General Linear Model:

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

GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

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

linearity, homoskedasticity (constant variance), normality, and independence.

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

The GLM coefficients only show the multiplicative change in odds ratio. so if p1 is the risk of getting a high score for black defendants and p0 is the risk of getting a high score for white defendants, then exp(0.47721) shows (p1/(1-p1))/(p0/(1-p0)).

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

Univariate analysis is the analysis of one variable. Multivariate analysis is the analysis of more than one variable

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

The existence of an **interaction** means that the effect of one variable *depends* on the value of the other variable with which it interacts. If there isn't an interaction, then the value of the other variable *doesn't matter*.

Imagine we are looking at the adult height (say at 25) of a child based on the adult height of the father. We further include sex as an additional predictor variable, because men and women differ considerably in adult height. Let's imagine that there is no interaction between these two variables (which may be true, at least to a first approximation). We could then plot our model simply as two lines on a scatterplot of the data. We may want to use different colors or symbols / line styles for men vs. women, but at any rate we would see a football-ish (or rugby-ball-ish, depending on where you live) shaped cloud of points with two parallel lines going through it. The important part is that the lines are parallel; if someone asked you what the effect would be of the father being 1 inch (1 cm) taller, you would respond with 𝛽height�height. If they further asked you what the effect would be if the child were male or female, you would respond, 'that *doesn't matter*, you would expect them to be 𝛽height�height taller as an adult either way'. That is because the lines are parallel (with the same slope, 𝛽height�height) / there is no interaction. Imagine we are looking at the adult height (say at 25) of a child based on the adult height of the father. We further include sex as an additional predictor variable, because men and women differ considerably in adult height. Let's imagine that there is no interaction between these two variables (which may be true, at least to a first approximation). We could then plot our model simply as two lines on a scatterplot of the data. We may want to use different colors or symbols / line styles for men vs. women, but at any rate we would see a football-ish (or rugby-ball-ish, depending on where you live) shaped cloud of points with two parallel lines going through it. The important part is that the lines are parallel; if someone asked you what the effect would be of the father being 1 inch (1 cm) taller, you would respond with 𝛽height�height. If they further asked you what the effect would be if the child were male or female, you would respond, 'that *doesn't matter*, you would expect them to be 𝛽height�height taller as an adult either way'. That is because the lines are parallel (with the same slope, 𝛽height�height) / there is no interaction.

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

GLM supports both binary and multinomial classification. For binary classification, the response column can only have two levels; for multinomial classification, the response column will have more than two levels.

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

The design matrix is used in certain statistical models, e.g., the general linear model. It can contain indicator variables (ones and zeros) that indicate group membership in an ANOVA, or it can contain values of continuous variables.

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

we consider that if p-value < 0.05 for a certain variable then it is significant and has some relationship with your predictor

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

If interaction is present, then type II is inappropriate while type III can still be used, but results need to be interpreted with caution (in the presence of interactions, main effects are rarely interpretable). The anova and aov functions in R implement a sequential sum of squares (type I)

1. Explain the concept of deviance in a GLM.

Deviance is a goodness-of-fit metric for statistical models, particularly used for GLMs. It is defined as the difference between the Saturated and Proposed Models and can be thought as how much variation in the data does our Proposed Model account for. Therefore, the lower the deviance, the better the model.

Regression:

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

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

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

Simple linear regression has only one x and one y variable. Multiple linear regression has one y and two or more x variables. For instance, when we predict rent based on square feet alone that is simple linear regression.

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

a higher r-squared indicates more variability is explained by the model. For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model.

1. What is the difference between correlation and regression?

Correlation is a statistical measure that determines the association or co-relationship between two variables. Regression describes how to numerically relate an independent variable to the dependent variable. To represent a linear relationship between two variables.

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

The simple linear regression model is essentially a linear equation of the form y = c + b\*x; where y is the dependent variable (outcome), x is the independent variable (predictor), b is the slope of the line; also known as regression coefficient and c is the intercept; labeled as constant.

1. How do you handle outliers in regression analysis?

There are many possible approaches to dealing with outliers: removing them from the observations, treating them (for example, capping the extreme observations at a reasonable value), or using algorithms that are well-suited for dealing with such values on their own.

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

Ridge regression is a term used to refer to a linear regression model whose coefficients are estimated not by ordinary least squares (OLS), but by an estimator, called ridge estimator, that, albeit biased, has lower variance than the OLS estimator.

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

Heteroskedastic refers to a condition in which the variance of the residual term, or error term, in a regression model varies widely. The least squares method is a statistical technique to determine the line of best fit for a model, specified by an equation with certain parameters to observed data.

1. How do you handle multicollinearity in regression analysis?

Remove some of the highly correlated independent variables.

Linearly combine the independent variables, such as adding them together.

Partial least squares regression uses principal component analysis to create a set of uncorrelated components to include in the model.

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

A polynomial regression model is a machine learning model that can capture non-linear relationships between variables by fitting a non-linear regression line, which may not be possible with simple linear regression. It is used when linear regression models may not adequately capture the complexity of the relationship.

