Day 15 for 100 Data Science Interview Questions Series!!

Link to prev 65 Questions:
https://www.linkedin.com/posts/alaapdhall_65-of-100-data-science-interview-questions-activity-6722399764228423680-2VC_

Q 66.) What is the curse of dimensionality?

Generally more features the better, or so we've heard. Basically, More features can make it harder for your model to make any sense out of that clutter.

Here are some scenarios where you might consider this problem and apply some Dimensionality Reduction technique. When confronted with a ton of data, we can use dimensionality reduction algorithms to make the data "get to the point".

- If we have more features than observations then we run the risk of massively overfitting our model — this would generally result in terrible out-of-sample performance.
- When we have too many features, observations become harder to cluster — believe it or not, too many dimensions cause every observation in your dataset to appear equidistant from all the others. And because clustering uses a distance measure such as Euclidean distance to quantify the similarity between observations, this is a big problem.

Q 67.) How can you deal with different types of seasonality in time series modelling?

Seasonality in time series occurs when the time series shows the repeated patterns over time. E.g., stationary sales decreases during the holiday season, air conditioner sales increases during the summers, etc. are few examples of seasonality in time-series.

Seasonality makes your time series non-stationary because the average value of the variables at different time periods. Differentiating a

time series is generally known as the best method of removing seasonality from a time series.

Seasonal differencing can be defined as a numerical difference between a particular value and a value with a periodic lag (i.e. 12, if monthly seasonality is present)

Q 68.) Can you cite some examples where both false positive and false negatives are equally important?

In the banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses. Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

These days we hear many cases of players using steroids during sports competitions Every player has to go through a steroid test before the game starts. A false positive can ruin the career of a great sportsman and a false negative can make the game unfair. Both are equally important.

Q 69.) What do you understand by the statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, RF, etc.). Sensitivity is nothing but "Predicted TRUE events/ Totalevents". True events here are the events that were true and the model also predicted them as true.

Calculation of sensitivity is pretty straight forward - **Sensitivity** = True Positives /Positives in Actual Dependent Variable, Where, True positives are Positive events which are correctly classified as Positives.

Q 70.) Give some situations where you will use an SVM over a RandomForestMachine Learning algorithm and vice-versa

SVM and Random Forest are both used in **classification** problems.

1. If you are sure that your data is outlier free and clean then go for SVM.
2. And if your data might contain outliers then Random forest would be the better choice
3. Generally, SVM consumes more computational power than RandomForest, so if you are constrained with memory go for Random Forest machine learning algorithm.
4. Random Forest gives you a very good idea of variable importance in your data, so if you want to have variable importance then choose the Random Forest machine learning algorithm.
5. Random Forest machine learning algorithms are preferred for multi-class problems.
6. SVM is preferred in multi-dimensional problem set - like text classification but as a good data scientist, you should experiment with both of them and test for the accuracy, or rather you can use ensemble of many MachineLearning techniques.

- 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.