Machine Learning - 100 Answers

Pages 12+ Part 2: With Explanations

Nova IMS
Generated: January 16, 2026

LEVEL 1: ADVANCED CLASSIFICATION METRICS

Q1. Which metric should you optimize when false negatives are very costly (e.g., disease diagnosis)?

ANSWER: C

Explanation: When false negatives are very costly (missing a disease diagnosis could be fatal), you must optimize recall to catch as many true positives as possible, even at the cost of some false positives.

Q2. Which metric should you optimize when false positives are very costly (e.g., spam filtering important emails)?

ANSWER: B

Explanation: When false positives are very costly (marking important emails as spam is highly undesirable), you must optimize precision to ensure that when you predict positive, you're almost always right.

Q3. What does the F1-score represent?

ANSWER: B

Explanation: The F1-score is the harmonic mean (not arithmetic mean) of precision and recall: F1 = 2 × (precision × recall) / (precision + recall). It balances both metrics.

Q4. What is the formula for F1-score?

ANSWER: B

Explanation: F1 = 2 × (P × R) / (P + R). The harmonic mean penalizes extreme values more than arithmetic mean, requiring both precision and recall to be reasonably high.

Q5. When is F1-score most useful?

ANSWER: B

Explanation: F1-score is most useful for imbalanced datasets where you need to balance precision (avoiding false alarms) and recall (catching positives), rather than relying on accuracy alone.

Q6. What is the F-Beta score?

ANSWER: B

Explanation: F-Beta is a generalized F-score: $F\beta = (1 + \beta^2) \times (P \times R) / (\beta^2 \times P + R)$, where β controls the weight given to recall vs precision.

Q7. What does $\beta > 1$ in F-Beta score emphasize?

ANSWER: B

Explanation: $\beta > 1$ (e.g., $\beta=2$) weights recall more heavily than precision. F2-score considers recall twice as important as precision, useful when catching positives matters more.

Q8. What does $\beta < 1$ in F-Beta score emphasize?

ANSWER: B

Explanation: $\beta < 1$ (e.g., $\beta=0.5$) weights precision more heavily than recall. F0.5-score considers precision twice as important, useful when false positives are more costly.

Q9. What is specificity (True Negative Rate)?

ANSWER: B

Explanation: Specificity (True Negative Rate) = TN / (TN + FP), measuring the proportion of actual negatives correctly identified. It's the complement of false positive rate.

Q10. What does ROC stand for?

ANSWER: B

Explanation: ROC stands for Receiver Operating Characteristic, a curve plotting True Positive Rate (recall) against False Positive Rate at various classification thresholds.

Q11. What does the ROC curve plot?

ANSWER: B

Explanation: The ROC curve plots TPR (sensitivity/recall) on y-axis vs FPR (1-specificity) on x-axis for all possible classification thresholds from 0 to 1.

Q12. What does each point on the ROC curve represent?

ANSWER: B

Explanation: Each point on the ROC curve represents model performance at a specific classification threshold. Moving the threshold changes the balance between TPR and FPR.

Q13. What does the diagonal line in ROC space represent?

ANSWER: B

Explanation: The diagonal line (y=x) represents a random classifier with no discriminative power (AUC=0.5). Points above the diagonal are better than random.

Q14. What does AUC stand for?

ANSWER: B

Explanation: AUC stands for Area Under the (ROC) Curve, quantifying overall model performance across all thresholds. It ranges from 0 to 1.

Q15. What does an AUC of 0.5 indicate?

ANSWER: B

Explanation: AUC of 0.5 means the model has no discriminative power - it performs no better than random guessing, unable to distinguish between classes.

Q16. What does an AUC of 1.0 indicate?

ANSWER: B

Explanation: AUC of 1.0 indicates perfect classification - the model can perfectly separate the two classes at some threshold with 100% TPR and 0% FPR.

Q17. What is considered "good" AUC performance?

ANSWER: B

Explanation: AUC between 0.8-0.9 is generally considered good performance in real-world applications. 0.9-1.0 is excellent, 0.7-0.8 is acceptable, below 0.7 is poor.

Q18. What is Matthews Correlation Coefficient (MCC)?

ANSWER: B

Explanation: MCC (Matthews Correlation Coefficient) is a balanced metric that considers all four confusion matrix categories (TP, TN, FP, FN), performing well even with imbalanced classes. Range: [-1, 1].

Q19. What MCC value indicates perfect prediction?

