Unit 9 · Machine Learning

ANSWER KEY · FOR TEACHERS

Full marks 33

Give the marks for the points in the marking scheme; an answer in other words that says the same is right.

1 Key words

1. features
2. label
3. k nearest neighbors
4. Euclidean distance
5. scaling
6. test set
7. baseline
8. noise

Each carries 0.5 mark. **K · 4**

2 Multiple choice

i. **b.** Mountain, because two of the three are Mountain **U · 1**

With k = 3 the three neighbors vote, and Mountain has 2 votes to Hill's 1. The nearest one only decides when the votes are tied (as when k = 2 gives one each).

ii. **b.** 0.5 **A · 1**

(2600 − 100) ÷ (5100 − 100) = 2500 ÷ 5000 = 0.5: exactly halfway between the lowest and the highest.

iii. **b.** Rain doesn't change much with the belt, so it adds gaps to the distances that have nothing to do with the label. **U · 1**

A feature that doesn't change with the label is noise. It brings in neighbors that are alike in rain but not in belt.

iv. **b.** The model has learned something: it beats the simplest guess by a lot. **U · 1**

A baseline is the score you get without learning anything. 70 of 77 (91%) against 40 of 77 (52%) shows the features really help. A model that only matched the baseline would have learned nothing.

v. **c.** Recognizing which of 50 crop diseases a leaf has, from a photo **A · 1**

A photo has millions of numbers (its pixels) and nobody can choose the useful features by hand: that is what neural networks are for, given enough labeled photos. A fare and a bill follow fixed rules; the belts need only a few numbers.

3 Calculations

ANSWER **A · 3**

The rows must add up to 30: 7 + 17 = 24 in the first two rows, so the Terai row has 6, all said Terai: **? = 6.**
Right answers are on the diagonal: 5 + 13 + 6 = 24. Accuracy = 24 ÷ 30 × 100% = **80%.**
Its most common mistake is calling Hill districts Mountain (3 times).

MARKING SCHEME

✓ ? = 30 − 24 = 6.
✓ Accuracy = (5 + 13 + 6) ÷ 30 × 100% = 80%.
✓ The commonest mistake: Hill districts called Mountain, 3 times.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

4 What does it print?

MARKING SCHEME **A · 3**

- ✓ The three distances: 0.5 (Hill), 0.32 (Mountain), 0.4 (Terai).
- ✓ The tuples sorted by their first item, nearest first: `[(0.32, 'Mountain'), (0.4, 'Terai'), (0.5, 'Hill')]`.
- ✓ `k = 1: Mountain`, the nearest example's belt.

5 Trace the program

ANSWER **A · 4**

`d`	`belt`	`votes`	**Output**
(before)		{}	
0.1	Hill	{'Hill': 1}	
0.17	Terai	{'Hill': 1, 'Terai': 1}	
0.24	Terai	{'Hill': 1, 'Terai': 2}	
0.24	Hill	{'Hill': 2, 'Terai': 2}	
0.25	Terai	{'Hill': 2, 'Terai': 3}	
			{'Hill': 2, 'Terai': 3} Terai

MARKING SCHEME

- ✓ `d` and `belt` taken from each tuple in order, all five (k = 5).
- ✓ `votes` growing by one each round: Hill 1, then Terai 1, 2, then Hill 2, then Terai 3.
- ✓ `best` starts as Hill (the nearest), and changes to Terai, which has more votes.
- ✓ Output `{'Hill': 2, 'Terai': 3} Terai`, once, at the end.

6 Find and fix the bug

MARKING SCHEME **A · 3**

- ✓ Each district is among its own examples: its nearest is itself, at distance 0, so it is always right.
- ✓ Line 6: the examples passed to `predict()` include `rows[i]`.
- ✓ `others = rows[:i] + rows[i + 1:]`, then `predict(rows[i], others, features, 1)`.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking



7 Write a program

ANSWER A · 4

```
labels = ["Hill", "Terai", "Hill", "Mountain", "Hill", "Hill", "Terai", "Hill",
          "Hill", "Terai", "Mountain", "Hill", "Hill", "Terai", "Hill", "Hill",
          "Mountain", "Hill", "Terai", "Hill"]
counts = {}
for belt in labels:
    counts[belt] = counts.get(belt, 0) + 1
best = ""
for belt in counts:
    if best == "" or counts[belt] > counts[best]:
        best = belt
print(f"Always {best}: {round(100 * counts[best] / len(labels))}%")
```

```
OUTPUT
Always Hill: 60%
```

Other right ways: `max(counts, key=counts.get)` for the commonest, or `labels.count(belt)` for each belt.

MARKING SCHEME

- ✓ A tally of each belt in a dictionary (`counts.get(belt, 0) + 1`, or `labels.count()`).
- ✓ The commonest belt found (a loop keeping the biggest so far, or `max`).
- ✓ The percentage worked out: its count ÷ the number of labels × 100.
- ✓ Prints the belt and the percentage, `Always Hill: 60%`.

8 Short answers

a. **MARKING SCHEME** U · 2

- ✓ The pocket money's gaps are far bigger (thousands, against at most 50 cm), so it decides the distance almost alone and height hardly counts.
- ✓ Scale both features to the same range first, such as min–max scaling, (value − lowest) ÷ (highest − lowest), from 0 to 1.

b. **MARKING SCHEME** A · 3

- ✓ A first question with its reason, such as: who is in the data and who is missing (village farmers, women), since few examples mean unfair answers for them.
- ✓ A second, such as: do the old decisions (the labels) carry unfairness the model would learn as right?
- ✓ A third, such as: could a feature like a home address stand in for caste or wealth; is 20-year-old data still true; did customers agree, and is it kept safe?

9 Think

ANSWER HA · 2

No. Calling Bhaktapur a Terai district harms nobody; calling a sick person "healthy" could cost a life. For the X-ray model you must ask which mistakes it makes (a confusion table: how many sick people does it call healthy?), test it on patients like the ones it will meet (in Nepal, not only abroad), and keep a doctor checking its answers. How good a model must be depends on what its mistakes cost, not on one accuracy number.

MARKING SCHEME

- ✓ No: its mistakes cost far more (a sick person sent home), so one accuracy number isn't enough.
- ✓ Check which mistakes it makes (a confusion table), test it on patients like those it will meet, and keep a doctor checking.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

10 Activity: Seventy-seven districts

(not marked)

WHAT STUDENTS SHOULD FIND

With the first split: 1 neighbor gets 11 of 15 (Bhaktapur and Kathmandu called Terai, Gorkha Mountain, Nawalpur Hill), 3 neighbors 14 of 15 (only Bhaktapur wrong), 5 neighbors 12 of 15. With new splits the score jumps about, from about 11 to 15 of 15, because each split tests different districts, some easy (flat Terai districts) and some hard (Gorkha, Myagdi, Bhaktapur, Sindhupalchok). A test set of 15 gives only a rough score, which is why leave-one-out testing is better for 77 districts.

Marks by level	Knowledge	Understanding	Application	Higher ability	**Total**
Marks	4	5	22	2	**33**

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking