**Assignment -4**

General Linear Model:

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

Ans: The General Linear Model (GLM) is a useful framework for comparing how several variables affect different [continuous variables](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/discrete-vs-continuous-variables/). In its simplest form, GLM is described as:

Data = Model + Error

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

Ans: There are four assumptions associated with a linear regression model

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

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

## Ans: The formula for the general linear model is: [general linear model formula](https://www.statisticshowto.com/wp-content/uploads/2016/11/general-linear-model-formula.png)

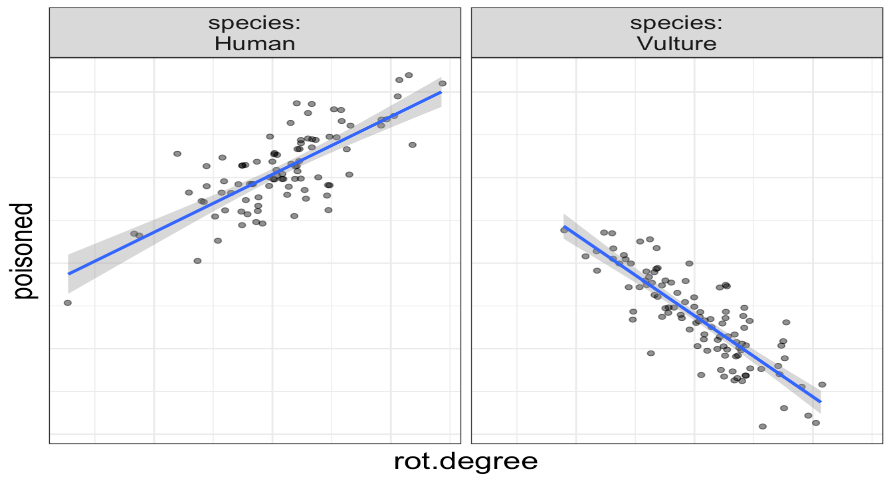
Where:

* [yhat General Linear Model ](https://www.statisticshowto.com/wp-content/uploads/2016/11/yhat.png) = the [dependent variable](https://www.statisticshowto.com/dependent-variable-definition/) (also called the predicted, [explanatory, or response variable](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/explanatory-variable/)).
* β0 = the intercept — always a constant (i.e. the value never changes within the model).
* β1 = a weight or [slope](https://www.statisticshowto.com/what-is-a-slope/) (also called a coefficient). Determines how much weight one variable contributes to the model. If everything in the equation holds constant, β0 gives the predicted change in Y for a unit change in X.
* X = a variable.

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

Ans: The GLM Univariate procedure provides regression analysis and analysis of variance for one dependent variable by one or more factors and/or variables. The GLM Multivariate procedure provides regression analysis and analysis of variance for multiple dependent variables by one or more factor variables or covariates.

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

Ans: An interaction occurs when the effect of one variable (e.g., the effect of waking early on productivity) depends on another variable (what time you go to bed).   


if the slopes are not parallel, there is an interaction present in the data.And, it makes sense. If a slope represents the form of the relationship (e.g., increased rot means more poisoning), and if that slope changes from one group to the next, then it’s clear that the form of the relationship depends on group.

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

Ans: categorical predictors in a GLM is handled by logistics regression. In regression analyses, categorical predictors are represented using 0 and 1 for dichotomous variables or using indicator (or dummy) variables for ordinal or categorical variables.

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

Ans: Using the General Linear Model (GLM), the statistical model specified in a design matrix is **compared with the measured time course at each voxel**. The comparison of the model and the data is expressed as an R or F value for each voxel which tells how good the overall model fits or explains the data.

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

Ans: You can either do an **asymptotic chi-square test**  vs a chi-square or use anova .  The last column of the coefficients matrix is called "Pr(>|t|)" and holds the pvalues of the factors used in the model.

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

Ans: Type I SS Type I SS are order-dependent (hierarchicalEach effect is adjusted for all other effects that appear earlier in the model, but not for any effects that appear later in the model. TypeII SS are the reduction in the SSE as a result of adding the effect to a model that contains all other effects except effects that contain the effect being tested. Type III SS are identical to those of Type II SS when the design is balanced.

10. Explain the concept of deviance in a GLM.

Ans: [Deviance is a **measure of goodness of fit of a generalized linear model**](https://www.bing.com/ck/a?!&&p=7ac10ce2d717202bJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTU5Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=9.%09.+Explain+the+concept+of+deviance+in+a+GLM.&u=a1aHR0cHM6Ly93d3cudGhlYW5hbHlzaXNmYWN0b3IuY29tL3ItZ2xtLW1vZGVsLWZpdC8&ntb=1)[1](https://www.bing.com/ck/a?!&&p=437bb6ff5e943d3eJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTU5OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=9.%09.+Explain+the+concept+of+deviance+in+a+GLM.&u=a1aHR0cHM6Ly93d3cudGhlYW5hbHlzaXNmYWN0b3IuY29tL3ItZ2xtLW1vZGVsLWZpdC8&ntb=1). [It is a measure of badness of fit, with higher numbers indicating worse fit](https://www.bing.com/ck/a?!&&p=ecc473b24f13f49dJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTU5OQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=9.%09.+Explain+the+concept+of+deviance+in+a+GLM.&u=a1aHR0cHM6Ly93d3cudGhlYW5hbHlzaXNmYWN0b3IuY29tL3ItZ2xtLW1vZGVsLWZpdC8&ntb=1)[1](https://www.bing.com/ck/a?!&&p=9f84656fb39a481cJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYwMA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=9.%09.+Explain+the+concept+of+deviance+in+a+GLM.&u=a1aHR0cHM6Ly93d3cudGhlYW5hbHlzaXNmYWN0b3IuY29tL3ItZ2xtLW1vZGVsLWZpdC8&ntb=1). [R reports two forms of deviance – the null deviance and the residual deviance](https://www.bing.com/ck/a?!&&p=ec04973dc59943cdJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYwMQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=9.%09.+Explain+the+concept+of+deviance+in+a+GLM.&u=a1aHR0cHM6Ly93d3cudGhlYW5hbHlzaXNmYWN0b3IuY29tL3ItZ2xtLW1vZGVsLWZpdC8&ntb=1)[1](https://www.bing.com/ck/a?!&&p=f6acc221c9fa980eJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYwMg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=9.%09.+Explain+the+concept+of+deviance+in+a+GLM.&u=a1aHR0cHM6Ly93d3cudGhlYW5hbHlzaXNmYWN0b3IuY29tL3ItZ2xtLW1vZGVsLWZpdC8&ntb=1). [The residual deviance tells us how well the response variable can be predicted by the specific model that we fit with p predictor variables](https://www.bing.com/ck/a?!&&p=e8dc51a273e81b3fJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYwMw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=9.%09.+Explain+the+concept+of+deviance+in+a+GLM.&u=a1aHR0cHM6Ly93d3cuc3RhdG9sb2d5Lm9yZy9pbnRlcnByZXQtZ2xtLW91dHB1dC1pbi1yLw&ntb=1). [The lower the value, the better the model is able to predict the value of the response variable](https://www.bing.com/ck/a?!&&p=178f641c156c7d8fJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYwNQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=9.%09.+Explain+the+concept+of+deviance+in+a+GLM.&u=a1aHR0cHM6Ly93d3cuc3RhdG9sb2d5Lm9yZy9pbnRlcnByZXQtZ2xtLW91dHB1dC1pbi1yLw&ntb=1)

Regression:

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

Ans: [Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables](https://www.bing.com/ck/a?!&&p=abcfd72f045d9e45JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcyNw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+regression+analysis+and+what+is+its+purpose%3f&u=a1aHR0cHM6Ly9jb3Jwb3JhdGVmaW5hbmNlaW5zdGl0dXRlLmNvbS9yZXNvdXJjZXMvZGF0YS1zY2llbmNlL3JlZ3Jlc3Npb24tYW5hbHlzaXMv&ntb=1). [The purpose of regression analysis is to describe the relationship between variables, and to estimate or predict the value of one variable using the known values of other va](https://www.bing.com/ck/a?!&&p=ffba20c442a3d984JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcyOQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+regression+analysis+and+what+is+its+purpose%3f&u=a1aHR0cHM6Ly93d3cuZzIuY29tL2FydGljbGVzL3JlZ3Jlc3Npb24tYW5hbHlzaXM&ntb=1)riable.

