Linear and Logistic Regression Q&A

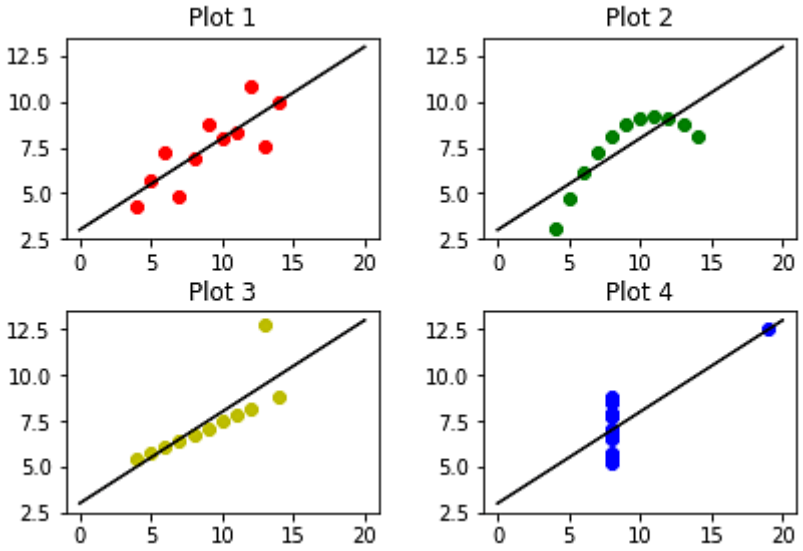
1. Describe a well-known law or natural phenomena that you could model with linear regression.

Many common laws have proportional relations and can be described by linear relationships. These can therefore be modeled with linear regression. A few famous examples are Ohm's law-related by V=IR or Newton's second law given by F=ma. On close observation, it is apparent that each of these laws has the form y=mx+c where c=0, or are essentially described by lines that pass through the origin.

2. Is it necessary to visualize the data when you have fitted a line? Why or why not?

It is crucial to visualize the data when you fit a line because numerically fitting a line is easy with pure numerical analysis or methods like Least Squares regression. But determining whether these fitted lines make any sense requires further analysis of which visualizing the data is one of the easiest ways.

To underscore this point, let us consider Anscombe's quartet, the four datasets with almost identical simple statistical properties but appear wildly different in distribution when plotted.



These four datasets result in a regression line with a slope of 0.50 and an intercept of 3.00. However, on plotting this line with the datasets, it becomes apparent that although fitted for datasets with almost identical results through quantitative analysis, the resultant line does not make any sense.

3. Does correlation imply causation? Why or why not?

No, while correlation is popularly used to provide information on the extent and direction of the linear relationship between two variables and can be used to determine whether a variable can be used to predict another, a high correlation does not imply causation.

For instance, you might find a correlation between umbrella malfunctions and a carpenter's income. As you can imagine, it is unlikely that there is a direct relation between the two, except that people tend to open up their umbrellas during the rainy season and that wooden doors swell due to high humidity. In this case, there is a hidden cause, rain, that causes both the phenomena as mentioned above and consequently the high correlation between them.

4. Is linear regression suitable for time series analysis?

While linear regression can be used for time series analysis and generally yield workable results, the performance is not particularly remarkable. The two main factors for this are :

- Time series generally have seasonal or periodic trends (such as peak seasons or even peak hours), which might be treated as outliers in linear regression and hence not appropriately accounted for.
- Future prediction is a generally sought-after use case in time series analysis, which will require extrapolation and rarely results in good predictions.

ARIMA, ARCH, and LSTM are widely used and better performing algorithms for time series analysis.

5. Is Feature Engineering necessary for even simple linear regression? Explain with an example.

Yes, Feature Engineering could be helpful even with the most straightforward linear regression problems. Say, for instance, you are trying to predict the Cost of a chocolate bar given the following

Length	Breadth	Cost
3.0	2.0	7.5
3.5	2.0	7.6
3.5	2.5	8.0
5.0	3.0	9.0

Here you might find a workable relationship between the Length and the Cost or the Breadth and the Cost. However, on multiplying the Length and the Breadth to derive the Size, you will see that this is a much better indicator of the Cost and will fit the resulting linear regression model better.

6. Are there any risks to extrapolation? Explain with an example when you would and would not use this.

Extrapolation is essentially predicting values of the target function for parameter values outside the range of those observed during training. While extrapolation could reasonably work well in some cases, such as predicting the voltage in Ohm's law, it can also result in meaningless results.

One easy example to explain this can be to extrapolate the decreasing rainfall trend at the end of the rainy season. If the extrapolation is done unchecked, the model could predict negative rain after a period that is about as nonsensical as it gets!

7. Can linear regression be used for representing quadratic equations?

Yes, paradoxically, a multiple linear regression model can be used to represent quadratic equations. For more complex linear regression models, we use multiple independent variables to predict the dependent variable. Such a linear regression model is called a multiple linear regression model.

A linear model with multiple dependent variables x1, x2, ..., xn can be written as

y=1x1+2x2+...+nxn.

For a quadratic function given by, say, y=ax2+bx+c, we can use x1=x2, x2=x, and x3=1, effectively representing the desired quadratic equation. Similarly, linear regression models can be used to describe higher-order polynomials as well.

8. Give an example scenario where a multiple linear regression model is necessary.

Consider an example where you are considering customer satisfaction for a particular brand of cereal. This would usually be decided by several factors, including cost, nutritional value, and taste. Say you are given all the above parameters and choose x1, x2, and x3 to represent them.

If these are the only three dependent variables, then your linear regression model, in this case, would be a multiple linear regression model that can be represented in the form

y=1x1+2x2+3x3

9. Is Overfitting a possibility with linear regression?

Yes, Overfitting is possible even with linear regression. This happens when multiple linear regression is used to fit an extremely high-degree polynomial. When the parameters of such a model are learned, they will fit too closely to the training data, fitting even the noise, and thereby fail to generalize on test data.

