Machine Learning - 100 Questions

Pages 12+ Part 3

Nova IMS
Generated: January 16, 2026

LEVEL 1: RANDOM FOREST BASICS

Q1. What type of ensemble method is Random Forest?

A) Boosting

B) Bagging (Bootstrap Aggregating)

C) Stacking

D) Blending

Q2. What is bootstrap sampling in Random Forest?

A) Sampling without replacement

B) Sampling with replacement to create datasets of same size

C) Splitting data sequentially

D) Removing outliers

Q3. What percentage of data does each bootstrap sample typically contain?

A) 100% unique samples

B) ~63% of original data (~37% out-of-bag)

C) 50% of data

D) 10% of data

Q4. What are the two sources of randomness in Random Forest?

A) Random labels and features

B) Bootstrap sampling of data AND random feature subsampling at splits

C) Random algorithms and parameters

D) Random data and models

Q5. How many features does each split consider in Random Forest?

A) All features

B) A random subset (typically sqrt(n) for classification, n/3 for regression)

C) Only one feature

D) Exactly 10 features

Q6. Why does Random Forest use feature subsampling at splits?

A) Speeds up training

B) Decorrelates trees to reduce variance

C) Reduces memory

D) Improves interpretability

Q7. How does Random Forest make predictions for classification?

A) Uses the first tree only

B) Majority vote across all trees

C) Uses the best tree

D) Random selection

Q8. How does Random Forest make predictions for regression?

A) Uses median of tree predictions

B) Average of all tree predictions

C) Uses maximum prediction

D) Uses minimum prediction

Q9. What is Out-of-Bag (OOB) error?

A) Error on training set

B) Validation error using samples not in bootstrap sample for each tree

C) Error on test set

D) Training error

Q10. What is an advantage of OOB error?

A) More accurate than all methods

B) Built-in validation without separate validation set

C) Requires extra data

D) Only for classification

Q11. What is the n_estimators hyperparameter?

A) Number of features

B) Number of trees in the forest

C) Number of samples

D) Tree depth

Q12. What is a typical range for n_estimators?

A) 5-10

B) 100-500

C) 1-5

D) 10,000+

Q13. Does Random Forest typically overfit with more trees?

A) Yes, always overfits with more trees

B) No, performance plateaus but rarely degrades

C) Only with small datasets

D) Always degrades with more trees

Q14. What is the max_features hyperparameter?

A) Maximum number of trees

B) Number of features to consider at each split

C) Maximum depth

D) Maximum samples

Q15. What is the max_depth hyperparameter in Random Forest?

A) Number of trees

B) Maximum depth of each tree (often unlimited/None)

C) Number of features

D) Training time limit

Q16. Do Random Forest trees need pruning?

A) Yes, always required

B) No, deep trees are fine due to ensemble averaging

C) Only for classification

D) Only for regression

Q17. What is an advantage of Random Forest?

A) Perfect interpretability

B) Excellent performance, minimal tuning, reduces overfitting

C) Always fastest training

D) Requires no data

Q18. What is a limitation of Random Forest?

A) Always overfits

B) Less interpretable than single tree, slower prediction, large memory

C) Cannot handle missing values

D) Only works for binary classification

Q19. Does Random Forest require feature scaling?

A) Yes, always required

B) No, tree-based splits don't depend on scale

C) Only for classification

D) Only for regression

Q20. How does Random Forest handle feature importance?

A) Cannot calculate importance

B) Based on average decrease in impurity or permutation importance

C) All features equally important

D) Random assignment

LEVEL 2: GRADIENT BOOSTING FUNDAMENTALS

Q21. What type of ensemble method is Gradient Boosting?

A) Bagging

B) Boosting (sequential training)

C) Stacking

D) Voting

Q22. How are models trained in boosting?

A) In parallel independently

B) Sequentially, each correcting previous errors

C) Randomly

D) All at once

Q23. What does each new tree in Gradient Boosting learn?

A) The original target variable

B) The residual errors (gradients) of the current ensemble

C) Random patterns

D) Feature importances

Q24. What is the general Gradient Boosting algorithm?

A) Train all trees independently

B) Initialize with simple model, iteratively add trees to reduce residuals

C) Use only one tree

D) Random tree addition

Q25. What does Gradient Boosting minimize?

