General Linear Model:

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

>>> The general linear model is a statistical model which em]ncompasses various techniques of ANOVA,ANOCVA,simple linear regression,or multiple linear regression.this is a framework that provides the analysis of the data and relationship between the dependent and independent variable.

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

>>>the assumptions are:

* It is homogenous
* There is linearity of data.
* Independence of the data.it is assumed that the data is independent of each other
* The error or residual is assumed to have followed the normal distribution.

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

>>>the coefficients represent the estimated predictor effect on the response variable.

The positive coefficient will represent the positive relation and the negative will give the negative relationship between the variables.

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

>>>here will be the one single response or the predicted variable.

And in the multivariate GLM there will be multiple response variable,there will be one or more predicted values based on the data provided.

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

>>>Interaction effects in a GLM refer to situations where the effect of one predictor on the response variable depends on the value of another predictor. In other words, the impact of one predictor is not constant across different levels of another predictor.

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

>>>the categorical predictors are handled by the binary balancing with initiating the values as 0 and 1.the approach can be dummy coding or the sum coding to solve the categorical predictors in GLM.

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

>>>The design matrix in a GLM is a matrix representation of the predictor variables used in the model.

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

>>>To test the significance of predictors in a GLM, statistical tests are used, such as t-tests or F-tests, which assess the null hypothesis that the corresponding coefficient is zero. The p-value associated with the test indicates the probability of observing the estimated coefficient if the null hypothesis is true.

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

>>>Type I sums of squares assess the unique contribution of each predictor, sequentially considering the predictors in a specified order. The order of the predictors can affect the results.

Type II sums of squares assess the contribution of each predictor after accounting for the effects of other predictors. It is generally preferred when there are categorical predictors or interactions in the model, as it avoids confounding effects.

10. Explain the concept of deviance in a GLM.

>>>Deviance in a GLM represents the measure of lack of fit or discrepancy between the observed data and the fitted model.

Regression:

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

>>>The regression analysis is statistical concept which determines the relationship between the data that dependent and independent variable.

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

>>>in the simple linear regression there is only one feature and the predictions to be done only on one related feature .but in the multiple linear regression there are more features on the basis of them we have to predict the result.

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

>>>the r squared is the performance matrix.this is used to assess the goodness of the regression model is whether good or bad fit.the regression model is biased by the features and the correlation for this model cannot be determined by the accuracy level of the model.

14. What is the difference between correlation and regression?

>>>correlation determines the strength and direction of linear relationship between two or more variables.its usually quantifies how two variables are relatively closely to each other.

And the regression models the relationship and allows for prediction and estimation of the effect of independent variables on the dependent variable.

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

>>>the coefficient are also known as the slope which represent the estimated change in the dependent variable with every single unit of change in the independent variable.

The intercept means also known as the constant term or the y-intercept, is the value of the dependent variable when all independent variables are set to zero.

16. How do you handle outliers in regression analysis?

>>>the outliers can be handled by z-score standardisation,we can remove the outliers,transformation of outliers (square,dividing,root ,log values can be taken into considerations.)

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

>>>OLS regression assumes that the independent variables are not highly correlated with each other, as multicollinearity can lead to unstable and unreliable coefficient estimates. In contrast, ridge regression is specifically designed to handle multicollinearity.

OLS regression aims to find the "best-fitting" line that minimises the sum of squared residuals. In ridge regression, the coefficient estimates are obtained by minimising the sum of squared residuals with the addition of a penalty term.

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

>>>it is refers to the situation where the variability of the errors (residuals) in a regression model is not constant across the range of the independent variables.

19. How do you handle multicollinearity in regression analysis?

>>>Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other.this can be handled by the PCA,T-sne,feature selection,feature merging ,regularization,data transformation.

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

>>>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. In other words, it extends the linear regression model by including higher-order terms of the independent variable(s) in the equation.

Loss function:

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

>>>loss basically assess to the difference of actual point of the data and the predicted point that is assumed to be present on the slope of the line.the cost function is such that to be minimizes for the better performance of the regression model.