Loss function:

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

 A loss function is a measure of how good your prediction model does in terms of being able to predict the expected outcome(or value). We convert the learning problem into an optimization problem, define a loss function and then optimize the algorithm to minimize the loss function.

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

A convex function is one in which a line drawn between any two points on the graph lies on the graph or above it. There is only one requirement. A non-convex function is one in which a line drawn between any two points on the graph may cross additional points. It was described as “wavy.”

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

Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function. Fig: Regression Line. A larger MSE indicates that the data points are dispersed widely around its central moment (mean), whereas a smaller MSE suggests the opposite

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

MAE is calculated as the sum of absolute errors divided by the sample size: It is thus an arithmetic average of the absolute errors , where is the prediction and. the true value. Alternative formulations may include relative frequencies as weight factors.

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

Also called logarithmic loss, log loss or logistic loss. Each predicted class probability is compared to the actual class desired output 0 or 1 and a score/loss is calculated that penalizes the probability based on how far it is from the actual expected value.

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

Most machine learning algorithms use some sort of loss function in the process of optimization or finding the best parameters (weights) for your data. Importantly, the choice of the loss function is directly related to the activation function used in the output layer of your neural network.

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

Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.

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

the Huber loss is a loss function used in robust regression, that is less sensitive to outliers in data than the squared error loss. A variant for classification is also sometimes used.

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

A quantile is a value below which a fraction of samples in a group falls. Machine learning models work by minimizing (or maximizing) an objective function. As the name suggests, we apply the quantile regression loss function to predict quantiles. For a set of predictions, the loss will be its average.

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

For square loss, you will choose the estimated mean of y0, as the true mean minimizes square loss on average (where the average is taken across random samples of y0 subject to x=x0). For absolute loss, you will choose the estimated median

Optimizer (GD):

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

An optimizer is an algorithm or function that adapts the neural network's attributes, like learning rate and weights. Hence, it assists in improving the accuracy and reduces the total loss. But it is a daunting task to choose the appropriate weights for the model.

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

Gradient descent is an optimization algorithm which is commonly-used to train machine learning models and neural networks. Training data helps these models learn over time, and the cost function within gradient descent specifically acts as a barometer, gauging its accuracy with each iteration of parameter updates.

1. What are the different variations of Gradient Descent?

Batch Gradient Descent, Mini-batch Gradient Descent, and Stochastic Gradient Descent

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

Learning rate (also referred to as step size or the alpha) is the size of the steps that are taken to reach the minimum. This is typically a small value, and it is evaluated and updated based on the behavior of the cost function. High learning rates result in larger steps but risks overshooting the minimum.

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

Learning rate (also referred to as step size or the alpha) is the size of the steps that are taken to reach the minimum. This is typically a small value, and it is evaluated and updated based on the behavior of the cost function. High learning rates result in larger steps but risks overshooting the minimum.

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

Stochastic Gradient Descent is a drastic simplification of GD which overcomes some of its difficulties. Each iteration of SGD computes the gradient on the basis of one randomly chosen partition of the dataset which was shuffled, instead of using the whole part of the observations

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

Batch size is important because it affects both the training time and the generalization of the model. A smaller batch size allows the model to learn from each individual example but takes longer to train. A larger batch size trains faster but may result in the model not capturing the nuances in the data.

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

Momentum is a strategy for accelerating the convergence of the optimization process by including a momentum element in the update rule. This momentum factor assists the optimizer in continuing to go in the same direction even if the gradient changes direction or becomes zero.

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

Batch Gradient Descent can be used for smoother curves. SGD can be used when the dataset is large. Batch Gradient Descent converges directly to minima. SGD converges faster for larger datasets.

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

A too high learning rate will make the learning jump over minima but a too low learning rate will either take too long to converge or get stuck in an undesirable local minimum.

Regularization:

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

Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

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

L1 regularization penalizes the sum of absolute values of the weights, whereas L2 regularization penalizes the sum of squares of the weights.

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

ridge regression puts constraint on the coefficients (w). The penalty term (lambda) regularizes the coefficients such that if the coefficients take large values the optimization function is penalized. So, ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity.

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

In statistics and, in particular, in the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods.

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

L2 regularization, also known as Ridge Regression, adds a penalty term proportional to the square of the model's parameters. This encourages the model to use all of the parameters but to reduce their values, resulting in a model that is less complex and less prone to overfitting.

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

In Regularization by Early Stopping, we stop training the model when the performance of the model on the validation set is getting worse-increasing loss or decreasing accuracy or poorer values of the scoring metric.

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

Dropout is a regularization method approximating concurrent training of many neural networks with various designs. During training, some layer outputs are ignored or dropped at random

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

on the training set, we estimate several different Ridge regressions, with different values of the regularization parameter; on the validation set, we choose the best model (the regularization parameter which gives the lowest MSE on the validation set);

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

Feature selection, also known as feature subset selection, variable selection, or attribute selection. This approach removes the dimensions (e.g. columns) from the input data and results in a reduced data set for model inference. Regularization, where we are constraining the solution space while doing optimization.