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

Ans: [Simple linear regression has only one x and one y variable, while multiple linear regression has one y and two or more x variables](https://www.bing.com/ck/a?!&&p=6cd23ae2998bfdfbJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc2OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+simple+linear+regression+and+multiple+linear+regression%3f&u=a1aHR0cHM6Ly9ibG9nLnV3Z2IuZWR1L2JhbnNhbGcvc3RhdGlzdGljcy1kYXRhLWFuYWx5dGljcy9saW5lYXItcmVncmVzc2lvbi93aGF0LWlzLWRpZmZlcmVuY2UtYmV0d2Vlbi1zaW1wbGUtbGluZWFyLWFuZC1tdWx0aXBsZS1saW5lYXItcmVncmVzc2lvbnMv&ntb=1). [Simple linear regression is used when you have only one predictor, or X variable, predicting the response or Y variable](https://www.bing.com/ck/a?!&&p=f7e42dff3691db0cJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc3MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+simple+linear+regression+and+multiple+linear+regression%3f&u=a1aHR0cHM6Ly93d3cuY291cnNlYmIuY29tLzIwMTcvMDcvMjQvc2ltaWxhcml0aWVzLWRpZmZlcmVuY2VzLXNpbXBsZS1saW5lYXItcmVncmVzc2lvbi1hbmFseXNpcy1tdWx0aXBsZS1yZWdyZXNzaW9uLWFuYWx5c2lzLw&ntb=1). Multiple linear regression is used when you have multiple X predictors that [all contribute to predicting Y](https://www.bing.com/ck/a?!&&p=1831ec5a2daa77bcJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc3Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+simple+linear+regression+and+multiple+linear+regression%3f&u=a1aHR0cHM6Ly93d3cuY291cnNlYmIuY29tLzIwMTcvMDcvMjQvc2ltaWxhcml0aWVzLWRpZmZlcmVuY2VzLXNpbXBsZS1saW5lYXItcmVncmVzc2lvbi1hbmFseXNpcy1tdWx0aXBsZS1yZWdyZXNzaW9uLWFuYWx5c2lzLw&ntb=1) .[Simple linear regression occurs in 2 dimensions, while multiple linear regression can occur in an infinite number of dimensions](https://www.bing.com/ck/a?!&&p=389c17030eedcf65JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc3NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+simple+linear+regression+and+multiple+linear+regression%3f&u=a1aHR0cHM6Ly93d3cuY291cnNlYmIuY29tLzIwMTcvMDcvMjQvc2ltaWxhcml0aWVzLWRpZmZlcmVuY2VzLXNpbXBsZS1saW5lYXItcmVncmVzc2lvbi1hbmFseXNpcy1tdWx0aXBsZS1yZWdyZXNzaW9uLWFuYWx5c2lzLw&ntb=1)

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

Ans: [R-squared is a statistical measure that indicates how well the regression model fits the observed data values](https://www.bing.com/ck/a?!&&p=50358cf598ce301cJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1MA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+interpret+the+R-squared+value+in+regression%3f&u=a1aHR0cHM6Ly9jb3Jwb3JhdGVmaW5hbmNlaW5zdGl0dXRlLmNvbS9yZXNvdXJjZXMvZGF0YS1zY2llbmNlL3Itc3F1YXJlZC8&ntb=1). [The most common interpretation of r-squared is how well the regression model explains observed data. For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model. Generally, a higher r-squared indicates more variability is explained by the model](https://www.bing.com/ck/a?!&&p=c87848ff8e9e9aadJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+interpret+the+R-squared+value+in+regression%3f&u=a1aHR0cHM6Ly9jb3Jwb3JhdGVmaW5hbmNlaW5zdGl0dXRlLmNvbS9yZXNvdXJjZXMvZGF0YS1zY2llbmNlL3Itc3F1YXJlZC8&ntb=1).

14. What is the difference between correlation and regression?

Ans: [Regression and correlation are both used to describe the relationship between two variables](https://www.bing.com/ck/a?!&&p=19ce8b40a19af3b1JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTg1MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+correlation+and+regression%3f&u=a1aHR0cHM6Ly93d3cuY2FwdGVycmEuY29tL3Jlc291cmNlcy9jb3JyZWxhdGlvbi12cy1yZWdyZXNzaW9uLw&ntb=1). [R**egression shows a cause-and-effect relationship and can predict the value of one variable based on another**](https://www.bing.com/ck/a?!&&p=8eeceecfaf7cbdbdJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTg1Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+correlation+and+regression%3f&u=a1aHR0cHM6Ly93d3cuc3RhdG9sb2d5Lm9yZy9jb3JyZWxhdGlvbi12cy1yZWdyZXNzaW9uLw&ntb=1). [Correlation only shows the degree and direction of the relationship and does not imply causation](https://www.bing.com/ck/a?!&&p=e49965795d8aa499JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTg1Ng&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+correlation+and+regression%3f&u=a1aHR0cHM6Ly93d3cuc3RhdG9sb2d5Lm9yZy9jb3JyZWxhdGlvbi12cy1yZWdyZXNzaW9uLw&ntb=1). [Regression is represented by a line and the order of the variables matters](https://www.bing.com/ck/a?!&&p=c59c4aaad49b0a77JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTg1OQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+correlation+and+regression%3f&u=a1aHR0cHM6Ly9ibG9nLmNhcHRlcnJhLmNvbS9jb3JyZWxhdGlvbi12cy1yZWdyZXNzaW9uLw&ntb=1). [Correlation is represented by a single data point and the order of the variables does not matter](https://www.bing.com/ck/a?!&&p=b92785740978c8dcJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTg2Mg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+the+difference+between+correlation+and+regression%3f&u=a1aHR0cHM6Ly9ibG9nLmNhcHRlcnJhLmNvbS9jb3JyZWxhdGlvbi12cy1yZWdyZXNzaW9uLw&ntb=1)

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

Ans: Regression describes the relationship between independent variable (x) and dependent variable (y), **Beta zero (intercept) refer to a value of Y when X=0, while Beta one (regression coefficient, also we call it the slope) refer to the change in variable Y when the variable X change one unit**.

16. How do you handle outliers in regression analysis?

Ans: Outlier Handling:

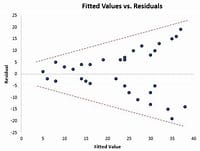
Trimming/Remove the outliers In this technique, we remove the outliers from the dataset. Quantile based flooring and capping- In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value. Mean/Median imputation : As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value.

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

Ans: [**The difference between ordinary least squares and ridge regression is that** ridge regression uses a biased estimator that reduces the variance of the coefficients, while ordinary least squares uses an unbiased estimator that minimizes the mean squared error](https://www.bing.com/ck/a?!&&p=4596e19c3a7828f6JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY0Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+the+difference+between+ridge+regression+and+ordinary+least+squares+regression%3f&u=a1aHR0cHM6Ly93d3cubGlua2VkaW4uY29tL3B1bHNlL2xpbmVhci1yZWdyZXNzaW9uLW9yZGluYXJ5LWxlYXN0LXNxdWFyZXMtdnMtcmlkZ2UtYWZuYW4tYWJvdS1lbC13YWZh&ntb=1).

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

Ans:

[[](https://www.bing.com/images/search?q=imgurl%3ahttps%3a%2f%2ffourpillarfreedom.com%2fwp-content%2fuploads%2f2019%2f02%2fhet2.jpg&s=30&view=detailv2&iss=sbi&idpp=imgqna&vt=1&idpview=singleimage&idpbck=1)](https://www.bing.com/images/search?q=imgurl%3ahttps%3a%2f%2ffourpillarfreedom.com%2fwp-content%2fuploads%2f2019%2f02%2fhet2.jpg&s=30&view=detailv2&iss=sbi&idpp=imgqna&vt=1&idpview=singleimage&idpbck=1)

[Heteroscedasticity refers to the unequal variance of the residuals in a regression analysis](https://www.bing.com/ck/a?!&&p=eccbf27d7c81b7d3JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY5OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+heteroscedasticity+in+regression+and+how+does+it+affect+the+model%3f&u=a1aHR0cHM6Ly93d3cuc3RhdG9sb2d5Lm9yZy9oZXRlcm9zY2VkYXN0aWNpdHktcmVncmVzc2lvbi8&ntb=1). [It violates the assumption of the classical linear regression model and causes the regression coefficient estimates to be unreliable and have higher variance](https://www.bing.com/ck/a?!&&p=0aa2898e03b1489dJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+heteroscedasticity+in+regression+and+how+does+it+affect+the+model%3f&u=a1aHR0cHM6Ly93d3cuc3RhdG9sb2d5Lm9yZy9oZXRlcm9zY2VkYXN0aWNpdHktcmVncmVzc2lvbi8&ntb=1). [This can lead to higher risk of type I error, which is rejecting a true null hypothesis](https://www.bing.com/ck/a?!&&p=b8bbfff63c47b456JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwNQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+heteroscedasticity+in+regression+and+how+does+it+affect+the+model%3f&u=a1aHR0cHM6Ly93d3cuc3RhdG9sb2d5Lm9yZy9oZXRlcm9zY2VkYXN0aWNpdHktcmVncmVzc2lvbi8&ntb=1). [Heteroscedasticity can also affect the inclusion or exclusion of observations, especially when the sample size is small](https://www.bing.com/ck/a?!&&p=20b0d5428a113be6JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwOA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+heteroscedasticity+in+regression+and+how+does+it+affect+the+model%3f&u=a1aHR0cHM6Ly9pdGZlYXR1cmUuY29tL2NvcnJlbGF0aW9uLWFuZC1yZWdyZXNzaW9uLWFuYWx5c2lzL2ludHJvZHVjdGlvbi1yZWFzb25zLWFuZC1jb25zZXF1ZW5jZXMtb2YtaGV0ZXJvc2NlZGFzdGljaXR5&ntb=1).