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

>>>In the convex loss function:

* There is one global minima
* Optimization of loss

In the non convex loss function thre will be

* Many local minimas

Gradient descent limitations

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

>>>MSE is a loss function that calculates the average of the squared differences between the predicted values (ŷ) and the true values (y) of the dependent variable. It penalizes larger errors more heavily due to the squaring operation. The formula for MSE is:

MSE = (1/n) \* Σ(y - ŷ)²

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

>>>MAE is a loss function that calculates the average of the absolute differences between the predicted values (ŷ) and the true values (y) of the dependent variable. It treats all errors equally, regardless of their magnitude. The formula for MAE is:

MAE = (1/n) \* Σ|y - ŷ|

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

>>>It quantifies the dissimilarity between predicted probabilities and true binary labels. Log loss is designed to penalize models that are less confident in their predictions and assign high probabilities to incorrect classes.

Log loss = -[y \* log(p) + (1 - y) \* log(1 - p)]

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

>>>the choice of loss function is not always fixed and may involve experimentation and iteration. It is often combined with other techniques, such as hyperparameter tuning and model selection

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

>>>Loss function with regularization = Loss function + λ \* Penalty term

Regularization helps to shrink the coefficients towards zero, effectively reducing the impact of less important features or irrelevant variables. This helps in feature selection and promotes sparsity in the model

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

>>>Huber loss basically is the combination of MAE & MSE

It provides a compromise between the two and is less sensitive to outliers compared to squared loss. Huber loss has a linear region for smaller errors and a quadratic (squared) region for larger errors.

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

>>>The quantile loss function measures the discrepancy between the predicted quantiles and the true quantiles.

Quantile regression is robust to Quantile loss and quantile regression provide a flexible and robust framework for analyzing data beyond the mean and provide insights into the conditional distribution of the dependent variable.

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

>>>mean squared error, calculates the average of the squared differences between the predicted values (ŷ) and the true values (y).

the average of the absolute differences between the predicted values (ŷ) and the true values (y).

Optimizer (GD):

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

>>>the purpose of optimizer is to make the minimum loss of the value in the machine learning model.it itervatively updates the values of the points upto it possess to the global minima .the famous optimizers are gradient descent,stochastic gradient descent.minibatch GD

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

>>>The algorithm starts with an initial set of parameter values and repeatedly adjusts them until convergence is reached.

33. What are the different variations of Gradient Descent?

>>>the famous optimizers are gradient descent,stochastic gradient descent.minibatch GD

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

>>>The learning rate is a hyperparameter in Gradient Descent that determines the size of the steps taken during parameter updates. It controls the magnitude of the parameter adjustments based on the gradient information.

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

>>>we can multiple restart the training of th e data or we can also schedule the new learning rates until that gives us the optimized GD.

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

>>>(SGD) is an optimization algorithm that performs updates using the gradient computed from a single randomly selected training example at each iteration.SGD is particularly useful when dealing with large datasets and in scenarios where frequent updates are desirable, such as in online learning or real-time applications.

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

>>>the batch size refers to the number of training examples used in each iteration to compute the gradient and update the model parameters. The choice of batch size impacts both the computational efficiency and the quality of the parameter updates during training.

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

>>>Momentum is controlled by a hyperparameter called the momentum coefficient, typically denoted by β. A higher value of β increases the influence of the accumulated momentum, leading to faster convergence but potentially overshooting the optimal solution. Careful tuning of the momentum coefficient is necessary to ensure optimal convergence and avoid instability.

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

>>>BGD computes the gradient of the loss function using the entire training dataset

SGD updates the parameters using the gradient computed from a single randomly selected training example at each iteration.

It computes the gradient using a small randomly selected subset (mini-batch) of the training data at each iteration.

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

>>>while converging if the learning rate will be higher then the GD will not be stable and this overshoot the minimum.while the smaller one will take the slowly directing towards GD and which will give the optimized GD.

Regularization:

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

>>>Regularization is a technique used in machine learning to prevent overfitting and improve the generalization performance of models. It involves adding a penalty term to the loss function during training

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

>>>L1 regularization adds the absolute values of the coefficients as the penalty term. The L1 penalty encourages sparsity by shrinking less important coefficients towards zero, effectively performing feature selection. L1 regularization can yield sparse models, with only a subset of features having non-zero coefficients.

