IAE ACADEMY INSTITUTE FOR ENGINEERS

CORE JOBS · INTERVIEW QUESTIONS

IAE Academy free study material · Updated 28 October 2026

Machine Learning Interview Questions

INSIDE

- 25 questions with short answers
- 5 numericals with the working
- Grouped by topic

Companies do not publish their interview questions; these are ours.

Online, with links to the notes: **iaeacademy.in/study/machine-learning/interview-questions**

FREE COUNSELLING

Preparing for GATE or a core job? Talk to IAE Academy.

Or call **+91 84204 36524**

Every day, 10:30 am to 7:30 pm

Chat on WhatsApp

Free. Not for sale. Pass it on.

Simple, multiple and ridge regression

1. What does ordinary least squares minimise, and what are the normal equations?

Ordinary least squares minimises the sum of squared residuals. Setting the gradient to zero gives the normal equations $X^T X\, w = X^T y$, so $w = (X^T X)^{-1} X^T y$, provided $X^T X$ is invertible, which needs the feature columns to be linearly independent.

2. Why does ridge regression help when the features are highly correlated?

Ridge adds λI to $X^T X$, giving $w = (X^T X + \lambda I)^{-1} X^T y$. With correlated features $X^T X$ is nearly singular and plain least-squares weights swing wildly. The penalty keeps the matrix well conditioned and shrinks the weights, trading a little bias for a large drop in variance.

3. Why are features usually scaled before running gradient descent on a regression problem?

Features on very different scales make the loss surface a long, narrow valley, so gradient descent needs a small learning rate and zigzags slowly. Standardising the features makes the contours rounder and speeds convergence. It also matters for ridge, because the penalty treats every weight alike.

4. A line $y = w_1 x + w_0$ is fitted by least squares to the points (1, 2), (2, 4) and (3, 5). Find the slope.

The least-squares slope is S_{xy}/S_{xx}. With means $\bar{x} = 2$ and $\bar{y} = 11/3$, the sum of cross deviations is $S_{xy} = 3$ and the sum of squared x deviations is $S_{xx} = 2$, so the slope is $3/2 = 1.5$.

WORKING

1. Means: $\bar{x} = 2$, $\bar{y} = 11/3 \approx 3.667$.
2. $S_{xy} = (-1)(2 - 3.667) + 0 + (1)(5 - 3.667) = 3$.
3. $S_{xx} = 1 + 0 + 1 = 2$.
4. Slope $= S_{xy}/S_{xx} = 3/2 = 1.5$.

Logistic regression, k-nearest neighbour and naive Bayes

5. How does logistic regression model a class probability, and what loss is it trained on?

It models $P(y = 1 \mid x) = \sigma(w^T x + b)$ with $\sigma(z) = 1/(1 + e^{-z})$, and is trained by minimising cross-entropy, usually with gradient descent. Its decision boundary $w^T x + b = 0$ is linear in the features.

6. A logistic regression model has weights 2 and -1 and a bias of -1. For the input (1, 2), find the predicted probability of class 1.

The linear score is $z = w^T x + b = 2(1) + (-1)(2) - 1 = -1$. The sigmoid turns this score into the probability of class 1: $P(y = 1) = \sigma(-1) = 1/(1 + e) \approx 0.27$. A negative score always gives a probability below 0.5.

WORKING

1. $z = w^T x + b = 2(1) + (-1)(2) - 1 = -1$.
2. $P(y = 1) = \sigma(-1) = 1/(1 + e^{1})$.
3. $e \approx 2.718$, so $P = 1/3.718 \approx 0.27$.

7. What is the difference between logistic regression and k-nearest neighbour as classifiers?

Logistic regression is parametric: training learns a weight vector, prediction is fast and the boundary is linear. k-NN is non-parametric: it only stores the data and predicts from distances to stored points. Its boundary is

flexible, but prediction is slow and it is sensitive to feature scale.

