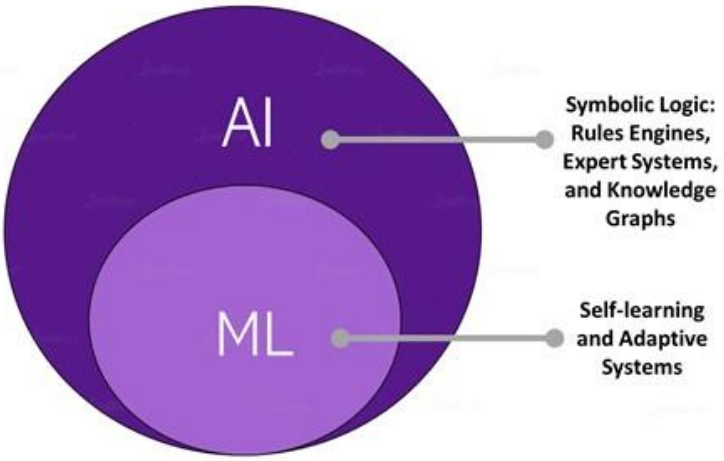
Day 9: Data Science interview questions

Q1: How would you define Machine Learning?

Ans: **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.

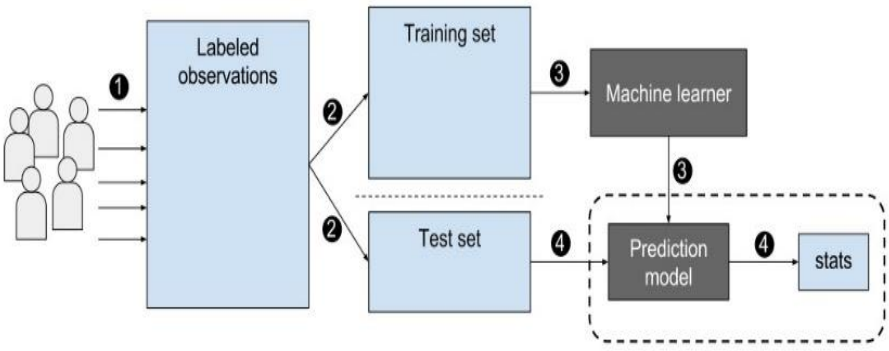


The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.

Q2. What is a labelled training set?

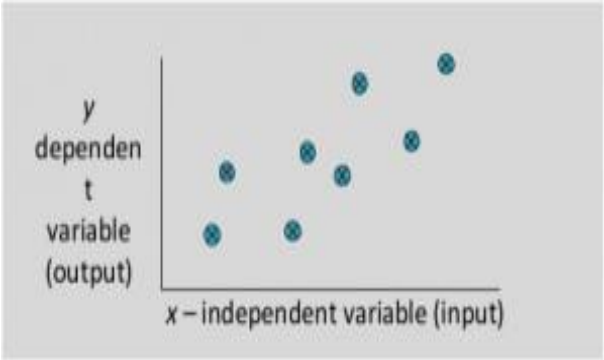
Ans: **Machine learning** is derived from the availability of the labelled data in the form of a ***training set*** and ***test set*** that is used by the learning algorithm. The separation of data into the training portion and a test portion is the way the algorithm learns.

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 are satisfied with the accuracy.



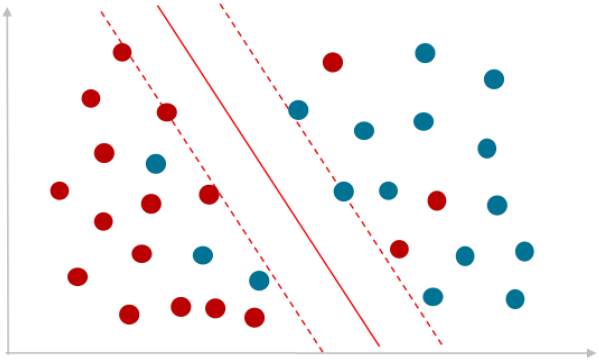
Q3. What are the two common supervised tasks?

