General Linear Model:

1. What is the purpose of the General Linear Model (GLM)?

2. What are the key assumptions of the General Linear Model?

3. How do you interpret the coefficients in a GLM?

4. What is the difference between a univariate and multivariate GLM?

5. Explain the concept of interaction effects in a GLM.

6. How do you handle categorical predictors in a GLM?

7. What is the purpose of the design matrix in a GLM?

8. How do you test the significance of predictors in a GLM?

9. What is the difference between Type I, Type II, and Type III sums of squares in a GLM?

10. Explain the concept of deviance in a GLM.

Regression:

11. What is regression analysis and what is its purpose?

12. What is the difference between simple linear regression and multiple linear regression?

13. How do you interpret the R-squared value in regression?

14. What is the difference between correlation and regression?

15. What is the difference between the coefficients and the intercept in regression?

16. How do you handle outliers in regression analysis?

17. What is the difference between ridge regression and ordinary least squares regression?

18. What is heteroscedasticity in regression and how does it affect the model?

19. How do you handle multicollinearity in regression analysis?

20. What is polynomial regression and when is it used?

Loss function:

21. What is a loss function and what is its purpose in machine learning?

22. What is the difference between a convex and non-convex loss function?

23. What is mean squared error (MSE) and how is it calculated?

24. What is mean absolute error (MAE) and how is it calculated?

25. What is log loss (cross-entropy loss) and how is it calculated?

26. How do you choose the appropriate loss function for a given problem?

27. Explain the concept of regularization in the context of loss functions.

28. What is Huber loss and how does it handle outliers?

29. What is quantile loss and when is it used?

30. What is the difference between squared loss and absolute loss?

Optimizer (GD):

31. What is an optimizer and what is its purpose in machine learning?

32. What is Gradient Descent (GD) and how does it work?

33. What are the different variations of Gradient Descent?

34. What is the learning rate in GD and how do you choose an appropriate value?

35. How does GD handle local optima in optimization problems?

36. What is Stochastic Gradient Descent (SGD) and how does it differ from GD?

37. Explain the concept of batch size in GD and its impact on training.

38. What is the role of momentum in optimization algorithms?

39. What is the difference between batch GD, mini-batch GD, and SGD?

40. How does the learning rate affect the convergence of GD?

Regularization:

41. What is regularization and why is it used in machine learning?

42. What is the difference between L1 and L2 regularization?

43. Explain the concept of ridge regression and its role in regularization.

44. What is the elastic net regularization and how does it combine L1 and L2 penalties?

45. How does regularization help prevent overfitting in machine learning models?

46. What is early stopping and how does it relate to regularization?

47. Explain the concept of dropout regularization in neural networks.

48. How do you choose the regularization parameter in a model?

49. What

is the difference between feature selection and regularization?

50. What is the trade-off between bias and variance in regularized models?

SVM:

51. What is Support Vector Machines (SVM) and how does it work?

52. How does the kernel trick work in SVM?

53. What are support vectors in SVM and why are they important?

54. Explain the concept of the margin in SVM and its impact on model performance.

55. How do you handle unbalanced datasets in SVM?

56. What is the difference between linear SVM and non-linear SVM?

57. What is the role of C-parameter in SVM and how does it affect the decision boundary?

58. Explain the concept of slack variables in SVM.

59. What is the difference between hard margin and soft margin in SVM?

60. How do you interpret the coefficients in an SVM model?

Decision Trees:

61. What is a decision tree and how does it work?

62. How do you make splits in a decision tree?

63. What are impurity measures (e.g., Gini index, entropy) and how are they used in decision trees?

64. Explain the concept of information gain in decision trees.

65. How do you handle missing values in decision trees?

66. What is pruning in decision trees and why is it important?

67. What is the difference between a classification tree and a regression tree?

68. How do you interpret the decision boundaries in a decision tree?

69. What is the role of feature importance in decision trees?

70. What are ensemble techniques and how are they related to decision trees?

Ensemble Techniques:

71. What are ensemble techniques in machine learning?

72. What is bagging and how is it used in ensemble learning?

73. Explain the concept of bootstrapping in bagging.

74. What is boosting and how does it work?

75. What is the difference between AdaBoost and Gradient Boosting?

76. What is the purpose of random forests in ensemble learning?

77. How do random forests handle feature importance?

78. What is stacking in ensemble learning and how does it work?

79. What are the advantages and disadvantages of ensemble techniques?

80. How do you choose the optimal number of models in an ensemble?

ANSWERS

General Linear Model:

1. The General Linear Model (GLM) is a statistical framework used to analyze the relationship between a dependent variable and one or more independent variables. It provides a flexible approach to modeling various types of data and can handle both continuous and categorical predictors.

2. The key assumptions of the General Linear Model include linearity, independence of errors, homoscedasticity (constant variance of errors), and normally distributed errors. These assumptions ensure the validity of the model and the reliability of the statistical inferences.

3. In a GLM, the coefficients represent the estimated effects of the independent variables on the dependent variable. They indicate the change in the expected value of the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant. The coefficients can be interpreted as the slope of the relationship between the independent variable and the dependent variable.

4. A univariate GLM involves a single dependent variable and one or more independent variables, while a multivariate GLM involves multiple dependent variables simultaneously modeled with one or more independent variables. Univariate GLM focuses on analyzing the relationship between one dependent variable and predictors, whereas multivariate GLM explores the relationships between multiple dependent variables and predictors.

5. Interaction effects in a GLM occur when the relationship between the dependent variable and one independent variable depends on the value of another independent variable. It signifies that the effect of one predictor on the dependent variable is influenced by the presence or absence of another predictor. Interaction effects provide insights into more complex relationships and can help capture non-linear or conditional effects in the model.

6. Categorical predictors in a GLM are typically represented using dummy variables or indicator variables. Each category of the categorical predictor is converted into a binary variable, indicating the presence or absence of that category. These binary variables are then included as independent variables in the GLM. The coefficient associated with each dummy variable represents the difference in the mean response between that category and the reference category.

7. The design matrix in a GLM represents the matrix of predictor variables used in the model. Each row corresponds to an observation, and each column corresponds to an independent variable (including any interaction terms or categorical variables). The design matrix is used to estimate the coefficients and perform statistical inference in the GLM.

8. The significance of predictors in a GLM can be tested using hypothesis tests or confidence intervals. A commonly used test is the t-test, which assesses whether the estimated coefficient significantly differs from zero. The p-value associated with the t-test indicates the probability of observing the estimated coefficient if the null hypothesis (no effect) is true. A small p-value (typically less than a predefined significance level, e.g., 0.05) suggests a significant effect of the predictor.

9. Type I, Type II, and Type III sums of squares are methods used to partition the variation in the dependent variable explained by different predictors in the GLM. Type I sums of squares evaluate the unique contribution of each predictor, sequentially adding predictors to the model. Type II sums of squares assess the contribution of each predictor, adjusting for the presence of other predictors. Type III sums of squares evaluate the contribution of each predictor, considering all other predictors in the model.

10. Deviance in a GLM measures the discrepancy between the observed data and the model's predicted values. It is a measure of the goodness-of-fit and can be used to compare different models. Lower deviance values indicate a better fit of the model to the data.

Regression:

11. Regression analysis is a statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables. Its purpose is to understand how changes in the independent variables are associated with changes in the dependent variable and to make predictions or infer causal relationships.

12. Simple linear regression involves a single dependent variable and one independent variable, while multiple linear regression involves a single dependent variable and two or more independent variables. Simple linear regression aims to model a linear relationship between the dependent variable and the independent variable, while multiple linear regression extends this to incorporate multiple predictors.

13. The R-squared value in regression represents the proportion of the variance in the dependent variable that is explained by the independent variables. It measures the goodness-of-fit of the regression model. R-squared ranges from 0 to 1, with a value of 1 indicating that the model explains all the variance in the dependent variable.