ANSWER: B

Explanation: MCC = 1 indicates perfect prediction, MCC = 0 indicates random prediction, and MCC = -1 indicates total disagreement (inverse prediction).

Q20. What is the Precision-Recall curve useful for?

ANSWER: B

Explanation: The Precision-Recall curve is more informative than ROC for highly imbalanced datasets because it focuses on the positive class performance without being inflated by the large number of true negatives.

LEVEL 2: REGRESSION METRICS

Q21. What does MAE stand for?

ANSWER: B

Explanation: MAE stands for Mean Absolute Error, the average of absolute differences between predicted and actual values.

Q22. What is the formula for MAE?

ANSWER: B

Explanation: MAE = (1/n) Σ |y■ - ■■|, taking the mean of absolute prediction errors. It treats all errors equally (linear penalty).

Q23. What is an advantage of MAE?

ANSWER: B

Explanation: MAE is easy to interpret (same units as target), robust to outliers (not squared), and provides a linear penalty. However, it's not differentiable at zero (can complicate optimization).

Q24. What does MSE stand for?

ANSWER: B

Explanation: MSE stands for Mean Squared Error, the average of squared differences between predicted and actual values.

Q25. What is the formula for MSE?

ANSWER: B

Explanation: MSE = (1/n) Σ (y■ - ■■)², squaring each error before averaging. Squaring heavily penalizes large errors.

Q26. What is a characteristic of MSE?

ANSWER: B

Explanation: MSE heavily penalizes large errors due to squaring. A prediction off by 10 contributes 100 to MSE, while one off by 1 contributes only 1. It's differentiable everywhere (good for optimization).

Q27. What does RMSE stand for?

ANSWER: B

Explanation: RMSE stands for Root Mean Squared Error, the square root of MSE.

Q28. What is the formula for RMSE?

ANSWER: B

Explanation: RMSE = √[(1/n) Σ (y■ - ■■)²], which is √MSE. Taking the square root returns the metric to the original units of the target variable.

Q29. Why is RMSE preferred over MSE?

ANSWER: B

Explanation: RMSE is in the same units as the target variable (unlike MSE which is squared units), making it more interpretable. For example, if predicting prices in dollars, RMSE is in dollars.

Q30. What does R^2 (R-squared) represent?

ANSWER: B

Explanation: R^2 (R-squared or coefficient of determination) represents the proportion of variance in the dependent variable explained by the model. It measures goodness of fit.

Q31. What is the range of R^2?

ANSWER: B

Explanation: R^2 ranges from -∞ to 1. $R^2 = 1$ is perfect, $R^2 = 0$ means the model is no better than predicting the mean, $R^2 < 0$ means the model is worse than predicting the mean.

Q32. What does $R^2 = 1$ indicate?

ANSWER: B

Explanation: $R^2 = 1$ indicates perfect fit - the model explains 100% of the variance in the target variable, with all predictions exactly matching actual values.

Q33. What does $R^2 = 0$ indicate?

ANSWER: B

Explanation: $R^2 = 0$ indicates the model performs as well as simply predicting the mean for all inputs. The model provides no explanatory power beyond the baseline mean.

Q34. What does negative R^2 indicate?

ANSWER: B

Explanation: Negative R^2 indicates the model performs worse than the baseline (mean). This suggests a fundamentally flawed model, perhaps from severe overfitting or using test data the model wasn't trained for.

Q35. What is adjusted R^2?

ANSWER: B

Explanation: Adjusted R^2 modifies R^2 to account for the number of predictors, penalizing models with unnecessary features: Adj $R^2 = 1 - [(1-R^2)(n-1)/(n-p-1)]$ where p is number of predictors.

LEVEL 3: BIAS-VARIANCE TRADEOFF & OVERFITTING

Q36. What is bias in machine learning?

ANSWER: B

Explanation: Bias is the error from wrong assumptions in the learning algorithm. High bias causes the model to miss relevant patterns (underfitting).

Q37. What does high bias indicate?

ANSWER: B

Explanation: High bias indicates the model is too simple for the data complexity - it makes strong assumptions that don't hold, causing systematic errors (underfitting).

Q38. What is variance in machine learning?

ANSWER: B

Explanation: Variance is the error from sensitivity to small fluctuations in training data. High variance means predictions vary significantly with different training sets.

Q39. What does high variance indicate?

ANSWER: B

Explanation: High variance indicates the model is too complex - it fits noise and random fluctuations in the training data rather than true patterns (overfitting).

