100 Data Science Interview Questions Series!!

Here are the first 50 questions.

First 25 Question, (Q1 to Q25) can be found here:
https://www.linkedin.com/posts/alaapdhall_25-of-100-data-science-interview-questions-activity-6711212704985624576-0XqH

Q 26.) How can you use eigenvalue or eigenvector?

It is difficult to understand and visualize data with more than 3 dimensions, let alone a dataset of over 100+ dimensions. Hence, it would be ideal to somehow compress/transform this data into a smaller dataset. This is where we can use this concept.

We can utilize Eigenvalues and Eigenvectors to reduce the dimension space ensuring most of the key information is maintained.

Eigenvalues are the directions along which a particular linear transformation acts by flipping, compressing, or stretching.

Eigenvectors are for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix.

Please view this article which has explained this concept better than I ever could!
https://medium.com/fintechexplained/what-are-eigenvalues-and-eigenvectors-a-must-know-concept-for-machine-learning-80d0fd330e47

Q 27.) What is lemmatization and Stemming, Which one should I use in Sentimental Analysis, and which one should I use in QnA bot?

They are used as Text Normalization techniques in NLP for preprocessing text.

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language."

Follow: Alaap Dhall on LinkedIn for more!

Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma.

- Stemming is a better option for Sentimental Analysis as the meaning of the word is not necessary for understanding sentiments, and stemming is a little faster than Lemmatization.
- Lemmatization is better for QnA bot as the word should have a proper meaning while conversing with a human subject.

Q 28.) What are some common Recommendation System Types, where can I use them?

Recommendation systems are used to recommend or generate some outputs based on previous inputs that were given by users. Recommendation system can be built through Deep Learning, like Deep Belief networks, RBM, AutoEncoder, etc or some traditional techniques.

Some common types are:

1. Collaborative Recommender system
2. Content-based recommender system
3. Demographic-based recommender system
4. Utility-based recommender system
5. Knowledge-based recommender system
6. Hybrid recommender system.

- DL based Recommendation systems can be used for dimensionality reduction and generating similar output.
- RS can also be used for suggestions of similar items based on the user's past choices and item's content.
- RS can also be used for suggestions of similar products based on a group of users with similar features as you.

Q 29.) What is bias, variance trade-off?

Bias is the error introduced in your model due to oversimplification of the machine learning algorithm.” It can lead to underfitting.

Follow: Alaap Dhall on LinkedIn for more!

- Low bias machine learning algorithms — Decision Trees, k-NN and SVM
- High bias machine learning algorithms — Linear Regression, Logistic Regression

Variance is the error introduced in your model due to the complex machine learning algorithm, your model learns noise also from the training data set and performs badly on test data set. It can lead to high sensitivity and overfitting.

Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens toill a particular point. **As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance. Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.** This is Bias-Variance Trade-Off.

Q 30.) What are vanishing/exploding gradients?

Gradient is the direction and magnitude calculated during the training of a neural network that is used to update the network weights in the right direction and by the right amount.

- Exploding gradient is a problem where large error gradients accumulate and result in very large updates to neural network model weights during training.
- Vanishing gradient is a problem whereas more layers are added to neural networks, the gradients of the loss function approach zero, making the network hard to train. This occurs in large models with many layers. **Models like ResNet, that have skip connections, are a good solution to this problem.**

Q 31.) What are Entropy and Information gain in the 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/

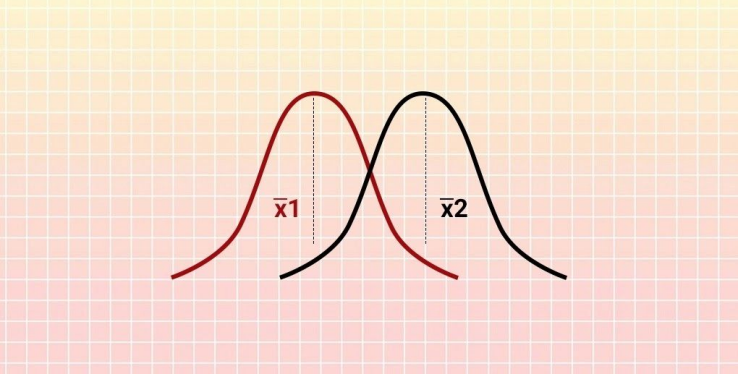
Q 33.) When do you use T-test in Data **Science**?

