100 Data Science Interview Questions Series!!

Here are the first 25 questions.

Q1.) What is Regularization, What is the difference between L1 and L2 regularization?

Intuitively, both are l1 and l2 are penalty terms that are added to the loss function in order to decrease overfitting. A linear regression model that implements the L1 regularization technique is called Lasso Regression whereas L2 is called Ridge Regression. The key difference between these two is the penalty term. L1 is the absolute sum of coefficients whereas L2 is the squared sum coefficients. L1 tends to shrink coefficients to zero whereas L2 tends to shrink coefficients evenly. L1 is therefore useful for feature selection, as we can drop any variables associated with coefficients that go to zero. L2, on the other hand, is useful when you have collinear/codependent features.

Q2.) What are some Model Evaluation techniques you know?

Area Under Curve (AUC) Receiver operating characteristic (ROC) Akaike Information Criterion (AIC) Cumulative Accuracy Profile (CAP) Confusion Matrix

Q.3) What is Selection Bias? Any Examples?

The Selection bias is an experimental error that occurs when the population being studied does not provide the data that we require to make conclusions. Basically, our Sample is not representative of the target population

For eg if you want to study college student's behavior and you only selected students from 1 college, then that sample of students will not be a good representative of all the college students in India. Thus introducing Sample/Selection Bias.

Q.4) Define terms used for evaluation such as Specificity, Sensitivity, Recall, Precision, f1 score, f2 score, etc.

I wrote a blog on it myself. I still have more to add in it, but you can read current explanations of the question here:

https://www.aiunquote.com/post/evaluate-a-covid-19-ml-model-to-learn-about-evaluation-matrices

Q.5) What is Codependence? Give an Example.

Codependence is when 1 feature has some 1-1 mapping with another. An example pair of codependent features is gender and ispregnant since, at the current level of medical technology, only females can be ispregnant. Codependence tends to increase coefficient variance, making coefficients unreliable/unstable, which hurts model generality. L2 reduces the variance of these estimates, which counteracts the effect of codependencies.

Q 6.) Are Analysis and Analytics the same thing?

There's a big difference between analysis and analytics, and it is imp for every Data Scientist to know. Analysis deals with the past, which means you have previous data of something and you perform analysis on the past data to understand it better.

Analytics on the other hand deals with the future, which means you perform future predictions on data. Analytics is also defined as the method of logical analysis, this is where all machine learning thing comes into play.

Q. 7) What are the assumptions of Linear Regression?

1. Linearity: Data is Linearly Separable

2. Normality: Data is Normal Distributed, that is CLT is applicable.

3. Independence: No AutoCorrelation.

4. Homoscedasticity: The variance of residual[error] is the same for any value of X.

5. mean of residuals should be 0.

Q.8) What do you understand by Central Limit Theorem?(Imp**)

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n > 30).

Intuitively, if you take means of multiple samples, you can assume that mean of all the samples combines will be equal to the mean of the population. This is really important to understand as this allows us to do significance tests like z-tests.

Q.9) How is Random forest different from XGBoost and other CART Algorithms. What are some advantages of CART Algorithms?

Decision trees are a series of sequential steps designed to answer a question and provide probabilities, costs, or other consequences of making a particular decision. This divides the query in tree-like fashion, making decisions at every node, with leaf nodes being the target.

Decision Tree generally leads to overfitting because of their nature and are prone to bias, to resolve this Random Forest was introduced. It is a collection of decision trees with a single, aggregated result. Random forest combines multiple decision trees by aggregating their result at the 'End'. whereas boosting algorithms like xgboost are also ensemble algorithms like RF but they combine their result in the beginning instead of at the end.

Gradient Boosting build trees one at a time, where each new tree helps to correct errors made by the previously trained tree. Boosting generally performs better than RF.

Advantages of CART are:

> they work with both continuous and categorical data.

> Normalization is not required

> They can be used with very less pre-processing and can yield good results.

Q.10) Why is Collinearity a problem for Logistic Regression?

Multicollinearity is a common problem with linear models, and that is because it leads to unstable estimators of regression coefficients. This can lead to overfitting as well in some cases.

Tho there are cases when you can ignore multicollinearity, and in order to learn them please read from this Blog. I found it full of important information.

https://statisticalhorizons.com/multicollinearity

Q 11.) What is the Significance of correlation in data? What do you mean by Pearson correlation? what value of Pearson correlation is significant acc. to you?

Correlation explains how one or more variables are related to each other. A +ve relation means that if the value of X increases then the value of Y also increases. Correlation is imp as it'll help us reduce our features.

If 2 columns are highly correlated, it is better to remove 1 of them as both will provide same info to model. Just finding a correlation that tells you whether the relation is + or -ve is not enough. We need a metric, Pearson correlation is just that, It is a measure of the strength of a linear association between two variables. The value lies between -1 to 1, 1 being strong +ve relation 0 means no relation and -1 means strong -ve relation. values above 0.9 and below -0.9 are

generally discarded. values close to 0 are also discarded in some cases.

Q 12.) Difference between correlation and covariance? When should we prefer one over the other?

