# Machine learning

International Baccalaureate · IB Diploma · Computer Science · HL

## Scope and prerequisites

This handout teaches Theme A4, Machine learning, for the course first assessed in 2027: what machine learning is, its main types, how models are judged, and its ethical issues. Sections marked **AHL** are for Higher Level only.

Prerequisites: reading a scatter graph and a percentage. The datasets here are small and fictional.

## What is machine learning?

### A program that learns from examples

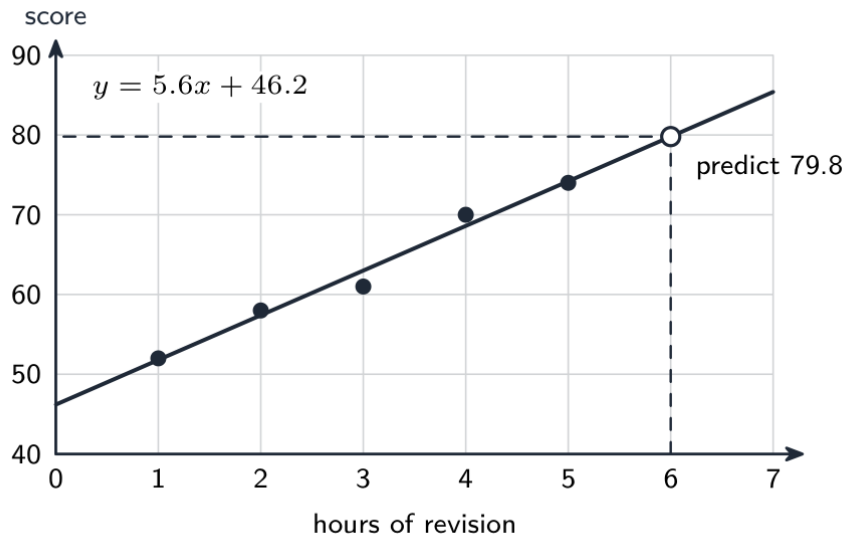
A spam filter is not given a list of rules such as "block any e-mail containing 'prize'". Instead it is shown thousands of e-mails labelled "spam" or "not spam" and finds the patterns itself. When spammers change their wording, it is retrained on new examples. This is **machine learning** (ML): a program improves at a task by learning patterns from data rather than following rules written by a programmer.

ML is one part of **artificial intelligence** (AI), the wider field of making computers do tasks that seem to need human intelligence. **Deep learning** is a kind of ML that uses neural networks with many layers.

### Types of machine learning

| Type | Learns from | Typical task | Example |
|---|---|---|---|
| **supervised learning** | labelled data: inputs with the right answers | classify, predict a number | spam or not; a house price |
| **unsupervised learning** | unlabelled data | find groups or patterns | grouping shoppers by habits |
| **reinforcement learning** | rewards and penalties from trial and error | choose actions | a robot learning to walk; a game player |

Supervised learning splits into **classification** (the answer is a category: cat or dog) and **regression** (the answer is a number: a score or a price).




*Supervised learning: the model learns a line from labelled examples, then predicts a new case*

### Training and testing

1. Collect and label data.
2. Split it: most for **training** the model, the rest held back as **test data**.
3. Train the model on the training data.
4. Measure it on the test data, which it has never seen.