A) Training time

B) A loss function using gradient descent in function space

C) Number of trees

D) Memory usage

Q26. What is the learning rate (shrinkage) in Gradient Boosting?

A) Speed of training

B) Weight applied to each tree's contribution (typically 0.01-0.3)

C) Number of trees

D) Tree depth

Q27. What is the effect of a low learning rate?

A) Faster convergence with fewer trees

B) Needs more trees but better generalization

C) Always worse performance

D) No effect on performance

Q28. What is the typical tree depth in Gradient Boosting?

A) Unlimited depth

B) Shallow trees (3-8 levels)

C) Always depth 1

D) Always depth 100+

Q29. Why use shallow trees in Gradient Boosting?

A) Faster only

B) Prevents individual trees from overfitting; ensemble handles complexity

C) Required technically

D) Uses less memory only

Q30. What is the subsample hyperparameter?

A) Number of trees

B) Fraction of samples to use for each tree (e.g., 0.8 = 80%)

C) Number of features

D) Learning rate

Q31. What does subsample < 1.0 do?

A) Increases training time

B) Adds randomness to reduce overfitting (stochastic gradient boosting)

C) Always reduces accuracy

D) Required for algorithm to work

Q32. What is XGBoost?

A) A new ML paradigm

B) Optimized Gradient Boosting with regularization and parallel processing

C) A neural network

D) A clustering algorithm

Q33. What are key innovations in XGBoost?

A) Uses random forests

B) Regularization, parallelization, handling missing values, tree pruning

C) No innovations

D) Only uses linear models

Q34. What is LightGBM?

A) Same as XGBoost

B) Faster gradient boosting using leaf-wise growth and histogram-based learning

C) A neural network

D) A dimensionality reduction method

Q35. What is unique about LightGBM's tree growth?

A) Depth-wise growth

B) Leaf-wise growth (splits leaf with max delta loss)

C) Random growth

D) No growth

Q36. What is CatBoost?

A) For image classification

B) Gradient boosting designed for categorical features

C) Only for regression

D) A neural network

Q37. What is an advantage of Gradient Boosting?

A) Highly interpretable

B) State-of-art performance for tabular data

C) Requires no tuning

D) Always fastest training

Q38. What is a limitation of Gradient Boosting?

A) Always underfits

B) Computationally expensive (sequential), prone to overfitting without regularization

C) Too simple

D) Cannot handle numerical data

Q39. What is early stopping in Gradient Boosting?

A) Starting training early

B) Stopping when validation error stops improving

C) Stopping at fixed iterations

D) Never stopping

Q40. When is Gradient Boosting preferred over Random Forest?

A) When interpretability is critical

B) When maximum accuracy is needed and computation time acceptable

C) When data is very small

D) Never preferred

LEVEL 3: NEURAL NETWORKS ARCHITECTURE

Q41. What are the three types of layers in neural networks?

A) Fast, medium, slow

B) Input layer, hidden layers, output layer

C) Small, medium, large

D) Linear, nonlinear, mixed

Q42. What does the input layer do?

A) Makes predictions

B) Receives feature values (one neuron per feature)

C) Trains the model

D) Calculates loss

Q43. What do hidden layers do?

A) Store data

B) Learn hierarchical feature representations

C) Only connect layers

D) Calculate accuracy

Q44. What does the output layer do?

A) Receives inputs

B) Produces predictions (neurons depend on task)

C) Stores weights

D) Only for visualization

Q45. How many output neurons for binary classification?

A) 2 neurons with softmax

B) 1 neuron with sigmoid

C) 10 neurons

D) Depends on features

Q46. How many output neurons for multiclass classification with 10 classes?

A) 1 neuron

B) 10 neurons with softmax

C) 2 neurons

D) 100 neurons

Q47. How many output neurons for regression?

A) 0 neurons

B) 1 neuron (typically linear activation)

C) 10 neurons

D) Depends on classes

Q48. What is a neuron's computation?

A) z = x + b

B) z = Σ(w■x■) + b, then a = activation(z)

C) z = x × w

D) z = activation only

Q49. What are weights in neural networks?

A) Input values

B) Learnable parameters connecting neurons

C) Activation functions

D) Output values

Q50. What is bias in a neuron?