14. Correlation measures the strength and direction of the linear relationship between two variables, while regression aims to model the relationship between a dependent variable and one or more independent variables. Correlation does not involve predicting one variable from another, whereas regression focuses on predicting the dependent variable based on the independent variables.

15. The coefficients in regression represent the estimated effects of the independent variables on the dependent variable. They indicate the change in the expected value of the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant. The intercept is the value of the dependent variable when all independent variables are zero.

16. Outliers in regression analysis are extreme observations that deviate significantly from the pattern or trend in the data. They can have a strong influence on the estimated regression coefficients and may adversely affect the model's performance. Outliers should be carefully examined and their impact on the model should be evaluated. Techniques such as robust regression or data transformation can help handle outliers.

17. Ridge regression is a variant of ordinary least squares regression that includes a penalty term (L2 regularization) to shrink the coefficients and reduce overfitting. It is particularly useful when dealing with multicollinearity, as it can mitigate the impact of highly correlated predictors. Ordinary least squares regression, on the other hand, does not include a regularization term and aims to minimize the sum of squared residuals.

18. Heteroscedasticity in regression refers to the situation where the variance of the residuals (errors) is not constant across different levels of the independent variables. It violates one of the assumptions of regression, which assumes constant variance of errors (homoscedasticity). Heteroscedasticity can affect the reliability of the regression coefficients and the significance of statistical tests.

19. Multicollinearity in regression occurs when two or more independent variables are highly correlated with each other. It can lead to unstable or unreliable estimates of the regression coefficients, making it difficult to interpret the individual effects of the predictors. Techniques for handling multicollinearity include removing one of the correlated variables, combining the variables into a single composite variable, or using regularization methods like ridge regression.

20. Polynomial regression is a form of regression analysis where the relationship between the independent variable(s) and the dependent variable is modeled as an nth-degree polynomial. It is used when the relationship between the variables is nonlinear and can capture more complex patterns in the data. Polynomial regression allows for curved or nonlinear relationships to be modeled.

Loss function:

21. A loss function is a mathematical function that quantifies the discrepancy between the predicted values of a model and the true values of the target variable. Its purpose in machine learning is to guide the model towards minimizing the prediction errors and learning the optimal set of parameters.

22. A convex loss function is one that has a single global minimum and is bowl-shaped, meaning that any two points on the curve lie below the line segment connecting them. Optimization of convex loss functions can guarantee convergence to the global minimum. Non-convex loss functions, on the other hand, can have multiple local minima and can make the optimization problem more challenging.

23. Mean Squared Error (MSE) is a commonly used loss function that measures the average squared difference

between the predicted values and the true values. It penalizes larger errors more than smaller errors, making it sensitive to outliers. It is calculated by taking the average of the squared differences between the predicted and true values.

24. Mean Absolute Error (MAE) is a loss function that measures the average absolute difference between the predicted values and the true values. It is less sensitive to outliers compared to MSE, as it does not square the differences. MAE is calculated by taking the average of the absolute differences between the predicted and true values.

25. Log loss, also known as cross-entropy loss, is a loss function commonly used in classification problems. It quantifies the difference between the predicted probabilities and the true class labels. Log loss penalizes confident and incorrect predictions more than less confident predictions. It is calculated by taking the negative logarithm of the predicted probability for the true class.

26. The choice of the appropriate loss function depends on the specific problem and the characteristics of the data. Mean Squared Error (MSE) is commonly used for regression tasks, while log loss (cross-entropy loss) is used for binary classification problems. The selection of the loss function should align with the evaluation metric and the desired behavior of the model.

27. Regularization in the context of loss functions is a technique used to prevent overfitting and improve the generalization of the model. It introduces additional constraints or penalties on the model's parameters during the optimization process. Regularization can help reduce the complexity of the model and prevent the coefficients from taking on excessively large values.