8. Why is naive Bayes called naive, and what does the assumption buy?

It assumes the features are conditionally independent given the class, so $P(x \mid c) = \prod_j P(x_j \mid c)$. This needs only one distribution per feature and class (simple counts for discrete features) instead of a joint distribution over all features, so training is fast and works with little data, even when the assumption holds only roughly.

LDA, SVMs and decision trees

9. What does linear discriminant analysis assume about the classes?

LDA models each class as Gaussian with its own mean but one covariance matrix shared by all classes. This makes the log-odds between classes linear in x, so the boundary is a hyperplane. If the true class covariances differ a lot, the linear boundary fits poorly.

10. What is the margin of a support vector machine, and what does it maximise?

The margin is the gap between the two classes on either side of the separating hyperplane. With labels ± 1 and constraints $y_i(w^T x_i + b) \geq 1$, its width is $2/\|w\|$, so maximising it means minimising $\|w\|$. Only the support vectors on the margin fix the plane.

11. What does the parameter C control in a soft-margin SVM, and why does it matter?

C sets the penalty for margin violations. A large C punishes violations heavily, giving a narrow margin that fits the training data closely and may overfit. A small C tolerates more violations, giving a wider margin and a smoother boundary that may underfit.

12. A node holds 8 samples, 4 of each class. A split sends 4 samples to the left child (3 of class A, 1 of class B) and 4 to the right child (1 of class A, 3 of class B). Find the information gain of the split, using entropy in bits.

The parent node, with 4 samples of each class, has entropy 1 bit. Each child holds a 3 to 1 split, with entropy $H(3/4, 1/4) \approx 0.811$ bits. Both children hold half the samples, so the weighted child entropy is 0.811 and the information gain is $1 - 0.811 \approx 0.19$ bits.

WORKING

1. Parent entropy: $H = 1$ bit (4 and 4 out of 8).
2. Each child: $H(3/4, 1/4) = -0.75 \log_2 0.75 - 0.25 \log_2 0.25 \approx 0.811$.
3. Weighted child entropy $= 0.5(0.811) + 0.5(0.811) = 0.811$.
4. Gain $= 1 - 0.811 \approx 0.19$ bits.

13. Why does a deep decision tree overfit, and how does pruning help?

A tree grown until every leaf is pure memorises noise, so it has low bias but high variance. Pruning, or limiting depth or minimum leaf size, removes splits that reduce impurity very little. The simpler tree usually generalises better to unseen data.

Bias–variance trade-off and cross-validation

14. What are the three parts of expected squared test error in the bias-variance decomposition?

Expected squared error equals bias squared plus variance plus irreducible noise. Bias is the error from wrong model assumptions. Variance is how much the fitted model changes from one training set to another. Noise is randomness in the target, which no model can remove.

15. Why is training error a poor estimate of test error?

The model was fitted to those same points, so training error is optimistic and keeps falling as the model grows more complex, even when it is memorising noise. Error on held-out data rises again past some complexity, so only held-out or cross-validated error reveals overfitting.

16. What is the difference between k-fold cross-validation and leave-one-out cross-validation?

In k-fold cross-validation the data is split into k parts, each used once for validation, so the model is trained k times. Leave-one-out is the case $k = n$: n fits, each on $n - 1$ points. It uses almost all data per fit but costs far more computation.

17. What are precision and recall, and why are they used instead of accuracy on imbalanced data?

Precision is $TP/(TP + FP)$, the share of predicted positives that are truly positive. Recall is $TP/(TP + FN)$, the share of actual positives that were found. Accuracy can look high on imbalanced classes while missing the rare class, so precision, recall and their harmonic mean F1 are reported.

Multi-layer perceptrons and feed-forward neural networks

18. Why must the activation functions in a neural network be non-linear?

Without a non-linearity each layer is a linear map, and a composition of linear maps is itself one linear map, so any depth collapses to a single linear model. A non-linear activation such as ReLU or sigmoid lets the network represent non-linear functions.