10. Is it necessary to remove outliers? Why or why not?

Yes, it is necessary to remove outliers as they can have a huge impact on the model's predictions. Take, for instance, plots 3 and 4 for the Anscombe's quartet provided above. It is apparent from these plots that the outliers have caused a significant change in the best fit line in comparison to what it would have been in their absence.

11. How do you identify outliers?

An outlier is an observation that is unlikely to have occurred in a data set under ordinary circumstances. Its values are so widely different from the other observations that it is most likely a result of noise or a rare exceptional case.

Box plots are a simple, effective, and hence commonly used approach to identifying outliers. Besides this, scatter plots, histograms and Z-scores are other methods used whenever feasible.

12. Is the vertical offset, horizontal offset, or the perpendicular offset minimized for least-square fitting, assuming that the vertical axis is the dependent variable? Why is this so?

In most cases, the vertical offsets from a line (or a surface) are minimized instead of the perpendicular offsets (or the horizontal offset). The resultant fitting function for the independent variables that predict the dependent variable allows uncertainties of the data points to be represented in a simple manner. Further, compared to a fit based on perpendicular offsets, this practice allows for a much simpler analytic form for the fitting parameters.

13. How do residuals help in determining the quality of a model?

Residuals are the deviations of the observed values from a fitted line. Checking the residuals is an important step to ascertain whether our assumptions of the regression model are valid. Suppose there is no apparent pattern in the plot of residuals versus fitted values, and the ordered residuals result in an almost normal distribution. In that case, we can conclude that there are no apparent violations of assumptions. On the other hand, if there is a relationship between the residuals and our fitted values, it is an indicator that the model is not good.

14. What is scaling? When is it necessary?

The technique to standardize the features in the data set to fit within a fixed range is called Feature Scaling. It is performed during the preprocessing stage and helps avoid the dominance of certain features due to high magnitudes.

When using the analytical solution for Ordinary Least Square, feature scaling is almost useless. However, when using gradient descent as an optimization technique, the data scaling results can be valuable. It can help to ensure that the gradient descent moves smoothly towards the global minimum and that the gradient descent steps update at the same rate for all the features.

15. If your training error is 10% and your test error is 70%, what do you infer?

A low error in training error while the test data yields a significantly higher error is a strong indicator of Overfitting. Such an observation strongly suggests that the model has learned so well over the training set that it hardly makes any mistakes during prediction over training data but cannot generalize over the unseen test set.

16. If you have two choices of hyperparameters, one resulting in a training and test error of 10% and another with a training and test error of 20%, which one of the two would you prefer and why?

Given that both the training and the test set are yielding an error of 10% in case 1 and an error of 20% in case 2, it is pretty easy to opt for the hyperparameters of case 1 for our machine learning problem as it is always desirable to have a lower error in predictions.

17. If your training error is high despite adjusting the hyperparameter values and increasing the number of iterations, what is most likely to be the issue? How can you resolve this problem?

High training error despite hyperparameter adjustment and a significant number of iterations strongly indicates that the model is unable to learn the problem it is presented with despite its best effort, or in other words, that it is underfitting. Reducing the regularisation and using more complex models can be some ways used to address this problem.

18. If the deviations of the residuals from your model are extremely small, does it suggest that the model is good or bad?

Residuals are essentially how much the actual data points vary from the fitted line and are hence indicators of deviation or error. Therefore, the smaller the deviation of the residuals from the fitted line, the better the model is likely to be.

19. What scenario would you prefer to use Gradient Descent instead of Ordinary Least Square Regression and why?

Ordinary Least Square Regression is computationally very expensive. Therefore, while it performs well with small data sets, it is infeasible to use this approach for significant machine learning problems. Consequently, for problems with larger data sets, Gradient Descent is the preferred optimization algorithm.

20. If you observe that the test error is increasing after a certain number of iterations, what do you infer is most likely to be occurring? How do you address this problem?

Observing an increase in error on the validation set after a certain number of iterations can indicate that the model is Overfitting. We can arrive at this diagnosis because we expect the error to decrease with more optimized parameters. While simplifying the model is one way to address this problem, early stopping is another commonly used solution.

Early stopping is probably one of the most commonly used forms of regularization. Unlike a weight decay used in the cost function, which helps to arrive at less complex models by explicit regularization, early stopping can be considered as a form of implicit regularization.

21. What are the odds?

Odds are defined as the ratio of the probability of an event occurring to the probability of the event not occurring.

For Example, let’s assume that the probability of winning a game is 0.02. Then, the

probability of not winning is 1- 0.02 = 0.98.

- The odds of winning the game= (Probability of winning)/(probability of not winning)
- The odds of winning the game= 0.02/0.98
- The odds of winning the game are 1 to 49, and the odds of not winning the game are 49 to 1.

22. What factors can attribute to the popularity of Logistic Regression?

Logistic Regression is a popular algorithm as it converts the values of the log of odds which can range from -inf to +inf to a range between 0 and 1.

Since logistic functions output the probability of occurrence of an event, they can be applied to many real-life scenarios therefore these models are very popular.

23. Is the decision boundary Linear or Non-linear in the case of a Logistic Regression model?

The decision boundary is a line or a plane that separates the target variables into different classes that can be either linear or nonlinear. In the case of a Logistic Regression model, the decision boundary is a straight line.

Logistic Regression model formula = $\boldsymbol{\alpha+1X_1+2X_2+\ldots.+kX_k}$. This clearly represents a straight line.

It is suitable in cases where a straight line is able to separate the different classes. However, in cases where a straight line does not suffice then nonlinear algorithms are used to achieve better results.

24. What is the Impact of Outliers on Logistic Regression?

The estimates of the Logistic Regression are sensitive to unusual observations such as outliers, high leverage, and influential observations. Therefore, to solve the problem of outliers, a sigmoid function is used in Logistic Regression.

25. What do you mean by the Logistic Regression?

It's a classification algorithm that is used where the target variable is of categorical nature. The main objective behind Logistic Regression is to determine the relationship between features and the probability of a particular outcome.