28. Huber loss is a loss function that combines the best properties of squared loss (MSE) and absolute loss (MAE). It is less sensitive to outliers than MSE and provides a more robust measure of error. Huber loss is calculated differently for values below a certain threshold (delta), where it behaves like absolute loss, and for values above the threshold, where it behaves like squared loss.

29. Quantile loss is a loss function used to estimate quantiles (percentiles) of the predicted distribution. It measures the discrepancy between the predicted quantile and the true value. Quantile loss is particularly useful when modeling the entire distribution of the target variable rather than just the mean.

30. Squared loss and absolute loss differ in the way they penalize prediction errors. Squared loss (MSE) penalizes larger errors more than absolute loss (MAE) due to the squaring of the differences. Squared loss is more sensitive to outliers, as they contribute significantly to the overall loss. Absolute loss is less sensitive to outliers as it treats all errors equally. The choice between squared loss and absolute loss depends on the specific problem and the desired behavior of the model.

Optimizer (GD):

31. An optimizer is an algorithm or method used to adjust the parameters of a model in order to minimize the loss function and improve the model's performance. It plays a crucial role in the training of machine learning models.

32. Gradient Descent (GD) is an optimization algorithm used to minimize the loss function by iteratively adjusting the model parameters in the direction of steepest descent. It calculates the gradient of the loss function with respect to the parameters and updates the parameters in proportion to the negative gradient.

33. There are different variations of Gradient Descent, including Batch Gradient Descent, Stochastic Gradient Descent (SGD), and Mini-Batch Gradient Descent. Batch Gradient Descent computes the gradient using the entire training dataset in each iteration, while SGD uses a single randomly selected sample. Mini-Batch Gradient Descent computes the gradient using a subset (mini-batch) of the training dataset.

34. The learning rate in Gradient Descent determines the step size taken in each iteration when updating the parameters. It controls how quickly or slowly the model learns. Choosing an appropriate learning rate is important, as a too small value may result in slow convergence, while a too large value may cause overshooting and instability.

35. Gradient Descent handles local optima in optimization problems by iteratively updating the parameters in the direction of steepest descent. Although it can get stuck in local optima, the use of stochasticity in SGD or the exploration of different solutions through random initialization or multiple restarts can help escape local optima.

36. Stochastic Gradient Descent (SGD) is a variation of Gradient Descent that randomly selects a single sample from the training dataset in each iteration. This reduces computational complexity and memory requirements, making it suitable for large datasets. However, the stochastic nature of SGD introduces more noise in the optimization process, which can result in more erratic convergence behavior.

37. Batch size in Gradient Descent refers to the number of training samples used to compute the gradient in each iteration. In Batch Gradient Descent, the batch size is equal to the size of the entire training dataset. In Mini-Batch Gradient Descent, the batch size is a small subset of the training dataset. The choice of batch size affects the trade-off between computational efficiency and optimization accuracy.

38. Momentum in optimization algorithms, including Gradient Descent, introduces a velocity term that accumulates the past gradients to help navigate flat or ravine-like loss surfaces. It helps accelerate convergence by allowing the algorithm to retain some memory of previous updates. Momentum reduces oscillations and can help overcome local optima.

39. Batch Gradient Descent uses the entire training dataset to compute the gradient in each iteration, which can be computationally expensive for large datasets. Mini-Batch Gradient Descent strikes a balance by using a subset (mini-batch) of the training dataset. Stochastic Gradient Descent uses a single randomly selected sample. The choice of the algorithm depends on the computational resources, convergence speed, and noise tolerance required.

40. The learning rate affects the convergence of Gradient Descent. A too small learning rate may result in slow convergence, as the updates are too conservative. A too large learning rate may cause the algorithm to overshoot and fail to converge. Appropriate learning rate tuning or adaptive learning rate methods can help achieve faster convergence and stable optimization.

Regularization:

41. Regularization is a technique used in machine learning to prevent overfitting and improve the generalization performance of models. It introduces additional constraints or penalties on the model's parameters during training to reduce complexity and control the trade-off between model fit and model complexity.