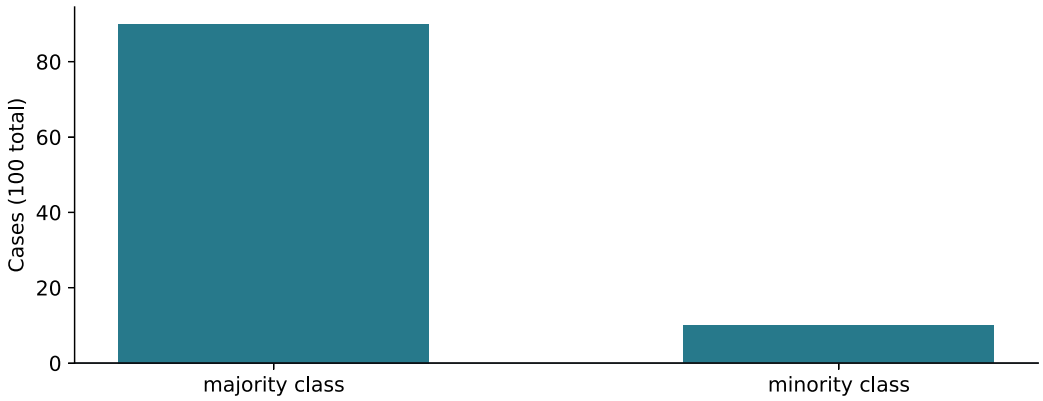
Testing on unseen data shows whether the model has learned the general pattern. A model that memorises its training data but fails on new data is **overfitting**. A model too simple to capture the pattern is **underfitting**.

### Measuring a classifier

**Accuracy** is the share of predictions that are correct:

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{total predictions}} \times 100\%$$

A model correct on 90 of 100 test cases has 90% accuracy. Accuracy can mislead when the classes are unbalanced:




*With 90 of 100 cases in one class, always predicting that class scores 90% yet misses every minority case*

A "model" that always answers "healthy" for a disease affecting 10% of patients scores 90% and finds no one who is ill. So check how the model does on each class, not just overall.

### Hardware for machine learning

Training repeats the same arithmetic on huge amounts of data, so it runs on **GPUs** and special chips with thousands of cores working in parallel, often rented in the cloud. Using a trained model (**inference**) needs far less power and can run on a phone.

## Ethics of machine learning

### Who is affected?

| Issue | What can go wrong | Example |
|---|---|---|
| **bias** | training data under-represents a group, so the model works worse for them | face recognition less accurate for darker skin |
| privacy | personal data collected or used without consent | photos scraped from social media |
| transparency | no one can explain why the model decided | a loan refused with no reason given |
| accountability | unclear who is responsible for a harmful decision | a self-driving car accident |
| jobs | tasks automated | checkout staff, translators |
| misuse | realistic fake images, audio or text | deepfakes spread as news |

Ways to reduce harm: train on data that represents everyone the model affects; test accuracy separately for each group; keep a human in the loop for serious decisions; collect data with consent and remove what is not needed; explain decisions to the people affected.

## AHL: Preparing data

### Data preprocessing

Real data is messy. **Data preprocessing** prepares it before training:

| Step | What it does | Example |
|---|---|---|
| cleaning | fix or remove missing, duplicate and wrong values | delete a row with age 250 |
| encoding | turn categories into numbers | red, green, blue → 0, 1, 2 |
| **normalisation** (scaling) | put features on a similar range, such as 0 to 1 | income in thousands and age in years |
| **feature selection** | keep only the inputs that help prediction | drop a student's ID number |
| dimensionality reduction | combine many features into fewer | thousands of pixel values → a few patterns |

A **feature** is one input variable, such as hours of revision; the **label** is the answer to be predicted, such as the score.

Watch for **data leakage** : a feature that would not be known when the prediction is made, such as using "days in hospital" to predict whether a patient will be admitted. The model then looks excellent in testing and fails in real use.