For Example, when we need to predict whether a student passes or fails in an exam given the number of hours spent studying as a feature, the target variable comprises two values i.e. pass and fail.

Therefore, we can solve classification problem statements which is a supervised machine learning technique using Logistic Regression.

26. What are the different types of Logistic Regression?

Three different types of Logistic Regression are as follows:

1. Binary Logistic Regression: In this, the target variable has only two 2 possible outcomes.

For Example, 0 and 1, or pass and fail or true and false.

2. Multinomial Logistic Regression: In this, the target variable can have three or more possible values without any order.

For Example, Predicting preference of food i.e. Veg, Non-Veg, Vegan.

3. Ordinal Logistic Regression: In this, the target variable can have three or more values with ordering.

For Example, Movie rating from 1 to 5.

27. Explain the intuition behind Logistic Regression in detail.

Given:

By using the training dataset, we can find the dependent(x) and independent variables(y), so if we can determine the parameters w (Normal) and b (y-intercept), then we can easily find a decision boundary that can almost separate both the classes in a linear fashion.

Objective:

In order to train a Logistic Regression model, we just need w and b to find a line(in 2D), plane(3D), or hyperplane(in more than 3-D dimension) that can separate both the classes point as perfect as possible so that when it encounters with any new unseen data point, it can easily classify, from which class the unseen data point belongs to.

For Example, Let us consider we have only two features as x_1 and x_2.

Let's take any of the +ve class points (figure below) and find the shortest distance from that point to the plane. Here, the shortest distance is computed using:

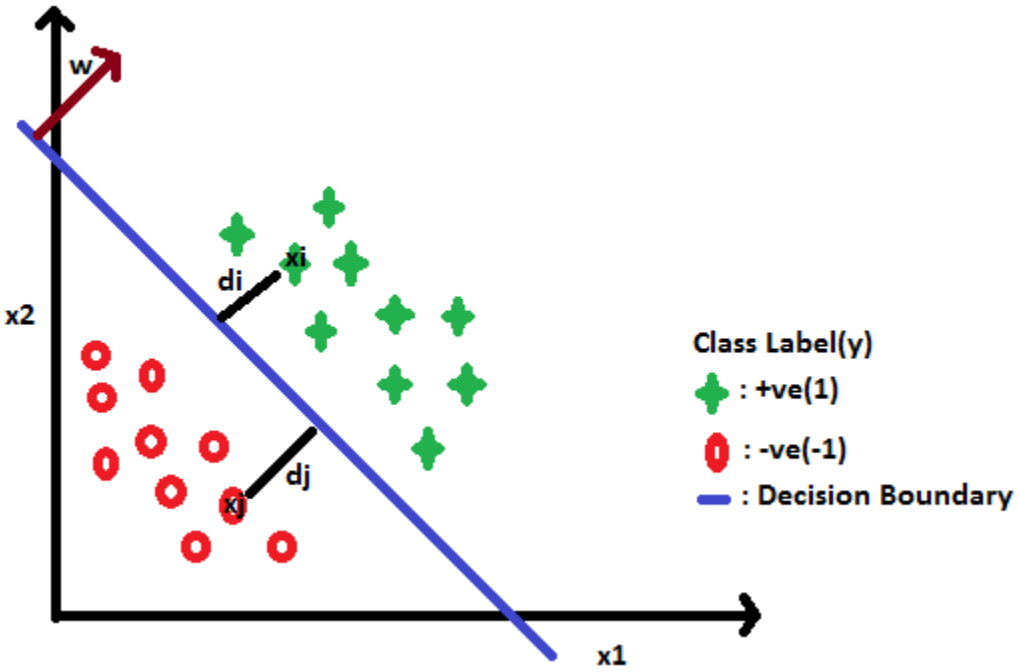
$$\mathbf{d_i = w^T * xi \,/\, ||w||}$$

If weight vector is a unit vector i.e, ||w||=1. Then,

$$\mathbf{d_i = w^T * xi}$$

Since w and x_i are on the same side of the decision boundary therefore distance will be +ve. Now for a negative point, we have to compute $d_i = w^T * xj$. For point x_i, distance will be -ve since this point is the opposite side of w.

Thus we can conclude, points that are in the same direction of w are considered as +ve points and the points which are in the opposite direction of w are considered as -ve points.



Now, we can easily classify the unseen data points as -ve and +ve points. If the value of $w^T*x_i>0$, then y =+1 and if value of $w^T*x_i < 0$ then y = -1.

- If $y_i = +1$ and $w^T*x_i > 0$, then the classifier classifies it as+ve points. This implies if $y_i*w^T*x_i > 0$, then it is a correctly classified point because multiplying two +ve numbers will always be greater than 0.
- If $y_i = -1$ and $w^T*x_i < 0$, then the classifier classifies it as -ve point. This implies if $y_i * w^T*x_i > 0$ then it is a correctly classified point because multiplying two -ve numbers will always be greater than zero. So, for both +ve and -ve points the value of $y_i* w^T*x_i$ is greater than 0. Therefore, the model classifies the points x_i correctly.
- If $y_i = +1$ and $w^T*x_i < 0$, i.e, y_i is +ve point but the classifier says that it is -ve then we will get -ve value. This means that point is classified as -ve but the actual class label is +ve, then it is a miss-classified point.
- If $y_i = -1$ and $w^T*x_i > 0$, this means actual class label is -ve but classified as +ve, then it is miss-classified point($y_i*w^T*x_i < 0$).

Now, by observing all the cases above now our objective is that our classifier minimizes the miss-classification error, i.e, we want the values of $y_i*w^T*x_i$ to be greater than 0.

In our problem, x_i and y_i are fixed because these are coming from the dataset.