42. L1 and L2 regularization are two common types of regularization used in machine learning. L1 regularization adds a penalty term proportional to the absolute value of the model's coefficients (L1 norm), promoting sparsity and leading to feature selection. L2 regularization adds a penalty term proportional to the squared magnitude of the model's coefficients (L2 norm), promoting smaller but non-zero coefficients.

43. Ridge regression is a regression technique that incorporates L2 regularization. It adds a penalty term proportional to the sum of squared coefficients to the ordinary least squares regression objective function. Ridge regression shrinks the coefficients towards zero, reducing the impact of individual predictors and mitigating the effects of multicollinearity.

44. Elastic Net regularization is a regularization technique that combines L1 and L2 penalties. It adds a linear combination of the L1 and L2 norms to the objective function, allowing for simultaneous feature selection and coefficient shrinkage. Elastic Net provides a balance between L1 and L2 regularization and can handle highly correlated predictors.

45. Regularization helps prevent overfitting in machine learning models by penalizing large coefficients and reducing model complexity. It discourages the model from learning intricate details of the training data that may not generalize well to unseen data. Regular

ization can improve the model's ability to generalize by finding a balance between fitting the training data and avoiding overfitting.

46. Early stopping is a regularization technique where the training of the model is stopped before convergence if the performance on a validation set starts to deteriorate. It prevents the model from overfitting the training data by selecting the iteration that achieves the best performance on the validation set. Early stopping allows the model to generalize better and avoid over-optimization on the training data.

47. Dropout regularization is a technique commonly used in neural networks. It randomly selects a subset of neurons in each training iteration and sets their outputs to zero, effectively dropping them out. Dropout helps prevent overfitting by reducing co-adaptation among neurons and encouraging the network to learn more robust and generalizable representations.

48. The regularization parameter in a model determines the strength of the regularization penalty. A higher regularization parameter increases the penalty, leading to stronger regularization and potentially smaller coefficients. The choice of the regularization parameter depends on the specific problem and can be determined through techniques such as cross-validation.

49. Feature selection and regularization are related concepts but differ in their approaches. Feature selection explicitly aims to identify a subset of relevant features and eliminate irrelevant or redundant ones. Regularization, on the other hand, implicitly encourages sparse coefficients or smaller coefficient values, which can lead to feature selection as a byproduct. Feature selection can be considered as a more explicit form of regularization.

50. There is a trade-off between bias and variance in regularized models. Increasing the regularization strength (higher penalty) reduces the model's variance by shrinking the coefficients towards zero and reducing overfitting. However, it also increases the model's bias by potentially underfitting the training data. The optimal balance between bias and variance depends on the specific problem and the available data.

SVM:

51. Support Vector Machines (SVM) is a supervised learning algorithm used for classification and regression tasks. It finds an optimal hyperplane that separates different classes or predicts the value of a continuous target variable.

52. The kernel trick in SVM allows nonlinear transformations of the input features to be implicitly applied, enabling SVM to handle complex decision boundaries. It maps the data into a higher-dimensional space where it can be linearly separable.

53. Support vectors in SVM are the data points that lie closest to the decision boundary (hyperplane). They play a crucial role in defining the decision boundary and determining the margins.

54. The margin in SVM is the distance between the decision boundary and the support vectors. SVM aims to maximize the margin to achieve better generalization and robustness.

55. SVM can handle imbalanced datasets by incorporating class weights or using techniques such as oversampling the minority class or undersampling the majority class. The choice of technique depends on the specific problem and the characteristics of the data.

56. Linear SVM finds a linear decision boundary to separate classes, while non-linear SVM uses the kernel trick to map the data into a higher-dimensional space where it becomes linearly separable.

57. The C-parameter in SVM controls the trade-off between achieving a larger margin and allowing more training samples to be misclassified. A smaller C-value produces a wider margin but allows more misclassifications, while a larger C-value produces a narrower margin with fewer misclassifications.

58. Slack variables in SVM are introduced to handle non-separable datasets by allowing some training samples to be misclassified or fall within the margin. They relax the strictness of the margin constraint, allowing for more flexible decision boundaries.