19. How do you handle multicollinearity in regression analysis?

Ans: Following ways to handle multicollinearity in regression analysis.

* Linearly combine the independent variables, such as adding them together.
* Remove one or more variables showing a high correlation.
* Use techniques such as principal components analysis (PCA) or partial least squares regression (PLS).
* Center the predictors, that is subtract the mean of one series from each value.
* Use advanced forms of regression analysis such as LASSO and Ridge regression

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

Ans: [Polynomial Regression is used when the relationship between the dependent and independent variables is not linear](https://www.bing.com/ck/a?!&&p=4f959d6fdb3001fdJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+polynomial+regression+and+when+is+it+used%3f&u=a1aHR0cHM6Ly9tZWRpdW0uY29tL2FuYWx5dGljcy12aWRoeWEvdW5kZXJzdGFuZGluZy1wb2x5bm9taWFsLXJlZ3Jlc3Npb24tNWFjMjViOTcwZTE4&ntb=1). [It is generally used when the points in the data are not captured by the Linear Regression Model and the Linear Regression fails in describing the best result clearly](https://www.bing.com/ck/a?!&&p=c0bfd545181f25c8JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc2Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+polynomial+regression+and+when+is+it+used%3f&u=a1aHR0cHM6Ly9tZWRpdW0uY29tL2FuYWx5dGljcy12aWRoeWEvdW5kZXJzdGFuZGluZy1wb2x5bm9taWFsLXJlZ3Jlc3Npb24tNWFjMjViOTcwZTE4&ntb=1). [Polynomial regression is used to fit a regression model that describes the relationship between one or more predictor variables and a numeric response variable](https://www.bing.com/ck/a?!&&p=e4bf22a251230ec5JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc2NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+polynomial+regression+and+when+is+it+used%3f&u=a1aHR0cHM6Ly93d3cuc3RhdG9sb2d5Lm9yZy90eXBlcy1vZi1yZWdyZXNzaW9uLw&ntb=1). [It involves fitting a polynomial function to the data points to obtain a curve that represents the relationship between the variables](https://www.bing.com/ck/a?!&&p=15269992000e8e1eJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc2Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+polynomial+regression+and+when+is+it+used%3f&u=a1aHR0cHM6Ly9zcGFya2J5ZXhhbXBsZXMuY29tL21hY2hpbmUtbGVhcm5pbmcvZGlmZmVyZW5jZS1iZXR3ZWVuLWxpbmVhci1yZWdyZXNzaW9uLWFuZC1wb2x5bm9taWFsLXJlZ3Jlc3Npb24v&ntb=1).

Loss function:

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

Ans: [The loss function is a **method of evaluating how well a machine learning algorithm models a featured data set**](https://www.bing.com/ck/a?!&&p=a8091f0064f90c84JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc0OQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+a+loss+function+and+what+is+its+purpose+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cuYW5hbHl0aWNzdmlkaHlhLmNvbS9ibG9nLzIwMjIvMDgvYmFzaWMtaW50cm9kdWN0aW9uLXRvLWxvc3MtZnVuY3Rpb25zLw&ntb=1) [The loss function indicates how inaccurate the model is at determining the relationship between x and y](https://www.bing.com/ck/a?!&&p=38b5aec8db2a725eJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+a+loss+function+and+what+is+its+purpose+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cuYW5hbHl0aWNzdmlkaHlhLmNvbS9ibG9nLzIwMjIvMDgvYmFzaWMtaW50cm9kdWN0aW9uLXRvLWxvc3MtZnVuY3Rpb25zLw&ntb=1). [It is a metric that the model utilizes to put a number to its performance, which is how close or far the model has made its prediction to the actual label](https://www.bing.com/ck/a?!&&p=3b59f27f390c624eJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+a+loss+function+and+what+is+its+purpose+in+machine+learning%3f&u=a1aHR0cHM6Ly9jb2RlZGFtbi5jb20vbmV3cy9tYWNoaW5lLWxlYXJuaW5nL2xvc3MtZnVuY3Rpb25zLWluLW1hY2hpbmUtbGVhcm5pbmc&ntb=1).

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

Ans: A**convex** optimization problem is a problem where all of the constraints are**convex functions,** and the objective is**a convex function** if minimising, or a concave function if maximising. A**convex function** can be described as a smooth surface with a single global minimum while non-convex function has many local minimum .Making the algorithm difficult to learn.

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

## Ans: Mean Squared Error : the average Squared error between actual and predicted values. MSE formula = (1/n) \* Σ(yactual – yforecast)2 Where:n = number of items,Σ = [summation notation](https://www.statisticshowto.com/calculus-definitions/summation-notation-sigma-function/),Actual = original or observed y-value,Forecast = y-value from regression.

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

## Ans: Mean Absolute Error:The Mean Absolute Error(MAE) is the [average](https://www.statisticshowto.com/average-value-of-a-function/#def)  the difference between predicted value and “true” value. The formula is: MSE formula = (1/n) \* Σ|(yactual – yforecast) | **Where**:n = the number of errors,Σ = [summation symbol](https://www.statisticshowto.com/what-is-sigma-summation-notation/) (which means “add them all up”),| actual – forecast | = the absolute errors.

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

Ans: Binary Cross Entropy is the negative average of the log of corrected predicted probabilities which can be either 0 or 1. It then calculates the score that penalizes the probabilities based on the distance from the expected value. That means how close or far from the actual value.

**loss = -(y \* log (y) + (1- y) \* log (1 — y))**.

* Predict the probabilities
* we will calculate the log value for each of the corrected probabilities
* Then cal use a negative average of the values.
* The value of the negative average of corrected probabilities is our Log loss or Binary cross-entropy .

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

Ans: Loss functions are classified into two classes based on the type of learning task

* Regression Models: predict continuous values.
* Classification Models: predict the output from a set of finite categorical values.

**Regression Losses**

* **Mean Squared Error (MSE)** **/ Quadratic Loss / L2 Loss:** When the cost function is far away from its minimal value, squaring the error will penalize the model more and thus helping in reaching the minimal value faster.It is convex function,but sensitive to outliers.
* **Mean Absolute Error (MAE) / La Loss:** It is non-convex,MAE is that is robust to outliers.
* Huber Loss / Smooth Mean Absolute Error:It is the combination of MSE and MAE. It takes the good properties of both the loss functions by being less sensitive to outliers and differentiable at minima.When the error is smaller, the MSE part of the Huber is utilized and when the error is large, the MAE part of Huber loss is used.

### **Classification Losses**

**Cross-Entropy Loss/** **Negative Log Likelihood**. It is the commonly used loss function for classification. Cross-entropy loss progress as the predicted probability diverges from the actual label.

**Hinge Loss**:  Also known as Multi-class**SVM Loss**. Hinge loss is applied for maximum-margin classification, prominently for support vector machines.

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

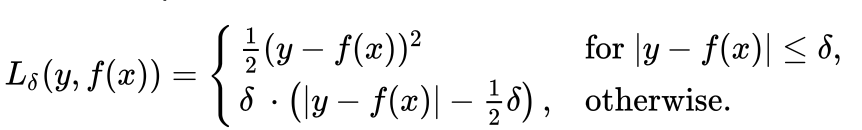
Ans: Regularization means restricting a model to avoid overfitting by shrinking the coefficient estimates to zero. When a model suffers from overfitting, we should control the model's complexity. Technically, regularization avoids overfitting by adding a penalty to the model's loss function: **Regularization = Loss Function + Penalty**

The commonly used [regularization techniques](https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/) are :

1. Lasso Regularization – L1 Regularization
2. Ridge Regularization – L2 Regularization
3. Elastic Net Regularization – L1 and L2 Regularization

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

Ans:



The Huber loss function is a compromise between the MSE and L1 loss functions. Like the MSE loss, it penalizes large errors more than small errors, but like the L1 loss, it is less sensitive to outliers than the MSE loss.

By choosing an appropriate value for delta, we can control the balance between the quadratic and linear regions of the loss function. For small values of delta, the Huber loss behaves like the MSE loss and is more sensitive to outliers. For large values of delta, the Huber loss behaves like the L1 loss and is less sensitive to outliers.