As we change the values of the parameters w, and b the sum will change and we want to find that w and b that maximize the sum given below. To calculate the parameters w and b, we

can use the Gradient Descent optimizer. Therefore, the optimization function for logistic regression is:

$$\text{argmax}_w \sum_{i=0}^{n} yi(w^t xi) + b$$

28. What is Logistic Regression?

Logistic Regression is a type of statistical analysis that is used to predict the probability of an event occurring. This event can be something like whether or not a customer will make a purchase, or if a patient will develop a certain disease. Logistic Regression is a type of regression analysis, which means that it is used to predict a dependent variable based on one or more independent variables.

29. How does logistic regression work in the context of machine learning?

Logistic regression is a type of supervised machine learning algorithm that is used for classification tasks. The algorithm works by using a linear function to map input values to a set of predicted probabilities, and then uses a sigmoid function to map those probabilities to a binary output. The output of the logistic regression algorithm is either a 0 or a 1, which represents the two classes that the algorithm is trying to predict.

30. Can you explain what a logit function is?

A logit function is a mathematical function that is used to model the probability of an event occurring. The function takes in a value between 0 and 1, and outputs a value between -infinity and +infinity. The logit function is used in logistic regression, which is a type of statistical analysis that is used to predict the probability of an event occurring.

31. Why do you think linear regression cannot be used for binary classification problems?

Linear regression is not appropriate for binary classification problems because the output of a linear regression model is a continuous value, not a discrete value. In binary classification, we are looking to predict a class label (e.g. 0 or 1), not a continuous value. Therefore, a linear regression model would not be able to accurately predict class labels in a binary classification problem.

32. What are some common applications of logistic regression models?

Logistic regression models are commonly used in fields such as medicine, criminology, and marketing. In medicine, logistic regression models can be used to predict the likelihood of a patient developing a certain disease. In criminology, logistic regression models can be used to predict the likelihood of a person committing a crime. In marketing, logistic regression models can be used to predict the likelihood of a customer making a purchase.

33. What are the strengths and weaknesses of using logistic regression?

The main strength of logistic regression is that it is a very versatile tool that can be used for a variety of tasks, such as classification, prediction, and estimation. Additionally, logistic regression is relatively simple to implement and can be run on most standard statistical software packages.

The main weakness of logistic regression is that it can be prone to overfitting, especially when working with small datasets. Additionally, logistic regression makes a number of assumptions about the data that may not always hold true in practice.

34. When should logistic regression be preferred over KNN or decision trees?

Logistic regression should be preferred over KNN or decision trees when you have a linear decision boundary. Logistic regression will be more accurate than KNN or decision trees in this case.

8. Is it possible to use the same approach as linear regression to predict probabilities? If yes, then how?

Yes, it is possible to use the same approach as linear regression to predict probabilities. This can be done by using the sigmoid function to transform the linear regression output into a probability between 0 and 1.

35. What's an odds ratio?

The odds ratio is a statistical measure that is used to compare the odds of an event occurring in one group to the odds of the same event occurring in another group. The odds ratio can be used to compare the likelihood of two different outcomes, or to compare the risk of a certain event occurring in two different groups.

36. Can you explain the training process followed by logistic regression?

The training process for logistic regression is similar to that of other supervised learning algorithms. First, the algorithm is given a set of training data, which includes a set of input vectors and corresponding output labels. The algorithm then learns a model that maps the input vectors to the output labels. Finally, the algorithm is tested on a set of test data to see how well the model generalizes to new data.

37. How are different datasets handled when building a logistic regression model?

When building a logistic regression model, different datasets are handled according to their size and complexity. If the dataset is small and simple, then the model can be built using only that dataset. However, if the dataset is large and complex, then the model may be built using a subset of the data, or using a different technique altogether.

38. Do all features need to be included in the dataset while building a logistic regression model? If not, then which ones should be chosen?

No, all features do not need to be included in the dataset while building a logistic regression model. In fact, it is often better to not include all features, as this can lead to overfitting. Instead, you should carefully select the features that you believe will be most predictive of the target variable.

39. Is there a way to assess the importance of each feature in a logistic regression model? If yes, then how?

Yes, there are a few ways to assess the importance of each feature in a logistic regression model. One way is to look at the coefficients of the features and see how they compare to each other. Another way is to look at the p-values of the features and see which features have a statistically significant impact on the outcome.

40. Why is Overfitting a problem in logistic regression?

Overfitting is a problem in logistic regression because it can cause the model to inaccurately represent the underlying data. This can lead to poor predictions and suboptimal results.

41. What are some good ways to prevent overfitting in logistic regression?

Some ways to prevent overfitting in logistic regression include using cross-validation, using regularization methods such as L1 or L2 regularization, or by pruning the model.

42. What can be done if we have categorical variables with multiple values?

One option would be to create dummy variables for each value of the categorical variable. Another option would be to use a technique called binning, which would group the values of the categorical variable into a smaller number of bins.

43. What are multinomial logistic regressions?

Multinomial logistic regressions are a type of logistic regression that is used when there are more than two possible outcomes. This can be used, for example, to predict the likelihood of someone voting for a particular candidate in an election, where there are three or more candidates.

44. Can you explain what regularization is?

Regularization is a technique used to avoid overfitting in machine learning models. Overfitting occurs when a model is too closely fit to the training data, and does not generalize well to new data. Regularization helps to avoid overfitting by adding a penalty term to the model that discourages the model from fitting too closely to the training data.

45. What are the various types of hyperparameters that we can tune using cross-validation in logistic regression?

The various types of hyperparameters that we can tune using cross-validation in logistic regression are the regularization parameter, the learning rate, and the number of iterations.

46. Can you give me some examples of real-world applications where logistic regression has been successfully applied?

Logistic regression can be used for a variety of classification tasks, such as predicting whether a given email is spam or not, whether a given patient has a certain disease, and so on. In each case, we are trying to predict a binary outcome (yes/no, spam/not spam, etc.) based on a set of input features. Logistic regression is a powerful tool for modeling these types of problems.

47. What is the difference between the outputs of the Logistic model and the Logistic function?

The Logistic model outputs the logits, i.e. log-odds; whereas the Logistic function outputs the probabilities.