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 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.

Q 36.) Given a data set of features X and labels y , what assumptions are made when using Naive Bayes methods?

The Naive Bayes algorithm assumes that the features of X are conditionally independent of each other for the given Y.

The idea that each feature is independent of each other may not always be true, but we assume it to be true to apply Naive Bayes. This "naive" assumption is where the namesake comes from.

Q 37.) What is a Box-Cox Transformation?

A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

The residuals could either curve as the prediction increases or follow the skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions.

A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality.

Q 38.) Where do you use TF/IDF vectorization?

Follow: Alaap Dhall on LinkedIn for more!

The tf-idf is short for term frequency-inverse document frequency. It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Q 39.) Tell me about Pattern Recognition and what areas in which it is used?

Pattern recognition is the process of recognizing patterns by using machine learning algorithm. Pattern recognition can be defined as the classification of data based on knowledge already gained or on statistical information extracted from patterns and/or their representation.

Pattern Recognition can be used in

- Computer Vision
- Speech Recognition
- Data Mining
- Statistics
- Informal Retrieval
- Bio-Informatics

Q 40.) What is the difference between Type I vs Type II error?

A type I error occurs when the null hypothesis (H0) is true but is rejected. It is asserting something that is absent, a false hit. A type I error may be likened to a so-called false positive (a result that indicates that a given condition is present when it actually is not present).

A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected. It is failing to assert what is present, a miss.
A type II error may be compared with a so-called false negative (where an actual 'hit' was disregarded by the test and seen as a 'miss') in a test checking for a single condition with a definitive result of true or false.

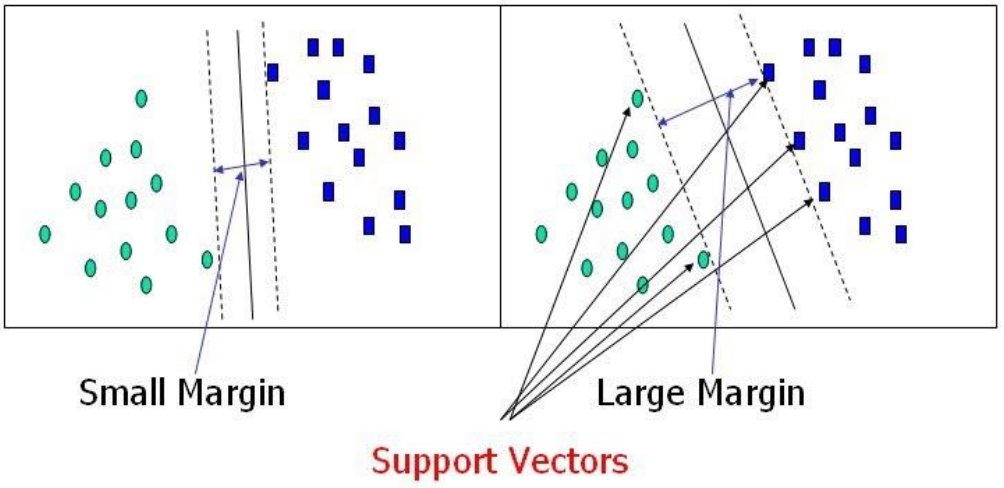
Follow: Alaap Dhall on LinkedIn for more!

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision About Null Hypothesis (H_0)	**Reject**	Type I error (False Positive)	Correct inference (True Positive)
	Fail to reject	Correct inference (True Negative)	Type II error (False Negative)

Q 41.) Describe how the support vector machine (SVM) algorithm works, or any other algorithm that you've used.

The objective of the support vector machine algorithm is to find a hyperplane in N-dimensional space(N — the number of features) that distinctly classify the data points.

SVM attempt to find a hyperplane that separates classes by maximizing the margin.