59. The difference between hard margin and soft margin in SVM is the strictness of the margin constraint. Hard margin SVM requires that all training samples be correctly classified and lies outside the margin, which is only possible for linearly separable datasets. Soft margin SVM allows for some misclassifications and samples to fall within the margin, which makes it more robust to noise and outliers.

60. The coefficients in an SVM model represent the weights assigned to the input features. They indicate the contribution of each feature in determining the class boundary or predicting the target variable.

Decision Trees:

61. A decision tree is a supervised learning algorithm used for classification and regression tasks. It builds a tree-like model of decisions and their possible consequences.

62. Splits in a decision tree are determined by evaluating different features and selecting the one that provides the most information gain or decrease in impurity.

63. Impurity measures, such as the Gini index and entropy, quantify the impurity or disorder of a set of samples. They are used in decision trees to determine the quality of splits and the overall purity of the resulting subsets.

64. Information gain measures the reduction in entropy or impurity achieved by splitting a dataset based on a particular feature. It helps determine which feature provides the most useful information for making decisions in the tree.

65. Missing values in decision trees can be handled by assigning them to the majority class or using imputation techniques such as mean or median imputation. Alternatively, decision trees can treat missing values as a separate category.

66. Pruning in decision trees is a technique used to reduce overfitting. It involves removing or collapsing branches that do not contribute significantly to the overall improvement in impurity or information gain.

67. A classification tree is used for predicting categorical or discrete class labels, while a regression tree is used for predicting continuous numerical values.

68. Decision boundaries in a decision tree are defined by the splits at each node. The tree branches represent different decisions based on the feature values, leading to different class predictions or target variable values.

69. Feature importance in decision trees indicates the relative importance of different features in making decisions. It can be measured based on the information gain or the Gini index achieved by each feature.

70. Ensemble techniques, such as Random Forests, use multiple decision trees to make predictions. They combine the predictions of individual trees to achieve better overall performance and reduce the risk of overfitting.

Ensemble Techniques:

71. Ensemble techniques in machine learning combine multiple models to make predictions or decisions. They leverage the wisdom of crowds and aim to improve overall performance and generalization.

72. Bagging is an ensemble technique that involves training multiple models independently on different subsets of the training data and combining their predictions through voting or averaging. It helps reduce variance and improve the stability of the model.

73. Bootstrapping in bagging is a technique where random subsets of the training data are created by sampling with replacement. This allows for the generation of multiple subsets that may contain duplicate or similar samples.

74. Boosting is an ensemble technique that iteratively trains weak models and adjusts the weights of the training samples based on their performance. It focuses on difficult-to-classify samples and combines the predictions of multiple weak models to create a stronger ensemble.

75. AdaBoost (Adaptive Boosting) is a specific boosting algorithm that assigns weights to training samples and adjusts them based on misclassification errors. It assigns higher weights to misclassified samples in subsequent iterations, allowing subsequent weak models to focus on those samples.

76. Random Forests is an ensemble technique that combines the predictions of multiple decision trees. It builds individual decision trees on randomly selected subsets of features and combines their predictions through voting or averaging. Random Forests reduce overfitting and improve performance by introducing randomness in feature selection and training data sampling.

77. Random Forests can measure feature importance based on the average decrease in impurity or information gain achieved by each feature across multiple trees. This provides insight into the relative importance of different features in making predictions.

78. Stacking is an ensemble technique that combines the predictions of multiple models through a meta-model. It involves training multiple base models on the training data, using their predictions as input features for the meta-model, and making final predictions based on the meta-model's output.

79. Ensemble techniques offer several advantages, including improved generalization, reduced overfitting, increased stability, and the ability to capture complex relationships in the data. However, they can be computationally expensive and require careful tuning of hyperparameters.

80. The optimal number of models in an ensemble depends on the specific problem, the available data, and the ensemble technique used. It is typically determined through experimentation and validation, balancing the trade-off between model complexity and generalization performance.