Data Science Interview Questions

1. Where is the confusion matrix used? Which module would you use to show it?

In machine learning, a confusion matrix is one of the easiest ways to summarize the performance of your algorithm. At times, it is difficult to judge the accuracy of a model by just looking at the accuracy because of problems like unequal distribution. So, a better way to check how good your model is to use a confusion matrix. First, let's look at some key terms.

Classification accuracy – This is the ratio of the number of correct predictions to the number of predictions made
True positives – Correct predictions of true events
False positives – Incorrect predictions of true events
True negatives – Correct predictions of false events
False negatives – Incorrect predictions of false events.

The confusion matrix is now simply a matrix containing true positives, false positives, true negatives, false negatives.

Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment)	
		positives	negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

2. What is Accuracy?

It is the most intuitive performance measure and it simply a ratio of correctly predicted to the total observations. We can say as, if we have high accuracy, then our model is best. Yes, we could say that accuracy is a great measure but only when you have symmetric datasets where false positives and false negatives are almost the same.

$$\textbf{Accuracy} = \frac{True\ Positive + True\ Negative}{(True\ Positive + False\ Positive + False\ Negative + True\ Negative)}$$

3. What is Precision?

It is also called the positive predictive value. A number of correct positives in your model that predicts compared to the total number of positives it predicts. Precision = True Positives / (True Positives + False Positives) Precision = True Positives / Total predicted positive It is the number of positive elements predicted properly divided by the total number of positive elements predicted. We can say Precision is a measure of exactness, quality, or accuracy. High precision Means that more or all of the positive results you predicted are correct.

4. What is Recall?

Recall what we can also call sensitivity or true positive rate. It is several positives that our model predicts compared to the actual number of positives in our data. Recall = True Positives / (True Positives + False Positives) Recall = True Positives / Total Actual Positive Recall is a measure of completeness. High recall means that our model classified most or all of the possible positive elements as positive.

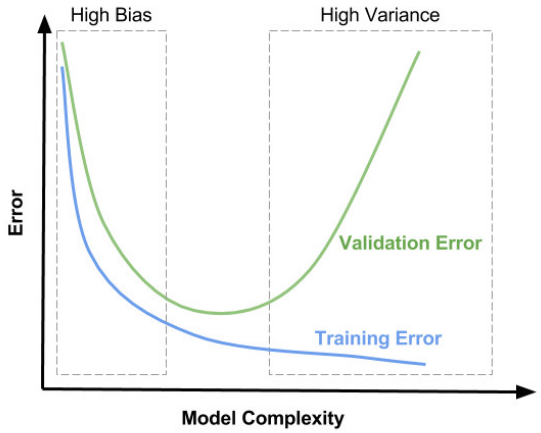
5. What is F1 Score?

We use Precision and recall together because they complement each other in how they describe the effectiveness of a model. The F1 score combines these two as the weighted harmonic mean of precision and recall.

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

6. What is Bias and Variance trade-off?

Bias means it's how far are the predicted values from the actual values. If the average predicted values are far off from the actual values, then we called this one have high bias. When our model has a high bias, then it means that our model is too simple and does not capture the complexity of data, thus underfitting the data. Variance It occurs when our model performs good on the trained dataset but does not do well on a dataset that it is not trained on, like a test dataset or validation dataset. It tells us that actual value is how much scattered from the predicted value. Because of High variance it cause overfitting that implies that the algorithm models random noise present in the training data. When model have high variance, then model becomes very flexible and tune itself to the data points of the training set.



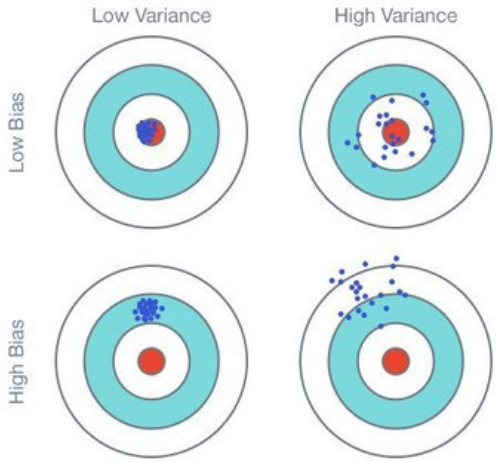


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

7. Why is normalization required before applying any machine learning model? What module can you use to perform normalization?