## AHL: Machine learning algorithms

### Linear regression

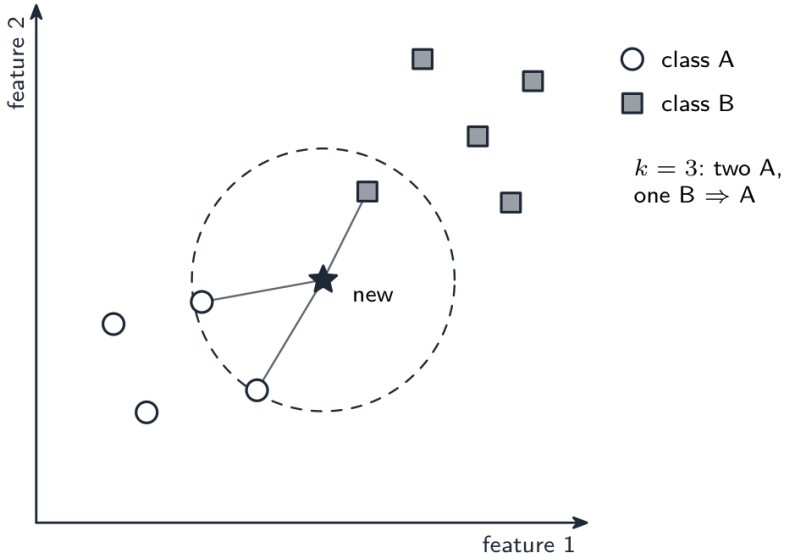
**Linear regression** fits the straight line $y = mx + c$ that best matches the data, then uses it to predict. For the revision data above, the best-fit line is $y = 5.6x + 46.2$:

$$y = 5.6 \times 6 + 46.2 = 79.8$$

so 6 hours of revision predicts a score of about 80. Predicting far outside the data, such as for 30 hours, is unreliable.

### k-nearest neighbours

**k-nearest neighbours** (k-NN) classifies a new item by the classes of the $k$ training items closest to it.




*With $k = 3$ the new point is classed A, although its single nearest neighbour is B*

1. Work out the distance from the new item to every training item.
2. Take the $k$ nearest.
3. Give the new item the most common class among them.

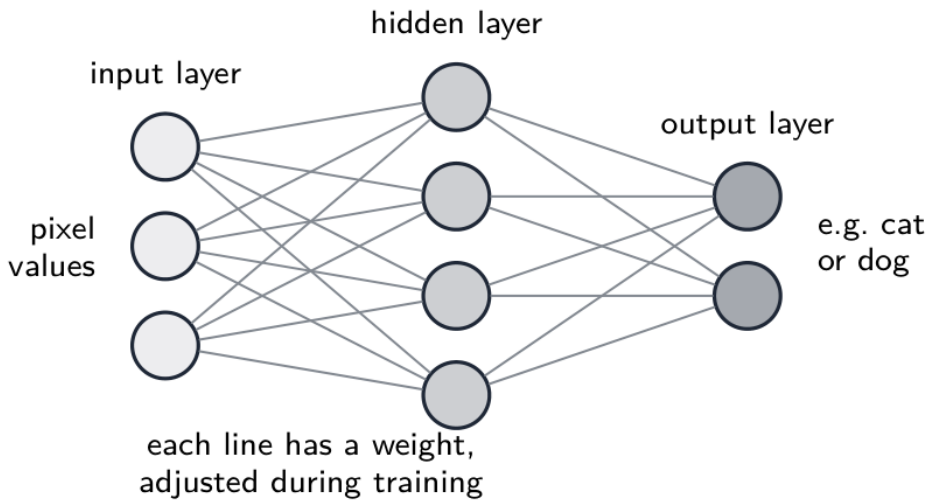
An odd $k$ avoids ties between two classes. k-NN needs no training, but each prediction compares with every stored item, so it is slow on large datasets.

### Decision trees and clustering

- A **decision tree** asks a sequence of yes/no questions about the features, such as "Is the e-mail from a known contact?", ending at a leaf that gives the class. It is easy for people to follow and explain.
- **k-means clustering** is unsupervised: choose $k$ centres, put each item in the cluster of its nearest centre, move each centre to the mean of its cluster, and repeat until nothing changes. Used to group customers or documents.

### Neural networks

An **artificial neural network** is made of layers of simple units (nodes). Each connection has a **weight** ; each node adds up its weighted inputs and passes the result on.




