Commonly Asked ML Interview Questions

1. What is the difference between Parametric and Non Parametric Algorithms?
2. Difference between convex and non-convex cost function; what does it mean when a cost function is non-convex?
3. How do you decide when to go for deep learning for a project?
4. Give an example of when False positive is more crucial than false negative and vice versa
5. Why is “Naive” Bayes naive?
6. Give an example where the median is a better measure than the mean
7. What do you mean by the unreasonable effectiveness of data?
8. Why KNN is known as a lazy learning technique?
9. What do you mean by semi supervised learning?
10. What is an OOB error and how is it useful?
11. In what scenario decision tree should be preferred over random forest?
12. Why Logistic Regression is called regression?
13. What is Online Machine Learning? How is it different from Offline machine learning? List some of it’s applications
14. What is No Free Lunch Theorem?
15. Imagine you are woking with a laptop of 2GB RAM, how would you process a dataset of 10GB?
16. What are the main differences between Structured and Unstructured Data?
17. What are the main points of difference between Bagging and Boosting?
18. What are the assumptions of linear regression?
19. How do you measure the accuracy of a Clustering Algorithm?
20. What is Matrix Factorization and where is it used in Machine Learning?
21. What is an Imbalanced Dataset and how can one deal with this problem?
22. How do you measure the accuracy of a recommendation engine?
23. What are some ways to make your model more robust to outliers?
24. How can you measure the performance of a dimensionality reduction algorithm on your dataset?
25. What is Data Leakage? List some ways using which you can overcome this problem.
26. What is Multicollinearity? How to detect it? List some techniques to overcome Multicollinearity.
27. List some ways using which you can reduce overfitting in a model.
28. What are the different types of bias in Machine Learning?
29. How do you approach a categorical feature with high cardinality?
30. Explain Pruning in Decision Trees and how it is done
31. What is ROC-AUC curve? List some of it’s benefits.
32. What are kernels in SVM? Can you list some popular SVM kernels.
33. What is the difference between Gini Impurity and Entropy? Which one is better and why?
34. Why does L2 regularization give sparse coefficients?
35. List some ways using which you can improve a model’s performance.
36. Can PCA be used to reduce the dimensionality of a highly nonlinear dataset?
37. What’s the difference between probability and likelihood?
38. What cross-validation technique would you use on a time series data set.
39. Once a dataset’s dimensionality has been reduced, is it possible to reverse the operation? If so, how? If not, why?
40. Why do we always need the intercept term in a regression model??
41. When Your Dataset Is Suffering From High Variance, How Would You Handle It?
42. Which Among These Is More Important Model Accuracy Or Model Performance?
43. What is active learning and where is it useful?
44. Why is Ridge Regression called Ridge?
45. State the differences between causality and correlation?
46. Does it make any sense to chain two different dimensionality reduction algorithms?
47. Is it possible to speed up training of a bagging ensemble by distributing it across multiple servers?
48. If a Decision Tree is underfitting the training set, is it a good idea to try scaling the input features?
49. Say you trained an SVM classifier with an RBF kernel. It seems to underfit the training set: should you increase or decrease γ (gamma)? What about C?
50. What is cross validation and it's types?
51. How do we interpret weights in linear models?
52. Which Gradient Descent algorithm (among those we discussed) will reach the vicinity of the optimal solution the fastest? Which will actually converge?
53. Why is it important to scale the inputs when using SVMs?
54. What is p value and why is it important?
55. What is OvR and OvO for multiclass classification and which machine learning algorithm supports this
56. How will you do feature selection using Lasso Regression?
57. What is the difference between loss function and cost function?
58. What are the common ways to handle missing data in a dataset?
59. What is the difference between standard scaler and minmax scaler? What you will do if there is a categorical variable?
60. What types of model tend to overfit?
61. What are some advantages and Disadvantages of regression models and tree based models.
62. What are some important hyperparameters for XGBOOST
63. Can you tell the complete life cycle of a data science project?
64. What are the properties of a good ML model?
65. What are the different evaluation metrices for a regression model?
66. What are the different evaluation metrices for a classification model?
67. Difference between R2 and adjusted R2? Why do you preffer adjusted r2?
68. List some of the drawbacks of a Linear model
69. What do you mean by Curse of Dimensionality?
70. What do you mean by Bias variance tradeoff?
71. Explain Kernel trick in SVM
72. What is the main difference between Machine Learning and Data Mining?
73. Why sometimes it is needed to scale or normalise features?
74. What is the difference between Type 1 and Type 2 error?
75. What is the difference between a Generative model vs a Discriminative model?