Normalization is a process that is required when an algorithm uses something like distance measures. Examples would be clustering data, finding cosine similarities, creating recommender systems. Normalization is not always required and is done to prevent variables that are on higher scale from affecting outcomes that are on lower levels. For example, consider a dataset of employees' income. This data won't be on the same scale if you try to cluster it. Hence, we would have to normalize the data to prevent incorrect clustering. A key point to note is that normalization does not distort the differences in the range of values. A problem we might face if we don't normalize data is that gradients would take a very long time to descend and reach the

global maxima/ minima. For numerical data, normalization is generally done between the range of 0 to 1. The general formula is:

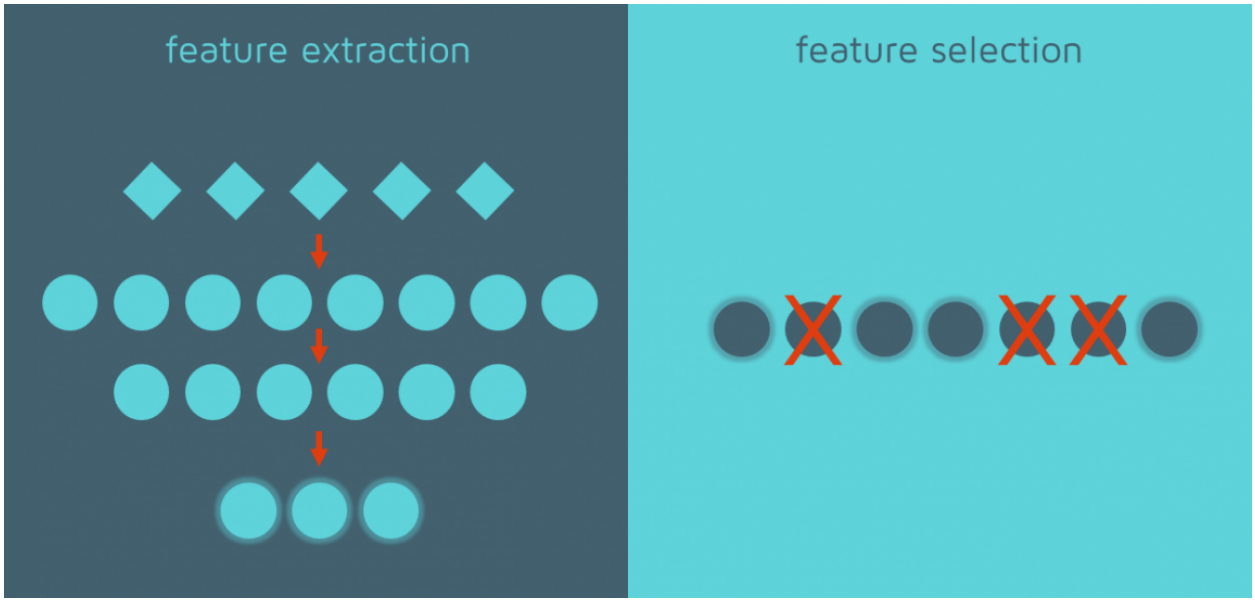
Xnew = (x-xmin)/(xmax-xmin)

8. What is the difference between feature selection and feature extraction?

Feature selection and feature extraction are two major ways of fixing the curse of dimensionality

1. Feature selection - Feature selection is used to filter a subset of input variables on which the attention should focus. Every other variable is ignored. This is something which we, as humans, tend to do subconsciously. Many domains have tens of thousands of variables out of which most are irrelevant and redundant. Feature selection limits the training data and reduces the amount of computational resources used. It can significantly improve a learning algorithms performance. In summary, we can say that the goal of feature selection is to find out an optimal feature subset. This might not be entirely accurate, however, methods of understanding the importance of features also exist. Some modules in python such as Xgboost help achieve the same.

2. Feature extraction - Feature extraction involves transformation of features so that we can extract features to improve the process of feature selection. For example, in an unsupervised learning problem, the extraction of bigrams from a text, or the extraction of contours from an image are examples of feature extraction. The general workflow involves applying feature extraction on given data to extract features and then apply feature selection with respect to the target variable to select a subset of data. In effect, this helps improve the accuracy of a model.