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

If the algorithm is too simple (hypothesis with linear equation) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex (hypothesis with high degree equation) then it may be on high variance and low bias

SVM:

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

SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems

1. How does the kernel trick work in SVM?

The “trick” is that kernel methods represent the data only through a set of pairwise similarity comparisons between the original data observations **x**(with the original coordinates in the lower dimensional space), instead of explicitly applying the transformations *ϕ*(**x**) and representing the data by these transformed coordinates in the higher dimensional feature space.

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

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier.

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

it is the distance between the hyperplane and the observations closest to the hyperplane (support vectors). In SVM large margin is considered a good margin. There are two types of margins hard margin and soft margin

1. How do you handle unbalanced datasets in SVM?

We can use the make\_classification() function to define a synthetic imbalanced two-class classification dataset. We will generate 10,000 examples with an approximate 1:100 minority to majority class ratio. Once generated, we can summarize the class distribution to confirm that the dataset was created as we expected.

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

When we can easily separate data with hyperplane by drawing a straight line is Linear SVM. When we cannot separate data with a straight line we use Non – Linear SVM. In this, we have Kernel functions. They transform non-linear spaces into linear spaces.

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

C parameter adds a penalty for each misclassified data point. If c is small, the penalty for misclassified points is low so a decision boundary with a large margin is chosen at the expense of a greater number of misclassifications

1. Explain the concept of slack variables in SVM.

Slack variables are introduced to allow certain constraints to be violated. That is, certain train- ing points will be allowed to be within the margin. We want the number of points within the margin to be as small as possible, and of course we want their penetration of the margin to be as small as possible.

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

Hard Margin Classification only works if the data is linearly separable also Hard Margins are very sensitive to outliers. We can use soft margin classifications to avoid these issues. To avoid issues it is recommended to use a more flexible model with soft margin classifications

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

The weights obtained from svm. coef\_ represent the vector coordinates which are orthogonal to the hyperplane and their direction indicates the predicted class. The absolute size of the coefficients in relation to each other can then be used to determine feature importance for the data separation task.

Decision Trees:

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

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.

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

For each split, individually calculate the entropy of each child node.

Calculate the entropy of each split as the weighted average entropy of child nodes.

Select the split with the lowest entropy or highest information gain.

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

The gini impurity measures the frequency at which any element of the dataset will be mislabelled when it is randomly labeled. The minimum value of the Gini Index is 0. This happens when the node is pure, this means that all the contained elements in the node are of one unique clas

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

Information gain is the basic criterion to decide whether a feature should be used to split a node or not. The feature with the optimal split i.e., the highest value of information gain at a node of a decision tree is used as the feature for splitting the node.

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

One way to handle missing values is to add a preprocessing step to treat for them in our machine learning pipeline. In fact, many machine learning algorithms will require such a step to be added in order to prevent software failures during operations.

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

pruning is a technique associated with decision trees. Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances

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

The major difference between a classification tree and a regression tree is the nature of the variable to be predicted. In a regression tree, the variable is continuous rather than categorical.

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

The reason for this is that a Decision tree splits the data based on a feature value and this value would remain constant throughout for one decision boundary e.g., x=2 or y=3 where x and y are two different features. Whereas in a linear classifier, a decision boundary could be for instance: y=mx+c.

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

Feature importance scores play an important role in a predictive modeling project, including providing insight into the data, insight into the model, and the basis for dimensionality reduction and feature selection that can improve the efficiency and effectiveness of a predictive model on the problem.

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

Ensemble techniques address the limitations of individual decision trees by combining multiple models to make more accurate predictions. The fundamental idea is that the collective wisdom of several models is often superior to that of a single model

Ensemble Techniques:

1. What are ensemble techniques in machine learning?

In this ensemble technique, machine learning professionals use a number of models for making predictions about each data point. The predictions made by different models are taken as separate votes. Subsequently, the prediction made by most models is treated as the ultimate prediction

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

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.

1. Explain the concept of bootstrapping in bagging.

Bootstrapping is a sampling method, where a sample is chosen out of a set, using the replacement method. The learning algorithm is then run on the samples selected

1. What is boosting and how does it work?

Boosting creates an ensemble model by combining several weak decision trees sequentially. It assigns weights to the output of individual trees. Then it gives incorrect classifications from the first decision tree a higher weight and input to the next tree

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

AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

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

Random forest algorithm is an ensemble learning technique combining numerous classifiers to enhance a model's performance. Random Forest is a supervised machine-learning algorithm made up of decision trees. Random Forest is used for both classification and regression problems.

1. How do random forests handle feature importance?

The final feature importance, at the Random Forest level, is it's average over all the trees. The sum of the feature's importance value on each trees is calculated and divided by the total number of trees: RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model

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

Stacking is one of the most popular ensemble machine learning techniques used to predict multiple nodes to build a new model and improve model performance. Stacking enables us to train multiple models to solve similar problems, and based on their combined output, it builds a new model with improved performance.

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

Ensemble methods offer several advantages over single models, such as improved accuracy and performance, especially for complex and noisy problems. They can also reduce the risk of overfitting and underfitting by balancing the trade-off between bias and variance, and by using different subsets and features of the data.

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

The method consists of building multiple models independently and returning the average of the prediction of all the models. In general, the combined output is better than an individual output because variance is reduced