Correlation is a function of the covariance. Covariance is used to determine how much two random variables vary together, whereas correlation is used to determine when a change in one variable can result in a change in another. Covariance gives the extent to which two random variables change. Correlation represents how strongly two random variables are related.

Tho not a rule, but is preferred to use the covariance matrix when the variable are on similar scales and the correlation matrix when the scales of the variables differ.

Q 13.) When should we do chi sq test? what is it used for?

A Chi-Square test is a test of statistical significance for categorical variables. We typically use it to find how the observed value of a given event is significantly different from the expected value. Intuitively, we see whether our categorical feature is relevant/significant to predict the output variable. We use it for feature selection. We remove values that are not so significant acc to test.

Q 14.) If both filter kernels and pooling layers extract features from an image in CNN, then what is the difference between them?

Filters extract information from the image like Edges, Patters, etc. This info that needs to be extracted are learned during training. It depends largely on Data what info needs to be extracted from filters. Whereas Pooling does not extract features, rather it takes the avg value of nearby n pixels so that the most important info in the image is retained reducing the image size. Which is req to make large models.

Q 15.) Difference between vectors and scalers?

Scalars are quantities that are fully described by a magnitude (or numerical value) alone. Vectors are quantities that are fully described by both a magnitude and a direction.

Q 16.) If a dataset has 1000 columns, how will you deal with them? Can we perform DR with Deep Learning?

When we have huge datasets with many features it is hard to visualize and group the data, and also we need to see if 2 features are highly correlated.
Thus, we use some Techniques to map multiple dimensions into small space, such as 2 dimensions for plottings. This is called Dimensionality Reduction.

Some Common Techniques are:

- Linear Discriminant Analysis (LDA)
- Principal component analysis (PCA)
- A high correlation between two columns
- Backward/Forward Elimination
- t-distributed Stochastic Neighbor Embedding (t-SNE)

Yes, we can use techniques such as AutoEncoders, Self Organizing Maps(SOM) to perform DR using Deep Learning.

Q 17.) How will you find if a feature is significant to your model?

There is a saying in DS: "Garbage in Garbage out". When we have a large no. of features, it is a good idea to see if 2 or more features have 1:1 mapping btw them or are some features too correlated or are they even significant to predict our output?
In order to see their significance to predict output, we use 2 common Elimination Techniques.
Forward and Backward elimination.

Backward elimination is a feature selection technique while building a machine learning model.

It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output.

- We fit the model with all features first and find the p-value for all columns.
- Then we remove the column with the highest p-value, provided it is more than our significance level, typically 0.05.
- Then we fit the model again and do the same until only columns with p-value<SL are left.

Forward Elimination is just the opposite of that.
There are many other ways as well to find imp. Features as we will discuss in further questions.

Q 18.) How can you avoid the overfitting your model?

Overfitting is when the model learns the training data too much and doesn't generalize well.
In order to prevent overfitting we can take these steps:
Reduce the complexity of the model.
Add Dropout in case of ANN
Use cross-validation techniques, such as k folds cross-validation
Use Regularization techniques such as L1,L2 regularization.
I gave an explanation on this in one of my prev questions:
Here is the link:
https://www.linkedin.com/posts/alaapdhall_datascience-python-machinelearning-activity-6708287741261606912-xllZ

Q 19.) What are some assumptions or pre-requisite for Pearson correlation?

For the Pearson r correlation, both variables should be normally distributed.

There should be no significant outliers.
Each variable should be continuous
The two variables have a linear relationship.
The observations are paired observations.
Homoscedascity- Homoscedascity simply refers to 'equal variances'.

I talked about the Pearson correlation in prev question.

Here→
https://www.linkedin.com/posts/alaapdhall_datascience-python-machinelearning-activity-6708985598704705536-ayAM/

Q 20.) How do you see if your model's inputs are explaining the outcome properly? Can you use r^2 and adjusted r^2?

Yes, These metrics can help us judge the goodness of the model.
R-squared explains the degree to which your input variables explain the variation of your output / predicted variable. So, if R-square is 0.8, it means 80% of the variation in the output variable is explained by the input variables.
The problem with R-squared is that it will either stay the same or increase with the addition of more variables, even if they do not have any relationship with the output variables. This is where "Adjusted R square" comes to help. Adjusted R-square penalizes you for adding variables that do not improve your existing model.

Q 21.) What if your dataset has 1 column related to multiple columns of your dataset, how will you handle your data?

In this case, if we draw a correlation heat map, we will see a lot of collinearity and won't be able to decide the columns we want to remove.

This problem is known as the MultiCollinearity Problem. One good way to deal with Multicollinearity is the Variance Inflation Factor (VIF).

Variance inflation factors (VIF) measure how much one column is related to all others in terms of their dependency. It is obtained by regressing each independent variable, say X on the remaining independent variables (say Y and Z) and checking how much of it (of X) is explained by these variables.

Intuitively, we just perform linear regression with all the columns as target variable one by one, in eat iteration we keep 1 column as target and the rest as features and see how it is explained by other variables. We do this with all the columns.