### Why binary\_crossentropy and categorical\_crossentropy give different performances for the same problem?

### [Why does one hot encoding improve machine learning performance?](https://stackoverflow.com/questions/17469835/why-does-one-hot-encoding-improve-machine-learning-performance)

1. Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?
2. Differentiate between wide and tall data formats?
3. What is the difference between inductive machine learning and deductive machine learning?
4. How will you know which machine learning algorithm to choose for your classification problem?
5. What is the difference between Covariance and Correlation
6. How will you find the correlation between a categorical variable and a continuous variable ?
7. What are the differences between “Bayesian” and “Frequentist” approach for Machine Learning?
8. What is the difference between stochastic gradient descent (SGD) and gradient descent ?
9. What is the difference between Gaussian Mixture Model and K-Means Algorithm?

#### Is more data always better?

#### How can you determine which features are the most im- portant in your model?

1. Which hyper-parameter tuning strategies (in general) do you know?
2. How to select K for K-means?
3. Describe the differences between and use cases for box plots and histograms

#### How would you differentiate between Multilabel and MultiClass classification?

#### What is KL divergence, how would you define its usecase in ML?

#### Can you define the concept of Undersampling and Oversampling?

### Considering a Long List of Machine Learning Algorithms, given a Data Set, How Do You Decide Which One to Use?

### Explain the difference between Normalization and Standardization

### List the most popular distribution curves along with scenarios where you will use them in an algorithm.

### List all types of popular recommendation systems?

### Which metrics can be used to measure correlation of categorical data?

### Which type of sampling is better for a classification model and why?

Here are brief answers to the commonly asked machine learning interview questions:

1. \*\*Parametric vs Non-Parametric Algorithms\*\*:

   - \*\*Parametric\*\*: Make assumptions about the data distribution (e.g., linear regression). They have a fixed number of parameters.

   - \*\*Non-Parametric\*\*: Do not make strong assumptions about data distribution (e.g., KNN, decision trees). They can adapt to data complexity.

2. \*\*Convex vs Non-Convex Cost Function\*\*:

   - \*\*Convex\*\*: A function where any line segment between two points on the curve lies above or on the curve (e.g., quadratic cost function). Easier to optimize.

   - \*\*Non-Convex\*\*: Can have multiple local minima or maxima (e.g., neural networks). More challenging to find the global optimum.

3. \*\*Deep Learning Decision\*\*: Use deep learning if the problem involves complex patterns and large datasets, especially in image, speech, or text analysis.

4. \*\*False Positive vs False Negative\*\*:

   - \*\*False Positive More Crucial\*\*: Medical screening for a rare disease where false positives could lead to unnecessary stress and treatment.

   - \*\*False Negative More Crucial\*\*: Fraud detection where missing a fraudulent transaction could be costly.

5. \*\*Naive Bayes\*\*: Called "naive" because it assumes that all features are independent given the class label, which is often not the case in real-world data.

6. \*\*Median vs Mean\*\*: Median is better when data has outliers or is skewed (e.g., income data).

7. \*\*Unreasonable Effectiveness of Data\*\*: Refers to the observation that with enough data, simple models can perform exceptionally well, often surpassing more complex models.

8. \*\*KNN and Lazy Learning\*\*: KNN is a lazy learner because it doesn't learn a model but instead stores training instances and makes decisions during the query phase.

9. \*\*Semi-Supervised Learning\*\*: Uses a combination of labeled and unlabeled data for training. Useful when labeling data is expensive.

10. \*\*OOB Error\*\*: Out-Of-Bag error estimates the generalization error of a random forest by evaluating on data not used in training a particular tree.

11. \*\*Decision Tree vs Random Forest\*\*: Use a decision tree for simplicity and interpretability; use random forest for better performance and to handle overfitting.

12. \*\*Logistic Regression\*\*: It's called regression because it models the probability of a binary outcome using a logistic function, despite being used for classification.

13. \*\*Online vs Offline Machine Learning\*\*:

    - \*\*Online\*\*: Models are updated continuously as new data arrives. Suitable for dynamic environments.

    - \*\*Offline\*\*: Models are trained in batches. Suitable for static datasets.

    - \*\*Applications\*\*: Real-time recommendations, financial trading.

14. \*\*No Free Lunch Theorem\*\*: No single algorithm works best for all problems. The effectiveness of an algorithm depends on the specific problem and data.

15. \*\*Processing 10GB Dataset on 2GB RAM\*\*: Use data sampling, batch processing, or distributed computing frameworks like Spark to handle large datasets.

16. \*\*Structured vs Unstructured Data\*\*:

    - \*\*Structured\*\*: Data organized in tables with rows and columns (e.g., SQL databases).