19. What is the perceptron update rule, and when does it converge?

With labels $y = \pm 1$, the perceptron predicts $\operatorname{sign}(w^T x + b)$ and, on a misclassified point, updates $w \leftarrow w + \eta\, y\, x$ and $b \leftarrow b + \eta\, y$. It converges in a finite number of steps if and only if the classes are linearly separable. XOR is the standard example where it never settles.

20. A fully connected feed-forward network has 4 inputs, two hidden layers of 5 and 3 neurons, and 1 output neuron. Every neuron has a bias. How many trainable parameters does it have?

Each layer has one weight per input for every neuron, plus one bias per neuron. The layers give $4 \times 5 + 5 = 25$, $5 \times 3 + 3 = 18$ and $3 \times 1 + 1 = 4$, so the network has $25 + 18 + 4 = 47$ trainable parameters.

WORKING

1. Input to first hidden layer: $4 \times 5 + 5 = 25$.
2. First to second hidden layer: $5 \times 3 + 3 = 18$.
3. Second hidden layer to output: $3 \times 1 + 1 = 4$.
4. Total $= 25 + 18 + 4 = 47$.

21. Why is backpropagation efficient for computing gradients in a neural network?

Backpropagation applies the chain rule layer by layer from the output backwards, reusing each layer's gradient for the layer before it. It gives the loss gradient for every weight in one backward pass, costing about as much as a forward pass, instead of one pass per weight.

Clustering

22. What is the difference between k-means and k-medoids?

k-means uses the mean of a cluster as its centre, which need not be a data point and is pulled by outliers. k-medoids uses an actual data point as the centre, so it is more robust to outliers and works with any distance measure, but it costs more to compute.

23. How does the k-means algorithm work, and what does it minimise?

k-means alternates two steps: assign each point to its nearest centre, then move each centre to the mean of its points. Neither step can increase the within-cluster sum of squared distances, so it converges, but only to a local minimum, so the result depends on the starting centres.

Dimensionality reduction and principal component analysis

24. What does principal component analysis find, and how is it computed?

PCA finds orthogonal directions of maximum variance in centred data. They are the eigenvectors of the covariance matrix, and each eigenvalue is the variance along its direction. Projecting onto the top k eigenvectors reduces the dimension while keeping as much variance as possible.

25. The covariance matrix of centred two-dimensional data has eigenvalues 6 and 2. What percentage of the total variance does the first principal component explain?

In PCA each eigenvalue of the covariance matrix is the variance along its principal component, and the total variance is their sum, $6 + 2 = 8$. The first component explains $6/8 = 0.75$ of it, which is 75% of the total variance.

> **WORKING**
>
> 1. Total variance $=$ sum of eigenvalues $= 6 + 2 = 8$.
> 2. Fraction explained $= 6/8 = 0.75$.
> 3. As a percentage, 75%.

FROM IAE ACADEMY

Preparing for GATE or a core job?

IAE Academy teaches GATE in seven papers in Ballygunge, Kolkata, and online. Counselling is free: ask us about batches, fees or your preparation.

FREE COUNSELLING

Chat on WhatsApp

Call +91 84204 36524

Open every day, 10:30 am to 7:30 pm.

Scan to chat on WhatsApp

GATE COURSES

Classroom batches in Ballygunge and online batches. Fees and details are on the course pages.

Scan: courses

See the courses

FREE ON IAEACADEMY.IN

Notes and solved previous-year questions for Machine Learning, topic by topic.

Scan: the notes

Open the notes

Found this useful? Send it to a friend preparing for core-job interviews.

Send to a friend on WhatsApp

IAE Academy free study material. Companies do not publish their interview questions; these are ours. Prepared and published by IAE Academy, Kolkata. Free to read and free to share, unaltered and with attribution.

IAE ACADEMY INSTITUTE FOR ENGINEERS

GATE coaching in Kolkata since 2012 · EC, EE, IN, CS and IT, DA, BT and CE · in the classroom and online

227A Rash Behari Ave, Ballygunge, Kolkata 700019

+91 84204 36524 · iaeacademy.in · Updated 28 October 2026

Get directions