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

Ans: [The quantile loss or pinball loss function, is a **metric used to assess the accuracy of a quantile forecast**](https://www.bing.com/ck/a?!&&p=eee005d4440b7003JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYzMg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+quantile+loss+and+when+is+it+used%3f&u=a1aHR0cHM6Ly93d3cubG9rYWQuY29tL3BpbmJhbGwtbG9zcy1mdW5jdGlvbi1kZWZpbml0aW9u&ntb=1)[.](https://www.bing.com/ck/a?!&&p=6be92a07c32a65d7JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYzMw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+quantile+loss+and+when+is+it+used%3f&u=a1aHR0cHM6Ly93d3cubG9rYWQuY29tL3BpbmJhbGwtbG9zcy1mdW5jdGlvbi1kZWZpbml0aW9u&ntb=1) [A quantile is the value below which a fraction of observations in a group falls](https://www.bing.com/ck/a?!&&p=e593c3008db33f2dJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYzNQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+quantile+loss+and+when+is+it+used%3f&u=a1aHR0cHM6Ly93d3cuZXZlcmdyZWVuaW5ub3ZhdGlvbnMuY28vYmxvZy1xdWFudGlsZS1sb3NzLWZ1bmN0aW9uLWZvci1tYWNoaW5lLWxlYXJuaW5nLw&ntb=1). [The quantile regression loss function is applied to predict quantiles](https://www.bing.com/ck/a?!&&p=568000b845e9effeJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYzNw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+quantile+loss+and+when+is+it+used%3f&u=a1aHR0cHM6Ly93d3cuZXZlcmdyZWVuaW5ub3ZhdGlvbnMuY28vYmxvZy1xdWFudGlsZS1sb3NzLWZ1bmN0aW9uLWZvci1tYWNoaW5lLWxlYXJuaW5nLw&ntb=1). [It is an objective function that translates the problem we are trying to solve into a mathematical formula to be minimized by the model](https://www.bing.com/ck/a?!&&p=007dd73c363cff58JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYzOQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+What+is+quantile+loss+and+when+is+it+used%3f&u=a1aHR0cHM6Ly93d3cuZXZlcmdyZWVuaW5ub3ZhdGlvbnMuY28vYmxvZy1xdWFudGlsZS1sb3NzLWZ1bmN0aW9uLWZvci1tYWNoaW5lLWxlYXJuaW5nLw&ntb=1).

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

Ans: **The squared loss function results in an arithmetic mean - unbiased estimator, and the absolute-value loss function results in a median -unbiased estimator**

Optimizer (GD):

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

Ans: [Optimizers in machine learning are **algorithms or functions that adjust the attributes of the neural network, such as weights and learning rates, to minimize the loss function and improve the accuracy**](https://www.bing.com/ck/a?!&&p=4aed2d0c9fdc9307JmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcxNA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+an+optimizer+and+what+is+its+purpose+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cudXBncmFkLmNvbS9ibG9nL3R5cGVzLW9mLW9wdGltaXplcnMtaW4tZGVlcC1sZWFybmluZy8&ntb=1). [Optimizers are used in various stages of a machine learning project, such as data preparation, hyperparameter tuning, and model selection](https://www.bing.com/ck/a?!&&p=d46fffe4ba05c67fJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcxNw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+an+optimizer+and+what+is+its+purpose+in+machine+learning%3f&u=a1aHR0cHM6Ly9tYWNoaW5lbGVhcm5pbmdtYXN0ZXJ5LmNvbS93aHktb3B0aW1pemF0aW9uLWlzLWltcG9ydGFudC1pbi1tYWNoaW5lLWxlYXJuaW5nLw&ntb=1)

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

Ans: [Gradient descent is an **algorithm that finds a local minimum of a function by taking repeated steps in the opposite direction of the gradient**](https://www.bing.com/ck/a?!&&p=024ef7d7f30c797aJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+Gradient+Descent+(GD)+and+how+does+it+work%3f&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvR3JhZGllbnRfZGVzY2VudA&ntb=1).[It is often used in machine learning and neural networks to optimize the parameters of a model and minimize a cost function](https://www.bing.com/ck/a?!&&p=a828b95614c0c51bJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1NA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+Gradient+Descent+(GD)+and+how+does+it+work%3f&u=a1aHR0cHM6Ly9idWlsdGluLmNvbS9kYXRhLXNjaWVuY2UvZ3JhZGllbnQtZGVzY2VudA&ntb=1)[.](https://www.bing.com/ck/a?!&&p=09ef1e088694d5cbJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+Gradient+Descent+(GD)+and+how+does+it+work%3f&u=a1aHR0cHM6Ly9idWlsdGluLmNvbS9kYXRhLXNjaWVuY2UvZ3JhZGllbnQtZGVzY2VudA&ntb=1) [The gradient of a function points in the direction of steepest ascent, while the opposite direction is the steepest](https://www.bing.com/ck/a?!&&p=0945cdcbf3533f0dJmltdHM9MTY4OTAzMzYwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc1OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+Gradient+Descent+(GD)+and+how+does+it+work%3f&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvR3JhZGllbnRfZGVzY2VudA&ntb=1)descent.

33. What are the different variations of Gradient Descent?

Ans:

* Mini-batch gradient descent: a small batch of training examples is used to compute the gradient and update the parameters at each iteration.
* Stochastic Gradient Descent (SGD):  only one training example is used to compute the gradient and update the parameters at each iteration.
* Batch Gradient Descent: n batch gradient descent, To update the model parameter values like weight and bias, the entire training dataset is used to compute the gradient and update the parameters at each iteration.
* **Momentum-based Gradient Descent:**In momentum-based gradient descent, Momentum is a variant of gradient descent that incorporates information from the previous weight updates to help the algorithm converge more quickly to the optimal solution.
* Adagrad: In this variant, the learning rate is adaptively adjusted for each parameter based on the historical gradient information. This allows for larger updates for infrequent parameters and smaller updates for frequent parameters.
* RMSprop: In this variant, the learning rate is adaptively adjusted for each parameter based on the moving average of the squared gradient. This helps the algorithm to converge faster in the presence of noisy gradients.
* Adam: Adam stands for adaptive moment estimation, it combines the benefits of Momentum-based Gradient Descent, Adagrad, and RMSprop the learning rate is adaptively adjusted for each parameter based on the moving average of the gradient and the squared gradient, which allows for faster convergence and better performance on non-convex optimization problems

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

Ans:  No value of the learning rate that will guarantee convergence to the minimum value of the error function (assuming global) value of a function. A learning rate that is too big or too small affects the loss during training. If learning rate (α ) is too large you may "shoot over" your target and end up bouncing around the search space without converging. If α is too small your convergence will be slow and you could end up stuck on a plateau or a local minimum.The standard gradient descent procedure uses a fixed learning rate (e.g. 0.01) that is determined by trial and error. . Learning rate annealing entails starting with a high learning rate and then gradually reducing the learning rate linearly during training. The learning rate can decrease to a value close to 0.The idea behind this method is to quickly descend to a range of acceptable weights, and then do a deeper dive within this acceptable range.

A graph of loss and loss

Description automatically generated

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

Ans: A local optimum refers to a point within the domain of a function where the function attains the lowest (or highest) value in its local neighborhood. In the Non-convex function since the function can take multiple shapes hence it can have multiple valleys so in this case it becomes hard to find the global minima.

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

Ans: It addresses the computational inefficiency of traditional Gradient Descent methods when dealing with large datasets in machine learning projects.

In SGD, instead of using the entire dataset for each iteration, only a single random training example (or a small batch) is selected to calculate the gradient and update the model parameters. This random selection introduces randomness into the optimization process, hence the term “stochastic” in stochastic Gradient Descent

The advantage of using SGD is its computational efficiency, especially when dealing with large datasets

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

Ans: n Mini-batch gradient descent a small batch of training examples is used to compute the gradient and update the parameters at each iteration. This can be a good compromise between batch gradient descent and Stochastic Gradient Descent, as it can be faster than batch gradient descent and less noisy than Stochastic Gradient Descent.

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

Ans: In momentum-based gradient descent, Momentum is a variant of gradient descent that incorporates information from the previous weight updates to help the algorithm converge more quickly to the optimal solution. Momentum adds a term to the weight update that is proportional to the running average of the past gradients, allowing the algorithm to move more quickly in the direction of the optimal solution. The updates to the parameters are based on the current gradient and the previous updates. This can help prevent the optimization process from getting stuck in local minima and reach the global minimum faster.

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

Ans:

* Mini-batch gradient descent: a small batch of training examples is used to compute the gradient and update the parameters at each iteration.
* Stochastic Gradient Descent (SGD):  only one training example is used to compute the gradient and update the parameters at each iteration.
* Batch Gradient Descent: n batch gradient descent, To update the model parameter values like weight and bias, the entire training dataset is used to compute the gradient and update the parameters at each iteration.

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

Ans: If learning rate (α ) is too large you may "shoot over" your target and end up bouncing around the search space without converging. If α is too small your convergence will be slow and you could end up stuck on a plateau or a local minimum.

Regularization:

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

Ans: Regularization is a technique used to reduce errors by fitting the function appropriately on the given training set ,avoiding overfitting and outlier removal.

In machine learning noramlly overfitting and underfitting occurs.

Overfitting is a phenomenon that occurs when a [Machine Learning](https://www.geeksforgeeks.org/machine-learning/) model is constrained to the training set and not able to perform well on unseen data.