Q40. What is the total error decomposition?

ANSWER: B

Explanation: Total Expected Error = Bias² + Variance + Irreducible Error. This decomposition shows the fundamental tradeoff between underfitting (bias) and overfitting (variance).

Q41. What is irreducible error?

ANSWER: B

Explanation: Irreducible error is the noise inherent in the data that no model can eliminate. It comes from unmeasured factors, random noise, and inherent randomness in the phenomenon.

Q42. What characterizes underfitting?

ANSWER: B

Explanation: Underfitting shows poor performance on both training and test sets. The model is too simple to capture patterns, resulting in high training error that doesn't improve with more training.

Q43. What are symptoms of underfitting?

ANSWER: B

Explanation: In underfitting, learning curves show both training and validation error plateauing at a high value early in training. The curves converge but at poor performance.

Q44. What are solutions to underfitting?

ANSWER: B

Explanation: Solutions to underfitting: Increase model complexity (more layers, more neurons), add more relevant features, reduce regularization, train longer, use more powerful algorithms.

Q45. What characterizes overfitting?

ANSWER: B

Explanation: Overfitting shows excellent training performance but poor test performance. The model memorizes training data including noise rather than learning generalizable patterns.

Q46. What are symptoms of overfitting?

ANSWER: B

Explanation: In overfitting, there's a large gap between training and validation curves. Training error continues decreasing while validation error increases or plateaus, indicating poor generalization.

Q47. What are solutions to overfitting?

ANSWER: B

Explanation: Solutions to overfitting: Get more training data, reduce model complexity, add regularization (L1/L2), use cross-validation, early stopping, dropout, data augmentation, ensemble methods.

Q48. What is regularization?

ANSWER: B

Explanation: Regularization adds a penalty term to the loss function proportional to model complexity, discouraging the model from fitting noise and encouraging simpler models that generalize better.

Q49. What does L1 regularization (Lasso) do?

ANSWER: B

Explanation: L1 regularization (Lasso) adds penalty λΣ|w■|, driving some coefficients exactly to zero. This performs automatic feature selection by eliminating less important features.

Q50. What does L2 regularization (Ridge) do?

ANSWER: B

Explanation: L2 regularization (Ridge) adds penalty λΣw■², shrinking all coefficients toward zero but keeping all features. It's preferred when all features are potentially relevant.

Q51. What does the regularization parameter (α or λ) control?

ANSWER: B

Explanation: The regularization parameter (λ or α) controls penalty strength. Higher values = more regularization = simpler model = higher bias, lower variance. Lower values = less regularization = more complex model.

Q52. What is early stopping?

ANSWER: B

Explanation: Early stopping monitors validation performance during training and stops when it stops improving. This prevents overfitting by stopping before the model memorizes training noise.

Q53. What is data leakage?

ANSWER: B

Explanation: Data leakage occurs when information from outside the training dataset (especially from test set) influences the training process, causing overly optimistic performance estimates that don't generalize.

Q54. What is an example of data leakage?

ANSWER: B

Explanation: Fitting a scaler on the entire dataset before splitting means test set statistics (mean, std) influence the transformation applied to training data, causing leakage and invalidating evaluation.

Q55. Why is data leakage critical to avoid?

ANSWER: B

Explanation: Data leakage causes inflated performance estimates during development that don't translate to production. The model appears to work great but fails in the real world because it had unfair information.

LEVEL 4: LINEAR & LOGISTIC REGRESSION

Q56. What does linear regression predict?

ANSWER: B

Explanation: Linear regression predicts continuous numerical values (e.g., price, temperature, age) on a continuous scale from -∞ to +∞.

Q57. What is the hypothesis function for linear regression?

ANSWER: B

Explanation: The linear regression hypothesis is h(x) = β■ + β■x■ + β■x■ + ... + β■x■, a weighted sum of features plus intercept.

Q58. What does linear regression minimize?

ANSWER: B

Explanation: Linear regression minimizes Mean Squared Error (MSE) or equivalently the sum of squared residuals. This is the ordinary least squares (OLS) criterion.

Q59. What is the Normal Equation in linear regression?

ANSWER: B

Explanation: The Normal Equation provides a closed-form solution: β = (X^T X)^(-1) X^T y. It solves for optimal coefficients directly without iteration (for small datasets).

Q60. What are key assumptions of linear regression?

ANSWER: B

Explanation: Key assumptions: Linearity (relationship is linear), Independence (observations independent), Homoscedasticity (constant variance of residuals), Normality (residuals normally distributed).

