorites 🔥

I'll keep the explanations simple, clear, and practical, with tiny examples so they stick.

1. 📊 What are Overfitting and Underfitting?

- ♦ Underfitting

Model is too simple

Cannot learn the pattern in the data

📈 Performance

Poor on training data

Poor on testing data

📌 Example:

Using a straight line to fit a curved relationship

🧠 Cause:

Very simple model

Too few features

Insufficient training

- ♦ Overfitting

Model is too complex

Learns noise + memorizes data

📈 Performance

Very good on training data

Poor on testing/unseen data

 Example:

Decision tree grown too deep memorizing every data point

 Cause:

Too complex model

Too many features

Too little data

2 Difference between Supervised and Unsupervised Learning

- ◆ Supervised Learning

Data is labeled

We know input (X) and output (y)

 Used for:

Classification

Regression

 Examples:

Spam vs Not Spam (label = spam / not spam)

House price prediction (label = price)

Algorithms:

Linear Regression

Logistic Regression

SVM

Decision Tree

KNN

- ◆ Unsupervised Learning

Data is not labeled

No target variable

📌 Used for:

Pattern discovery

Grouping

Dimensionality reduction

📖 Examples:

Customer segmentation

Topic modeling

Algorithms:

K-Means

Hierarchical Clustering

DBSCAN

PCA

3 📖 Training Dataset & Testing Dataset

- ◆ Training Dataset

Used to train the model

Model learns patterns from this data

- ◆ Testing Dataset

Used to evaluate performance

Simulates unseen real-world data

- ♦ Why is Data Splitting Important?

Data splitting means dividing data into training and testing sets.

✓ Why it is important:

To check how well the model performs on unseen data

To detect overfitting

To measure real-world performance

✗ Without data splitting:

Model may memorize data

High accuracy but poor generalization

- 4 ♦ What is Feature Scaling and why is it needed?

✓ Feature Scaling

Feature scaling is the process of bringing all features to a similar numerical range.

- ♦ Why Feature Scaling is Needed:

Prevents features with large values from dominating

Improves model convergence

Makes distance-based algorithms work correctly

📌 Needed for:

KNN

K-Means

SVM

Linear Regression

PCA

✗ Not needed for:

Decision Tree

Random Forest

Common Methods:

Min-Max Scaling

Standardization (Z-score)

5  How does Linear Regression work?

- ♦ Goal

Find a best-fit straight line that predicts output y from input x

- ♦ Equation

y

=

m

x

+

b

$y = mx + b$

Where:

$m \rightarrow$ slope (weight)

$b \rightarrow$ intercept

- ♦ How It Learns

Start with random m and b

Predict output

Calculate error (actual – predicted)

Update m and b to minimize error

Repeat until error is minimum

 Error is me