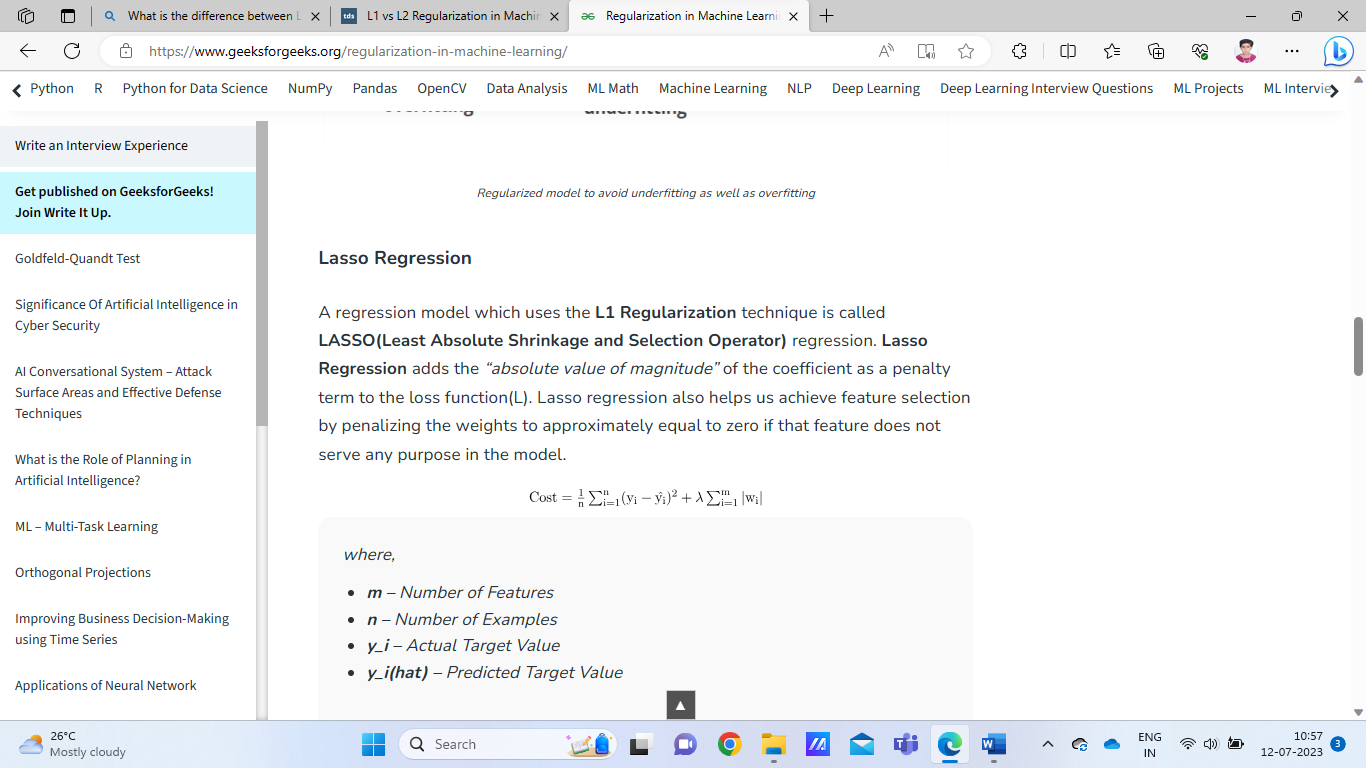
Underfitting on the other hand is the case when our model is not able to learn even the basic patterns available in the dataset. In the case of the underfitting model is unable to perform well even on the training data hence we cannot expect it to perform well on the validation data. This is the case when we are supposed to increase the complexity of the model or add more features to the feature set.

The commonly used [regularization techniques](https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/) are:

1. Lasso Regularization – L1 Regularization
2. Ridge Regularization – L2 Regularization
3. Elastic Net Regularization – L1 and L2 Regularization

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

Ans **L1 Regularization**- **LASSO(Least Absolute Shrinkage and Selection Operator)** regression adds the “absolute value of magnitude” of the coefficient as a penalty term to the loss function(L). Lasso regression also helps us achieve feature selection by penalizing the weights to approximately equal to zero if that feature does not serve any purpose in the model. Effective for outlier removal



here,

* m – Number of Features
* n – Number of Examples
* y\_i – Actual Target Value
* y\_i(hat) – Predicted Target Value

**L2 regularization** - **Ridge regression** adds the “squared magnitude” of the coefficient as a penalty term to the loss function(L).Mainly used for avoiding overfitting.

A math equation with black text

Description automatically generated

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

Ans: **L2 regularization** - **Ridge regression** adds the “squared magnitude” of the coefficient as a penalty term to the loss function(L).Mainly used for avoiding overfitting.

A math equation with black text

Description automatically generated

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

Ans: This model is a combination of L1 as well as L2 regularization. That implies that we add the absolute norm of the weights as well as the squared measure of the weights. With the help of an extra [hyperparameter](https://www.geeksforgeeks.org/hyperparameter-tuning/) that controls the ratio of the L1 and L2 regularization. 

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

Ans: Overfitting happens when model is too complicated to generalise for new data. When your model fits your data perfectly, it is unlikely to fit new data well.

Underfit happens when your model is not complicated enough. This introduces a bias in the model, such that there is systematic deviation from the true underlying estimator.

Regularization attemts to reduce the variance of the estimator by simplifying it, something that will increase the bias, in such a way that the expected error decreases. Often this is done in cases when the problem is ill-posed, e.g. when the number of parameters is greater than the number of samples.

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

Ans: Regularization by early stopping can be done either by dividing the dataset into training and test sets and then using cross-validation on the training set or by dividing the dataset into training, validation and test sets, in which case cross-validation, is not required. In early stopping, the algorithm is trained using the training set, and the point at which to stop training is determined from the validation set. Training error and validation error are analyzed. The training error steadily decreases while the validation error decreases until a point, after which it increases. This is because, during training, the learning model starts to overfit the training data. This causes the training error to decrease while the validation error increases. So a model with better validation set error can be obtained if the parameters that give the least validation set error are used. Each time the error on the validation set decreases, a copy of the model parameters is stored. When the training algorithm terminates, these parameters which give the least validation set error are finally returned and not the last modified parameters.

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

Ans: When a fully-connected layer has a large number of neurons, co-adaptation is more likely to happen. Co-adaptation refers to when multiple neurons in a layer extract the same, or very similar, hidden features from the input data. This can happen when the connection weights for two different neurons are nearly identical.

Problem is:

* Wastage of machine’s resources when computing the same output.
* If many neurons are extracting the same features, it adds more significance to those features for our model. This leads to overfitting if the duplicate extracted features are specific to only the training set.

We use dropout while training the NN to minimize co-adaptation. In dropout, we randomly shut down some fraction of a layer’s neurons at each training step by zeroing out the neuron values. The fraction of neurons to be zeroed out is known as the dropout rate(rd). The remaining neurons have their values multiplied by (1/1-rd)so that the overall sum of the neuron values remains the same.

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

Ans:  with train-val-test splits, we can choose the regularization parameter $lambda $ as follows:

* 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);
* on the test set, we check how much [overfitting](https://statlect.com/machine-learning/overfitting) we have done by doing [model selection](https://statlect.com/fundamentals-of-statistics/model-selection-criteria) on the validation set.

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

Ans: The goal of feature selection techniques in machine learning is to find the best set of features that allows one to build optimized models of studied phenomena. Regularization is a technique used to reduce errors by fitting the function appropriately on the given training set and avoiding overfitting.

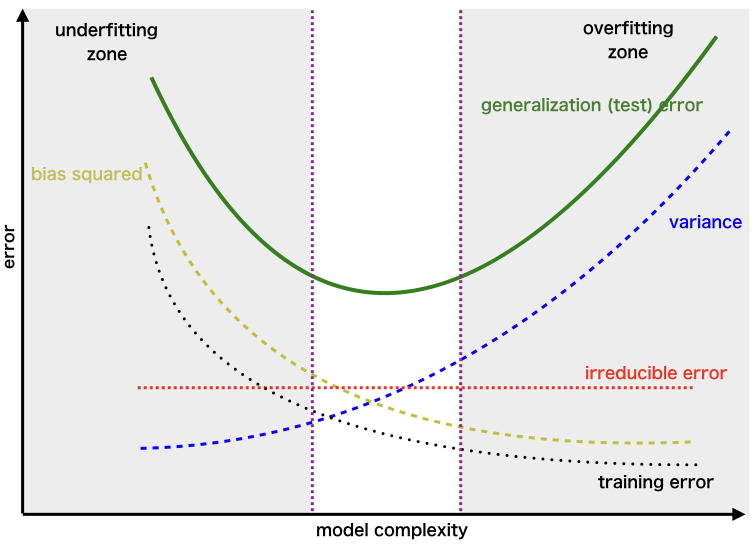
50. What is the trade-off between bias and variance in regularized models?

## Ans: **Bias Variance Tradeoff**

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. In the latter condition, the new entries will not perform well. Bias Variance Trade-off in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.

We try to optimize the value of the total error for the model by using the [Bias-Variance](https://www.geeksforgeeks.org/bias-vs-variance-in-machine-learning/) Tradeoff. 

The best fit will be given by the hypothesis on the tradeoff point. The error to complexity graph to show trade-off is given as –



Region for the Least Value of Total Error

 This is referred to as the best point chosen for the training of the algorithm which gives low error in training as well as testing data.

