

## 4.2 Machine learning algorithms

---

Name: \_\_\_\_\_ Class: \_\_\_\_\_ Date: \_\_\_\_\_

**Total: 22 marks**

**KEY WORDS** data preprocessing • data leakage  
• linear regression • k-nearest neighbours  
• artificial neural network • confusion matrix

### Objective

---

Build the skills to answer HL exam questions on **machine learning algorithms** – data preprocessing, linear regression, k-nearest neighbours, neural networks, and precision and recall.

**You must be able to:**

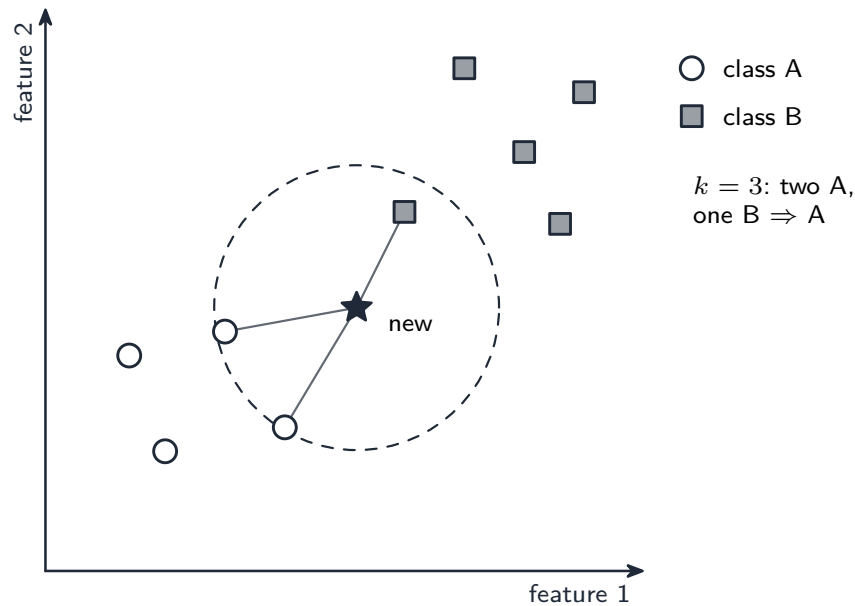
- describe the steps of **data preprocessing** and spot **data leakage**
- predict with a **linear regression** line
- classify a point by **k-nearest neighbours**
- describe the structure and training of an **artificial neural network**
- calculate precision and recall from a **confusion matrix**

### 1 Key knowledge and worked examples

---

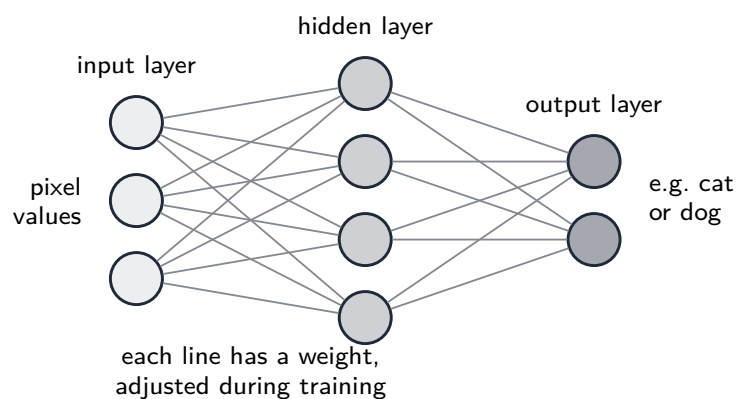
Study these first. Each one shows the method for a question type used later.

### ■ Worked example: k-nearest neighbours



1. Find the distance from the new point to every training point.
2. Take the  $k$  nearest: here B (0.89), A (1.12), A (1.17).
3. Majority class: **A**. With  $k = 1$  it would be **B**: the choice of  $k$  matters; an odd  $k$  avoids ties.

