1. Recognize the differences between supervised, semi-supervised, and unsupervised learning.

|  |  |  |
| --- | --- | --- |
| **Supervised Learning** | **Unsupervised Learning** | **Semi supervised Learning** |
| Supervised Learning can be used for 2 different types of problems i.e. regression and classification | Unsupervised Learning can be used for 2 different types of problems i.e. clustering and association. | Semi-Supervised learning is a type of Machine Learning algorithm that represents the intermediate ground between Supervised and Unsupervised learning algorithms. It uses the combination of labeled and unlabeled datasets during the training period. |
| Input Data is provided to the model along with the output in the Supervised Learning. | Only input data is provided in Unsupervised Learning. | Although Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. |

2. Describe in detail any five examples of classification problems.

Logistic regression.

Decision trees.

Random forest.

XGBoost.

Light GBM.

Voting classifiers.

Artificial neural networks.

3. Describe each phase of the classification process in detail.

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

4. Go through the SVM model in depth using various scenarios.

“Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

5. What are some of the benefits and drawbacks of SVM?

Advantages:

SVM works relatively well when there is a clear margin of separation between classes.

SVM is more effective in high dimensional spaces.

SVM is effective in cases where the number of dimensions is greater than the number of samples.

SVM is relatively memory efficient

Disadvantages:

SVM algorithm is not suitable for large data sets.

SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.

In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.

As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

6. Go over the kNN model in depth.

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

7. Discuss the kNN algorithm's error rate and validation error.

By observing validation error rate we can interpret that At K=1, we were over fitting the boundaries. In Validation graph Error rate initially decreases and reaches a minima. After the minima point, it then increase with increasing K. This value of K where error reaches minima should be used for all predictions.

8. For kNN, talk about how to measure the difference between the test and training results.

This is a supervised machine learning algorithm, which is generally used to solve classification problems but sometimes it was also used in regression problems too. The main aim of KNN is to find the nearest neighbours of our query point. This algorithm believes that similar things are in close proximity, in other words, we can say that suppose X is +ve in a group of points so there is a high chance that the point nearer to X is also +ve. Let us talk about the concept of nearest neighbour for better understanding.

9. Create the kNN algorithm.

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the

1. What is a decision tree, exactly? What are the various kinds of nodes? Explain all in depth.

A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a treelike shape. There are three different types of nodes: chance nodes, decision nodes, and end nodes.

11. Describe the different ways to scan a decision tree.

There are three different types of nodes: chance nodes, decision nodes, and end nodes. A chance node, represented by a circle, shows the probabilities of certain results. A decision node, represented by a square, shows a decision to be made, and an end node shows the final outcome of a decision path.

12. Describe in depth the decision tree algorithm.

A decision tree is a very specific type of probability tree that enables you to make a decision about some kind of process. For example, you might want to choose between manufacturing item A or item B, or investing in choice 1, choice 2, or choice 3.

13. In a decision tree, what is inductive bias? What would you do to stop overfitting?

The inductive bias (also known as learning bias) of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered. In machine learning, one aims to construct algorithms that are able to learn to predict a certain target output.

14.Explain advantages and disadvantages of using a decision tree?

Advantages:

Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.

A decision tree does not require normalization of data.

A decision tree does not require scaling of data as well.

Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.

A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.

Disadvantage:

A small change in the data can cause a large change in the structure of the decision tree causing instability.

For a Decision tree sometimes calculation can go far more complex compared to other algorithms.

Decision tree often involves higher time to train the model.

Decision tree training is relatively expensive as the complexity and time has taken are more.

The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

15. Describe in depth the problems that are suitable for decision tree learning.

Instances are represented by attribute-value pairs.

The target function has discrete output values.

Disjunctive descriptions may be required.

The training data may contain errors.

The training data may contain missing attribute values.

16. Describe in depth the random forest model. What distinguishes a random forest?

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

17. In a random forest, talk about OOB error and variable value.

Out of bag (OOB) score is a way of validating the Random forest model.

The feature importance (variable importance) describes which features are relevant. It can help with better understanding of the solved problem and sometimes lead to model improvements by employing the feature selection.