SVM:

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

Ans Supervised Machine Learning Algorithm used for classification and/or regression. The main idea behind SVMs is to find a hyperplane that maximally separates the different classes in the training data. This is done by finding the hyperplane that has the largest margin, which is defined as the distance between the hyperplane and the closest data points from each class. Once the hyperplane is determined, new data can be classified by determining on which side of the hyperplane it falls. In 2-dimensional space, this hyper-plane is nothing but a line. In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. Support Vector Machine for Multi-Class Problems To perform SVM on multi-class problems, we can create a binary classifier for each class of the data. The two results of each classifier will be :

* The data point belongs to that class OR
* The data point does not belong to that class.

52. How does the kernel trick work in SVM?

Ans: [The kernel trick is a technique that transforms low dimensional input space into a higher dimensional space, where non-separable problems become separable](https://www.bing.com/ck/a?!&&p=8e1cc1e6ada8a124JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY5Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+kernel+trick+work+in+SVM%3f&u=a1aHR0cHM6Ly9kYXRhYXNwaXJhbnQuY29tL3N2bS1rZXJuZWxzLw&ntb=1). [The kernel trick allows the SVM algorithm to deal with non-linear separation problems by using a kernel function that computes the similarity between pairs of points in the higher dimensional feature space](https://www.bing.com/ck/a?!&&p=3ab86a601e939391JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+kernel+trick+work+in+SVM%3f&u=a1aHR0cHM6Ly93d3cuZ2Vla3Nmb3JnZWVrcy5vcmcvaW50cm9kdWN0aW9uLXRvLXN1cHBvcnQtdmVjdG9yLW1hY2hpbmVzLXN2bS8&ntb=1)[2](https://www.bing.com/ck/a?!&&p=6c32600ea9d0eb7fJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+kernel+trick+work+in+SVM%3f&u=a1aHR0cHM6Ly93d3cuZ2Vla3Nmb3JnZWVrcy5vcmcvaW50cm9kdWN0aW9uLXRvLXN1cHBvcnQtdmVjdG9yLW1hY2hpbmVzLXN2bS8&ntb=1). [The kernel function does not need to explicitly calculate the transformed feature representation, which saves computation time and memory](https://www.bing.com/ck/a?!&&p=16f9466039a7cf91JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+kernel+trick+work+in+SVM%3f&u=a1aHR0cHM6Ly93d3cuZ2Vla3Nmb3JnZWVrcy5vcmcvaW50cm9kdWN0aW9uLXRvLXN1cHBvcnQtdmVjdG9yLW1hY2hpbmVzLXN2bS8&ntb=1)

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

Ans: [The kernel trick is a technique that transforms low dimensional input space into a higher dimensional space, where non-separable problems become separable](https://www.bing.com/ck/a?!&&p=8e1cc1e6ada8a124JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY5Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+kernel+trick+work+in+SVM%3f&u=a1aHR0cHM6Ly9kYXRhYXNwaXJhbnQuY29tL3N2bS1rZXJuZWxzLw&ntb=1). [The kernel trick allows the SVM algorithm to deal with non-linear separation problems by using a kernel function that computes the similarity between pairs of points in the higher dimensional feature space](https://www.bing.com/ck/a?!&&p=3ab86a601e939391JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+kernel+trick+work+in+SVM%3f&u=a1aHR0cHM6Ly93d3cuZ2Vla3Nmb3JnZWVrcy5vcmcvaW50cm9kdWN0aW9uLXRvLXN1cHBvcnQtdmVjdG9yLW1hY2hpbmVzLXN2bS8&ntb=1)[2](https://www.bing.com/ck/a?!&&p=6c32600ea9d0eb7fJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+kernel+trick+work+in+SVM%3f&u=a1aHR0cHM6Ly93d3cuZ2Vla3Nmb3JnZWVrcy5vcmcvaW50cm9kdWN0aW9uLXRvLXN1cHBvcnQtdmVjdG9yLW1hY2hpbmVzLXN2bS8&ntb=1). [The kernel function does not need to explicitly calculate the transformed feature representation, which saves computation time and memory](https://www.bing.com/ck/a?!&&p=16f9466039a7cf91JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+kernel+trick+work+in+SVM%3f&u=a1aHR0cHM6Ly93d3cuZ2Vla3Nmb3JnZWVrcy5vcmcvaW50cm9kdWN0aW9uLXRvLXN1cHBvcnQtdmVjdG9yLW1hY2hpbmVzLXN2bS8&ntb=1)

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

Ans: [The purpose of margins in SVM is to maximize the distance between the hyperplane and the closest data points](https://www.bing.com/ck/a?!&&p=f7bbbb2a814c2d9eJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+the+margin+in+SVM+and+its+impact+on+model+performance.&u=a1aHR0cHM6Ly93d3cuamF2YXRwb2ludC5jb20vbWFjaGluZS1sZWFybmluZy1zdXBwb3J0LXZlY3Rvci1tYWNoaW5lLWFsZ29yaXRobQ&ntb=1) [The distance between the vectors and the hyperplane is called the margin](https://www.bing.com/ck/a?!&&p=20b592780544b3ebJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+the+margin+in+SVM+and+its+impact+on+model+performance.&u=a1aHR0cHM6Ly93d3cuamF2YXRwb2ludC5jb20vbWFjaGluZS1sZWFybmluZy1zdXBwb3J0LXZlY3Rvci1tYWNoaW5lLWFsZ29yaXRobQ&ntb=1). [The hyperplane with maximum margin is called the optimal hyperplane](https://www.bing.com/ck/a?!&&p=ca0171fca35e2742JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2OQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+the+margin+in+SVM+and+its+impact+on+model+performance.&u=a1aHR0cHM6Ly93d3cuamF2YXRwb2ludC5jb20vbWFjaGluZS1sZWFybmluZy1zdXBwb3J0LXZlY3Rvci1tYWNoaW5lLWFsZ29yaXRobQ&ntb=1). [A large margin can avoid the effect of random noise and reduce overfitting](https://www.bing.com/ck/a?!&&p=e17664621d145144JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+the+margin+in+SVM+and+its+impact+on+model+performance.&u=a1aHR0cHM6Ly9tZWRpdW0uY29tL2dlZWtjdWx0dXJlL3N2bS13aHktbG9vay1mb3ItbWF4aW11bS1tYXJnaW4tOWY2NTBlYjI5Y2Ux&ntb=1). [A larger margin will lead to a smaller VC dimension, reduce the number of potential classifiers, and, therefore, reduce the possibility of generalization error](https://www.bing.com/ck/a?!&&p=08b9effad4d38696JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+the+margin+in+SVM+and+its+impact+on+model+performance.&u=a1aHR0cHM6Ly9tZWRpdW0uY29tL2dlZWtjdWx0dXJlL3N2bS13aHktbG9vay1mb3ItbWF4aW11bS1tYXJnaW4tOWY2NTBlYjI5Y2Ux&ntb=1)[.](https://www.bing.com/ck/a?!&&p=f4ed62f093e28ddbJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3NA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+the+margin+in+SVM+and+its+impact+on+model+performance.&u=a1aHR0cHM6Ly9tZWRpdW0uY29tL2dlZWtjdWx0dXJlL3N2bS13aHktbG9vay1mb3ItbWF4aW11bS1tYXJnaW4tOWY2NTBlYjI5Y2Ux&ntb=1)

55. How do you handle unbalanced datasets in SVM?

Ans: The strength of SVM is the robustness of its algorithm and the capability to integrate with kernel-based learning that results in a more flexible analysis and optimized solution.

Class-weighted SVM is designed to deal with unbalanced data by assigning higher misclassification penalties to training instances of the minority class.

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

Ans: [The main difference between a linear SVM and a non-linear SVM is that the linear SVM follows a simple rule, where the dot product between two features of its input is equal to the linear combination of its input](https://www.bing.com/ck/a?!&&p=3d4e0846e8ff85efJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY5Ng&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+the+difference+between+linear+SVM+and+non-linear+SVM%3f&u=a1aHR0cHM6Ly93d3cuYmFlbGR1bmcuY29tL2NzL3N2bS12cy1uZXVyYWwtbmV0d29yaw&ntb=1). [Linear SVM is used when data can be easily separated with a hyperplane by drawing a straight line](https://www.bing.com/ck/a?!&&p=e09ab8f80d78a095JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY5OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+the+difference+between+linear+SVM+and+non-linear+SVM%3f&u=a1aHR0cHM6Ly93d3cuYWl0dWRlLmNvbS9zdm0tZGlmZmVyZW5jZS1iZXR3ZWVuLWxpbmVhci1hbmQtbm9uLWxpbmVhci1tb2RlbHMv&ntb=1). Linear SVM is used for linearly separable data, where the decision boundary is a straight line. Linear SVM finds the hyperplane that separates the data points with the largest margin. The hyperplane is defined by the following equation: W^Tx+b=0

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

Ans:  This parameter controls the amount of regularization applied to the data. Large values of C mean low regularization which in turn causes the training data to fit very well (may cause overfitting). Lower values of C mean higher regularization which causes the model to be more tolerant of errors (may lead to lower accuracy).

