

ARTIFICIAL INTELLIGENCE (Subject Code: 843) — CLASS XI

Unit 6: Machine Learning Algorithms

Application-Based Questions | Scenarios, Numericals and Case Studies

Instructions: Read each situation carefully. Show all steps for numerical questions. Marks are given against each question. Max. Marks: 50

Section A: Short Application Questions (2 or 3 marks each)

Q1. Identify the type of machine learning (Supervised, Unsupervised or Reinforcement) used in each situation. Give one reason for each. **[3]**

- (a) A model is trained on thousands of chest X-ray images, each labelled 'Pneumonia' or 'Normal', to diagnose new X-rays.
- (b) An online store groups its customers by their buying behaviour. No groups or labels are given to the system in advance.
- (c) A robot vacuum cleaner gets points for cleaning new areas and loses points when it bumps into furniture. Over time it learns the best path.

Q2. Identify the type of correlation (positive, negative or zero) between the pairs of variables below. For which pair would Pearson's r be closest to 0? **[2]**

- (a) Outside temperature and number of ice-creams sold in a day.
- (b) Speed of a car and the time taken to cover a fixed distance.
- (c) Shoe size of a student and the marks scored in an AI exam.

Q3. Riya is building a linear regression model to predict flat prices in her city from the area of the flat. Her dataset has 200 ordinary flats and 3 luxury penthouses whose prices are 40 times higher than the rest. Her predictions for ordinary flats turn out to be very inaccurate. **[3]**

- (a) What is the likely cause of the problem?
- (b) Which disadvantage of linear regression does this show?
- (c) Suggest one way Riya can improve her model.

Q4. A farmer wants to predict crop yield (in quintals per acre). **[3]**

- (a) In Model 1 he uses only the amount of rainfall. In Model 2 he uses rainfall, fertiliser used and average temperature. Name the type of linear regression used in each model.
- (b) Identify the dependent variable and the independent variable(s) in Model 2.

Q5. A supermarket wants to divide its customers into 3 groups based on monthly spending and number of visits, so that it can send a different offer to each group. The customers have no labels. **[3]**

- (a) Which machine learning algorithm would you recommend? What value of K would you use?
- (b) State two limitations of this algorithm that the supermarket should keep in mind.

Q6. A music app wants to group its songs by similarity, but the team does not know how many groups to make. They would also like to see how smaller groups combine into bigger ones. **[2]**

- (a) Which clustering method is most suitable? Give a reason.
- (b) What is the name of the tree-like diagram produced by this method?

Section B: Numerical Application Questions (4 marks each)

Q7. A teacher recorded the hours studied per week and the marks scored by 5 students. Calculate Pearson's correlation coefficient (r) and interpret the result. [4]

Student	S1	S2	S3	S4	S5
Hours studied (x)	2	4	6	8	10
Marks scored (y)	50	60	65	80	90

$$\text{Use: } r = [n\sum xy - (\sum x)(\sum y)] / \sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$

Q8. A company found the following regression line between its advertising spend and sales: Sales (units) = 120 + 8 × Advertising spend (in ₹ thousand) [4]

- (a) Predict the sales when the advertising spend is ₹15 thousand.
- (b) What do the values 120 and 8 represent in this equation?
- (c) If the actual sales for ₹15 thousand of advertising were 250 units, find the error (residual).

Q9. Find the line of best fit ($y = a + bx$) for the data below using the least squares method, and predict y when $x = 6$. [4]

x	1	2	3	4	5
y	2	4	5	4	5

$$\text{Use: } b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \text{ and } a = \bar{y} - b \cdot \bar{x}$$

Q10. A clothing store uses the KNN algorithm to recommend a T-shirt size from a customer's height and weight. Using $K = 3$ and Euclidean distance, predict the T-shirt size for a new customer with height 162 cm and weight 58 kg. Show the distance calculations. [4]

Customer	Height (cm)	Weight (kg)	T-shirt Size
C1	150	45	Medium
C2	160	55	Medium
C3	170	70	Large
C4	175	75	Large
C5	155	48	Medium
C6	180	80	Large
C7	165	62	Large

$$\text{Euclidean distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Q11. Apply one iteration of the K-Means algorithm with $K = 2$ to the six points below. Take A and D as the initial centroids. (i) Assign each point to its nearest centroid. (ii) Calculate the new centroids. [4]

Point	A	B	C	D	E	F
(x, y)	(2, 3)	(3, 4)	(4, 3)	(8, 8)	(9, 7)	(8, 9)

Section C: Case-Based Questions (5 marks each)

Q12. Read the case and answer the questions that follow. [5]

City Care Hospital has collected records of 10,000 patients. Each record contains age, BMI, blood pressure, daily exercise time and a label showing whether the patient is Diabetic or Non-diabetic. The hospital wants

to: (1) predict whether a new patient is likely to be diabetic, (2) predict a patient's exact blood-sugar level, and (3) form groups of patients with similar lifestyles to run wellness camps.

