Day 4 of 100 Data Science Interview Questions Series!!

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.

- 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.