Ans: 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.

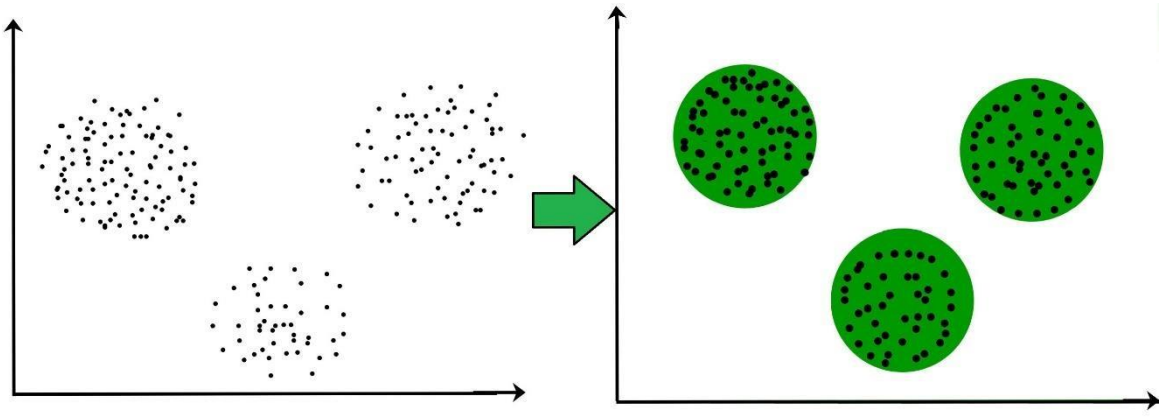
Q4. Can you name four common unsupervised tasks?

Ans: The common unsupervised tasks include clustering, visualization, dimensionality reduction, and association rule learning.

Clustering

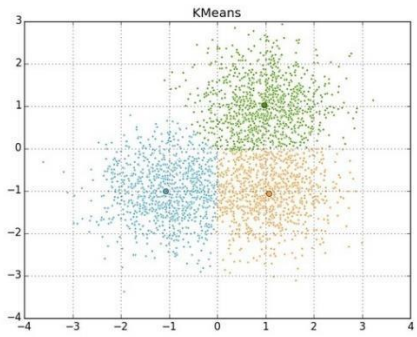
It is a Machine Learning technique that involves the grouping of the data points. Given a set of data points, and we can use a clustering algorithm to classify each data point into the specific group.

In theory, data points that lie in the same group should have similar properties and/ or features, and data points in the different groups should have high dissimilar properties and/or features. Clustering is the method of unsupervised learning and is a common technique for statistical data analysis used in many fields.



Visualization

Data visualization is the technique that uses an array of static and interactive visuals within the specific context to help people to understand and make sense of the large amounts of data.



The data is often displayed in the story format that visualizes patterns, trends, and correlations that may go otherwise unnoticed.

It is extensively used as an avenue to monetize data as the product. An example of using monetization and data visualization is Uber. The app combines visualization with real-time data so that customers can request a ride.

Q5. What type of Machine Learning algorithm we use to allow a robot to walk in various unknown terrains?

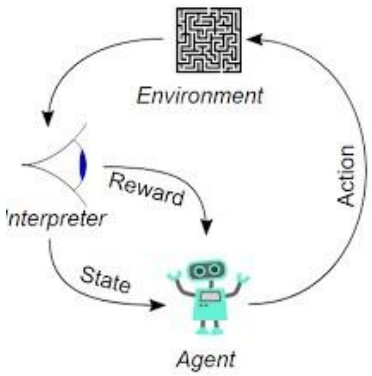
Ans: Reinforcement Learning is likely to perform the best if we want a robot to learn how to walk in the various unknown terrains since this is typically the type of problem that the reinforcement learning tackles.

