Day 7 of 100 Data Science Interview Questions Series!!

Q 31.) What are Entropy and Information gain in Decision tree algorithm?

The core algorithm for building a decision tree is called ID3. ID3 uses Entropy and Information Gain to construct a decision tree.

Entropy: A decision tree is built top-down from a root node and involves the partitioning of data into homogeneous subsets. ID3 uses entropy to check the homogeneity of a sample.

If the sample is completely homogeneous then entropy is zero and if the sample is equally divided it has an entropy of one.

Information Gain: The Information Gain is based on the decrease in entropy after a dataset is split on an attribute.

Constructing a decision tree is all about finding attributes that return the highest information gain.

Check this great article out:
https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134

Q 32.) What is Ensemble Learning?

The ensemble is the art of combining a diverse set of individual models together to improvise on the stability and predictive power of the model.

Ensemble learning has many types but two more popular ensemble learning techniques are mentioned below.

Bagging: It tries to implement similar learners on small sample populations and then takes a mean of all the predictions.

Boosting: It is an iterative technique that adjusts the weight of an observation based on the last classification.

If an observation was classified incorrectly, it tries to increase the weight of this observation and vice versa.

A rather good article I found for you:
https://www.analyticsvidhya.com/blog/2015/08/introduction-ensemble-learning/

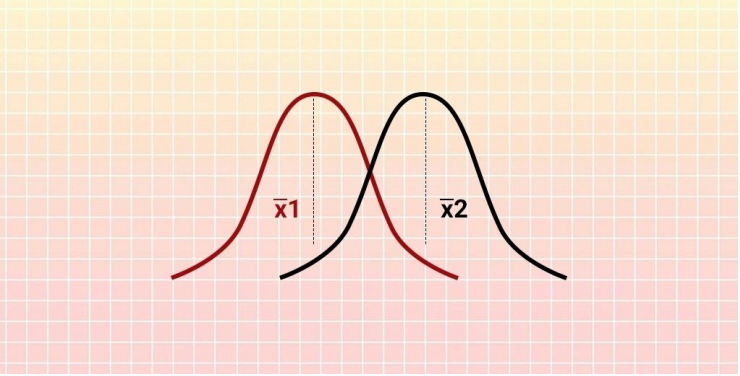
Q 33.) When do you use T-test in Data **Science**?

It helps us understand if the difference between two sample means is actually real or simply due to chance.

Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement by assuming a null hypothesis that the two means are equal. Based on the applicable formulas, certain values are calculated and compared against the standard values, and the assumed null hypothesis is accepted or rejected accordingly.

If the null hypothesis is rejected, it indicates that data readings are strong and are probably not due to chance. The t-test is just one of many tests used for this purpose.

The link you must go through:
https://www.analyticsvidhya.com/blog/2019/05/statistics-t-test-introduction-r-implementation/



Q 34.) How do you deal with Unbalanced Data?

Unbalanced data is very common in the real-world data. Let's say we have 2 classes with 1 having 5000 eg and the other having 500.

- The most common way to deal with this is to Resample, i.e take 50-50 proportion from both the classes.[500-500 in our case]
- Another way is that you can improve the balance of classes by Upsampling the minority class or by Downsampling the majority class.
- Another method to improve unbalanced binary classification is by increasing the cost of misclassifying the minority class with your Loss function. By increasing the penalty of such, the model should classify the minority class more accurately.

Q 35.) What cross-validation technique would you use on a time series data set.

We can't use k-fold cross-validation with **TimeSeries** as time series is not randomly distributed data, and has temporal info. It is inherently ordered by chronological order, so we can not split randomly.

In the case of time-series data, you should use techniques like forward chaining — Where you will be model on past data then look at forward-facing data.

We can use TimeSeriesSplit from sklearn to do split data in train-test.

- 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.