Logistic Regression Interview Questions

We believe that you have learned both theoritical and practical knowledge on Naive Bayes classification algorithm through your assignment.

So let's test your knowledge here. This will help you to be prepared for interviews too!

Best with Quest

1. What is a logistic function? What is the range of values of a logistic function? Why is logistic regression very popular? ¶

```
f(z) = 1/(1+e -z )
The values of a logistic function will range from 0 to 1. The values of Z will vary from -
infinity to +infinity.

Logistic regression is famous because it can convert the values of logits (logodds), which
can range from -infinity to +infinity to a range between 0 and 1. As logistic functions
output the probability of occurrence of an event, it can be applied to many real-life
scenarios. It is for this reason that the logistic regression model is very popular.
```

2. How can the probability of a logistic regression model be expressed as conditional probability?

```
P(Discrete value of Target variable | X1, X2, X3….Xk).  It is the probability of the
target variable to take up a discrete value (either 0 or 1 in case of binary
classification problems) when the values of independent variables are given. For example,
the probability an employee will attrite (target variable) given his attributes such as
his age, salary, KRA’s, etc.
```

3. What are the outputs of the logistic model and the logistic function? What are odds?

```
The logistic model outputs the logits, i.e. log odds; and the logistic function outputs
the probabilities.
   Logistic model = α+1X1+2X2+….+kXk. The output of the same will be logits.
   Logistic function = f(z) = 1/(1+e-(α+1X1+2X2+….+kXk)). The output, in this case, will
be the probabilities.

Odds is 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 lottery is
0.01. Then, the probability of not winning is 1- 0.01 = 0.99.
The odds of winning the lottery = (Probability of winning)/(probability of not winning)
The odds of winning the lottery = 0.01/0.99
The odds of winning the lottery is 1 to 99, and the odds of not winning the lottery is 99
to 1.
```

4. How to interpret the results of a logistic regression model? Or, what are the meanings of alpha and beta in a logistic regression model?

Loading [MathJax]/extensions/Safe.js

Alpha is the baseline in a logistic regression model. It is the log odds for an instance when all the attributes (X1, X2,………….Xk) are zero. In practical scenarios, the probability of all the attributes being zero is very low. In another interpretation, Alpha is the log odds for an instance when none of the attributes is taken into consideration.

Beta is the value by which the log odds change by a unit change in a particular attribute by keeping all other attributes fixed or unchanged (control variables).

5. What is odds ratio? What is the formula for calculating odds ratio for 2X2 table

Odds ratio is the ratio of odds between two groups. For example, let’s assume that we are trying to ascertain the effectiveness of a medicine. We administered this medicine to the ‘intervention’ group and a placebo to the ‘control’ group.

Odds ratio (OR) = (odds of the intervention group)/(odds of the control group)

Interpretation

If odds ratio = 1, then there is no difference between the intervention group and the control group

If odds ratio is greater than 1, then the control group is better than the intervention group

If odds ratio is less than 1, then the intervention group is better than the control group.

The odds ratio (OR) is a popular measure of the strength of association between exposure and disease. In a cohort study, the odds ratio is expressed as the ratio of the number of cases to the number of noncases in the exposed and unexposed groups. The odds ratio and its familiar computation are attributed to Cornfield (1951), which is calculated as the ratio of the products of the pairs of diagonal elements in the 2 × 2 table:

OR = A × D/B × C

6. Why can’t linear regression be used in place of logistic regression for binary classification?

The reasons why linear regressions cannot be used in case of binary classification are as follows:

Distribution of error terms: The distribution of data in case of linear and logistic regression is different. Linear regression assumes that error terms are normally distributed. In case of binary classification, this assumption does not hold true.

Model output: In linear regression, the output is continuous. In 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 0 and 1. If we want the output in the form of probabilities, which can be mapped to two different classes, then its range should be restricted to 0 and 1. As the logistic regression model can output probabilities with logistic/sigmoid function, it is preferred over linear regression.

Variance of Residual errors: Linear regression assumes that the variance of random errors is constant. This assumption is also violated in case of logistic regression.

7. What is the importance of a baseline in a classification problem?

Most classification problems deal with imbalanced datasets. Examples include telecom churn, employee attrition, cancer prediction, fraud detection, online advertisement targeting, and so on. In all these problems, the number of the positive classes will be very low when compared to the negative classes. In some cases, it is common to have positive classes that are less than 1% of the total sample. In such cases, an accuracy of 99% may sound very good but, in reality, it may not be.

Loading [MathJax]/extensions/Safe.js

```
Here, the negatives are 99%, and hence, the baseline will remain the same. If the
algorithms predict all the instances as negative, then also the accuracy will be 99%. In
this case, all the positives will be predicted wrongly, which is very important for any
business. Even though all the positives are predicted wrongly, an accuracy of 99% is
achieved. So, the baseline is very important, and the algorithm needs to be evaluated
relative to the baseline.
```

8. How does logistic regression handle categorical variables?

```
The inputs to a logistic regression model need to be numeric. The algorithm cannot handle
categorical variables directly. So, they need to be converted into a format that is
suitable for the algorithm to process. The various levels of a categorical variable will
be assigned a unique numeric value known as the dummy variable. These dummy variables are
handled by the logistic regression model as any other numeric value.
```

9. How will you deal with the multiclass classification problem using logistic regression?

```
The most famous method of dealing with multiclass classification using logistic regression
is using the one-vs-all approach. Under this approach, a number of models are trained,
which is equal to the number of classes. The 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; the second model classifies the datapoint into class 2 or some other
class. This way, each data point can be checked over all the classes.
```

10. 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 wT*xi to classify the points.

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

11 . 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.
```

Loading [MathJax]/extensions/Safe.js

12. 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 = α+1X1+2X2+….+kXk. 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.
```

13. What is the Impact of Outliers on Logistic Regression? 10. Which algorithm is better in the case of outliers present in the dataset i.e., Logistic Regression or SVM?

```
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.

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.
```

14. 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.
```

Loading [MathJax]/extensions/Safe.js

15. 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 wT*xi to classify the points.

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

Now Rest with this Quest :)