It may be possible to express the problem as a supervised or semi supervised learning problem, but it would be less natural.

Reinforcement Learning-

It is about to take suitable actions to maximize rewards in a particular situation. It is employed by the various software and machines to find out the best possible behaviour/path it should take in specific situations.

Reinforcement learning is different from the supervised learning in a way that in supervised learning, training data has answer key with it so that the model is trained with the correct answer itself, but in reinforcement learning, there is no answer, and the reinforcement agent decides what to do to perform the given task. In the absence of the training dataset, it is bound to learn from its experience.



Q6. What type of algorithm would we use to segment your customers into multiple groups?

Ans: If we do not know how to define the groups, then we can use the clustering algorithm (unsupervised learning) to segment our customers into clusters of similar customers.

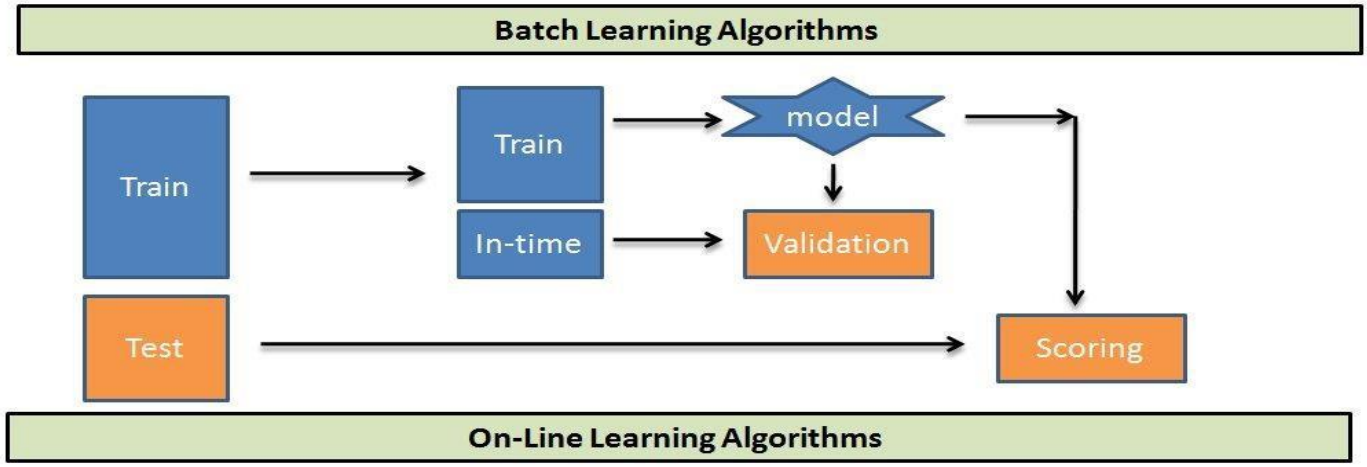
However, if we know what groups we would like to have, then we can feed many examples of each group to a classification algorithm (supervised learning), and it will classify all your customers into these groups.

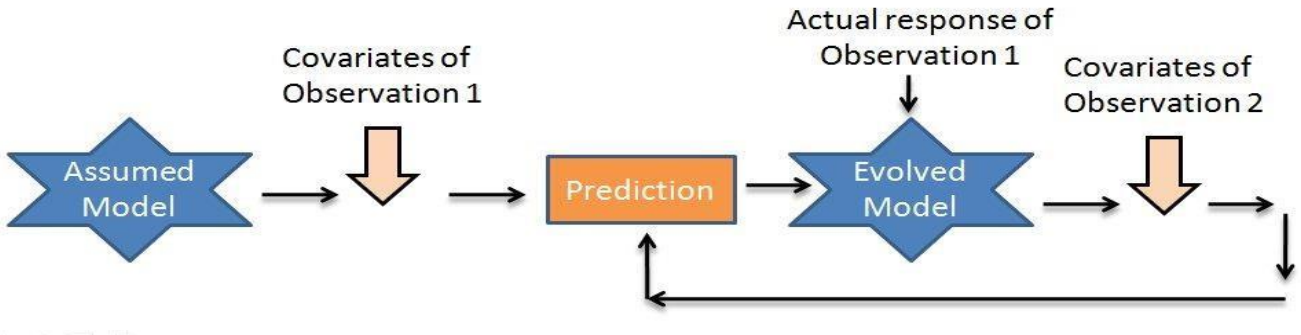
Q7: What is an online machine learning?