    - \*\*Unstructured\*\*: Data without a predefined format (e.g., text, images).

17. \*\*Bagging vs Boosting\*\*:

    - \*\*Bagging\*\*: Builds multiple models independently and combines their outputs (e.g., Random Forest).

    - \*\*Boosting\*\*: Sequentially builds models where each model corrects the errors of its predecessor (e.g., AdaBoost).

18. \*\*Linear Regression Assumptions\*\*:

    - Linearity

    - Independence of errors

    - Homoscedasticity (constant variance of errors)

    - Normality of errors

19. \*\*Clustering Accuracy\*\*: Use metrics like silhouette score, Davies-Bouldin index, or within-cluster sum of squares.

20. \*\*Matrix Factorization\*\*: Decomposes a matrix into product of matrices. Used in recommendation systems (e.g., collaborative filtering).

21. \*\*Imbalanced Dataset\*\*: Class imbalance where one class is significantly underrepresented. Use techniques like resampling, SMOTE, or cost-sensitive learning.

22. \*\*Recommendation Engine Accuracy\*\*: Use metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), or Precision@K.

23. \*\*Robust to Outliers\*\*: Use methods like robust scaling, trimming, or algorithms less sensitive to outliers (e.g., median).

24. \*\*Dimensionality Reduction Performance\*\*: Evaluate with explained variance ratio, visualization, or downstream task performance.

25. \*\*Data Leakage\*\*: Occurs when information from outside the training dataset is used to create the model. Prevent it by careful data splitting and feature selection.

26. \*\*Multicollinearity\*\*: High correlation among features. Detect with Variance Inflation Factor (VIF). Overcome with techniques like PCA or removing correlated features.

27. \*\*Reduce Overfitting\*\*: Use techniques like cross-validation, regularization, pruning, or ensemble methods.

28. \*\*Types of Bias\*\*:

    - \*\*Sampling Bias\*\*: Non-representative sample.

    - \*\*Selection Bias\*\*: Systematic difference between groups.

    - \*\*Confirmation Bias\*\*: Preference for information that confirms existing beliefs.

29. \*\*High Cardinality Categorical Feature\*\*: Use techniques like target encoding, dimensionality reduction, or feature hashing.

30. \*\*Pruning in Decision Trees\*\*: Reduces the size of the tree by removing branches that have little importance. Methods include cost complexity pruning.

31. \*\*ROC-AUC Curve\*\*: Measures the performance of a classification model across all thresholds. Higher AUC indicates better performance.

32. \*\*SVM Kernels\*\*: Functions that map data to higher dimensions. Popular kernels include linear, polynomial, and radial basis function (RBF).

33. \*\*Gini Impurity vs Entropy\*\*: Both measure node impurity in decision trees. Gini is often faster to compute, while entropy is more informative.

34. \*\*L2 Regularization\*\*: Penalizes large coefficients by adding a squared magnitude term. Results in smaller, non-zero coefficients rather than sparse coefficients (sparse coefficients are a feature of L1 regularization).

35. \*\*Improve Model Performance\*\*: Use feature engineering, hyperparameter tuning, cross-validation, or ensemble methods.

36. \*\*PCA on Nonlinear Data\*\*: PCA is linear and may not perform well on highly nonlinear data. Consider using techniques like kernel PCA.

37. \*\*Probability vs Likelihood\*\*:

    - \*\*Probability\*\*: Measures the chance of an event given a model.

    - \*\*Likelihood\*\*: Measures how well a model explains observed data.

38. \*\*Cross-Validation for Time Series\*\*: Use time series specific techniques like rolling window or expanding window cross-validation.

39. \*\*Reversing Dimensionality Reduction\*\*: In some cases, techniques like PCA allow for approximate reconstruction of the original data.

40. \*\*Intercept Term in Regression\*\*: Provides a baseline value when all predictors are zero, which is important for proper model interpretation.

41. \*\*High Variance\*\*: Use techniques like regularization, ensemble methods, or reducing model complexity.

42. \*\*Model Accuracy vs Performance\*\*: Accuracy may not always reflect performance if the model is imbalanced. Consider metrics like precision, recall, F1-score, etc.

43. \*\*Active Learning\*\*: A technique where the model selects the most informative data points to label. Useful when labeling is expensive.

44. \*\*Ridge Regression\*\*: Named "ridge" due to the "ridge" or penalty term added to the loss function (L2 regularization).

45. \*\*Causality vs Correlation\*\*:

    - \*\*Causality\*\*: One event causes another.