Logistic model = $\boldsymbol{\alpha+1X_1+2X_2+\ldots.+kX_k}$. Therefore, the output of the Logistic model will be logits.

Logistic function = f(z) = $\boldsymbol{1/(1+e-(\alpha+1X_1+2X_2+\ldots.+kX_k))}$. Therefore, the output of the Logistic function will be the probabilities.

48. Is logistic regression a generative or a descriptive classifier? Why?

Logistic regression is a descriptive model. Logistic regression learns to classify by knowing what features differentiate two or more classes of objects. For example, to classify between an apple and an orange, it will learn that the orange is orange in color and an apple is not. On the other hand, a generative classifier like a Naive Bayes will store all the classes' critical features and then classify based on the features the test case best fits.

49. Can you use logistic regression for classification between more than two classes?

Yes, it is possible to use logistic regression for classification between more than two classes, and it is called multinomial logistic regression. However, this is not possible to implement without modifications to the vanilla logistic regression model.

50. How do you implement multinomial logistic regression?

The multinomial logistic classifier can be implemented using a generalization of the sigmoid, called the softmax function. The softmax represents each class with a value in the range (0,1), with all the values summing to 1. Alternatively, you could use the one-vs-all or one-vs-one approach using multiple simple binary classifiers.

51. Suppose that you are trying to predict whether a consumer will recommend a particular brand of chocolate or not. Let us say your hypothesis function outputs h(x)=0.55 where h(x) is the probability that y=1 (or that a consumer recommends the chocolate) given any input x. Does this mean that the consumer will recommend the chocolate?

The answer to this question is 'cannot be determined.' And this will remain the case unless you are provided additional data on the decision boundary. Let us say that you set the decision boundary such that y=1 is $h(x) \geq 0.5$ and 0; otherwise, then the answer for this question would be a resounding YES. However, if you set the decision boundary (although this is not very common practice) such that y=1 is $h(x) \geq 0.6$ and 0, otherwise the answer will be a NO.

52. Why can't we use the mean square error cost function used in linear regression for logistic regression?

If we use mean square error in logistic regression, the resultant cost function will be non-convex, i.e., a function with many local minima, owing to the presence of the sigmoid function in h(x). As a result, an attempt to find the parameters using gradient descent may fail to optimize cost function properly. It may end up choosing a local minima instead of the actual global minima.

53. If you observe that the cost function decreases rapidly before increasing or stagnating at a specific high value, what could you infer?

A trend pattern of the cost curve exhibiting a rapid decrease before then increasing or stagnating at a specific high value indicates that the learning rate is too high. The gradient descent is bouncing around the global minimum but missing it owing to the larger than necessary step size.

54. What alternative could you suggest using a for loop (which is time-consuming) when using Gradient Descent to find the optimum parameters for logistic regression?

One commonly used efficient alternative to using for loop is vectorization, i.e., representing the parameter values to be optimized in a vector. By using this approach, all the vectors can be updated instead of iterating over them in a for loop.

55. Are there alternatives to find optimum parameters for logistic regression besides using Gradient Descent?

Yes, Gradient Descent is merely one of the many available optimization algorithms. Other advanced optimization algorithms can often help arrive at the optimum parameters faster and help with scaling for significant machine learning problems. A few such algorithms are Conjugate Gradient, BFGS, and L-BFGS algorithms.

56. How many binary classifiers would you need to implement one-vs-all for three classes? How does it work?

You would need three binary classifiers to implement one-vs-all for three classes since the number of binary classifiers is precisely equal to the number of classes with this approach. If

you have three classes given by y=1, y=2, and y=3, then the three classifiers in the one-vs-all approach would consist of **h(1)(x),** which classifies the test cases as 1 or not 1, **h(2)(x)** which classifies the test cases as 2 or not 2 and so on. You can then take the results together to arrive at the correct classification. For example, with three categories, Cats, Dogs, and Rabbits, to implement the one-vs-all approach, we need to make the following comparisons:

Binary Classification Problem 1: Cats vs. Dogs, Rabbits (or not Cats)

Binary Classification Problem 2: Dogs vs. Cats, Rabbits (or not Dogs)

Binary Classification Problem 3: Rabbits vs. Cats, Dogs (or not Rabbits)

57. How many binary classifiers would you need to implement one-vs-one for four classes? How does it work?

To implement one-vs-one for four classes, you will require six binary classifiers. This is because you will need to compare each class with each other class. In general, the formula for calculating the number of binary classifiers b is given as b=(no. of classes * (no. of classes - 1))/ 2.

Suppose we have four different categories into which we need to classify the weather for a particular day: Sun, Rain, Snow, Overcast. Then to implement the one-vs-one approach, we need to make the following comparisons:

Binary Classification Problem 1: Sun vs. Rain

Binary Classification Problem 2: Sun vs. Snow

Binary Classification Problem 3: Sun vs. Overcast

Binary Classification Problem 4: Rain vs. Snow

Binary Classification Problem 5: Rain vs. Overcast

Binary Classification Problem 6: Snow vs. Overcast

58. What is the importance of regularisation?

Regularisation is a technique that can help alleviate the problem of overfitting a model. It is beneficial when a large number of parameters are present, which help predict the target function. In these circumstances, it is difficult to select which features to keep manually.

Regularisation essentially involves adding coefficient terms to the cost function so that the terms are penalized and are small in magnitude. This helps, in turn, to preserve the overall trends in the data while not letting the model become too complex. These penalties, in effect, restrict the influence a predictor variable can have over the target by compressing the coefficients, thereby preventing overfitting.

59. Why is the Wald Test useful in logistic regression but not in linear regression?

The Wald test, also known as the Wald Chi-Squared Test, is a method to find whether the independent variables in a model are of significance. The significance of variables is decided by whether they contribute to the predictions or not. The variables that add no value to the model can therefore be deleted without risking severe adverse effects to the model. The Wald test is unnecessary in linear regression because it is easy to compare a more complicated model to a simpler model to check the influence of the added independent variables. After all, we can use the R2 value to make this comparison. However, this is not possible with logistic regression as we use Maximum Likelihood Estimate, which uses the previously mentioned method infeasible. The Wald test can be used for many different models, including those with binary variables or continuous variables, and has the added advantage that it only requires estimating one model.

