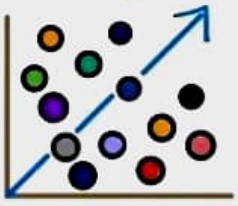
Data Science Interview Questions

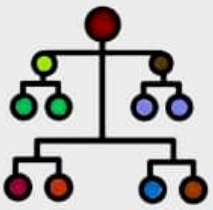
Essential Machine Learning Algorithm

Linear Regression



Logistic Regression

Decision Tree

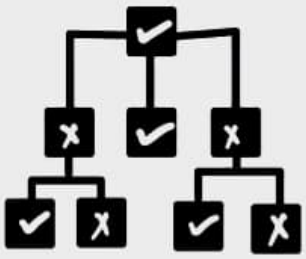


SVM

KNN

Dimensionality Reduction

Random Forest



K-Means

Naive Bayes

@codes-learning

Bias-variance trade off: The goal of any supervised machine learning algorithms is to have low bias and low variance to achieve good prediction performance.

1. The K-nearest neighbour algorithm has low bias and high variance, but the trade off can be changed by increasing the value of K which increases the no. of neighbours that contribute to the prediction and in turn increase the bias of the model.
2. The Support Vector Machine Algorithm has low bias and high variance, but the trade off can be changed by increasing the C parameter that influence the no. of violations of the margins allowed in the training data which increases the bias but decrease the variance.

What are the difference b/w overfitting & underfitting?

In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, so as to be able to make reliable predictions on general untrained data.

Under fitting

Over fitting

Can you site some examples where both false positive and false negative 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.
Bank don't want to loose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positive and false negative become very important.

What is 'Naive' in a Naive Bayes?

The 'Naive Bayes Algorithm' is based on the bayes theorem. It describes the probability of an event based on prior knowledge of condition, might be related to the event. It is naive because it makes assumptions that may or may not turn out to be correct.

What are the different Kernels in SVM?

1. Linear Kernel
2. Polynomial Kernel
3. Radia Basis Kernel
4. Sigmoid Kernel

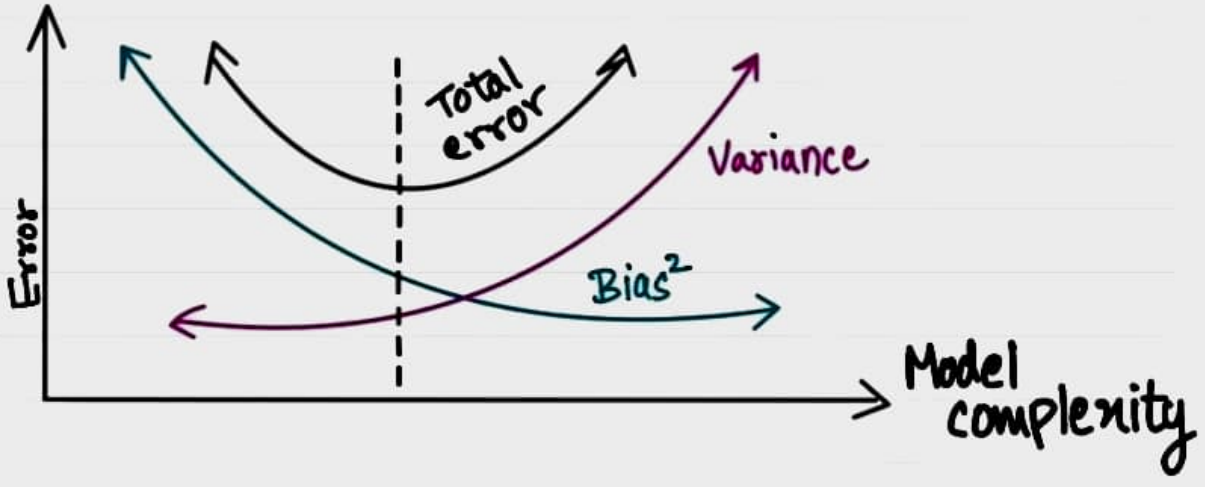
What is pruning in Decision Tree?

It is a technique in machine learning and search algorithm that reduces the size of decision trees by removing sections of the trees that provides little power to classify instance, so when we remove sub-nodes of a decision node. this process is pruning or opposite process of splitting.

What is bias-variance trade-off?

Bias :- Bias is an error introduced in your model due to oversimplification of the machine learning algorithm. It can lead to underfitting, when you train model at that time makes simplified assumptions to make the target function easier to understand.
Low bias machine learning algorithm - Decision trees, K-NN and SVM High bias machine learning algorithms - Linear Regression, Logistic Regression.

Variance :- Variance is error introduced in your model due to complex machine learning Algorithm, your model learns noise also from the training 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 until a particular point. As you continue to make your model more complex you end up over-fitting your model and hence your model will suffer high-violence.



In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as to many parameters relative to no. of observations. It has poor predictive performance, as it overreacts to minor fluctuations in the training data.
Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. It would occur, for ex: when fitting a linear model to non-linear data. It too have predictive performance.

What are the types of biases that can occur during sampling?

* Selection bias
* Under coverage bias
* Survivorship bias

What is selection bias?

Selection bias occurs when the sample obtained is not representative of the population intended to be analysed.

What is Tf/IDf Vectorization?

Tf-IDf is short for term frequency-inverse document frequency, is a numerical statistics that is intended to reflect how important a word is to document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

What is cluster sampling?

A technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied cluster sample is probability sample where each sampling unit is a collection or cluster of elements.

What is systematic sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best ex: systematic sampling is equal probability method.

What are Eigen Vector and Eighen Values?

Eigenvector are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along with a particular linear transformation acts by flipping, compressing or stretching. "Eigenvalue" can be referred to as the strength of the transformation in the direction of eigen vector or the factory by which compression occurs.

What is selection bias?

Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with the researcher where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is distortion of statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account then some conclusions of the study may not be accurate.

The types of selection bias include:

1) **Sampling Bias**: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
2) **Time Interval**: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have similar mean.
3) **Data**: When specific subsets of data are choosen to support a conclusion or rejection of bad data on arbitary grounds. Instead of according to previously stated or generally agreed criteria.
4) **Attrition**: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subject / tests that did not run to completion.