*Training adjusts the weights until the outputs match the labels*

Training compares the network's output with the right answer and adjusts every weight a little to reduce the error, over many thousands of examples. A **convolutional neural network** (CNN) has layers that scan small patches of an image to find edges, shapes and then objects, and is used for image recognition.

### ■ Precision and recall

For a classifier that flags positives (such as "has the disease"), compare predictions with reality in a **confusion matrix** :

| | **Actually positive** | **Actually negative** |
|---|---|---|
| predicted positive | true positive (TP) 4 | false positive (FP) 10 |
| predicted negative | false negative (FN) 1 | true negative (TN) 85 |

$$\text{accuracy} = \frac{TP + TN}{\text{total}} = \frac{4 + 85}{100} = 0.89$$

$$\text{precision} = \frac{TP}{TP + FP} = \frac{4}{14} \approx 0.29 \qquad \text{recall} = \frac{TP}{TP + FN} = \frac{4}{5} = 0.80$$

Recall asks "of the real positives, how many did we find?"; precision asks "of those we flagged, how many were right?". A disease screen wants high recall, so few sick patients are missed; a spam filter wants high precision, so real e-mails are not lost.

## Practice

1. Define machine learning, and explain how it differs from a program with written rules.
2. State the type of ML for each: (a) grouping songs by similarity with no labels; (b) predicting tomorrow's temperature from past records; (c) a program learning to play chess by playing itself.
3. State whether each is classification or regression: (a) a house price; (b) whether a photo shows a cat.
4. Explain why a model is tested on data it was not trained on.
5. A model makes 72 correct predictions out of 80. Calculate its accuracy.
6. A model for a rare disease (2% of patients) has 98% accuracy. Explain why this may be useless.
7. Explain **two** ethical issues of using ML to shortlist job applicants, and suggest a way to reduce each.
8. **AHL** Explain what data leakage is, with an example.
9. **AHL** Use $y = 5.6x + 46.2$ to predict the score for 3.5 hours of revision.
10. **AHL** In the k-NN figure, state the class given to the new point with $k = 1$ and with $k = 3$.
11. **AHL** A classifier has TP = 30, FP = 10, FN = 20, TN = 140. Calculate its accuracy, precision and recall.

**Answers.**

1. ML is a program that improves at a task by learning patterns from data. A rule-based program follows rules a programmer wrote; an ML model finds its own rules from examples.
2. (a) unsupervised; (b) supervised (regression); (c) reinforcement.
3. (a) regression; (b) classification.
4. To check that it has learned the general pattern rather than memorised the training examples (overfitting); unseen data shows how it will work in real use.
5. $\frac{72}{80} \times 100\% = 90\%$.
6. Always predicting "healthy" would also score 98% while finding no sick patients; the model must be judged on how many real cases it finds.
7. E.g. bias: past hiring data may favour one group, so the model repeats it; train on balanced data and test results for each group. Transparency: applicants cannot learn why they were rejected; use an explainable model and keep a human reviewer.
8. Using information that would not be available when the prediction is made, so test results are falsely good; e.g. predicting who will fail a course using their final exam mark.
9. $5.6 \times 3.5 + 46.2 = 19.6 + 46.2 = 65.8$.
10. $k = 1$: the nearest point is B, so **B**; $k = 3$: two A and one B, so **A**.
11. Accuracy $= \frac{30 + 140}{200} = 0.85$; precision $= \frac{30}{40} = 0.75$; recall $= \frac{30}{50} = 0.60$.

## Exam tips

- Name the type of learning from the data: labels → supervised; no labels → unsupervised; rewards → reinforcement.
- Say **why** test data is kept separate: to measure performance on unseen data and detect overfitting.
- Do not trust accuracy alone with unbalanced classes; mention per-class results.
- Ethics answers: name the issue, explain the harm **in the scenario**, then give a specific way to reduce it.
- **AHL** Show the formula before substituting for regression predictions, precision and recall.
- **AHL** For k-NN, state $k$, the neighbours found and the majority class.