58. Explain the concept of slack variables in SVM.

Ans: In an [optimization problem](https://en.wikipedia.org/wiki/Optimization_problem), a slack variable is a variable that is added to an [inequality constraint](https://en.wikipedia.org/wiki/Inequality_constraint) to transform it into an equality. Introducing a slack variable replaces an inequality constraint with an equality constraint and a non-negativity constraint on the slack variable.

* If a slack variable associated with a constraint is zero at a particular [candidate solution](https://en.wikipedia.org/wiki/Candidate_solution), the [constraint](https://en.wikipedia.org/wiki/Constraint_(mathematics)) is [binding](https://en.wikipedia.org/wiki/Binding_constraint) there, as the constraint restricts the possible changes from that point.
* If a slack variable is positive at a particular candidate solution, the constraint is [non-binding](https://en.wikipedia.org/wiki/Non-binding_constraint) there, as the constraint does not restrict the possible changes from that point.
* If a slack variable is negative at some point, the point is [infeasible](https://en.wikipedia.org/wiki/Feasible_region) (not allowed), as it does not satisfy the constraint.

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

Ans:

Hard Margin: The maximum-margin hyperplane or the hard margin hyperplane is a hyperplane that properly separates the data points of different categories without any misclassifications.

Soft Margin: When the data is not perfectly separable or contains outliers, SVM permits a soft margin technique. Each data point has a slack variable introduced by the soft-margin SVM formulation, which softens the strict margin requirement and permits certain misclassifications or violations. It discovers a compromise between increasing the margin and reducing violations.

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

Ans: The equation for the linear hyperplane can be written as:



The vector W represents the normal vector to the hyperplane. i.e the direction perpendicular to the hyperplane. The parameter **b** in the equation represents the offset or distance of the hyperplane from the origin along the normal vector **w**.

The distance between a data point x\_i and the decision boundary can be calculated as: A math equation with numbers and symbols

Description automatically generated

The equation for the linear hyperplane can be written as:

A group of black letters

Description automatically generated

The vector W represents the normal vector to the hyperplane. i.e the direction perpendicular to the hyperplane. The parameter **b** in the equation represents the offset or distance of the hyperplane from the origin along the normal vector **w**.

Decision Trees:

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

Ans: [A decision tree is a **diagrammatic approach to making a decision based on the outcomes of different choices**](https://www.bing.com/ck/a?!&&p=a6b1301f25ef7c5bJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwOA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+a+decision+tree+and+how+does+it+work%3f&u=a1aHR0cHM6Ly93d3cuaW5kZWVkLmNvbS9jYXJlZXItYWR2aWNlL2NhcmVlci1kZXZlbG9wbWVudC9kZWNpc2lvbi10cmVlLWhvdy1pdC13b3Jrcw&ntb=1). [It is drawn in a structure similar to a tree, where the choices branch out from nodes that depict various possibilities](https://www.bing.com/ck/a?!&&p=36d22b0b01e81fafJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcxMg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+a+decision+tree+and+how+does+it+work%3f&u=a1aHR0cHM6Ly93d3cuaW5kZWVkLmNvbS9jYXJlZXItYWR2aWNlL2NhcmVlci1kZXZlbG9wbWVudC9kZWNpc2lvbi10cmVlLWhvdy1pdC13b3Jrcw&ntb=1)[.](https://www.bing.com/ck/a?!&&p=960124672090bc3bJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcxMw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+a+decision+tree+and+how+does+it+work%3f&u=a1aHR0cHM6Ly93d3cuaW5kZWVkLmNvbS9jYXJlZXItYWR2aWNlL2NhcmVlci1kZXZlbG9wbWVudC9kZWNpc2lvbi10cmVlLWhvdy1pdC13b3Jrcw&ntb=1) [The decision-making process is carried out by choosing or discarding an option based on the statistical concept of probability](https://www.bing.com/ck/a?!&&p=de5b77a30c885955JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcxNQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+a+decision+tree+and+how+does+it+work%3f&u=a1aHR0cHM6Ly93d3cud2FsbHN0cmVldG1vam8uY29tL2RlY2lzaW9uLXRyZWUv&ntb=1).[A decision tree helps in concluding by allowing the interpretation of data visually](https://www.bing.com/ck/a?!&&p=549a09fe92798941JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcxOA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+a+decision+tree+and+how+does+it+work%3f&u=a1aHR0cHM6Ly93d3cud2FsbHN0cmVldG1vam8uY29tL2RlY2lzaW9uLXRyZWUv&ntb=1)

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

Ans: The decision tree operates by analyzing the data set to predict its classification. It commences from the tree’s root node, where the algorithm views the value of the root attribute compared to the attribute of the record in the actual data set. Based on the comparison, it proceeds to follow the branch and move to the next node.

The algorithm repeats this action for every subsequent node by comparing its attribute values with those of the sub-nodes and continuing the process further. It repeats until it reaches the leaf node of the tree.

We can use entropy or Gini Index.

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

#### Ans: Entropy:Entropy is the measure of the degree of randomness or uncertainty in the dataset. In the case of classifications, It measures the randomness based on the distribution of class labels in the dataset.

The entropy for a subset of the original dataset having K number of classes for the ith node can be defined as:



Where,

* S is the dataset sample.
* k is the particular class from K classes
* p(k) is the proportion of the data points that belong to class k to the total number of data points in dataset sample S. 
* Here p(i,k) should not be equal to zero.

**Important points related to Entropy:**

1. The entropy is 0 when the dataset is completely homogeneous, meaning that each instance belongs to the same class. It is the lowest entropy indicating no uncertainty in the dataset sample.
2. when the dataset is equally divided between multiple classes, the entropy is at its maximum value. Therefore, entropy is highest when the distribution of class labels is even, indicating maximum uncertainty in the dataset sample.
3. Entropy is used to evaluate the quality of a split. The goal of entropy is to select the attribute that minimizes the entropy of the resulting subsets, by splitting the dataset into more homogeneous subsets with respect to the class labels.
4. The highest information gain attribute is chosen as the splitting criterion (i.e., the reduction in entropy after splitting on that attribute), and the process is repeated recursively to build the decision tree.

#### Gini Impurity or index:

Gini Impurity is a score that evaluates how accurate a split is among the classified groups. The Gini Impurity evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes. In this case, we want to have a Gini index score as low as possible.



Here,

* pi is the proportion of elements in the set that belongs to the ith category.

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

Ans: Information gain measures the reduction in entropy or variance that results from splitting a dataset based on a specific property. It is used in decision tree algorithms to determine the usefulness of a feature by partitioning the dataset into more homogeneous subsets with respect to the class labels or target variable. The higher the information gain, the more valuable the feature is in predicting the target variable.

The information gain of an attribute A, with respect to a dataset S, is calculated as follows:



where

* A is the specific attribute or class label
* |H| is the entropy of dataset sample S
* |HV| is the number of instances in the subset S that have the value v for attribute A

Information gain measures the reduction in entropy or variance achieved by partitioning the dataset on attribute A. The attribute that maximizes information gain is chosen as the splitting criterion for building the decision tree.

Information gain is used in both classification and regression decision trees

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

Ans: For handling missing values C4.5 algorithm is used .

1. During feature selection, consider assigning the missing value to each possible branch.
2. Evaluate the effect of assigning the missing value to each branch by measuring the reduction in classification error.
3. Choose the feature and branch that result in the highest reduction in classification error when the missing value is assigned.

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

Ans: Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances. Decision trees are the most susceptible out of all the machine learning algorithms to overfitting and effective pruning can reduce this likelihood.

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

## Ans:

## **Classification tree:**

A classification tree is an algorithm where the target variable is categorical. The algorithm is then used to identify the “Class” within which the target variable is most likely to fall. Classification trees are used when the dataset needs to be split into classes that belong to the response variable(like yes or no)

## **Regression tree:**

A Regression tree is an algorithm where the target variable is continuous and the tree is used to predict its value. Regression trees are used when the response variable is continuous. For example, if the response variable is the temperature of the day.