60. Will the decision boundary be linear or non-linear in logistic regression models? Explain with an example.

The decision boundary is essentially a line or a plane that demarcates the boundary between the classes to which linear regression classifies the dependent variables. The shape of the decision boundary will depend entirely on the logistic regression model.

For logistic regression model given by hypothesis function h(x)=g(Tx)where g is the sigmoid function, if the hypothesis function is h(x)=g(1+2x2+3x3)then the decision boundary is linear. Alternatively, if h(x)=g(1+2x22+3x32)then the decision boundary is non-linear.

61. What are odds? Why is it used in logistic regression?

Odds are the ratio of the probability of success to the probability of failure. The odds serve to provide the constant effect a particular predictor or independent variable has on the output prediction. Expressing the effect of a predictor on the likelihood of the target having a particular value through probability does not describe this constant effect. In linear regression models, we often want to measure the unique effect of each independent variable on the output for which the odds are very useful.

62. Given fair die, what are the odds of occurrence of odd numbers?

The odds of occurrence of odd numbers is 1.

There are three odd and three even numbers in a fair die, and therefore, the probability of occurrence of odd numbers is 3/6 or 0.5. Similarly, the odds of occurrence of numbers that are not odd is 0.5. Since odds is the ratio of the probability of success and that of failure,

Odds = 0.5/0.5=1.

63. In classification problems like logistic regression, classification accuracy alone is not considered a good measure. Why?

Classification accuracy considers both true positives and false positives with equal significance. If this were just another machine learning problem of not too much consequence, this would be acceptable. However, when the problems involve deciding whether to consider a candidate for life-saving treatment, false positives might not be as bad

as false negatives. The opposite can also be true in some cases. Therefore, while there is no single best way to evaluate a classifier, accuracy alone may not serve as a good measure.

64. It is common practice that when the number of features or independent variables is larger in comparison to the training set, it is common to use logistic regression or support vector machine (SVM) with a linear kernel. What is the reasoning behind this?

It is common to use logistic regression or SVM with a linear kernel because when there are many features with a limited number of training examples, a linear function should be able to perform reasonably well. Besides, there is not enough training data to allow for the training of more complex functions.

65. Between SVM and logistic regression, which algorithm is most likely to work better in the presence of outliers? Why?

SVM is capable of handling outliers better than logistic regression. SVM is affected only by the points closest to the decision boundary. Logistic regression, on the other hand, tries to maximize the conditional likelihood of the training data and is therefore strongly affected by the presence of outliers.

66. Which is the most preferred algorithm for variable selection?

Lasso is the most preferred for variable selection because it performs regression analysis using a shrinkage parameter where the data is shrunk to a point, and variable selection is made by forcing the coefficients of not so significant variables to be set to zero through a penalty.

67. What according to you is the method to best fit the data in logistic regression?

Maximum Likelihood Estimation to obtain the model coefficients which relate to the predictors and target.

68. What are the disadvantages of Logistic Regression?

The disadvantages of the logistic regression are as follows:

1. Sometimes a lot of **Feature Engineering** is required.

2. If the independent features are correlated with each other it may affect the performance of the classifier.

3. It is quite sensitive to **noise** and **overfitting**.

4. Logistic Regression should not be used if the number of observations is lesser than the number of features, otherwise, it may lead to overfitting.

5. By using Logistic Regression, non-linear problems can't be solved because it has a linear decision surface. But in real-world scenarios, the linearly separable data is rarely found.

69. What are the advantages of Logistic Regression?

The advantages of the logistic regression are as follows:

1. Logistic Regression is very easy to understand.

2. It requires less training.

3. It performs well for simple datasets as well as when the data set is linearly separable.

4. It doesn't make any assumptions about the distributions of classes in feature space.

5. A Logistic Regression model is less likely to be over-fitted but it can overfit in high dimensional datasets. To avoid over-fitting these scenarios, One may consider regularization.

70. Why can't we use Linear Regression in place of Logistic Regression for Binary classification?

Linear Regressions cannot be used in the case of binary classification due to the following reasons:

1. Distribution of error terms: The distribution of data in the case of Linear and Logistic Regression is different. It assumes that error terms are normally distributed. But this assumption does not hold true in the case of binary classification.

2. Model output: In Linear Regression, the output is continuous(or numeric) while in the case of binary classification, an output of a continuous value does not make sense. For binary classification problems, Linear Regression may predict values that can go beyond the range between 0 and 1. In order to get the output in the form of probabilities, we can map these values to two different classes, then its range should be restricted to 0 and 1. As the Logistic Regression model can output probabilities with Logistic or sigmoid function, it is preferred over linear Regression.

3. The variance of Residual errors: Linear Regression assumes that the variance of random errors is constant. This assumption is also not held in the case of Logistic Regression.

71. Why can't we use Mean Square Error (MSE) as a cost function for Logistic Regression?

In Logistic Regression, we use the sigmoid function to perform a non-linear transformation to obtain the probabilities. If we square this nonlinear transformation, then it will lead to the problem of non-convexity with local minimums and by using gradient descent in such cases, it is not possible to find the global minimum. As a result, MSE is not suitable for Logistic Regression.

So, in the Logistic Regression algorithm, we used Cross-entropy or log loss as a cost function. The property of the cost function for Logistic Regression is that:

- The confident wrong predictions are penalized heavily
- The confident right predictions are rewarded less

By optimizing this cost function, convergence is achieved.

$$\text{Cost}(h_\Theta(x), Y(\text{actual})) = -\log(h_\Theta(x)) \text{ if } y=1$$
$$-\log(1-h_\Theta(x)) \text{ if } y=0$$

72. Discuss the Train complexity of Logistic Regression.

In order to train a Logistic Regression model, we just need w and b to find a line(in 2-D), plane(in 3-D), or hyperplane(in more than 3-D dimension) that can separate both the classes point as perfect as possible so that when it encounters with any new point, it can easily classify, from which class the unseen data point belongs to.

