

Exercise walk-through

4.2

Machine learning algorithms

International Baccalaureate · IB Diploma · Computer Science · HL

Computing Data · executable trace

- Trace the state changes and boundary guard.
- Run the boundary case and compare the printed output.

```
tp, fn, fp, tn = 4, 1, 10, 85
total = tp + fn + fp + tn
accuracy = (tp + tn) / total
recall = tp / (tp + fn) if tp + fn else None
print(accuracy, recall)
positives = 0
print(0 / positives if positives else "UNDEFINED")
```

Exercises 2.1·2.2

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]

Exercises 2.1·2.2

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]

2.1

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

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

2.2

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

Exercises 2.3-3.1

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

3.1 Which algorithm is unsupervised? [1]

☐

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

Exercises 2.3-3.1

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

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 Which algorithm is unsupervised? [1]

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

3.1 C

- no labels: it groups items

Exercises 3.2-3.3

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

Exercises 3.2-3.3

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

3.2 B

- weights are adjusted to reduce the error

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

3.3 B

- recall counts how many real cases are found

Exercise 3.4 (a)

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]

Exercise 3.4 (a)

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]

(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

Exercise 3.4 (b)

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

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

Exercise 3.4 (b)

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

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

(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

Exercise 3.4 (c)

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

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

[2]

Exercise 3.4 (c)

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

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

[2]

(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

Exercise 3.4 (d)

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

(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]

Exercise 3.4 (d)

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

(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]

(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