Q61. What does logistic regression predict?

ANSWER: B

Explanation: Logistic regression predicts probabilities between 0 and 1, representing the likelihood of belonging to the positive class. Output is P(y=1|x).

Q62. What function does logistic regression use?

ANSWER: B

Explanation: Logistic regression uses the sigmoid (logistic) function to map any real value to [0,1]: σ(z) = 1/(1+e^(-z)).

Q63. What is the sigmoid function formula?

ANSWER: B

Explanation: The sigmoid function is σ(z) = 1/(1+e^(-z)), which maps z∈(-∞,+∞) to σ(z)∈(0,1). As z→∞, σ→1; as z→-∞, σ→0.

Q64. What loss function does logistic regression use?

ANSWER: B

Explanation: Logistic regression uses log loss (binary cross-entropy): -[y log(■) + (1-y) log(1-■)], which heavily penalizes confident wrong predictions.

Q65. What is the decision boundary in logistic regression?

ANSWER: B

Explanation: The decision boundary occurs where P(y=1|x) = 0.5, which is when β^T x = 0. This forms a linear boundary in feature space.

Q66. What does the softmax function do?

ANSWER: B

Explanation: Softmax extends logistic regression to multiple classes by converting raw scores into a probability distribution: P(y=k|x) = e^(z■)/Σ■e^(z■).

Q67. What is an advantage of logistic regression?

ANSWER: B

Explanation: Advantages: Very fast training, probabilistic outputs (calibrated probabilities), highly interpretable coefficients, low computational requirements, works well with limited data.

Q68. What is a limitation of logistic regression?

ANSWER: B

Explanation: Logistic regression can only learn linear decision boundaries. If the true boundary is non-linear, logistic regression will underfit unless features are engineered.

Q69. When is logistic regression appropriate?

ANSWER: B

Explanation: Logistic regression is appropriate when: the decision boundary is roughly linear, interpretability is important, you need probability estimates, computational resources are limited.

Q70. What regularization is common in logistic regression?

ANSWER: B

Explanation: L1 (Lasso) drives coefficients to zero (feature selection), L2 (Ridge) shrinks coefficients (handles multicollinearity), ElasticNet combines both. Regularization prevents overfitting in logistic regression.

LEVEL 5: NAIVE BAYES & KNN

Q71. What theorem is Naive Bayes based on?

ANSWER: B

Explanation: Naive Bayes is based on Bayes' Theorem: P(A|B) = P(B|A)P(A)/P(B), which relates conditional probabilities.

Q72. What is Bayes' Theorem formula?

ANSWER: B

Explanation: Bayes' Theorem: P(y|X) = P(X|y)P(y)/P(X), where P(y) is prior, P(X|y) is likelihood, P(y|X) is posterior, P(X) is evidence.

Q73. What "naive" assumption does Naive Bayes make?

ANSWER: B

Explanation: Naive Bayes assumes features are conditionally independent given the class: P(X|y) = P(x■|y)P(x■|y)...P(x■|y). This 'naive' assumption rarely holds but often works well in practice.

Q74. What Naive Bayes variant is used for continuous features?

ANSWER: B

Explanation: Gaussian Naive Bayes assumes continuous features follow a normal distribution within each class. It models P(x■|y) as Gaussian with class-specific mean and variance.

Q75. What Naive Bayes variant is used for text classification (word counts)?

ANSWER: B

Explanation: Multinomial Naive Bayes is used for discrete count data, especially text classification where features represent word frequencies or TF-IDF values.

Q76. What is Laplace smoothing in Naive Bayes?

ANSWER: B

Explanation: Laplace smoothing adds a small constant (typically 1) to feature counts to avoid zero probabilities: P(x■|y) = (count + α)/(total + αn). Prevents zero probabilities from dominating.

Q77. What is an advantage of Naive Bayes?

ANSWER: B

Explanation: Advantages: Extremely fast training and prediction, works with small datasets, handles high-dimensional data well, performs surprisingly well despite naive assumption.

Q78. What is a limitation of Naive Bayes?

ANSWER: B

Explanation: The strong independence assumption rarely holds in practice - features are usually correlated. This causes Naive Bayes to underestimate probabilities, though relative rankings are often still correct.

Q79. What does KNN stand for?

ANSWER: B

Explanation: KNN stands for K-Nearest Neighbors, a non-parametric, instance-based learning algorithm.