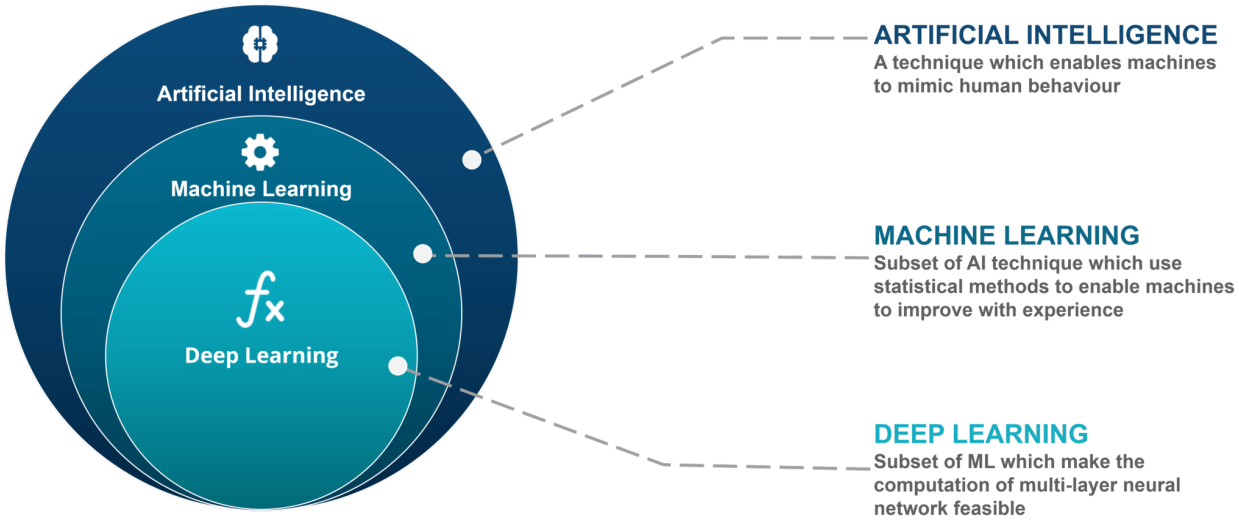


9. Why is polarity and subjectivity an issue?

Polarity and subjectivity are terms which are generally used in sentiment analysis. Polarity is the variation of emotions in a sentence. Since sentiment analysis is widely dependent on emotions and their intensity, polarity turns out to be an extremely important factor. In most cases, opinions and sentiment analysis are evaluations. They fall under the categories of emotional and rational evaluations. Rational evaluations, as the name suggests, are based on facts and rationality while emotional evaluations are based on non-tangible responses, which are not always easy to detect. Subjectivity in sentiment analysis, is a matter of personal feelings and beliefs which may or may not be based on any fact. When there is a lot of subjectivity in a text, it must be explained and analysed in context. On the contrary, if there was a lot of polarity in the text, it could be expressed as a positive, negative or neutral emotion.

10: How would you define Machine Learning?

Machine learning: It is an application of artificial intelligence (AI) that provides systems the ability to learn automatically and to improve from experiences without being programmed. It focuses on the development of computer applications that can access the data and used it to learn for themselves. The process of learning starts with the observations or data, such as examples, direct experience, or instruction, to look for the patterns in data and to make better decisions in the future based on examples that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.



11. What is a labeled training set?

Machine learning is derived from the availability of the labeled data in the form of a training set and test set that is used by the learning algorithm. The separation of data into the training portion and a test portion is the way the algorithm learns. We split up the data containing known response variable values into two pieces. The training set is used to train the algorithm, and then you use the trained model on the test set to predict the variable response values that are already known. The final step is to compare with the predicted responses against actual (observed) responses to see how close they are. The difference is the test error metric. Depending on the test error, you can go back to refine the model and repeat the process until you're satisfied with the accuracy

12. What are the two common supervised tasks?

The two common supervised tasks are regression and classification.

Regression - The regression problem is when the output variable is the real or continuous value, such as "salary" or "weight." Many different models can be used, and the simplest is linear regression. It tries to fit the data with the best hyper-plane, which goes through the points.

Classification - It is the type of supervised learning. It specifies the class to which the data elements belong to and is best used when the output has finite and discrete values. It predicts a class for an input variable, as well.

13. What type of algorithm would we use to segment your customers into multiple groups?

If we don't know how to define the groups, then we can use the clustering algorithm (unsupervised learning) to segment our customers into clusters of similar customers. However, if we know what groups we would like to have, then we can feed many examples of each group to a classification algorithm (supervised learning), and it will classify all your customers into these groups.

14. What is the Model Parameter?

Model parameter: It is a configuration variable that is internal to a model and whose value can be predicted from the data. While making predictions, the model parameter is needed. The values define the skill of a model on problems. It is estimated or learned from data. It is often not set manually by the practitioner. It is often saved as part of the learned model. Parameters are key to machine learning algorithms. They are part of the model that is learned from historical training data.

15. What is Model Hyperparameter?

Model hyperparameter: It is a configuration that is external to a model and whose values cannot be estimated from the data. It is often used in processes to help estimate model parameters. The practitioner often specifies them. It can often be the set using heuristics. It is tuned for the given predictive modeling problems. We cannot know the best value for the model hyperparameter on the given problem. We may use the rules of thumb, copy values used on other problems, or search for the best value by trial and error.