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

Ans: [Decision boundaries are the line or surface that separates the decision space into two or more regions](https://www.bing.com/ck/a?!&&p=7c63a5147cc171a4JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY0Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+interpret+the+decision+boundaries+in+a+decision+tree%3f&u=a1aHR0cHM6Ly93d3cuYWlmb3JhbnlvbmUub3JnL2dsb3NzYXJ5L2RlY2lzaW9uLWJvdW5kYXJ5&ntb=1). [Each region is associated with a particular class label. When a new data point is presented, the model predicts the class label of the point based on which region it falls into](https://www.bing.com/ck/a?!&&p=fb32707a72536db8JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY0OQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+interpret+the+decision+boundaries+in+a+decision+tree%3f&u=a1aHR0cHM6Ly93d3cuYWlmb3JhbnlvbmUub3JnL2dsb3NzYXJ5L2RlY2lzaW9uLWJvdW5kYXJ5&ntb=1)[1](https://www.bing.com/ck/a?!&&p=c83288bd2c242206JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1MA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+interpret+the+decision+boundaries+in+a+decision+tree%3f&u=a1aHR0cHM6Ly93d3cuYWlmb3JhbnlvbmUub3JnL2dsb3NzYXJ5L2RlY2lzaW9uLWJvdW5kYXJ5&ntb=1). [Decision tree does not learn to draw a decision boundary. It tries to split the tree based on the maximum information gain point](https://www.bing.com/ck/a?!&&p=380d2ee6e30271faJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+interpret+the+decision+boundaries+in+a+decision+tree%3f&u=a1aHR0cHM6Ly9zdGFja292ZXJmbG93LmNvbS9xdWVzdGlvbnMvNjA5NjA2OTIvZmluZC1kaXN0YW5jZS10by1kZWNpc2lvbi1ib3VuZGFyeS1pbi1kZWNpc2lvbi10cmVlcw&ntb=1)[2](https://www.bing.com/ck/a?!&&p=1bc43ef829c84c23JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1Mg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+interpret+the+decision+boundaries+in+a+decision+tree%3f&u=a1aHR0cHM6Ly9zdGFja292ZXJmbG93LmNvbS9xdWVzdGlvbnMvNjA5NjA2OTIvZmluZC1kaXN0YW5jZS10by1kZWNpc2lvbi1ib3VuZGFyeS1pbi1kZWNpc2lvbi10cmVlcw&ntb=1). [Decision boundary of a decision tree is determined by overlapping orthogonal half-planes (representing the result of each subsequent decision)](https://www.bing.com/ck/a?!&&p=b7637d9f91de07a1JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+interpret+the+decision+boundaries+in+a+decision+tree%3f&u=a1aHR0cHM6Ly9kYXRhc2NpZW5jZS5zdGFja2V4Y2hhbmdlLmNvbS9xdWVzdGlvbnMvMjA1NDgvd2h5LWRlY2lzaW9uLXRyZWUtYm91bmRhcnktZm9ybXMtYS1zcXVhcmUtc2hhcGUtYW5kLXN2bS1hLWNpcmN1bGFyLW92YWwtb25l&ntb=1).

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

Ans: [Decision tree algorithms provide feature importance scores based on reducing the criterion used to select split points](https://www.bing.com/ck/a?!&&p=ffbe470af29e4e99JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+the+role+of+feature+importance+in+decision+trees%3f&u=a1aHR0cHM6Ly93d3cuYmFlbGR1bmcuY29tL2NzL21sLWZlYXR1cmUtaW1wb3J0YW5jZQ&ntb=1). [The feature importance values of a single tree can be computed by initializing an array feature\_importances of all zeros with size n\_features, and traversing the tree to compute the error reduction of each internal node that splits on feature i, multiplied by the number of samples that were routed to the node](https://www.bing.com/ck/a?!&&p=8ac28f30288189ffJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3NA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+the+role+of+feature+importance+in+decision+trees%3f&u=a1aHR0cHM6Ly9zdGFja292ZXJmbG93LmNvbS9xdWVzdGlvbnMvMTU4MTAzMzkvaG93LWFyZS1mZWF0dXJlLWltcG9ydGFuY2VzLWluLXJhbmRvbWZvcmVzdGNsYXNzaWZpZXItZGV0ZXJtaW5lZA&ntb=1) and adding this values to feature importance.

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

Ans: Ensemble learning helps improve machine learning results by combining several models. It is mainly used for reducing overfitting , removing instability problem and local optimum problem in decision trees. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in orderto decrease variance (bagging), bias (boosting), or improve predictions (stacking). Bagging-Bagging is bootstrap aggregation. One way to reduce the variance of an estimate is to average together multiple estimates. For example, we can train M different trees on different subsets of the data (chosen randomly with replacement) and compute the ensemble:it is mostly used when our priority is decreasing variance instead of decreasing bias.

Ensemble Techniques:

71. What are ensemble techniques in machine learning?

Ans: Ensemble learning helps improve machine learning results by combining several models. It is mainly used for reducing overfitting , removing instability problem and local optimum problem in decision trees. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking).

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

Ans: Bagging stands for bootstrap aggregation. One way to reduce the variance of an estimate is to average together multiple estimates. For example, we can train M different trees on different subsets of the data (chosen randomly with replacement) and compute the ensemble:

A number of mathematical symbols

Description automatically generated

Bagging uses bootstrap sampling to obtain the data subsets for training the base learners. For aggregating the outputs of base learners, bagging uses voting for classification and averaging for regression. it is mostly used when our priority is decreasing variance instead of decreasing bias.

73. Explain the concept of bootstrapping in bagging.

Ans: [Bootstrapping is a sampling method used in bagging](https://www.bing.com/ck/a?!&&p=77771a159185784dJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+bootstrapping+in+bagging.&u=a1aHR0cHM6Ly9jb3Jwb3JhdGVmaW5hbmNlaW5zdGl0dXRlLmNvbS9yZXNvdXJjZXMvZGF0YS1zY2llbmNlL2JhZ2dpbmctYm9vdHN0cmFwLWFnZ3JlZ2F0aW9uLw&ntb=1). [It is a resampling method that generates different subsets of the training dataset by selecting data points at random and with replacement](https://www.bing.com/ck/a?!&&p=f43ea7eee5f59eafJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3NA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+bootstrapping+in+bagging.&u=a1aHR0cHM6Ly93d3cuaWJtLmNvbS90b3BpY3MvYmFnZ2luZw&ntb=1). [The bootstrapping technique uses sampling with replacements to make the selection procedure completely random](https://www.bing.com/ck/a?!&&p=2269b09e773438f9JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3Ng&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+bootstrapping+in+bagging.&u=a1aHR0cHM6Ly9jb3Jwb3JhdGVmaW5hbmNlaW5zdGl0dXRlLmNvbS9yZXNvdXJjZXMvZGF0YS1zY2llbmNlL2JhZ2dpbmctYm9vdHN0cmFwLWFnZ3JlZ2F0aW9uLw&ntb=1). [Bootstrapping is the method of randomly creating samples of data out of a population with replacement to estimate a population parameter](https://www.bing.com/ck/a?!&&p=9935e2bf5ca9773bJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+concept+of+bootstrapping+in+bagging.&u=a1aHR0cHM6Ly93d3cuc2ltcGxpbGVhcm4uY29tL3R1dG9yaWFscy9tYWNoaW5lLWxlYXJuaW5nLXR1dG9yaWFsL2JhZ2dpbmctaW4tbWFjaGluZS1sZWFybmluZw&ntb=1).

74. What is boosting and how does it work?

Ans: Boosting is a sequential method–it aims to prevent a wrong base model from affecting the final output. Instead of combining the base models, the method focuses on building a new model that is dependent on the previous one. A new model tries to remove the errors made by its previous one. Each of these models is called weak learners. The final model (aka strong learner) is formed by getting the weighted mean of all the weak learners.

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

Ans:

| Algorithms | Gradient Boosting | AdaBoost |
| --- | --- | --- |
| Year | – | 1995 |
| Handling Categorical Variables | May require preprocessing like one-hot encoding | No |
| Speed/Scalability | Moderate | Fast |
| Memory Usage | Moderate | Low |
| Regularization | NO | No |
| Parallel Processing | No | No |
| GPU Support | No | No |
| Feature Importance | Available | Available |

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

Ans: Random Forest Models can be thought of as **BAGG**ing, with a slight tweak. When deciding where to split and how to make decisions, BAGGed Decision Trees have the full disposal of features to choose from. Therefore, although the bootstrapped samples may be slightly different, the data is largely going to break off at the same features throughout each model. In contrary, Random Forest models decide where to split based on a random selection of features.

77. How do random forests handle feature importance?

Ans: In Random forest, generally the feature importance is computed based on **out-of-bag (OOB) error**. To compute the feature importance, the random forest model is created and then the OOB error is computed. This is followed by permuting (shuffling) a feature and then again the OOB error is computed. Like wise, all features are permuted one by one.

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

Ans: It is an ensemble method that combines multiple models (classification or regression) via meta-model (meta-classifier or meta-regression). The base models are trained on the complete dataset, then the meta-model is trained on features returned (as output) from base models. The base models in stacking are typically different. The meta-model helps to find the features from base models to achieve the best accuracy.

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

Ans:

|  |  |  |
| --- | --- | --- |
| **Ensemble Method** | **Advantages** | **Disadvantages** |
| **Bagging** | Reduces variance and improves accuracy, can turn weak learners into strong learners, and works well with high variance models | Can increase bias, may not work well with low-variance models, and can be computationally expensive for large datasets |
| **Boosting** | Improves accuracy and reduces bias, works well with high-bias models and imbalanced data | Can [overfit](https://www.analyticsfordecisions.com/overfitting-and-underfitting/) with noisy data and outliers, can be computationally intensive |
| [Stacking](https://www.analyticsfordecisions.com/stacking-in-machine-learning/) | Improves prediction accuracy by combining models with different strengths and weaknesses, and can build a more reliable meta-model | Can be complex and time-consuming to implement, especially with large datasets |

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

Ans: There are **no restrictions/guidelines on the number of models**. You can start even from 3 models. You can keep the number of models as a hyperparameter if the training cost is less. Typically, you will observe the slanted 'L' shaped curve for MSE vs # of models plot. You can take the elbow point as the final # of models.