Understanding Artificial Intelligence from Scratch

Lecture notes · AI Foundations for Everyone · Module 1: What AI is

Learning objectives

O1	Explain in plain words what artificial intelligence and machine learning are.
O2	Describe how training data turns into a model that makes predictions.
O3	Understand, at a simple level, how generative AI creates text.
O4	Recognise what AI does well, where it fails, and how to use it responsibly.

Prerequisites: None. Basic arithmetic (adding, dividing, percentages) is used in a few exercises.

Contents: 1. What AI is, and what it is not · 2. Rules versus learning · 3. Features, labels, training and the model · 4. Testing honestly: training data and test data · 5. Neural networks and deep learning · 6. Generative AI: how text is produced · Worked examples · Summary · References

1. What AI is, and what it is not

The OECD, whose definition many governments use, describes an AI system as a machine-based system that infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions. In plain words: it works out how to get from input to output.

AI is not the same as a robot. A robot is a machine with a body; most AI is software with no body at all, often running in a distant data centre.

Artificial intelligence is the broad field. Machine learning is the main way modern AI is built: instead of programmers writing every rule, the computer finds patterns in examples.

2. Rules versus learning

Rule-based programs follow instructions written by people: 'if an email contains free, mark it as spam'. They are predictable, but they break when the world changes or when cases are too varied to list.

Machine learning starts from examples with the correct answers attached (labels). The computer adjusts a model until its answers match the labels as closely as possible, then uses the model on new cases.

Both approaches are still used. Rules are best when the logic is simple and must be exact (tax rates); learning is best when patterns are complex and examples are plentiful (recognising faces, spam, speech).

3. Features, labels, training and the model

Features are the measurements that describe each example (a fruit's weight and colour; a house's size and location). The label is the answer we want the model to learn (apple or orange; the sale price).

Training is a loop: make a prediction, measure the error, adjust the model slightly, and repeat thousands of times. The result, the model, is a set of numbers called parameters. Our line has three;

large language models have billions.

Inference means using the trained model on new data to make a prediction. A good system also reports how confident it is: examples close to the dividing line are uncertain.

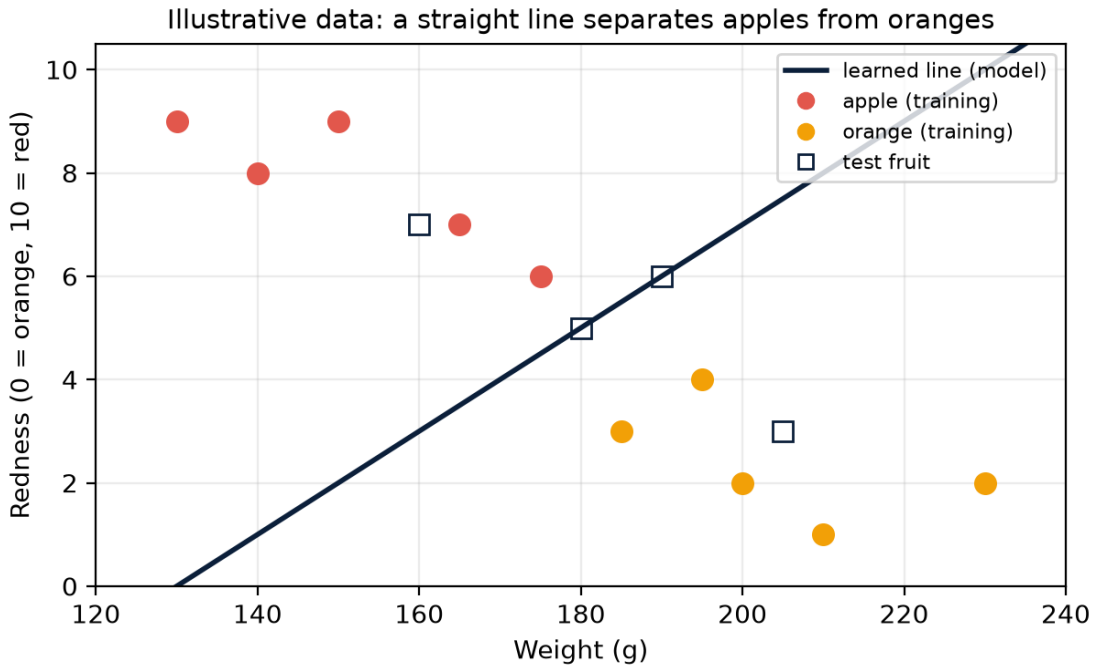


Figure 1. Training fruit (circles), test fruit (squares) and the learned rule: apple if redness >= 0.1 x weight - 13.

4. Testing honestly: training data and test data

A model must be tested on examples it never saw during training (test data). Testing on training data is like marking a student on last year's exam that they memorised.

Overfitting is when a model learns the training examples, including their noise, instead of the general pattern. It scores well on training data and badly on new data.