Ans: Online machine learning: It is a method of machine learning in which data becomes available in sequential order and to update our best predictor for the future data at each step, as opposed to batch learning techniques that generate the best predictor by learning on entire training data set at once.

Online learning is a common technique and used in the areas of machine learning where it is computationally infeasible to train over the datasets, requiring the need for Out-of-Core algorithms. It is also used in situations where the algorithm must adapt to new patterns in the data dynamically or when the data itself is generated as the function of time, for example, stock price prediction.

Online learning algorithms might be prone to catastrophic interference and problem that can be addressed by the incremental learning approaches.





Q8: What is out-of-core learning?

Ans: Out-of-core: It refers to the processing data that is too large to fit into the computer's main memory.

Typically, when the dataset fits neatly into the computer's main memory, randomly accessing sections of data has a (relatively) small performance penalty.

When data must be stored in a medium like a large spinning hard drive or an external computer network, it becomes awfully expensive to seek an arbitrary section of data randomly or to process the same data multiple times. In such a case, an out-of-core algorithm will try to access all the relevant data in a sequence.

However, modern computers have deep memory hierarchy, and replacing random access with the sequential access can increase the performance even on datasets that fit within memory.

Q9. What is the Model Parameter?

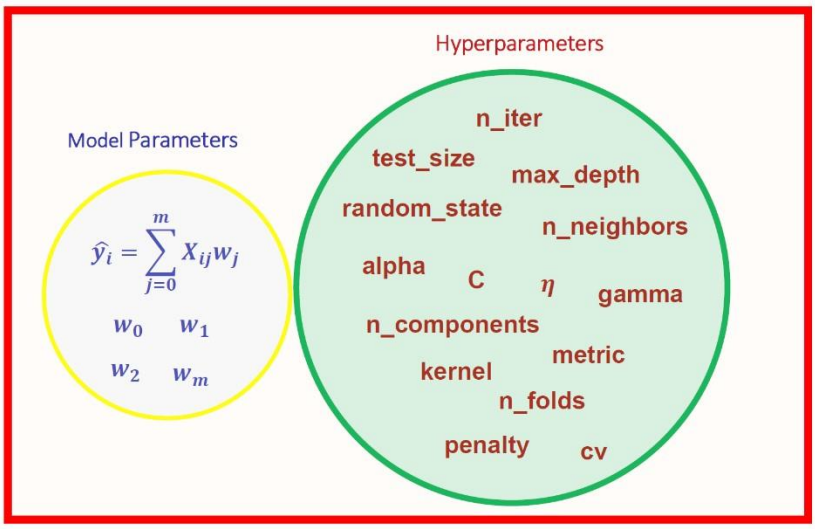
Ans: Model parameter: It is a configuration variable that is internal to a model and whose value can be predicted from the data.

- 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.

Q11: What is Model Hyperparameter?

Ans: 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.

Q12. What is cross-validation?

Ans: **Cross-validation**: It is a technique for evaluating Machine Learning models by training several Machine Learning models on subsets of available input data and evaluating them on the complementary subset of data.

Use cross-validation to detect overfitting, i.e., failing to generalize a pattern.

There are three steps involved in cross-validation are as follows:

- Reserve some portion of the sample dataset.
- Using the rest dataset and train models.
- Test the model using a reserve portion of the dataset.