L2 regularization (Ridge): L2 regularization adds the squared values of the coefficients as the penalty term. The L2 penalty encourages small values for all coefficients, effectively reducing the impact of all features. L2 regularization tends to produce models with small but non-zero coefficients for all features.

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

>>>this aims to mitigate the impact of multicollinearity and reduce the variance of the parameter estimates. Ridge regression introduces a tuning parameter, often denoted as λ (lambda), which controls the strength of the regularization. The larger the value of λ, the stronger the regularization effect.

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

>>>It provides a compromise between L1 regularization (Lasso) and L2 regularization (Ridge) by adding both penalties to the loss function.

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

>>>Regularization helps prevent overfitting in machine learning models by adding a penalty term to the loss function during training. Overfitting occurs when a model becomes too complex and fits the noise or random fluctuations in the training data

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

>>>early stopping monitors the model's performance during training and stops the training process when the performance on a validation dataset starts to deteriorate.

The concept behind early stopping is that, as the model continues to train, it becomes increasingly specialized to the training data and may start to overfit.

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

>>>Dropout regularization is a technique used in neural networks to prevent overfitting and improve generalization. It involves randomly "dropping out" (setting to zero) a proportion of the activations or connections in a layer during training.

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

>>>he choice of the regularization parameter should consider the problem complexity, the available data, and the desired trade-off between model complexity and generalization.

49. What is the difference between feature selection and regularization?

>>>Feature selection is the process of selecting a subset of relevant features from the available set of predictors. It aims to identify the most informative features and exclude irrelevant or redundant ones from the model

Regularization, on the other hand, is a technique that adds a penalty term to the loss function during model training. It discourages large coefficients and promotes simpler models by shrinking less important coefficients towards zero

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

>>>When the regularization strength is low, the model has more flexibility, lower bias, and higher variance. It can capture complex relationships in the data but may be prone to overfitting.

SVM:

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

>>>SVM aims to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class. The optimal hyperplane is selected based on support vectors, which are the data points that lie closest to the decision boundary.

52. How does the kernel trick work in SVM?

>>>The kernel trick in SVM is a technique that allows SVM to efficiently handle non-linearly separable data by implicitly mapping the data points to a higher-dimensional feature space. Instead of explicitly transforming the data, the kernel function calculates the dot product between the data points in the higher-dimensional space without explicitly computing the coordinates.

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

>>>because they determine the generalization ability of the model. They represent the critical data points that lie closest to the decision boundary and are most likely to influence the classification or regression predictions for new, unseen data. SVM relies on these support vectors to capture the underlying structure of the data and make accurate predictions.

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

>>>the distance between the decision boundary (hyperplane) and the closest data points from each class. SVM aims to find the hyperplane with the maximum margin, as this provides a larger separation between the classes and helps in achieving better generalization performance.

55. How do you handle unbalanced datasets in SVM?

>>>Class weighting: Assigning different weights to the classes to give higher importance to the minority class. This can be done by adjusting the penalty parameter (C) for each class during training.

Resampling: Modifying the dataset to balance the class distribution. This can involve oversampling the minority class by duplicating samples or undersampling the majority class by randomly selecting a subset of samples.

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

>>>Linear SVM: Linear SVM is suitable for datasets that can be separated by a linear hyperplane. It seeks to find the best linear decision boundary that maximizes the margin between the classes. Linear SVM performs well when the data is linearly separable or when a linear approximation is sufficient.

* Non-linear SVM: Non-linear SVM is designed to handle datasets that are not linearly separable. By utilizing the kernel trick, non-linear SVM transforms the data into a higher-dimensional feature space where a linear decision boundary can separate the data points effectively.

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

>>>The C-parameter in SVM is a regularization parameter that controls the trade-off between maximizing the margin and allowing misclassifications

58. Explain the concept of slack variables in SVM.

>>>SVM are introduced in soft margin SVM to handle non-linearly separable data or datasets with outliers. Slack variables allow for a certain degree of misclassification or margin violations. They represent the amount by which a data point is allowed to violate the margin or be misclassified.