Accuracy is the share of predictions that are correct: correct predictions divided by all predictions. Accuracy alone can mislead: if 95 out of 100 emails are genuine, a 'model' that always says 'genuine' is 95% accurate and useless at catching spam.

5. Neural networks and deep learning

A neural network is made of simple units arranged in layers. Each unit multiplies its inputs by weights, adds them up, and passes the result on. Training adjusts the weights with the same predict-measure-adjust loop.

Networks with many layers are called deep learning. They can learn very complex patterns, such as faces in millions of pixels, but they need a lot of data and computing power, and their reasoning is hard to inspect.

Despite the name, neural networks are only loosely inspired by the brain; they are mathematics running on computers.

6. Generative AI: how text is produced

A large language model is trained on huge amounts of text to predict the next token (a word or part of a word). For each possible next token it calculates a probability; the probabilities for all options add up to 1.

The model picks a token (usually one of the likely ones, with some randomness), adds it to the text and predicts again. Repeating this produces sentences, summaries and code.

Because it predicts likely words rather than checking facts, it can write fluent, confident statements that are false ('hallucinations').

Types of machine learning

Supervised learning: learning from labelled examples (spam / not spam). Unsupervised learning: finding groups or structure in unlabelled data (grouping customers with similar shopping habits). Reinforcement learning: learning by trial and reward (a program learning a game by playing it).

Most everyday AI products combine several of these with ordinary rule-based software.

Responsible use checklist

Accuracy: check important facts against reliable, independent sources. Privacy: do not enter personal, confidential or sensitive information you would not want stored. Fairness: ask whether the data might under-represent some groups. Accountability: keep a human responsible for decisions about health, money, law, employment or safety. Transparency: say when you used AI in your work, following your school's or employer's rules.

AI does not act on its own goals: people choose the data, the objective, the testing and the use, so people carry the responsibility.

Worked examples

Worked example 1: using a learned rule (inference)

The model learned the rule: apple if redness >= 0.1 x weight - 13. Classify a new fruit weighing 160 g with redness 7.

Step 1. Compute the threshold: 0.1 x 160 - 13 = 16 - 13 = 3.

Step 2. Compare: redness 7 >= 3, so the rule says apple.

Step 3. Check against the model function: apple.

Step 4. The fruit is well above the line (7 versus 3), so the prediction is a confident one.

Worked example 2: measuring accuracy on test data

Four test fruits: (160 g, 7, apple), (205 g, 3, orange), (180 g, 5, apple), (190 g, 6, orange). Apply the rule and calculate the accuracy.

Step 1. 160 g: threshold 3, redness 7, predicted apple: correct.

Step 2. 205 g: threshold 7.5, redness 3, predicted orange: correct.

Step 3. 180 g: threshold 5, redness 5, predicted apple (5 >= 5): correct.

Step 4. 190 g: threshold 6, redness 6, predicted apple, but it is an orange: wrong.

Step 5. Accuracy = 3 correct / 4 = 75%. On the training data the same rule scored 100%: the drop shows why we must test on new data.

Worked example 3: from probabilities to text

After 'The cat sat on the', a model gives: mat 0.40, floor 0.20, sofa 0.15, bed 0.12, chair 0.10, banana 0.03. Check the probabilities and explain what happens next.

Step 1. Add them: 0.40 + 0.20 + 0.15 + 0.12 + 0.10 + 0.03 = 1.00. Probabilities over all options must add up to 1.

Step 2. The most likely word is 'mat' (40%). With a little randomness the model may sometimes choose 'floor' or 'sofa'.

Step 3. The chosen word is added to the text, and the model predicts the following word from the longer text.

Step 4. Nothing in this process looks up whether a cat really sat anywhere: fluency is not the same as truth.

Summary

• An AI system works out from its input how to produce outputs such as predictions, content or decisions.

• Machine learning finds patterns in labelled examples instead of following hand-written rules.

• Training repeats predict, measure the error, adjust; the result is a model made of parameters.

• Always test on unseen data; beware overfitting and misleading accuracy.

• Language models predict the next token; they can be fluent and wrong, so check facts and keep humans in charge.

References

References: KN-GEN-AIN-C01-M01-T01-L01
OECD (2023/2024), updated definition of an AI system and its Explanatory Memorandum (OECD AI Papers No. 8, Feb 2024): "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (simplified in narration); the memorandum also notes that training means improving performance with machine-learning techniques. Standard introductory ML concepts (supervised learning, training/test split, overfitting) as in Russell & Norvig, Artificial Intelligence: A Modern Approach, 4th ed.; next-token prediction for language models as described in the original GPT papers (Radford et al., 2018/2019). The fruit data and next-word probabilities on screen are ILLUSTRATIVE (stated in narration), not measured from a real system.

All references were consulted for accuracy; no text or figures from them are reproduced in the video. Verify links before publication.
Fruit data and word probabilities are illustrative teaching examples, not measurements from a real system. Definitions follow the OECD (2024) and standard introductory texts (references).