Q80. How does KNN make predictions?

ANSWER: B

Explanation: For a new point, KNN finds the K training samples closest to it (using distance metric), then predicts the majority class (classification) or average value (regression) of those K neighbors.

Q81. Why is KNN called "lazy learning"?

ANSWER: B

Explanation: KNN is 'lazy learning' because there's no explicit training phase - it simply memorizes all training data and defers computation until prediction time.

Q82. What is the most common distance metric in KNN?

ANSWER: B

Explanation: Euclidean distance: $\sqrt{[\Sigma(x_i - y_i)^2]}$, the straight-line distance between points. Most commonly used in KNN.

Q83. How do you choose optimal K in KNN?

ANSWER: B

Explanation: Use cross-validation to test different K values and select the K that gives best validation performance. Typically test odd values to avoid ties in binary classification.

Q84. What happens with very small K (e.g., K=1)?

ANSWER: B

Explanation: Small K (K=1) creates complex, wiggly decision boundaries with high variance - very sensitive to noise and outliers in training data.

Q85. What is a major limitation of KNN?

ANSWER: B

Explanation: KNN requires storing all training data (memory intensive), computing distances to all points for each prediction O(n) (slow), and suffers from curse of dimensionality in high dimensions.

LEVEL 6: DECISION TREES

Q86. How do decision trees make predictions?

ANSWER: B

Explanation: Decision trees recursively partition the feature space using if-then-else rules (e.g., if age<30 then... else if income>50k then...) to create a tree-like decision structure.

Q87. What does Gini impurity measure?

ANSWER: B

Explanation: Gini impurity measures how often a randomly chosen element would be incorrectly classified: Gini = 1 - Σp■², where p■ is the fraction of samples in class i.

Q88. What is the Gini impurity range?

ANSWER: B

Explanation: Gini ranges from 0 (pure, all samples same class) to 0.5 (maximally impure for binary classification, equal mix of classes).

Q89. What does entropy measure in decision trees?

ANSWER: B

Explanation: Entropy measures disorder/uncertainty: Entropy = -Σp■log■(p■). Higher entropy = more mixed/uncertain node.

Q90. What is Information Gain?

ANSWER: B

Explanation: Information Gain = Entropy(parent) - Weighted_Average(Entropy(children)). It measures reduction in uncertainty from making a split.

Q91. What splitting criterion is sklearn's default for classification trees?

ANSWER: B

Explanation: Gini impurity is sklearn's default for classification trees because it's computationally faster than entropy (no logarithm) and gives similar results in practice.

Q92. What splitting criterion is used for regression trees?

ANSWER: B

Explanation: Regression trees use MSE or MAE as splitting criteria, choosing splits that minimize prediction error for continuous target values.

Q93. What is the max_depth hyperparameter?

ANSWER: B

Explanation: max_depth limits the maximum number of levels in the tree from root to leaf. It's the primary hyperparameter for controlling tree complexity and overfitting.

Q94. What is pruning in decision trees?

ANSWER: B

Explanation: Pruning removes branches/nodes after the tree is grown. Pre-pruning stops growth early (e.g., max_depth), post-pruning grows fully then removes branches that don't improve validation performance.

Q95. What are advantages of decision trees?

ANSWER: B

Explanation: Advantages: Highly interpretable (visualizable), requires no feature scaling, handles mixed data types (numerical + categorical), captures non-linear patterns, fast prediction.

Q96. What are limitations of decision trees?

ANSWER: B

Explanation: Decision trees are very prone to overfitting (high variance) - they can grow arbitrarily complex, memorizing training data. They're also unstable - small data changes cause large tree changes.

Q97. What does "greedy algorithm" mean for decision trees?

ANSWER: B

Explanation: Greedy algorithm means at each split, the tree chooses the locally optimal split without considering future splits. It doesn't backtrack or explore alternative paths that might be globally better.

Q98. Why are decision trees unstable?

ANSWER: B

Explanation: Trees are unstable because small changes in training data can produce completely different splits near the root, cascading through the entire tree structure.

Q99. Do decision trees require feature scaling?

ANSWER: B

Explanation: No, decision trees don't require feature scaling because splits are based on feature thresholds (e.g., age>30), which are scale-invariant.

Q100. What prediction does a regression tree leaf node contain?

ANSWER: B

Explanation: Regression tree leaves contain the mean (or median) of target values for all training samples that fall into that leaf node. This constant value is the prediction for any new sample reaching that leaf.