A) Model error

B) Additional learnable parameter (offset/intercept)

C) Activation function

D) Weight value

Q51. What is forward propagation?

A) Training the network

B) Passing inputs through network to get predictions

C) Calculating gradients

D) Updating weights

Q52. What is the universal approximation theorem?

A) All functions are linear

B) Neural networks can approximate any continuous function

C) Networks need infinite neurons

D) Only specific functions can be learned

Q53. How many hidden layers defines a "deep" neural network?

A) 1 hidden layer

B) 2+ hidden layers

C) 10+ hidden layers required

D) No hidden layers

Q54. What is a fully connected (dense) layer?

A) Only some neurons connected

B) Every neuron connected to all neurons in previous layer

C) No connections

D) Random connections

Q55. What architecture is used for images?

A) Fully connected only

B) Convolutional Neural Networks (CNN)

C) Recurrent Neural Networks

D) Linear regression

Q56. What architecture is used for sequential data?

A) Standard feedforward

B) Recurrent Neural Networks (RNN) or LSTM

C) Only CNNs

D) Decision trees

Q57. What are skip connections (residual connections)?

A) Removing layers

B) Direct connections that skip layers (e.g., ResNet)

C) Random connections

D) Only for small networks

Q58. Why use skip connections?

A) Reduce accuracy

B) Help gradients flow in very deep networks

C) Slow down training

D) Increase overfitting

Q59. What is network width?

A) Physical size

B) Number of neurons per layer

C) Number of layers

D) Training time

Q60. What is network depth?

A) Neuron size

B) Number of layers

C) Number of neurons

D) Training time

LEVEL 4: ACTIVATION FUNCTIONS & TRAINING

Q61. Why are activation functions necessary?

A) Speed up training

B) Introduce non-linearity to learn complex patterns

C) Reduce memory

D) Only for visualization

Q62. What happens without activation functions (or with only linear)?

A) Better performance

B) Network collapses to linear model regardless of depth

C) Faster training

D) More accurate predictions

Q63. What is the sigmoid activation function?

A) $\sigma(x) = x$

B) $\sigma(x) = 1 / (1 + e^{(-x)})$

C) $\sigma(x) = \max(0, x)$

D) $\sigma(x) = x^2$

Q64. What is sigmoid's output range?

A) $(-\infty, +\infty)$

B) (0, 1)

C) [-1, 1]

D) $[0, \infty)$

Q65. What is a problem with sigmoid?

A) Too fast

B) Vanishing gradient problem (gradients near 0 at extremes)

C) No problems

D) Cannot output probabilities

Q66. What is the tanh activation function?

A) tanh(x) = x

B) tanh(x) = (e^x - e^(-x)) / (e^x + e^(-x))

C) tanh(x) = max(0, x)

D) tanh(x) = 1/(1+e^(-x))

Q67. What is tanh's output range?

A) (0, 1)

B) (-1, 1)

C) $(0, \infty)$

D) $(-\infty, +\infty)$

Q68. What is ReLU (Rectified Linear Unit)?

A) ReLU(x) = x

B) ReLU(x) = max(0, x)

C) ReLU(x) = 1/(1+e^(-x))

D) $ReLU(x) = x^2$

Q69. What is ReLU's output range?

A) (0, 1)

B) $[0, \infty)$

C) (-∞, +∞)

D) [-1, 1]

Q70. Why is ReLU most popular for hidden layers?

A) Most complex

B) Simple, fast, mitigates vanishing gradient problem

C) Always best accuracy

D) No advantages

Q71. What is the "dying ReLU" problem?

A) Training too slow

B) Neurons output 0 and stop learning (gradient is 0 for negative inputs)

C) Too many neurons

D) Memory issues

Q72. What is Leaky ReLU?

A) Same as ReLU

B) Leaky ReLU(x) = max(αx, x) where α is small (e.g., 0.01)

C) Always outputs zero

D) Only for output layer

Q73. What is the softmax function used for?

A) Binary classification

B) Multiclass classification (converts to probability distribution)

C) Regression

D) Hidden layers

Q74. What does softmax output sum to?

A) 0

B) 1.0 (probability distribution)

C) 100

D) Variable

Q75. What is backpropagation?