The value of w and b should be such that it maximizes the sum $y_i * w^T * x_i > 0$.

Now, let's calculate its time complexity in terms of Big O notation:

- Performing the operation $y_i * w^T * x_i$ takes O(d) steps since w is a vector of size-d.
- Iterating the above step over n data points and finding the maximum sum takes n steps.

$$\textbf{argmax}_w \sum_{i=0}^{n} y_i (w^t x_i) + b$$

Therefore, the overall time complexity of the Logistic Regression during training is n(O(d))=O(nd).

73. Why is Logistic Regression termed as Regression and not classification?
The major difference between Regression and classification problem statements is that the target variable in the Regression is numerical (or continuous) whereas in classification it is categorical (or discrete).

Logistic Regression is basically a supervised classification algorithm. However, the Logistic Regression builds a model just like linear regression in order to predict the probability that a given data point belongs to the category numbered as “1”.

For Example, Let’s have a binary classification problem, and ‘x’ be some feature and ‘y’ be the target outcome which can be either 0 or 1.

The probability that the target outcome is 1 given its input can be represented as:

$$P(y = 1 \mid x)$$

If we predict the probability by using linear Regression, we can describe it as:

$$p(X) = \beta_0 + \beta_1 X.$$

where, p(x) = p(y=1|x)

Logistic regression models generate predicted probabilities as any number ranging from neg to pos infinity while the probability of an outcome can only lie between **0< P(x)<1**.

However, to solve the problem of outliers, a sigmoid function is used in Logistic Regression. The Linear equation is put in the sigmoid function.

$$g(x) = \frac{1}{1 + e^{-x}}$$

74. Discuss the Test or Runtime complexity of Logistic Regression.
At the end of the training, we test our model on unseen data and calculate the accuracy of our model. At that time knowing about runtime complexity is very important. After the training of Logistic Regression, we get the parameters w and b.

To classify any new point, we have to just perform the operation $w^T * xi$. If $w^T*xi>0$, the point is +ve, and if $w^T*xi < 0$, the point is negative. As w is a vector of size d, performing the operation w^T*xi takes O(d) steps as discussed earlier.

Therefore, the testing complexity of the Logistic Regression is **O(d)**.

Hence, Logistic Regression is very good for low latency applications, i.e, for applications where the dimension of the data is small.

75. Discuss the space complexity of Logistic Regression.
During training: We need to store four things in memory: x, y, w, and b during training a Logistic Regression model.

- Storing b is just 1 step, i.e, O(1) operation since b is a constant.
- x and y are two matrices of dimension (n x d) and (n x 1) respectively. So, storing these two matrices takes O(nd + n) steps.
- Lastly, w is a vector of size-d. Storing it in memory takes O(d) steps.

Therefore, the space complexity of Logistic Regression while training is **O(nd + n +d)**.

During Runtime or Testing: After training the model what we just need to keep in memory is w. We just need to perform w^T*x_i to classify the points.

Hence, the space complexity during runtime is in the order of d, i.e, **O(d**).

76. How can we express the probability of a Logistic Regression model as conditional probability?
We define probability **P(Discrete value of Target variable | X_1, X_2, X_3…., X_k)** as the probability of the target variable that takes up a discrete value (either 0 or 1 in the case of binary classification problems) when the values of independent variables are given.

For Example, the probability an employee will attain (target variable) given his attributes such as his age, salary, etc.

77. Can we solve the multiclass classification problems using Logistic Regression? If Yes then How?

Yes, in order to deal with multiclass classification using Logistic Regression, the most famous method is known as the one-vs-all approach. In this approach, a number of models are trained, which is equal to the number of classes. These models work in a specific way.

For Example, the first model classifies the datapoint depending on whether it belongs to class 1 or some other class(not class 1); the second model classifies the datapoint into class 2 or some other class(not class 2) and so-on for all other classes.

So, in this manner, each data point can be checked over all the classes.

78. What are the assumptions made in Logistic Regression?

Some of the assumptions of Logistic Regression are as follows:

1. It assumes that there is minimal or **no multicollinearity** among the independent variables i.e, predictors are not correlated.

2. There should be a linear relationship between the logit of the outcome and each predictor variable. The logit function is described as **logit(p) = log(p/(1-p))**, where p is the probability of the target outcome.

3. Sometimes to predict properly, it usually requires a **large sample size**.

4. The Logistic Regression which has **binary classification** i.e, two classes assume that the target variable is binary, and ordered Logistic Regression requires the target variable to be ordered.

For example, Too Little, About Right, Too Much.

5. It assumes there is **no dependency** between the observations.

79. Which algorithm is better in the case of outliers present in the dataset i.e., Logistic Regression or SVM?
SVM (Support Vector Machines) handles the outliers in a better manner than the Logistic Regression.

Logistic Regression: Logistic Regression will identify a linear boundary if it exists to accommodate the outliers. To accommodate the outliers, it will shift the linear boundary.

SVM: SVM is insensitive to individual samples. So, to accommodate an outlier there will not be a major shift in the linear boundary. SVM comes with inbuilt complexity controls, which take care of overfitting, which is not true in the case of Logistic Regression.

80. How do we handle categorical variables in Logistic Regression?
The inputs given to a Logistic Regression model need to be numeric. The algorithm cannot handle categorical variables directly. So, we need to convert the categorical data into a numerical format that is suitable for the algorithm to process.

Each level of the categorical variable will be assigned a unique numeric value also known as a **dummy variable.** These dummy variables are handled by the Logistic Regression model in the same manner as any other numeric value.