Hence,

In general, the Higher the VIF, the higher the R2 which means the variable X is collinear with Y and Z variables. Columns with values of VIF above 2 or 5 and in some cases 10, are removed.

Q 22.) How to handle skewness?]What are the mathematical properties of skewed data? How can you fix it?

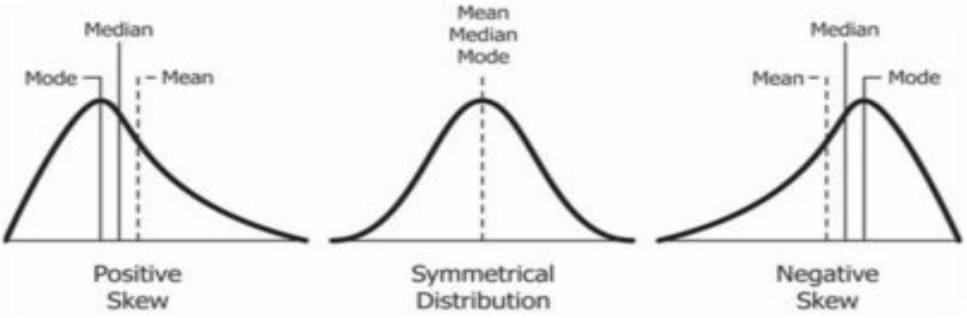
Skewness is when the distribution of data which is concentrated in 1 side(left or right) more, or it is sort of collected on one side. data can be left-skewed or right-skewed.

Right skewed is when data is concentrated on the left, and the tail is on right. This is also known as positive skewed. Left is -ve skewed when the tail is on left.

Skewness is caused by the presence of Outliers in the data.

In mathematical terms, the right skewness is when the mean is greater than the median.

We can remove Skewness by omitting Outliers or taking Log of the features.



Q 23.) Differentiate between univariate, bivariate, and multivariate analysis.

- Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.
- Bivariate data involves two different variables. The analysis of this type of data deals with causes and

relationships and the analysis is done to determine the relationship between the two variables.

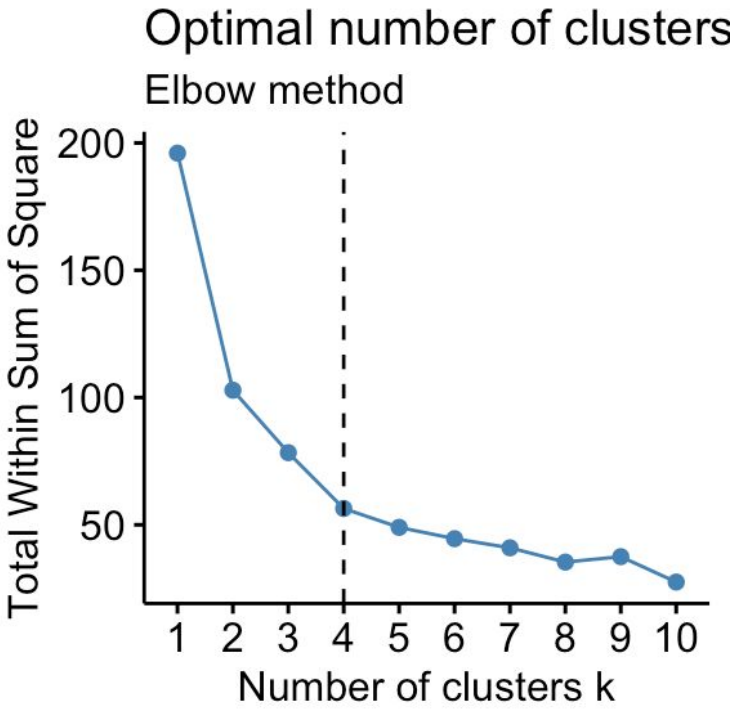
- Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable.

Q 24.) In K means clustering, how will you decide the number of clusters to chose? How do you deal with the bad initialization problem?

In K-means clustering we find clusters of similar data points. If we chose more clusters than what are needed, it'll show wrong clustering.

One Brute-Force way to fix is trying all possible clusters from a given range and see what works. The way is do that is calculating the sq distance btw them. This is called the Elbow method, we use within the sum of squares (WSS) to see which clusters work best. WSS is defined as the sum of the squared distance between each member of the cluster and its centroid.

Less number of clusters means more WSS. We continue to increase our number of clusters to the point where our square distance is not decreasing much.



Q 25.) 'People who bought this also bought…' recommendations seen on Amazon are a result of which algorithm?

Recommendation systems are built by 2 methods. Content-based filtering and collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.

It takes into account all the choices of other users and features in order to recommend it. The engine makes predictions on what might interest a person based on the preferences of other users. In this algorithm, item features are unknown.

Learn more about recommendation systems as they are an important Unsupervised Machine Learning ALgorithms.

I will build a song recommendation project using RBM and AutoEncoder on my website https://www.aiunquote.com soon this week. Stay updated. :)

- Alaap Dhall

Follow Alaap Dhall on LinkedIn for more insights in Data Science and Deep Learning!!

Visit https://www.aiunquote.com for a 100 project series in Deep Learning.