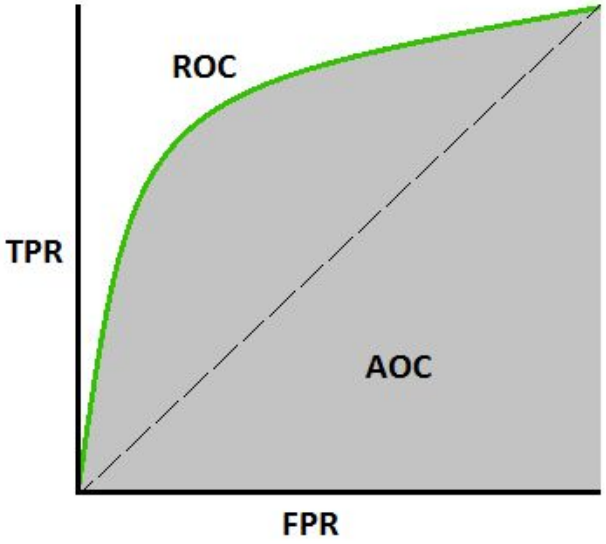
The Edge points in this diagram are the support vectors, against the decision hyperplane. These are the extreme values that represent the data and thus are used to do classification. They in a way support the data, thus known as support vector machine.

Here we show linear classification, but SVMs can perform nonlinear classification. SVMs can employ the kernel trick which can map linear non-separable inputs into a higher dimension where they become more easily separable.

Follow: Alaap Dhall on LinkedIn for more!

Q 42.) How and when can you use ROC Curve?

The ROC curve is a graphical representation of the contrast between true positive rates and false-positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity(true positive rate) and the false-positive rate. It tells how much the model is capable of distinguishing between classes. Higher the AUC(area under the curve of ROC), the better the model is at predicting 0s as 0s and 1s as 1s.



Intuitively, in a logistic regression we can have many thresholds, thus what we can do is check the model's performance on every threshold to see which works best. Calculate ROC at every threshold and plot it, this will give you a good measure of how your model is performing.

Q 43.) Give one scenario where false positive is more imp than false negative, and vice versa.

A false positive is an incorrect identification of the presence of a condition when it's absent.

A false negative is an incorrect identification of the absence of a condition when it's actually present.

An example of when false negatives are more important than false positives is when screening for cancer. It's much worse to say that

someone doesn't have cancer when they do, instead of saying that someone does and later realizing that they don't.
This is a subjective argument, but false positives can be worse than false negatives from a psychological point of view. For example, a false positive for winning the lottery could be a worse outcome than a false negative because people normally don't expect to win the lottery anyway.

Q 44.) Why we generally use Softmax non-linearity function in last layer but ReLU in rest? Can we switch?

We use Softmax because it takes in a vector of real numbers and returns a probability distribution between 0 and 1, which is useful when we want to do classification.

We use ReLU in all other layers because it keeps the original value and removes all the -ve, max(0,x). This performs better in general but not in every case and can easily be replaced by any other activation function such as tanh, sigmoid, etc.

Q 45.) What do you understand by p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.

Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. A high p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,

High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

Q 46.) How to check if the regression model fits the data well?

There are a couple of metrics that you can use:

- R-squared/Adjusted R-squared: Relative measure of fit. This was explained in a previous answer
- F1 Score: Evaluates the null hypothesis that all regression coefficients are equal to zero vs the alternative hypothesis that at least one doesn't equal zero
- RMSE: Absolute measure of fit.

Q 47.) Let's say you have a categorical variable with thousands of distinct values, how would you encode it?

This depends on whether the problem is a regression or a classification model.

- If it's a regression model, one way would be to cluster them based on the response variable by working backwards. You could sort them by the response variable, and then split the categorical variables into buckets based on the grouping of the response variable. This could be done by using a shallow decision tree to reduce the number of categories.
- For a binary classification, you can target encode the column by finding the conditional probability of the response variable being a one, given that the categorical column takes a particular value. Then replace the categorical column with this numerical value. For example if you have a categorical column of city in predicting loan defaults, and the probability of a person who lives in San Francisco defaults is 0.4, you would then replace "San Francisco" with 0.4.
- We could also try using a Louvain community detection algorithm. Louvain is a method to extract communities from large networks without setting a pre-determined number of clusters like K-means.

Q 48.) 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.

Q 49.) Why is mean square error a bad measure of model performance? What would you suggest instead?

Mean Squared Error (MSE) gives a relatively high weight to large errors — therefore, MSE tends to put too much emphasis on large deviations. A more robust alternative is MAE (mean absolute deviation) or Root MEan Square Error.

Q 50.) What is cross-validation?

Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will generalize to an independent dataset. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

- 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.