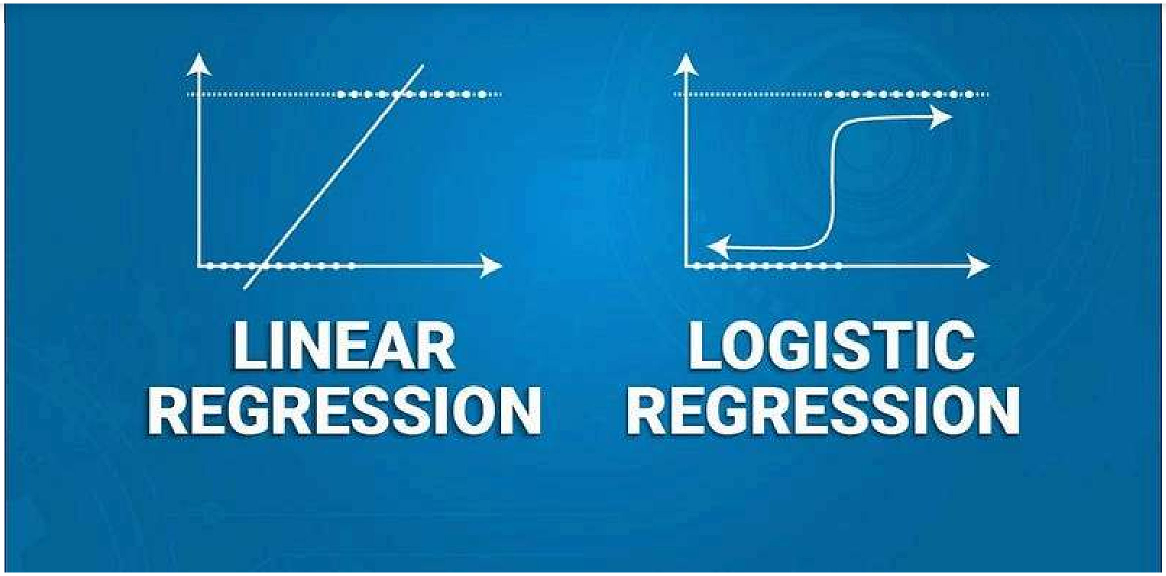
RG

Summary

The provided website content outlines key questions and answers

20 Interview Questions on Linear Regression and Logistic Regression



1. 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

2. What are the analytical challenges during model development?

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

3. 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 **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)

4. 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.

5. What are the assumptions of linear regression?

· Linearity of independent and dependent variable

· Errors have equal variance

· Errors are independent

6. 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

7. 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

8. What does odds-ratio signify?

· If OR = 1, then there is no association

· OR > 1, group in numerator has higher event

· OR < 1, group in denominator has higher event

9. 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

10. 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

· 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

11. 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

· 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

12. 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

13. 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

· 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.

15. 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

16. 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

17. 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

18. 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.

19. How do you correct for oversampling?

· Intercept needs to be corrected using an offset

· Offset = LN((p0 * P1) / (p1 * P0))

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

20. 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)

Interview Data Science Analytics

Recommended from ReadMedium

Fabiana Tortorelli

ROC Curve, Gini index, and KS statistic

Various approaches can be used to illustrate the performance of a binary classifier model. Here, I will explain the following:

6 min read

Amit Yadav

Top 11 Linear Regression Interview Questions (With Answers)

If you think you need to spend $2,000 on a 120-day program to become a data scientist, then listen to me for a minute.

10 min read

Samuel Ozechi

Stochastic Gradient Descent For Deep Learning

Understanding the parameter optimization process for deep learning models.

5 min read

Vikash Singh

If you're preparing for a data science interview, understanding ensemble methods is a must, and Bagging (Bootstrap Aggregating) is one of…

6 min read

Vasilis Kalyvas

"MLshorts" 23: What is Linear Regression?

A simple and clear explanation

4 min read

Be 10x Engineer

Staff Backend Engineer @ Google | Interview Experience

It was a crisp Tuesday morning when I first saw the email. My hands trembled slightly as I read the subject line: "Google Interview…

3 min read