(a) Which type of supervised learning task is goal (1)? Name one suitable algorithm from this unit.

(b) Which type of task is goal (2)? Why is this task different from goal (1)?

(c) Which ML technique suits goal (3)? Is it supervised or unsupervised?

(d) The model for goal (1) gives 99% accuracy on the training data but only 70% on new patients. What is this problem called?

(e) Only 5% of the patients in the data are diabetic. What type of classification problem is this?

Q13. Read the case and answer the questions that follow. [5]

A Smart City project wants to use AI to reduce traffic problems. The team plans to: use cameras to recognise each vehicle as car, bus, two-wheeler or truck; predict the travel time between two places from distance and time of day; find areas where many accidents happen close together on the map; and build a traffic-signal system that learns by itself to reduce waiting time, getting a reward whenever queues become shorter.

(a) Recognising vehicles as car, bus, two-wheeler or truck – which type of classification is this?

(b) Which ML technique should be used to predict travel time? Identify its dependent variable.

(c) Which clustering method is best for finding accident hotspots of irregular shape? Why?

(d) Which type of machine learning is used by the self-learning traffic signal? Explain briefly.

(e) Give one way the team could check whether 'distance' and 'travel time' are suitable for linear regression.

MARKING SCHEME / SUGGESTED ANSWERS

Q1.

- (a) Supervised learning – the model learns from labelled data (each X-ray is labelled Pneumonia/Normal). (1)
- (b) Unsupervised learning – the data is unlabelled and the system finds groups (clusters) by itself. (1)
- (c) Reinforcement learning – the robot learns by trial and error using rewards and penalties. (1)

Q2.

- (a) Positive correlation – both increase together. (b) Negative correlation – as speed increases, time decreases. (c) Zero correlation – no relationship. ($\frac{1}{2} \times 3$)
- Pair (c) would have Pearson's r closest to 0 because the variables are not related. ($\frac{1}{2}$)

Q3.

- (a) The 3 luxury penthouses are outliers – extreme values that pull the regression line towards them. (1)
- (b) Linear regression is sensitive to outliers. (1)
- (c) Identify and remove/handle the outliers (or build a separate model for luxury flats) before fitting the line. (1)

Q4.

- (a) Model 1 – Simple linear regression (one independent variable). Model 2 – Multiple linear regression (more than one independent variable). (1 + 1)
- (b) Dependent variable: crop yield. Independent variables: rainfall, fertiliser used and average temperature. (1)

Q5.

- (a) K-Means clustering (unsupervised, partitioning/centroid-based) with $K = 3$. (1)
- (b) Any two: results can change with the initial placement of centroids; the number of clusters must be known beforehand; outliers can distort clusters; assumes clusters are spherical; may reach a sub-optimal solution. (1 + 1)

Q6.

- (a) Hierarchical clustering – it does not need the number of clusters in advance, and shows how clusters merge level by level. (1)
- (b) Dendrogram. (1)

Q7.

$$\begin{aligned}n &= 5, \Sigma x = 30, \Sigma y = 345, \Sigma xy = 2270, \Sigma x^2 = 220, \Sigma y^2 = 24825 \quad (1) \\n\Sigma xy - (\Sigma x)(\Sigma y) &= 5 \times 2270 - 30 \times 345 = 11350 - 10350 = 1000 \\n\Sigma x^2 - (\Sigma x)^2 &= 5 \times 220 - 30^2 = 1100 - 900 = 200 \\n\Sigma y^2 - (\Sigma y)^2 &= 5 \times 24825 - 345^2 = 124125 - 119025 = 5100 \quad (1) \\r &= 1000 / \sqrt{(200 \times 5100)} = 1000 / \sqrt{1020000} = 1000 / 1009.95 \approx 0.99 \quad (1)\end{aligned}$$

Interpretation: r is very close to +1, so there is a strong positive correlation – students who study more hours tend to score higher marks. (1)

Q8.