    - \*\*Correlation\*\*: Two events occur together but one does not necessarily cause the other.

46. \*\*Chaining Dimensionality Reduction\*\*: It may be useful if different methods capture different aspects of data variance. Evaluate based on model performance.

47. \*\*Speeding Up Bagging\*\*: Yes, by parallelizing the training of different models across servers.

48. \*\*Scaling Features for Decision Trees\*\*: Typically, scaling is not necessary for decision trees as they are invariant to feature scaling.

49. \*\*SVM with RBF Kernel\*\*: If underfitting, increase γ (gamma) to make the model more complex. Increase C to reduce regularization.

50. \*\*Cross-Validation\*\*: A technique to evaluate model performance by partitioning data into training and test sets multiple times. Types include k-fold, stratified, and leave-one-out.

51. \*\*Interpreting Weights in Linear Models\*\*: Weights represent the impact of each feature on the target variable. Larger weights indicate greater importance.

52. \*\*Gradient Descent Algorithms\*\*: Variants like stochastic gradient descent (SGD) converge faster in practice but may have more variance in updates. Batch gradient descent is more stable but slower.

53. \*\*Scaling Inputs for SVM\*\*: Important because SVMs are sensitive to the scale of the data. Scaling ensures features contribute equally to the decision boundary.

54. \*\*P-value\*\*: Measures the probability of observing the data given that the null hypothesis is true. Important for hypothesis testing.

55. \*\*OvR vs OvO\*\*:

    - \*\*OvR (One-vs-Rest)\*\*: One classifier per class, separating each class from the rest.

    - \*\*OvO (One-vs-One)\*\*: One classifier per pair of classes. Suitable for algorithms like SVM.

56. \*\*Feature Selection with Lasso Regression\*\*: Lasso (L1 regularization) can drive some feature coefficients to zero, effectively selecting features.

57. \*\*Loss Function vs Cost Function\*\*:

    - \*\*Loss Function\*\*: Measures error for a single data

 point.

    - \*\*Cost Function\*\*: Measures error across the entire dataset (often average of losses).

58. \*\*Handling Missing Data\*\*: Techniques include imputation (mean, median, mode), deletion, or using algorithms that handle missing values.

59. \*\*Standard Scaler vs MinMax Scaler\*\*:

    - \*\*Standard Scaler\*\*: Scales features to have zero mean and unit variance.

    - \*\*MinMax Scaler\*\*: Scales features to a specified range (e.g., 0 to 1).

    - \*\*Categorical Variables\*\*: Use techniques like one-hot encoding.

60. \*\*Overfitting Models\*\*: Complex models like deep neural networks or high-degree polynomial regressions are prone to overfitting.

61. \*\*Regression Models vs Tree-Based Models\*\*:

    - \*\*Regression Models\*\*: Simple, interpretable but may not capture complex relationships.

    - \*\*Tree-Based Models\*\*: Handle complex relationships and interactions but can be prone to overfitting.

62. \*\*XGBoost Hyperparameters\*\*: Key parameters include learning rate, max depth, min child weight, gamma, and subsample.

63. \*\*Data Science Project Lifecycle\*\*: Includes problem definition, data collection, data cleaning, exploratory analysis, model building, evaluation, and deployment.

64. \*\*Properties of a Good ML Model\*\*: Accuracy, interpretability, robustness, generalization, and efficiency.

65. \*\*Regression Metrics\*\*: MAE, RMSE, R², Adjusted R².

66. \*\*Classification Metrics\*\*: Accuracy, Precision, Recall, F1-score, ROC-AUC.

67. \*\*R² vs Adjusted R²\*\*: Adjusted R² accounts for the number of predictors and adjusts for overfitting, while R² may be overly optimistic.

68. \*\*Drawbacks of Linear Models\*\*: Limited to linear relationships, may underperform with complex data, sensitive to outliers.

69. \*\*Curse of Dimensionality\*\*: Refers to the challenges and inefficiencies when working with high-dimensional data, including sparsity and increased computational cost.

70. \*\*Bias-Variance Tradeoff\*\*: The balance between model complexity (variance) and accuracy (bias). Too much variance leads to overfitting, too much bias leads to underfitting.

71. \*\*Kernel Trick in SVM\*\*: Allows SVM to perform well in high-dimensional spaces by implicitly mapping input features into higher dimensions.

72. \*\*Machine Learning vs Data Mining\*\*: Machine learning focuses on building models to make predictions, while data mining focuses on discovering patterns and relationships in data.

73. \*\*Scaling/Normalizing Features\*\*: Ensures that features contribute equally to model training and improves convergence.