A) Forward pass

B) Algorithm computing gradients using chain rule to update weights

C) Making predictions

D) Data preprocessing

Q76. What is the chain rule used for in backpropagation?

A) Connecting layers

B) Computing gradients of loss with respect to all weights

C) Forward propagation

D) Activation functions

Q77. What is gradient descent?

A) Making predictions

B) Iterative optimization: weights -= learning_rate × gradient

C) Forward propagation

D) Data cleaning

Q78. What is the learning rate in neural networks?

A) Training speed

B) Step size for weight updates (critical hyperparameter, e.g., 0.001-0.1)

C) Number of epochs

D) Batch size

Q79. What happens with too large learning rate?

A) Very slow convergence

B) Overshooting, unstable training, divergence

C) Perfect convergence

D) No effect

Q80. What happens with too small learning rate?

A) Fast convergence

B) Very slow training, may get stuck in local minima

C) Perfect results

D) Immediate divergence

LEVEL 5: NEURAL NETWORK REGULARIZATION & OPTIMIZATION

Q81. What is dropout?

A) Removing input features

B) Randomly dropping neurons during training (regularization)

C) Removing layers

D) Stopping training

Q82. What is a typical dropout rate?

A) 0.9 (90%)

B) 0.2-0.5 (20-50%)

C) 0.01 (1%)

D) 1.0 (100%)

Q83. When is dropout applied?

A) During testing only

B) During training only (disabled during inference)

C) Always active

D) Never applied

Q84. What is batch normalization?

A) Normalizing input data

B) Normalizing layer inputs during training (stabilizes/speeds training)

C) Normalizing outputs

D) Normalizing weights only

Q85. What problem does batch normalization address?

A) Overfitting only

B) Internal covariate shift (changing distributions in layers)

C) Only speeds training

D) Reduces accuracy

Q86. What is the vanishing gradient problem?

A) Gradients too large

B) Gradients become very small in deep networks, preventing learning

C) No gradients exist

D) Perfect gradients

Q87. What causes vanishing gradients?

A) Too much data

B) Repeated multiplication of small gradients through many layers

C) Large learning rates

D) Too few layers

Q88. What is the exploding gradient problem?

A) Gradients too small

B) Gradients become extremely large, causing instability

C) No gradients

D) Perfect gradients

Q89. What is gradient clipping?

A) Removing gradients

B) Limiting gradient magnitude to prevent exploding gradients

C) Increasing gradients

D) Random gradient changes

Q90. What is the Adam optimizer?

A) Simple gradient descent

B) Adaptive learning rate optimizer combining momentum and RMSprop

C) No optimization

D) Random updates

Q91. What does Adam stand for?

A) Advanced Data Analysis Method

B) Adaptive Moment Estimation

C) Automated Decision Algorithm Model

D) Advanced Deep Architecture Method

Q92. What is momentum in optimization?

A) Training speed

B) Using exponentially weighted average of past gradients

C) Number of epochs

D) Batch size

Q93. Why use momentum?

A) Slows training

B) Accelerates convergence and dampens oscillations

C) Always worse

D) Only for small networks

Q94. What is an epoch?

A) Single sample

B) One complete pass through entire training dataset

C) One weight update

D) One batch

Q95. What is batch size?

A) Dataset size

B) Number of samples processed before weight update

C) Number of epochs

D) Number of layers

LEVEL 6: SUPPORT VECTOR MACHINES & HYPERPARAMETER TUNING

Q96. What is the goal of SVM?

A) Minimize margin

B) Find hyperplane that maximizes margin between classes

C) Fit data exactly

D) Random classification

Q97. What is the margin in SVM?

A) Classification error

B) Distance from hyperplane to nearest data points (support vectors)

C) Number of support vectors

D) Training time

Q98. What are support vectors?

A) All data points

B) Data points on or within the margin that define the decision boundary

C) Outliers only

D) Test samples

Q99. What is the kernel trick?

A) Speeds training only

B) Implicitly maps data to higher dimensions for non-linear separation

C) Reduces dimensions

D) Only for linear problems

Q100. What is the RBF (Gaussian) kernel?

A) Linear kernel

B) $K(x,y) = \exp(-\gamma||x-y||^2)$ - most popular non-linear kernel

C) Polynomial kernel

D) No transformation