### ■ Key knowledge: neural networks



- Each connection has a **weight** ; each node sums its weighted inputs and passes the result on.
- Training: compare the output with the label, then adjust every weight slightly to reduce the error; repeat over many examples.

■ Worked example: precision and recall

|                    | Actually positive | Actually negative |
|--------------------|-------------------|-------------------|
| predicted positive | TP 4              | FP 10             |
| predicted negative | FN 1              | TN 85             |

$$\text{precision} = \frac{TP}{TP + FP} = \frac{4}{4 + 10} \approx 0.29$$

$$\text{recall} = \frac{TP}{TP + FN} = \frac{4}{4 + 1} = 0.80$$

Recall: of the real positives, how many were found. Precision: of those flagged, how many were right.

## 2 Practice

**2.1** State the preprocessing step for each: (i) removing a row where age is 250; (ii) scaling income and age to 0–1; (iii) dropping a customer ID column. [3]

---



---



---

**2.2** A best-fit line is  $y = 2.4x + 15$ . Predict  $y$  when  $x = 20$ . [2]

---



---

**2.3** A classifier has  $TP = 45$ ,  $FP = 5$ ,  $FN = 15$ ,  $TN = 135$ . Calculate (i) accuracy; (ii) precision; (iii) recall. [3]

---



---



---



---

### 3 Exam-style questions

---

3.1 Which algorithm is unsupervised? [1]

- ☐ A linear regression  
B k-nearest neighbours  
C k-means clustering  
D decision tree

3.2 What is changed during the training of a neural network? [1]

- ☐ A the number of input nodes  
B the weights of the connections  
C the training labels  
D the test data

3.3 A cancer screening test should miss as few real cases as possible. Which measure should be high? [1]

- ☐ A precision  
B recall  
C training speed  
D number of features

3.4 An app predicts whether a plant is healthy from leaf photos.

(a) Describe **three** preprocessing steps for the photos and their labels. [3]

---

---

---

(b) A developer adds “date the plant died” as a feature, and test accuracy jumps to 99%. Explain why this is a problem. [2]

---

---

(c) Explain why a convolutional neural network suits this task. [2]

---

---

(d) On 200 test photos the model gives  $TP = 50$ ,  $FP = 30$ ,  $FN = 10$ ,  $TN = 110$ , where positive means “diseased”. Calculate the precision and recall, and state which matters more to a gardener who wants to treat every sick plant. [4]

---

---

---

---

---

## Solutions

---

Each answer shows the steps, then the **result**.

### 2.1

- (i) **cleaning**
- (ii) **normalisation** (scaling)
- (iii) **feature selection**

### 2.2

- $y = 2.4 \times 20 + 15$
- **63**

### 2.3

- (i)  $\frac{45 + 135}{200} = \mathbf{0.90}$
- (ii)  $\frac{45}{45 + 5} = \mathbf{0.90}$
- (iii)  $\frac{45}{45 + 15} = \mathbf{0.75}$

### 3.1 C

- no labels: it groups items

### 3.2 B

- weights are adjusted to reduce the error

### 3.3 B

- recall counts how many real cases are found

### 3.4

(a) Any three, e.g.:

- remove blurred, duplicate or wrongly labelled photos
- resize all photos to the same size and scale pixel values to 0–1
- encode the labels as numbers, such as healthy 0, diseased 1, and balance the classes

(b)

- this is data leakage: the date of death is not known when the prediction is needed
- the model will look excellent in testing but fail in real use

(c)

- its layers scan small patches of the image
- so it finds local features such as spots and edges, wherever they appear on the leaf

(d)

- $\text{precision} = \frac{50}{50 + 30} = 0.625$
- $\text{recall} = \frac{50}{50 + 10} \approx 0.83$
- **recall** matters more
- a missed sick plant goes untreated; treating a healthy one by mistake costs little