81. What are the data challenges during model development?

· Observational data — missing values and outliers

· Mixed measurement scale — nominal, ordinal, interval and ratio

· High dimensionality — large number of predictors

· Rare target event — imbalanced dataset

82. What are the analytical challenges during model development?

· Non linearity — relationship between X and Y is non-linear. Hence difficult to model

· Model selection — selected the most accurate model but it may be an over-fit

83. What are the difference between linear regression and logistic?

· Outcome

o Linear regression — conditional mean of response is between –inf and +inf

o Logistic regression — conditional mean of response is between 0 and 1

· Relationship

o Linear regression — linear relationship between independent and dependent variable

o Logistic regression — linear relationship between independent and log-odds of dependent variable

· Error

o Linear regression — normal random error

o Logistic regression — does not have random normal error but binomial error (P * (1-P))

· Method of estimation

o Linear regression — method of ordinary least square (OLS)

o Logistic regression — method of maximum likelihood estimation (MLe)

84. What is stepwise selection method?

· Forward — Starts with zero variables. If a variable is added then it stays in the model even if it becomes insignificant later.

· Backward — Starts with all the variables. If a variable is eliminated then it cannot be included in the model.

· Stepwise — Includes aspects of forward and backward selection methods. It terminates when no variable can be added or removed from the model.

85. What are the assumptions of linear regression?

· Linearity of independent and dependent variable

· Errors should be normally distributed with mean of zero

· Errors have equal variance

· Errors are independent

86. How do you penalize the model for extra variables?

· Information value (AIC, BIC and SBS) — Each matrix has a different penalty for additional variables and tries to minimize the unexplained variance. Smallest information value is preferred.

· Adjusted R-Sq — R-Sq increased when more variables are added and Adj R-Sq takes into account the additional variables. Larger Adj R-Sq is preferred

87. What is the method of maximum likelihood?

· Estimated parameters that are most likely

· Logit (p) = b0 + b1 x1 + b2 x2 +… + bn xn

· Where, Logit (p) = LN(p / (1-p))

· If x1 is changed by 1 unit then change in odds is (e^b1) — 1

88. What does odds-ratio signify?

· The odds-ratio (OR) are always between 0 and infinity

· If OR = 1, then there is no association

· OR > 1, group in numerator has higher event

· OR < 1, group in denominator has higher event

89. How do you decide the cut-off for the output of logistic regression?

· Accuracy — cut-off such that the accuracy is maximum. Confusion matrix is used here, true negative (actual = 0 and predicted = 0), false negative (actual = 1 and predicted = 0), false positive (actual = 0 and predicted = 1), true positive (actual = 1 and predicted = 1).

· Business — cut-off such that the profit is maximum

90. What are the key matrices used to check the performance of logistic regression?

· C statistics — it represents the concordance of the model. It is the probability that an observation having event is more than the probability that an observation having non-event.

· Accuracy — (True positive + True negative) / Total cases

· Error Rate — (False positive + False negative) / Total cases

· Sensitivity — True positive / Total actual positive

· Specificity — True negative / Total actual negative

· Positive pred value — True positive / Total predicted positive

· Negative pred value — True negative / Total predicted negative

· KS — it measures the distance between cumulative good and cumulative bad. The maximum distance is KS.

· AUCROC — measures the performance of the model across all cut-offs. Sensitivity is on the y-axis and 1-specificty is on the x-axis

· Gain chart — positive prediction rate is on y-axis and percentage of cases allocated to event is on x-axis

91. How do you handling missing values?

· The goal of missing value imputation is to retain all original data and score new cases

· Numerical variable — impute with mean or median and create a missing value indicator

· Categorical variable — impute with a new label

· Regression imputation — does not involve target variable and can be used when two or more variables are highly correlated. However, it may lead to over-fitting, increase computation time and increased scoring efforts

· Cluster imputation — it is condition on other variables. The cluster mean is used to replace the missing data point

92. What is multi-colinearity?

· Co-linearity is the relationship between two variables. Multi-colinearity is the relationship between more than two variables.

· Variance inflation factor is used to identify presence of multi-colinearity. When multiple variables try to explain the variance it leads to inflated standard errors hence unstable model

93. How do you remove variable redundancy?

· Correlation matrix and variance inflation factor

· Variable clustering can be used and from each cluster one variable is selected such that the variable has high correlation with own cluster variables and low correlation with other cluster variables

94. What is an influential observation?

· An influential observation has large effect on some part of the model

· An outlier is an unusual data point

· To check for influential outliers, the data should be checked for errors and adequate modeling technique should be used.

95. What is the issue of high dimensionality?

· When a categorical variable has high number of labels, it leads to quasi complete separation

· It can affect the convergence of the model and can lead to incorrect decisions

· Solution — collapsing categories based on reduction in chi-square

96. What is the issue of non-linear relationship in logistic regression?

· Scatter plot is used. Logit (LN (p/(1-p))) on the y-axis and mean value of x (bins) on x-axis

· Use of polynomial models

· Use of a flexible multivariate function estimator

97. What is interaction?

· When two or more categorical variables are combined together

· If we have 3 categorical variables — A, B and C

· Interactions are — A*B, B*C, C*A and A*B*C

98. What is joint sampling and separate sampling?

· Joint sampling is done when there are equal number of events and non-events. Not appropriate for imbalanced data

· Separate sampling is done for imbalanced data. For rare event, all observations are kept when target = 1 and only few observations are kept when target = 0.

99. How do you correct for oversampling?

· Intercept needs to be corrected using an offset

· Offset = LN((p0 * P1) / (p1 * P0))

· Where, p0 is the proportion of non-event in population and p1 is the proportion of event in population

· P0 is the proportion of non-event in sample and P1 is the proportion of event in sample

· Over-sampling does not impact AUROC, sensitivity and specificity

· Over-sampling impacts the gain and lift charts

100. How do you correct for imbalanced data?

· Adjust the samples with weights

· 0 — n * p(0)

· 1 — n * p(1)

· If y = 1, then weight = p(1) / P(1)

· If y = 0, then weight = p(0) / P(1)