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?

>>>A decision tree is a supervised machine learning algorithm that can be used for both classification and regression tasks. It works by recursively splitting the data into smaller and smaller subsets until each subset is homogeneous (i.e., all of the data points in the subset belong to the same class). The splitting process is guided by a decision rule, which is a test on a feature of the data.

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

>>>The decision rule that is used to split the data is chosen based on an impurity measure. Impurity measures quantify how mixed the data is in a given subset. The most common impurity measures are the Gini index and entropy. The Gini index is a measure of how likely it is that a randomly selected data point from a subset will be misclassified. Entropy is a measure of how much information is contained in a subset.

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

>>>The entropy of a subset is a measure of how much information is contained in the subset. It is calculated as follows:The Gini index is a measure of how likely it is that a randomly selected data point from a subset will be misclassified.

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

>>>where entropy(parent) is the entropy of the parent node and entropy(child) is the entropy of the child nodes.

The decision rule that maximizes the information gain is chosen to split the data. This process is repeated recursively until each subset is homogeneous.

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

>>>Ignoring the values,replace the missing values.

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

>>>Pruning is a technique used to reduce the complexity of a decision tree. It is done by removing branches from the tree that are not very informative. Pruning can improve the accuracy of the decision tree by reducing overfitting.

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

>>>A classification tree is used to predict a categorical variable, such as the class of an object. A regression tree is used to predict a continuous variable, such as the price of a house.

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

>>>The splitting rules are based on the impurity measures, such as the Gini index or entropy.

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

>>>Feature importance can be used to understand which features are most important for predicting the class of an object. This information can be used to improve the accuracy of the decision tree by focusing on the most important features.

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

>>>Ensemble techniques are a way of combining multiple models to improve the accuracy of predictions. Decision trees are a popular type of model that can be used in ensemble techniques.

Ensemble Techniques:

71. What are ensemble techniques in machine learning?

>>>There are many different ensemble techniques, but some of the most popular include:

* Bagging
* Boosting
* Random forests
* AdaBoost
* Gradient boosting

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

>>>Bagging is a type of ensemble technique that creates multiple models by sampling the training data with replacement. This means that some data points may be included in multiple models, while other data points may not be included in any models.

73. Explain the concept of bootstrapping in bagging.

>>>Bootstrapping is a technique that is used to create multiple samples of data from a single data set. This is done by randomly sampling the data with replacement.

74. What is boosting and how does it work?

>>>Boosting is a type of ensemble technique that creates multiple models sequentially. The first model is trained on the entire training data. The second model is then trained on the residuals of the first model. The residuals are the errors that were made by the first model.

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

>>>AdaBoost works by iteratively training decision trees to correct the errors of the previous trees. Gradient Boosting works by iteratively training decision trees to fit the residuals of the previous trees.

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

>>>Random forests are a type of ensemble technique that uses decision trees to improve the accuracy of predictions. Random forests are created by training multiple decision trees on different subsets of the training data.

The different decision trees in a random forest are trained using a technique called bagging. Bagging helps to reduce overfitting by creating a diverse set of models.

77. How do random forests handle feature importance?

>>>Random forests can be used to calculate the importance of each feature in the dataset. Feature importance is a measure of how important a feature is in determining the class of an object.

Feature importance in random forests is calculated by measuring the reduction in impurity that is caused by splitting the data on a particular feature

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

>>>Stacking is a type of ensemble technique that combines the predictions of multiple models to make a final prediction. Stacking is often used to combine the predictions of different types of models, such as decision trees, support vector machines, and neural networks.

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

>>>Reduced overfitting: Ensemble techniques can help to reduce overfitting by creating a diverse set of models.

* Improved accuracy: Ensemble techniques can often achieve higher accuracy than single models.
* Robustness: Ensemble techniques can be more robust to noise and outliers than single models.

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

>>>The optimal number of models in an ensemble depends on the specific data set and the application. However, there are some general guidelines that can be followed.

A good starting point is to start with a small number of models, such as 5 or 10. Then, you can increase the number of models until you start to see diminishing returns.