74. \*\*Type 1 vs Type 2 Error\*\*:

    - \*\*Type 1\*\*: False positive (rejecting a true null hypothesis).

    - \*\*Type 2\*\*: False negative (failing to reject a false null hypothesis).

75. \*\*Generative vs Discriminative Models\*\*:

    - \*\*Generative\*\*: Models how data is generated (e.g., Gaussian Mixture Models).

    - \*\*Discriminative\*\*: Models the decision boundary between classes (e.g., Logistic Regression).

76. \*\*Binary Crossentropy vs Categorical Crossentropy\*\*: Binary crossentropy is for binary classification, while categorical crossentropy is for multi-class classification.

77. \*\*One-Hot Encoding\*\*: Converts categorical variables into a binary matrix, making them suitable for machine learning algorithms.

78. \*\*Choosing ML Algorithms\*\*: Consider data size, type, and problem complexity. Start with simpler models and move to complex ones as needed.

79. \*\*Wide vs Tall Data\*\*:

    - \*\*Wide\*\*: Many features, few instances.

    - \*\*Tall\*\*: Many instances, few features.

80. \*\*Inductive vs Deductive ML\*\*:

    - \*\*Inductive\*\*: Learning general rules from specific examples.

    - \*\*Deductive\*\*: Applying known rules to make specific predictions.

81. \*\*Choosing ML Algorithm for Classification\*\*: Consider data size, feature types, problem complexity, and model interpretability.

82. \*\*Covariance vs Correlation\*\*:

    - \*\*Covariance\*\*: Measures the joint variability of two variables.

    - \*\*Correlation\*\*: Normalized measure of covariance, indicating the strength and direction of the relationship.

83. \*\*Correlation Between Categorical and Continuous Variable\*\*: Use statistical tests like ANOVA or correlation ratios.

84. \*\*Bayesian vs Frequentist Approach\*\*:

    - \*\*Bayesian\*\*: Incorporates prior beliefs and updates them with data.

    - \*\*Frequentist\*\*: Focuses on the frequency of events and data without prior beliefs.

85. \*\*SGD vs Gradient Descent\*\*:

    - \*\*SGD\*\*: Updates parameters using a single data point at a time.

    - \*\*Gradient Descent\*\*: Uses the entire dataset to compute gradients.

86. \*\*Gaussian Mixture Model vs K-Means\*\*:

    - \*\*GMM\*\*: Probabilistic model assuming data comes from multiple Gaussian distributions.

    - \*\*K-Means\*\*: Clustering algorithm that partitions data into k clusters.

87. \*\*More Data\*\*: Generally, more data improves model performance, but only up to a point where the data becomes redundant.

88. \*\*Feature Importance\*\*: Use methods like feature importance scores from models, permutation importance, or SHAP values.

89. \*\*Hyperparameter Tuning Strategies\*\*: Grid search, random search, Bayesian optimization.

90. \*\*Selecting K for K-Means\*\*: Use methods like the elbow method, silhouette score, or cross-validation.

91. \*\*Box Plots vs Histograms\*\*:

    - \*\*Box Plots\*\*: Show distribution, median, and outliers.

    - \*\*Histograms\*\*: Show the frequency distribution of numerical data.

92. \*\*Multilabel vs Multiclass Classification\*\*:

    - \*\*Multilabel\*\*: Each instance can belong to multiple classes.

    - \*\*Multiclass\*\*: Each instance belongs to one class from multiple possible classes.

93. \*\*KL Divergence\*\*: Measures how one probability distribution diverges from a second, reference probability distribution. Used in tasks like distribution matching.

94. \*\*Undersampling vs Oversampling\*\*:

    - \*\*Undersampling\*\*: Reduces the number of majority class instances.

    - \*\*Oversampling\*\*: Increases the number of minority class instances.

95. \*\*Choosing ML Algorithm\*\*: Based on problem type, data characteristics, and performance metrics.

96. \*\*Normalization vs Standardization\*\*:

    - \*\*Normalization\*\*: Scales data to a specific range (e.g., 0 to 1).

    - \*\*Standardization\*\*: Scales data to have zero mean and unit variance.

97. \*\*Popular Distribution Curves\*\*: Normal, Poisson, Exponential, Binomial. Use based on data characteristics and problem requirements.

98. \*\*Recommendation Systems\*\*: Collaborative filtering, content-based filtering, hybrid methods.

99. \*\*Metrics for Categorical Data Correlation\*\*: Chi-square test, Cramér's V.

\*\*Sampling for Classification Models\*\*: \*\*Oversampling\*\* (like SMOTE) is often used to balance class distribution and improve model performance.