- (a) Sales = $120 + 8 \times 15 = 120 + 120 = 240$ units. (1)
- (b) 120 is the intercept (a) – expected sales when advertising spend is zero. 8 is the slope (b) – sales increase by 8 units for every ₹1 thousand increase in advertising. (2)
- (c) Error e = Actual – Predicted = $250 - 240 = 10$ units. (1)

Q9.

$$\bar{x} = 15/5 = 3, \bar{y} = 20/5 = 4 \text{ (}\frac{1}{2}\text{)}$$

x	y	x - $\bar{x}$	y - $\bar{y}$	(x - $\bar{x}$)(y - $\bar{y}$)	(x - $\bar{x}$) ²
1	2	-2	-2	4	4
2	4	-1	0	0	1
3	5	0	1	0	0
4	4	1	0	0	1
5	5	2	1	2	4
			Total	6	10

$$b = 6 / 10 = 0.6 \text{ (1)}$$

$$a = \bar{y} - b \cdot \bar{x} = 4 - 0.6 \times 3 = 4 - 1.8 = 2.2 \text{ (1)}$$

$$\text{Line of best fit: } y = 2.2 + 0.6x \text{ (}\frac{1}{2}\text{)}$$

$$\text{For } x = 6: y = 2.2 + 0.6 \times 6 = 2.2 + 3.6 = 5.8 \text{ (1)}$$

Q10.

Distance from new customer (162, 58): (2)

Customer	Calculation	Distance	Size
C1	$\sqrt{[(150-162)^2 + (45-58)^2]} = \sqrt{144 + 169} = \sqrt{313}$	17.69	Medium
C2	$\sqrt{[(160-162)^2 + (55-58)^2]} = \sqrt{4 + 9} = \sqrt{13}$	3.61	Medium
C3	$\sqrt{[(170-162)^2 + (70-58)^2]} = \sqrt{64 + 144} = \sqrt{208}$	14.42	Large
C4	$\sqrt{[(175-162)^2 + (75-58)^2]} = \sqrt{169 + 289} = \sqrt{458}$	21.40	Large
C5	$\sqrt{[(155-162)^2 + (48-58)^2]} = \sqrt{49 + 100} = \sqrt{149}$	12.21	Medium
C6	$\sqrt{[(180-162)^2 + (80-58)^2]} = \sqrt{324 + 484} = \sqrt{808}$	28.43	Large
C7	$\sqrt{[(165-162)^2 + (62-58)^2]} = \sqrt{9 + 16} = \sqrt{25}$	5.00	Large

3 nearest neighbours: C2 (3.61, Medium), C7 (5.00, Large), C5 (12.21, Medium). (1)

Majority vote: Medium = 2, Large = 1. Predicted T-shirt size = MEDIUM. (1)

Q11.

Initial centroids: $c_1 = A(2, 3)$, $c_2 = D(8, 8)$. Distances: (2)

Point	Distance to c_1 (2, 3)	Distance to c_2 (8, 8)	Cluster
A (2, 3)	0	$\sqrt{61} = 7.81$	1
B (3, 4)	$\sqrt{2} = 1.41$	$\sqrt{41} = 6.40$	1
C (4, 3)	$\sqrt{4} = 2.00$	$\sqrt{41} = 6.40$	1
D (8, 8)	$\sqrt{61} = 7.81$	0	2
E (9, 7)	$\sqrt{65} = 8.06$	$\sqrt{2} = 1.41$	2
F (8, 9)	$\sqrt{72} = 8.49$	$\sqrt{1} = 1.00$	2

Cluster 1 = {A, B, C}; Cluster 2 = {D, E, F}

$$\text{New } c_1 = ((2+3+4)/3, (3+4+3)/3) = (3, 3.33) \text{ (1)}$$

$$\text{New } c_2 = ((8+9+8)/3, (8+7+9)/3) = (8.33, 8) \text{ (1)}$$

Q12.

(a) Classification (binary – Diabetic / Non-diabetic). Suitable algorithm: K-Nearest Neighbours (or logistic regression / decision tree). (1)

- (b) Regression – the output (blood-sugar level) is a continuous numerical value, while goal (1) predicts a discrete class label. (1)
- (c) Clustering (e.g., K-Means); it is unsupervised learning because no group labels are given. (1)
- (d) Overfitting – the model has become too specialised on the training data and does not generalise to new data. (1)
- (e) Imbalanced classification – one class (diabetic) is a small minority. (1)

Q13.

- (a) Multi-class classification – more than two classes, each vehicle belongs to one class. (1)
- (b) Regression (multiple linear regression, since distance and time of day are both used). Dependent variable: travel time. (1)
- (c) Density-based clustering – it connects dense areas into clusters of arbitrary shape, separated by sparse areas. (1)
- (d) Reinforcement learning – the signal (agent) interacts with traffic (environment) and learns by trial and error to maximise rewards (shorter queues). (1)
- (e) Draw a scatter plot and/or calculate Pearson's r between distance and travel time; a strong linear correlation (r close to $+1$ or -1) with no outliers shows linear regression is suitable. (1)