

ES 335: Machine Learning

Quiz 1 — Foundations, Trees & Nearest Neighbours — ANSWER KEY

Date: 2026-08-28

Duration: 50 minutes

Maximum marks: 30

SET B

Instructions

- Write your name and roll number before starting.
- This paper has 12 questions in 3 parts, for a total of 30 marks.
- All logarithms are base 2 unless stated otherwise.
- Calculators are permitted; devices with network access are not.

MCQ Answer Grid

Q	1	2	3	4	5	6
Ans	A	B	D	C	B	AC

Part A — Multiple Choice

12 marks

Choose the single best answer unless a question says otherwise.

Q1. Increasing k in k -NN typically has which effect?

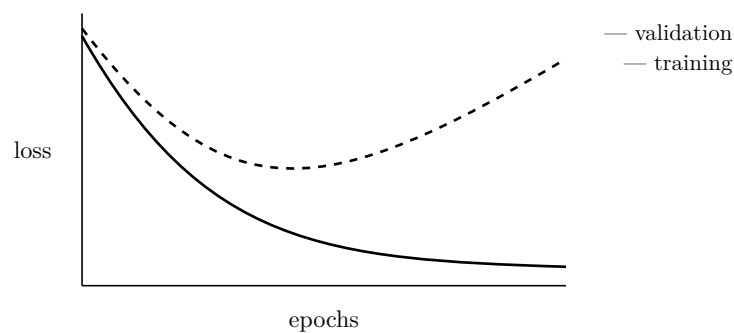
[2 marks]

- (A) Increases bias, decreases variance ✓
(B) Decreases bias, increases variance
(C) Increases both bias and variance
(D) None of the above

Explanation. Larger neighbourhoods average over more points: smoother (higher-bias), more stable (lower-variance) decision boundaries.

Q2. The curves below were logged during training.

[2 marks]



What is the most appropriate response?

- (A) Lower the learning rate — training has diverged
- (B) Stop earlier or regularize — the model is overfitting ✓**
- (C) Train longer — the model is underfitting
- (D) Increase model capacity — both losses are too high

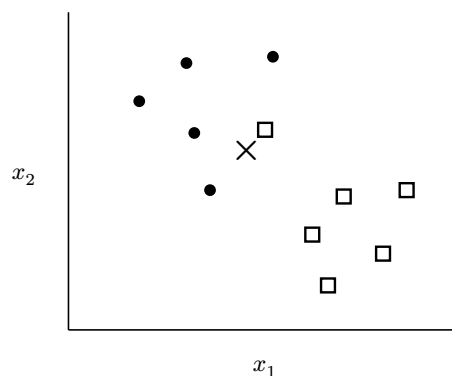
Explanation. Training loss keeps falling while validation loss rises after its minimum: the classic overfitting signature.

Q3. For a split that sends all positives left and all negatives right, the weighted entropy [2 marks] of the children is:

- (A) $\log_2 3$
- (B) 0.5
- (C) 1
- (D) 0 ✓**

Q4. The figure shows training points from two classes and a query point $\times$.

[2 marks]



Using Euclidean distance, how does the query point get classified by 1-NN and by 3-NN respectively?

- (A) 1-NN: class A (●), 3-NN: tie
- (B) 1-NN: class A (●), 3-NN: class A (●)
- (C) 1-NN: class B (□), 3-NN: class A (●) ✓**
- (D) 1-NN: class B (□), 3-NN: class B (□)

Explanation. The single nearest neighbour is the stray $\square$ just above the query; the 3-neighbourhood contains two $\bullet$ and one $\square$.

Q5. A binary classifier produces the confusion matrix below.

[2 marks]

	Predicted +	Predicted −
Actual +	40	10
Actual −	20	30

What is its precision?

- (A) 40/50
(B) 40/60 ✓
 (C) 40/100
 (D) 70/100

Explanation. Precision = $TP / (TP + FP) = 40 / (40 + 20)$. 40/50 is the recall and 70/100 the accuracy.

Q6. Which of the following objectives are convex in their parameters? (Select all that apply.) [2 marks]

- (A) Logistic regression negative log-likelihood ✓**
 (B) The k -means clustering objective (jointly in assignments and centroids)
(C) Linear regression with squared loss ✓
 (D) A two-layer neural network with squared loss

Explanation. Both linear-regression MSE and logistic NLL are convex; k -means and neural-network losses are not.

Part B — Fill in the Blanks

4 marks

Q7. Compared with L2, L1 regularization tends to produce weight vectors that are [1 mark]
sparse (many exact zeros).

Q8. For linear regression in matrix form, the least-squares solution is $\hat{\theta} =$ [2 marks]
 $(X^T X)^{-1} X^T \underline{y}$.

Q9. Computing inner products in a high-dimensional feature space without ever constructing the features is known as the kernel trick. [1 mark]

Part C — Long Answers

14 marks

Answer in the space provided; show your working.

Q10. The table records eight days of data for predicting whether a match is *Played*. [5 marks]

Day	Outlook	Windy	Played
1	Sunny	No	Yes
2	Sunny	Yes	No
3	Overcast	No	Yes
4	Overcast	Yes	Yes
5	Rainy	No	Yes
6	Rainy	Yes	No
7	Sunny	No	No
8	Rainy	No	Yes

Compute (i) the entropy of *Played*, and (ii) the information gain of splitting on *Windy*. Which single-feature split would a decision tree prefer if the gain for *Outlook* is 0.311?

Model answer. $H(\text{Played}) = H(5/8) = -\frac{5}{8} \log \frac{5}{8} - \frac{3}{8} \log \frac{3}{8} \approx 0.954$. $\text{Windy} = \text{Yes}$: (1 Yes, 2 No), $H \approx 0.918$; $\text{Windy} = \text{No}$: (4 Yes, 1 No), $H \approx 0.722$. Weighted: $\frac{3}{8} \cdot 0.918 + \frac{5}{8} \cdot 0.722 \approx 0.795$. $\text{IG}(\text{Windy}) \approx 0.954 - 0.795 = 0.159 < 0.311$, so the tree splits on *Outlook*.

Rubric. +1 $H(\text{Played})$; +2 conditional entropies; +1 gain; +1 correct comparison and conclusion.

Q11. Starting from $L(\theta) = \|y - X\theta\|_2^2$, derive the normal equation. State the condition [5 marks] under which the solution is unique, and give one practical remedy when it is not.

Model answer. $\nabla_{\theta} L = -2X^{\top}(y - X\theta) = 0 \Rightarrow X^{\top}X\theta = X^{\top}y$. Unique iff X has full column rank. Remedy: add ridge penalty λI (or remove collinear features).

Rubric. +2 gradient $-2X^{\top}(y - X\theta)$; +1 setting to zero $\rightarrow X^{\top}X\theta = X^{\top}y$; +1 uniqueness $\Leftrightarrow X^{\top}X$ invertible (full column rank); +1 remedy: ridge / drop collinear features.

Q12. Your colleague reports 99.2% accuracy on a fraud-detection dataset in which 0.8% of [4 marks] transactions are fraudulent. Explain why this number is not impressive, and name two evaluation metrics (or tools) better suited to this setting, justifying each in one line.

Model answer. A constant “not fraud” classifier already achieves 99.2% — accuracy is dominated by the majority class. Better: recall (catch rate of actual fraud) with precision or F1 / PR-AUC, which focus on the rare positive class rather than overall agreement.

Rubric. +2 base-rate argument (predicting “not fraud” gives 99.2%); +1 per metric with justification (max 2): precision/recall, F1, PR-AUC, recall at fixed FPR, cost-sensitive evaluation.

— End of paper —